Predicting and Improving Takeover Performance by A Context-Aware Deep Learning Based Assistive System

by

Erfan Pakdamanian

A dissertation submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

at the

School of Engineering and Applied Science

University of Virginia

Charlottesville, VA

May 2022

Committee Approval:

We, the undersigned committee members, certify that we have reviewed this dissertation and approve it in partial fulfillment of the requirements of the degree of Doctor of Philosophy in Systems Engineering.

Lu Feng, Ph.D. (School of Engineering and Applied Science) Date
Advisor

Laura E. Barnes, Ph.D. (School of Engineering and Applied Science) Date
Committee Chair

Donna Chen, Ph.D. (School of Engineering and Applied Science) Date
Committee Member

Sara Riggs, Ph.D. (School of Engineering and Applied Science) Date
Committee Member

Seongkook Heo, Ph.D. (School of Engineering and Applied Science) Date
Committee Member

Certified by:

Jennifer L. West, Dean, School of Engineering and Applied Science Date

Abstract

With the Level 3 of automation, drivers are no longer required to constantly drive or actively monitor their driving environments and may engage in activities other than driving. However, drivers will still be required to take control of vehicles as soon as automation reaches its limits. As a result of being decoupled from the operating task for a prolonged time, drivers have difficulty regaining the vehicle control in a timely manner. In order to counter the difficulty of takeovers, various factors affecting takeover performance have been evaluated. However, not all factors have been studied comprehensively, and the results of some factors have been contradictory. Additionally, there's a need for development of computational models that reliably predict drivers' takeover performance from their physiological and driving environment data, and utilize the outcome to inform drivers about the upcoming hazards.

This dissertation sought to address these shortcomings by (1) Examining the effect of cognitive load, situation awareness, stress, traffic density, and lead time on drivers' takeover behaviors (takeover time and quality) and psychophysiological responses (i.e. eye movements, electroencephalography, galvanic skin responses, and heart rate variability); (2) Developing neural network models for predicting drivers' attention and takeover performance by utilizing their physiological data, vehicle's status, and driving environment; (3) Designing an end-to-end context-aware in-vehicle alert system which notifies drivers in a real-time about the loss of situation awareness using multimodal modalities, and (4) Evaluating the system in critical conditions by conducting human-subject experiments.

Acknowledgments

The completion of this dissertation would not be possible without the support of a number of people.

First and foremost, I would like to express my gratitude towards my academic supervisor, Dr. Lu Feng. She has been an invaluable source of guidance, support, and encouragement throughout this journey from start to finish. I am incredibly grateful for all the trust and patience you have in me.

I am especially indebted to Dr. Seongkook Heo and Dr. Sarit Kraus for laying out the foundation of my doctoral studies, in particular creating an end-to-end system for predicting driver behavior and providing real-time context-aware feedback. I truly appreciate the invaluable time, wisdom and great support you offered, even during the unprecedented Covid-19 pandemic.

I would also like to thank my committee members, Dr. Laura Barnes, Dr. Donna Chen, and Dr. Sara Riggs, who have shared their time and wisdom with me, reviewing my proposal and providing valuable comments on this thesis.

Without the continuous emotional support of my parents, my brothers, and my considerate friends, I would not be where I am today. There were certainly a lot of up-and-downs during graduate school, but they are always the shining light that cheers me up when I met obstacles in research or have problems figuring out what to do after graduate school. This journey wouldn't have been possible without the encouragement they have given me throughout the years. I am blessed to have them all in my life.

There are so many wonderful people that I couldn't acknowledge. Thank you everyone so much for your positive influence in my life.

Finally, I am proud of myself because doing this PhD required leaving my country, stepping out of my comfort zone, studying extra-long hours, and writing up multiple papers amid the Covid-19 lockdown.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

ADAS	Advanced Driver Assistance Systems
ADS	Automated Driving System
ANOVA	Analysis Of VAriance
AOI	Areas of Interest
AV	Automated Vehicles
EEG	Electroencephalography
GSR	Galvanic Skin Responses
HRV	Heart Rate Variability
HR	Heart Rate
NDRT	Non-Driving Related Tasks
NHTSA	National Highway Traffic Safety Administration
SAE	Society of Automotive Engineers
SA	Situation Awareness
SD	Standard Deviation
TOR	Take-Over Request
TTC	Time To Collision

I

1 | Introduction & Motivation

1.1 Introduction

Automated vehicles (AVs) are considered as the next disruptive revolution in the transportation system. AVs are a collection of intelligent automation technologies which are designed to take some or all of the driving tasks from the human drivers. Several automotive manufacturers have committed to equip their commercialized vehicles with some degree of automation for decades. The rapid ongoing advancements in development of software and hardware technologies in recent years promises vehicles with autonomous capabilities in the near future. For example, Waymo's test fleet has driven over six million miles in autonomous mode on public roads [Schwall et al., 2020]. The predictions show that the shift towards higher levels of automation in driving continues to reach global sales of nearly 21 million autonomous vehicles in 2035 [Automotive, 2016, Urmson, 2015].

The advent of AVs could have a number of benefits for the individuals and society. On the individual level, AVs enhance the mobility access for users, including but not limited to elderly and physically impaired who cannot drive for medical reasons, by ensuring the secure participation of them in traffic [Yang and Coughlin, 2014, Rahman et al., 2019]. Additionally, eliminating the chauffeuring burdens increase productivity due to less nuisance tasks and leeway for action [Montgomery, 2018]. On the other hand, AVs could increase transport efficiency, allowing reduction of carbon dioxide emissions and fuel consumption by optimizing traffic flow [Anderson et al., 2014, Ntousakis et al., 2015]. More importantly, AVs equipped with full

	Level of Automation	Description of autonomy as defined in SAE
L0	No driving automation	The performance by the driver of the entire DDT*, even when enhanced by active safety systems.
L1	Driver assistance	The sustained and ODD*specific execution by a driving automation system of either the lateral or the longitudinal vehicle motion control sub task of the DDT (but not both simultaneously) with the expectation that the driver performs the remainder of the DDT.
L2	Partial driving automation	The sustained and ODD-specific execution by a driving automation system of both the lateral and longitudinal vehicle motion control subtasks of the DDT with the expectation that the driver completes the OEDR* subtask and supervises the driving automation system
L3	Conditional driving automation	The sustained and ODD-specific performance by an ADS* of the entire DDT with the expectation that the DDT fallback-ready user is receptive to ADS-issued requests to intervene, as well as to DDT performance-relevant system failures in other vehicle systems, and will respond appropriately.
L4	High driving automation	The sustained and ODD-specific performance by an ADS of the entire DDT and DDT fallback without any expectation that a user will respond to a request to intervene.
L5	Full driving automation	The sustained and unconditional performance by an ADS of the entire DDT and DDT fallback without any expectation that a user will respond to a request to intervene.

ODD = Operational Design Domain, DDT = Dynamic Driving task, ADS = Automated Driving System, OEDR = Object and Event Detection and Response

Table 1.1: Society of Automotive Engineers Levels of Automation

Automated Driving Systems (ADS) could significantly improve traffic safety, notably eliminate the primary cause of 90% of traffic fatalities, human-related errors including fatigue, inexperience, or even drug abuse [Olaverri-Monreal and Jizba, 2016, Shanker et al., 2013]. Nevertheless, it is important to understand the numerous considerations required to ensure a seamless integration of AVs in public road at a large scale, mainly in the incipient stages of their development. AV and human driver relationship are often deemed as the most important concerns and challenges of these vehicles which their occasional failures could dramatically shift the public perception and acceptance.

The relatively nascent concerns about AVs, including human resistance to change, distrust of automated vehicles, risk perception, accessibility, physical and cognitive workload need careful scrutiny. Many barriers still hinder the widespread, practical, and effective integration of AVs in traffic. For example, in countless scenarios other road users (i.e. cyclists and other manually-driven vehicles) can no longer interact with a driver and get information about the intentions of the vehicle as they would normally do. Furthermore, there is uncertainty about the consequence of the AVs introduction to the public at a scale as their greatest potential requires continuously monitoring of surroundings and making reliable decisions to avoid possibility of errors which human are prone to. Although autonomous vehicles should be able to outperform a human driver in reducing the fatal errors and no intentional violations of traffic regulations, recent fatal crashes indicate their failures to promptly and properly respond to unknown situations [Board, 2020]. Regardless of the technological advances in the automotive industry, we may argue that such developments are currently great only at performing regular task demands that are normally imposed on a human driver. But, the current technology limitations and the regular imperfections in automated systems illustrate the requirement for added supervision on the system. As a result, the integration of ADS is expected to be gradual until the full driving automation (level 5, Table 1.1) reaches mastery of the many technical complexities and challenges that pertain to their development and introduction to the public [Olaverri-Monreal, 2020]. So human drivers still play a prominent role in the human-system cooperative driving system.

The utmost aforementioned benefits of the AVs are still speculative and exist at the absence of human intervention in the control of vehicles [Olaverri-Monreal and Jizba, 2016]. Whilst most available AVs are currently at either level 2 or level 3 of automated driving (Table 1.1; [SAE, 2018])[1] which provide various forms of driver assistance, advanced monitoring systems, and control of the longitudinal and lateral vehicle kinematics on a sustained basis. Generally both Level 2 and Level 3 are ultimate human-system cooperation. At level 2 or "partial AV", human drivers are responsible for monitoring the surrounding where the system does steering

[1] In this dissertation AVs are mainly referred to the level 2 & 3 of automation

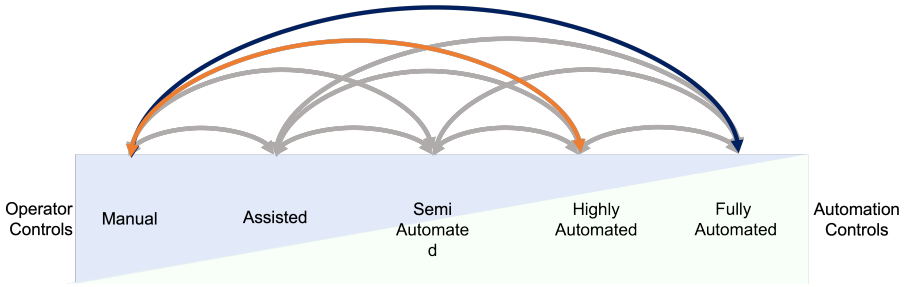

Figure 1.1: All possible transitions occurring between a human driver and automated driving system at different levels of automation

and acceleration/deceleration simultaneously. While at Level 3 or "conditional AV" human driver delegate control of the vehicle and monitoring of the road to an automated system but should react to uncertainties or request to intervene. Although in conditionally automated driving, drivers do not need to continuously monitor the driving environment, the automated system still needs to relinquish the control back and ask the human driver to resume the control in case of system failures, anticipated dangerous situation, or exceeding its operational limit via a so-called take-over request (TOR) [Bazilinskyy et al., 2018, Gold et al., 2013]. As human errors typically arises out of poor human-system interaction, new human factors challenges are ubiquitous in transition of control situation [Flemisch et al., 2012].

1.1.1 Transitions of control

AVs allow the human drivers to relinquish control of the vehicle, take their hands off the steering wheels, foot off the pedals, and instead engage in NDRTs under predefined conditions, but will still require that they maintain a relative vigilance and be prepared to resume control when requested or required to do so. The term *"transition"* refers to a transfer of driving responsibility [Saffarian et al., 2012], or alteration between different levels of autonomy [Flemisch et al., 2012]. [Flemisch et al., 2008] defines the level of involvement of human and ADS in the control of vehicle based on the principle of transition(see Figure 1.1). These principles outlines the possibility

of transfer of control, who initiates the transition, and who should be responsible of driving at the beginning and the end of the transition.

Presumably, a driver-initiated transition of control would be less complex, as the driver has the freedom to choose a safe time for a handover. Consequently, the driver-initiated transition is less of a concern from the Human Factors perspective as it less likely to present a significant threat to a driver. A recent study by [Boggs et al., 2020] investigates the role of the driver and the system in AV. The analysis of 159,840 transition reported on Autonomous Vehicle Tester Program revealed that the majority of the disengagements were driver-initiated. The results show the drivers not engaged in NDRTs are able to intervene before system issues a TOR. However, under normal driving circumstances, most of the TOR initiations are assumed to done by the ADS. Generally, ADS is programmed to initiate either *"unplanned"* or *"planned"* TOR. Unplanned TOR occurs once the ADS detects an unexpected events, such as construction zones, missing road marks, and an upcoming accident blocking the road. In *unplanned* takeover situation, the driver is expected to immediately take the vehicle control. Ideally, when the system perceives that it can no longer provide automated driving, a TOR should be issued as early as possible to allow the driver enough time to restore situational awareness and regain control of the vehicle. This situation raises many Human Factors concerns as drivers will experience a series of sub-processes for the takeover preparation, which includes: perception of TOR, cognitively process the information, gain situation awareness (SA), make decisions, and resume motor readiness [Zhang et al., 2019b, Zeeb et al., 2015]. In contrast, a *planned* TOR takes place once the system predicts a safety-concern situation, i.e., expectation-conform scenarios, forecasted severe weather [Holländer and Pfleging, 2018].There is no doubt that transition from automated to manual control is not a trivial task. [Vogelpohl et al., 2018] argued that drivers might require additional time and assistance in order to reach a level of situational awareness necessary to resume manual driving.

1.1.2 Internal factors influencing takeover performance

Despite extensive research on the effect of recent evolution of AVs, little is still known about psychological factors pertinent to driver behavior after TOR perception. In order to understand the main Human Factors concerns of AVs, it's essential to scrutinise why and how they impact driver's cognitive and physical abilities. As noted above, in Level 3, the role of drivers will transform from "operators" to a "fallback-ready" [Noy et al., 2018]. Once the system is set to the automated driving mode, human drivers are legally allowed to engaged in NDRTs. Consequently, drivers' perception, judgment, decision-making, and operation skill differ substantially between manually controlling the vehicle and delegating the driving responsibilities to an ADS. In any instance (e.g. system failure) whereby a TOR is issued, the driver is expected to be able to recapture control of the vehicle. However, as they become increasingly decoupled from the operational level of driving lead, higher level of automation potentially lead to the loss of SA [Stanton et al., 2011], increase in cognitive workload [De Winter et al., 2014, Winter et al., 2016], and overreliance [Stanton, 2015]. In response to such complex takeover, research has been conducted to understand main factors affecting driver's ability to takeover after a system-initiated TOR.

Takeover performance can be explained by both reaction time and post-takeover control [McDonald et al., 2019]. Despite many factors have been identified contributing to better reaction time and takeover control such as traffic density [Gold et al., 2016] and driver cognitive state [Sadeghian Borojeni et al., 2018, Van der Heiden et al., 2021] or emotion [Sanghavi et al., 2020], the impact of time budget ("lead time") [Eriksson and Stanton, 2017] and TOR modality [Borojeni et al., 2017] have been widely studied by researchers. For example, studies show that additional second of time budget lead to increase of reaction time by on average 0.27second [Zhang et al., 2019a, McDonald et al., 2019]. If drivers are given more time to gain sufficient SA, they could prepare for the upcoming transition of control. Gold et al. [Gold et al., 2013] has shown that shorter takeover times lead to faster responses but worse maneuvers. Furthermore, a study by Merat et al. [Merat et al., 2014] suggests 20-40second of

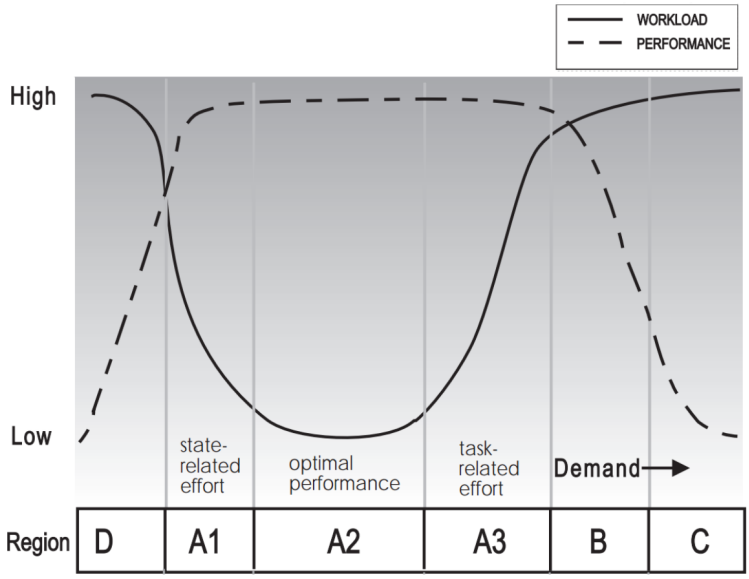

Figure 1.2: Workload and performance relationship.

time budget for a safe takeover to fully stabilised the vehicle after reclaiming control. As supplying such time budget may not be technologically feasible at the moment, researchers are required to study alternative approaches to enable drivers gaining enough SA as a function of available time [Lu et al., 2017].

Cognitive workload

Studies have shown that a sudden TOR to the driver about the upcoming potential hazards would incur higher stress and cognitive load [Shah et al., 2015]. Although there was a belief that automation could increase mental workload [Young and Stanton, 2002], a meta-analysis by [De Winter et al., 2014] showed controversial results. In fact, increasing the automation level reduced the mental workload. However, separate studies have shown the negative impact of driving with autonomous mode on mental workload, take-over performance, and reaction times [Strand et al., 2014, Zeeb et al., 2015, Bueno et al., 2016]. In order to apprehend the root cause of dissimilarities, we need to conceptualise driver's mental workload before and after

TOR perception within the framework of mental workload (MWL) theory defined by De Waard [De Waard and Brookhuis, 1996].

As can be seen in Figure 1.2, when demand increases, starting from the optimal operator state in region A2, the operator's capability of (effort) compensation will be exceeded at a certain moment and a transition from the A3 to the B region takes place. Performance in the B-region deteriorates, and when it reaches a minimum level, the C-region is entered. Task performance and workload as a function of demand are depicted in Figure 1.2. It is imperative to stress that demand on the x-axis in Figure 1.2 is not directly linked to region of performance. Task demands are determined by the goals that have to be reached by task performance and cannot be directly related to workload, which is subjective.

Physiological data has been used as an essential instrument for understanding and interpreting a driver's mental status. Applying neuropsychological and physiological measurements on drivers to investigate the relationship between mental behavior and performance while taking over could provide us a profound understanding of what modalities provide useful TOR for autonomous vehicles. Whilst previously mentioned research has explored the best TOR modality, relatively few studies have investigated the drivers' cognitive states at the time of transition [Izquierdo-Reyes et al., 2018, Sibi et al., 2016]. For the purpose of objectively obtaining the psychophysical state of the driver as accurate as possible, there is a need to shift from the simple questionnaires to a direct assessment of driver physiological responses and driver behavioral pattern.

1.1.3 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, and galvanic skin responses (GSRs).

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]. 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]. 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.

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 [Lohani et al., 2019, Mehler et al., 2012]. For instance, [Hidalgo-Muñoz et al., 2019] conducted a driving simulator study with eighteen 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. 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.

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 [Wintersberger et al., 2018, Mehler et al., 2012]. 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.

1.1.4 External factors impacting takeover performance

There are several environment-based factors indicating a safe takeover, including a takeover time, and the quality of takeover.

Takeover time In this paper, we consider the *takeover time* as the period of time from the initiation of TOR to the exact moment of the driver resuming manual control (see Figure 5.3), following the ISO standard definition in [ISO 21959:2020, 2020]. Note that the same concept has also sometimes been named as takeover reaction time or response time in the literature (e.g., [Johns et al., 2016, Kim and Yang, 2017, Petermeijer et al., 2017a, Eriksson and Stanton, 2017]). The empirical literature defines a large variety of takeover time from a mean of 0.87s to brake [Winter et al., 2016], to an average of 19.8s to response to a countdown TOR [Politis et al., 2018] and 40s to stabilize the vehicle [Merat et al., 2014]. This range is derived from influential factors impacting perception, cognitive processing, decision-making and resuming readiness [Gold et al., 2016, Zeeb et al., 2015]. A meta-analysis of 129 studies by Zhang et al. [Zhang et al., 2019a] found that a shorter takeover time is associated with the following factors: a higher urgency of the driving situation, the driver not performing a non-driving related task (NDRT) such as using a handheld device, the driver receiving an auditory or vibrotactile TOR rather than no TOR or a visual-only TOR. Recent studies by Mok et al. [Mok et al., 2017] and Eriksson et al. [Eriksson and Stanton, 2017] both confirmed that drivers occupied by NDRTs have higher responses to TORs. Similarly, [Feldhütter et al., 2017] found a significant increase in reaction time induced by NDRTs. It is further concluded that the visual distraction causes higher reaction time when it is loaded with cognitive tasks [Tang et al., 2020]. Studies have also revealed several driving environments, TOR modalities [van der Heiden et al., 2017, Tang et al., 2020], driving expectancy [Ruscio et al., 2015], age [Walch et al., 2017] and gender [Warshawsky-Livne and Shinar, 2002] associated with takeover time. The present study extend previous findings by considering various NDRTs, gender, and objective and subjective measurements of mental workload into the DeepTake framework.

Takeover quality In addition to takeover time, it is essential to assess the *takeover quality*, which is defined as the quality of driver intervention after resuming manual control [ISO 21959:2020, 2020]. There are a variety of takeover quality measures, depending on different takeover situations (e.g., collision avoidance, lane-keeping), including objective measures (e.g., mean lateral position deviation, steering

wheel angle deviation, metrics of distance to other vehicles or objects, minimum time to collision, frequency of emergency braking) and subjective measures (e.g., expert-based assessment, self-reported experience). Prior work has found that takeover quality can be influenced by factors such as drivers' cognitive load [Du et al., 2020a, Zeeb et al., 2016], emotions and trust [Dillen et al., 2020, Du et al., 2020c, Hergeth et al., 2017], and distraction of secondary NDRTs [Martelaro et al., 2019, Dogan et al., 2019]. Takeover time to an obstacle [Zeeb et al., 2016] has been used widely studies as an indicator of takeover performance [Eriksson and Stanton, 2017]. However, a study by Louw et al. [Louw et al., 2017] showed that takeover time and quality appear to be independent. This lack of consensus could be due to the fact that studies apply various time budget for takeover control.

1.1.5 Models for takeover performance prediction

While existing literature mostly focus on the empirical analysis of drivers' takeover time and quality, there are a few recent efforts on the predication of drivers' takeover behavior using machine learning (ML) approaches. Lotz and Weissenberger [Lotz and Weissenberger, 2018] applied a linear support vector machine (SVM) method to classify takeover time with four classes, using driver data collected with a remote eye-tracker and body posture camera; the results achieve an accuracy of 61%. Braunagel et al. [Braunagel et al., 2017] developed an automated system that can classify the driver's takeover readiness into two levels of low and high (labeled by objective driving parameters related to the takeover quality); their best results reached an overall accuracy of 79% based on a linear SVM classifier, using features including the traffic situation complexity, the driver's gazes on the road and NDRT involvement. Deo and Trivedi [Deo and Trivedi, 2019] proposed a Long Short Term Memory (LSTM) model for continuous estimation of the driver's takeover readiness index (defined by subjective ratings of human observers viewing the feed from in-vehicle vision sensors), using features representing the driver's states (e.g., gaze, hand, pose, foot activity); their best results achieve a mean absolute error (MAE) of 0.449 on a 5 point scale of the takeover readiness index. Du et al. [Du et al., 2020b, Du et al., 2020d] developed

random forest models for classifying drivers' takeover quality into two categories of good and bad (given by subjective self-reported ratings), using drivers' physiological data and environment parameters; their best model achieves an accuracy of 70%.

1.2 Motivation

As stated above, improving the infrastructure and user adaption of highly automated vehicles will not happen overnight. Consequently, a transition period between Level 3 and driverless vehicles could take decades. While these AVs let drivers to take their hands off the steering wheels, foot off the pedals, and instead engage in NDRTs such as reading or using mobile devices, a few sec permission to gain enough SA and prepare for a safe control of the vehicle are likely to be the cause for some safety concerns. Primarily, a human driver tend to lose visual attention to a source of information after about ten minutes. According to the multiple resources theory [Wickens, 2002], tasks demand specific resources. These resources can be categorized into four dichotomous dimensions: stages of information processing (perception/cognition, responding), perceptual modalities (visual, auditory), visual channels (focal, ambient), and processing codes (spatial, verbal). The distraction generated by the NDRTs means less attention being directed towards the road than in the case of manual driving. If several tasks build on the same resource dimension at the same time, task interferences might occur due to limited resources. Thus, it's enormously overwhelming for a driver to regain control of the vehicle in a short time while it's required to perceive the takeover request from the system, gain enough SA, and safely maneuver the vehicle while NDRTs simultaneously compete with the visual-attentional resources for the driving task [Naujoks et al., 2017].

Thus, an effective system should have to consider driver state, limitations, and abilities in mitigating a potential collision before relinquishing control. Humans are prone to distraction, as a result human drivers cannot be relied upon to guarantee they are sufficiently aware of the situation to ensure safe vehicle control. Accordingly, rather than using a capacitive sensing system or relying on torque placed on the

steering wheel, ADS system should be equipped with a driver monitoring system, which essentially makes the decision of when to initiate TOR, and how to relinquish driving control pertinent to the current state of the driver as well as the possible takeover performance. For example, if the pattern of drivers' visual attention in the lead up to a transition shows that they were completely disengaged from the driving task, then a TOR must be initiated to either accommodate the context of immersion or delay it until drivers' attention is back on the driving task. The latter may not be a safe option in in unplanned situation with an urgency of a handover. In addition, the ADS may be programmed to choose a minimum risk situation such as bringing the vehicle to a safe position on the road.

1.3 Contributions

To fill the aforementioned research gaps, this thesis contributions to the literature are as follows:

1. We investigate the effects of drivers' mental workload and type of TORs on their takeover performance (i.e. takeover time and takeover quality) and psychophysiological responses (gaze behavior, heart rate activities, GSR, and EEG).

2. We develop, to our knowledge, the first neural network model to predict drivers' takeover performance (i.e. takeover intention, takeover time, and takeover quality) by utilizing drivers' physiological data and driving environment.

3. We develop, to our knowledge, the first end-to-end in-vehicle alert system that informs drivers about the loss of SA using a context-aware warning system. To evaluate the system's practicality, we conduct a preliminary proof-of-concept human-subject experiments to study their takeover performance, perceived safety, acceptance, and preparedness in multiple traffic scenarios.

1.4 Organization

This thesis consists of 7 chapters which are arranged in five parts.

Chapters 2 and 3 examine to what extent driver's visual perception and mental workload can be relied on to choose a suitable type of TOR. Specifically, Chapter 2 presents the results of a driving simulator experiment in which participants were given haptic-feedback to analyze their SA and stress using EEG signals. Chapter 3 examines the extent to which the visual and auditory TOR will effect to drivers' behavioral and psychophysiological responses. To do so, the study investigated how well participants can perform a takeover after receiving a TOR and how they perceive each cue. The changes in participants' perception and the features reflect the takeover performance were measured. Then, the results were utilized for designing the end-to-end system in part III that help drivers in longitudinal control after a TOR.

Chapter 4 studies the a new approach in prediction of driver's visual attention. This chapter constitutes studies that explore how drivers quickly allocate their visual attention to the most important cue of the scene. Chapter 5 presents the state-of-the-art neural networks framework to predict multiple aspects of takeover, including takeover intention, takeover time, and takeover quality. We investigate the effectiveness of the presented model using various metrics.

Chapter 6 illustrates the design and evaluation of context-aware in-vehicle warning to alert driver about the loss of SA. The aim of this design was to assist drivers by informing about the immersion to a NDRT for a prolonged time. The design was based on the premise that a system giving context-warning pertinent to the task the driver is engaged in.

Chapter 7 provides a general discussion of the conducted studies and suggests opportunities for future research. Please note that each of the chapters is readable in isolation. That is, Chapters 2–6 each have their own introduction and literature review, methods, results, and discussion section.

II

2 | Understanding Driver' State

It remains uncertain regarding the safety of driving in autonomous vehicles that, after a long, passive control and inattention to the driving situation, how the drivers will be effectively informed to take-over the control in emergency. In particular, the active role of vehicle force feedback on the driver's risk perception on curves has not been fully explored. To investigate it, the current paper examined the driver's cognitive and visual responses to the whole-body haptic feedback during curve negotiations. The effects of force feedback on drivers' responses on curves were investigated in a high-fidelity driving simulator while measuring EEG and visual gaze over ten participants. The preliminary analyses of the first two participants revealed that pupil diameter and fixation time on the curves were significantly longer when the driver received whole-body feedback, compared to none. The findings suggest that whole-body feedback can be used as an effective "advance notification" of hazards.

2.1 Introduction

Although the future of car industries will be dominated by autonomous vehicles and the car will drive itself, there will still be a need for drivers to take over the car [Banks and Stanton, 2016]. Human intervention is particularly necessary to prevent tragic accidents when the autonomous vehicle encounters curves, bad weather and unpredictable pedestrian behavior [Wright et al., 2017]. Although autonomous vehicles will overall decrease the physical and mental workload of drivers by assigning these tasks to an automated system, human drivers would still play a critical role in car

safety responsibility [Parasuraman and Wickens, 2008]. However, it was been shown that a sudden alarm and notification to the driver about the upcoming potential hazards would incur higher stress and cognitive load [Shah et al., 2015]. Once drivers allow the automated system to control the car, meaning the driver tends to allocate his attention resources to non-driving tasks (e.g. video gaming, talking on the phone etc.), his or her attention will be taken away from the primary task of driving. In such circumstances, any simple form of visual, auditory or haptic signals would not be sufficient to communicate critical information about the vehicle conditions, only to startle and stress the driver in emergency [Petermeijer et al., 2017b]. One approach to cope with the stress from unexpected alarms is to examine the effects of signals on potentially safety-compromising situations, and accustom the driver to it. In this regard, this paper intends to investigate the effects of whole-body haptic feedback, delivering haptic cues to drivers' full body, on the drivers' visual perception and cognitive states during curve negotiation, as an alternative to its counterpart alarming signals. Assessing the drivers' cognitive states can help infer what type of haptic feedback the cars should provide to mitigate the stress of taking over during critical moments. In literature, vibrotactile haptic feedback was shown to enhance the reaction time of taking control back at life-threatening moments [Prewett et al., 2011]. It noted, however, that once the drivers were spatially aware, the vibrotactile "directional" cue may not be as effective as visual directional alternatives. Therefore, this study intends to focus on whole-body haptic feedback to complement this drawback. Morrell and Wasilewski [Morrell and Wasilewski, 2010] designed and developed a haptic-feedback seat for traditional vehicles that aimed to share spatial information, and improve situation awareness (SA). The drivers were informed about the location of car-following and close-by vehicles, through vibrotactile feedback from the seat back in a way that the closer the car is, the more sensors vibrated. Nonetheless, on the one hand, evaluating the time in the blind spot may not be the accurate measurement for the risk assessment. Nonetheless, on the one hand, evaluating the time in the blind spot may not be the accurate measurement for the risk assessment. On the other hand, as auto industries attend to autonomous technologies, alert systems need to become adaptive to vehicle speed and situation

but not particularly designed for a specific scenario. Petermeijer et al. [Petermeijer et al., 2017b] designed a vibrotactile feedback seat that contains static and dynamic vibration for automated vehicles. The authors aimed to analyze the accuracy of drivers' response rate and their reaction time to the requested time for maneuvering. After receiving tactile stimuli, drivers had to respond accordingly to the vibration direction by moving to the left or the right. However, in all the presented scenarios, there was not any additional warning cue. Furthermore, the participants reported difficulties in understanding whether the cue was to their left or right; alarms were only triggered about one second prior to an event occurrence, which was shorter than the realistic average reaction time needed (3.5 seconds) for a transition control in automated vehicles [Melcher et al., 2015]. This research aims to examine perceptual and cognitive effects of using whole-body force feedback on the control responses of the drivers. Through the controlled experiments in simulation setting, it is expected that the whole-body force feedback will be shown its values, in a way that does not only warn the driver when a takeover is required, but also assists the driver during the critical phases, including their lack of SA (shifting of attention) and cognitive processing. In this regard, we hypothesized that the whole-body haptic feedback would allow the drivers to be effectively aware of upcoming curves in a simulated driving environment.

2.2 Methodology

The experiment was conducted in a high-fidelity driving simulator (the 401cr motion system by Force Dynamics) equipped with three monitors. The simulator mimics various acceleration dynamics thereby creating a realistic response upon the driver's body. The motion-capable high-fidelity simulator was used with two configurations: (1) without whole-body motion feedback, (2) with whole-body motion feedback with approximately 18 inches of movement in 360 degrees. This also allowed six degrees of freedom to replicate the motions associated with driving in a way that vibration of the seat serves as an "intelligent messenger". It ensures human stays informed of

the vehicle safety. The study was approved by University of Virginia Institutional Review Board (Protocol# 2017-0296-00). The speedometer and the RPM gauge is located in the center of the middle monitor (Figure 2.1). Moreover, the implemented automation system had a longitudinal capability similar to common ACC systems, which allow drivers to follow the indicated speed limits as well as keep the car in the center of the lane. Data were recorded at a frequency of 100 Hz, including the vehicle's position, accelerations and steering wheel angle (they were not included in the preliminary study and will be reported in further analysis).

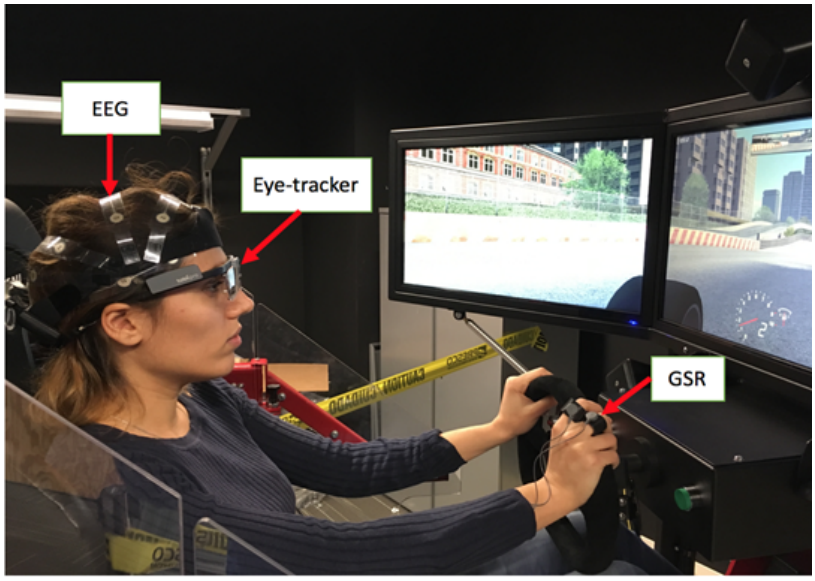

Figure 2.1: Experimental set-up for recording EEG and Eye movement.

2.2.1 Data Acquisition

A wearable eye-tracker glass (Tobii Pro-Glasses 2, Danderyd, Sweden; Tobii Pro-Glasses 2, 2017) was used to track the driver's gaze behavior at a sampling rate of 60 Hz (i.e., 60 gaze data points collected per second for each eye; 4 eye cameras, H.264 1920x1080 pixels at 25 fps) (see Figure 2.1). The Tobii Pro Glasses 2 eye-tracker is wireless with live view capability for insights in any real-world environment. Since

the driving simulator and curves are dynamic scenes, head-mounted eye tracker was required. Also, it ensures that the participant's full and complete range of motions for their head. A B-Alert X24 system with 24 channels was used with the sample rate of 256Hz to record the Electroencephalography (EEG) data (Figure 2.1). Wireless EEG signals were sent via Bluetooth to the data acquisition system. Also, in order to record the electoral activity of the brain, the sensor strip was placed according to the 10/20 extended standard. The sampled data was sent wirelessly to iMotion (biometric research platform) which allowed collection of the synchronized EEG and eye-tracker data [iMotions, 2015b].

2.2.2 Procedure

Two graduate students (both male, 22 and 35 years old) holding a driver's license voluntarily participated in this preliminary study (ten participants equally balanced between male and female aged between 18 to 40 will be recruited). None of the participants had visual impairments, or any other symptoms or diseases that could compromise their ability to drive. Once the participant arrived, the relevant information regarding gender, age and driving experiences was gathered. Subsequently, participants were verbally instructed regarding how to use the devices and simulator as well as their primary task of driving with their hands on the wheel by the experimenter. Furthermore, both drivers were told that they need to keep their speed under 60mph and drive as they would normally do. The experimenter allowed the participants to familiarize themselves with the system with 2-5 mins test drive. Once they showed that they were comfortable with all the devices and driving the simulator, there were asked to take 3mins break between the sessions in order to maximize the concentration level and minimize fatigue throughout the 18 min session. The experimenter started the three curve and force-feedback-free trials as the Baseline session. Afterwards, the participants drove through the counterbalanced designated scenario six times (three trials with force feedback and three without). Each scenario took approximately 3mins, depends on the speed.

Table 2.1: Mean and Standard Deviation for Metrics of Eye Movements.

	Independent Variables		
Dependent Variables	With force feedback	Without force feedback	p-value
Time spent (sec)	6.83 (1.92)	4.49 (0.28)	0.002
Fixation duration (sec)	3.45 (0.89)	3.04 (0.61)	<0.001
Pupil diameter (px)	35 (9.5)	27 (4.3)	0.008

2.2.3 Signal Preprocessing

256Hz sampled data was filtered using high and low band pass filter with a cut-off frequency between 0.5 Hz, to remove DC drift, and 80 Hz respectively to remove power-line noise and low frequencies separately [Gheorghe, 2017]. Also, a notch filter at 60Hz was used. EEG data pre-processing initiated by referencing to the left ear lobe channel as well as applying Fast Fourier transform (FFT) algorithm to filter the different frequency band. To analyze the EEG data, initially the blink artifacts were removed by using independent component analysis (ICA) and wavelet analyses were used to generate a continuous record of theta band by using Matlab (The MathWorks, Inc., Natick, Massachusetts, United States) and EEGLab toolbox. An electrode impedance test was performed to ensure proper conductivity of the electrodes. The impedance level threshold of 20 kω was used. Also, the EEG calibration procedure was implemented before data collection.

2.3 Results

Collected data were extracted using iMotion software. In order to perform a comparison analysis between three conditions (Baseline, with whole body force feedback and without), approximately four seconds before curves was analyzed following the approach taken by [Gheorghe, 2017]. Each trial consisted of twenty curves, including simple curves, compound curves, reverse curves and deviation. However, we were only interested in simple curves for our preliminary study.

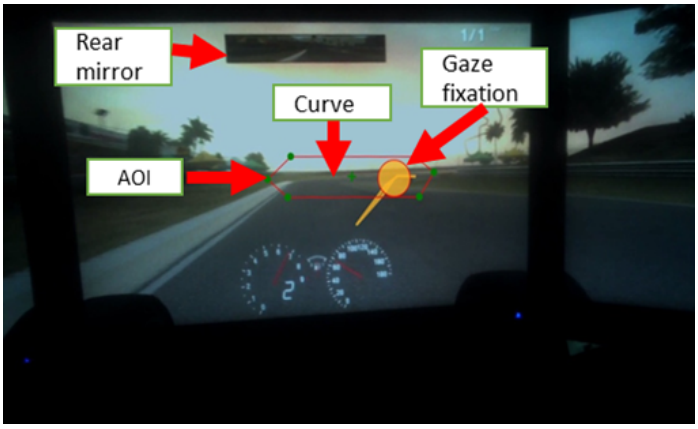

Figure 2.2: AOI and driver's view.

2.3.1 Analysis of visual attention

iMotion provides the following metric for analyzing eye movement: Time spent-fixation, fixation duration, Time to First Fixation (TTFF-F) and pupil diameter (Table 1). Table 1 summarizes the time spent and fixation duration on the AOI. In order to identify when curves as the critical section of the road on the visual display were fixed, AOI analysis was performed (see Figure 2.2). Comparing the TTFF values (see Figure 2.3) indicates both participants tended to concentrate slightly more on the curves at the presence of force feedback which indicates higher SA. Likewise, when the force feedback was applied to drivers, pupil diameter was larger approaching the curves (Table 1). Therefore, the drivers tend to fixed their gaze on curves significantly higher at the presence of the whole-body feedback. The differences between the two types of vibration patterns including force and none-force feedback was assessed using t-test. T-test yielded statistically difference between the force feedback and none in the dependent variables ($t(11) = 4.96$, $p = 0.002$; $t(11) = 12.38$, $p < 0.001$; $t(11) = 3.51$, $p = 0.008$, for time spent, fixation duration, and pupil diameter, respectively).

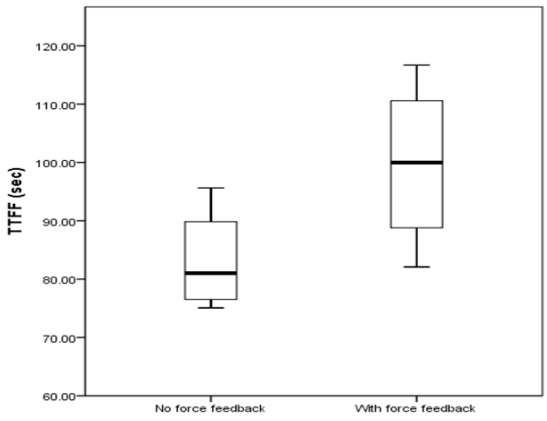

Figure 2.3: Eye- tracking results for fixation behaviors over different feedback condition.

2.3.2 Analysis of cognitive states

Analyzing three frequencies (Theta, Alpha and Beta) revealed that the Theta power increases in force feedback cases. Also, on the beta band, grown power was obtained. Still, the amount of power increasing on Theta band was higher, which may indicate the greater drivers' engagement while using haptic feedback. The findings represent that the force-feedback could correlate with higher ability in decision making and ultimately increase the capability of controlling the vehicle properly at the time of hazard encounter. It was initially expected to get the consistent results with Almahasneh et al. [Almahasneh et al., 2014] findings, however, the topographical map result (Figure 2.4) indicates that the difference between baseline and both cases is caused by more activity in corresponding brain region of the right frontal hemisphere near reaching the curves. Since most of cognitive activities occur at the frontal lobe, the findings are aligned with the role of frontal lobe in decision making and attention [Burgess et al., 2009]. The topographical analysis extracted from the scalp above the sensorymotor cortex indicates more activity on the bipolar channels C3 and C4 (Figure 2.4). Electrode C4 represents the highest activation throughout the six curves which may cause by Motor execution phase of driving. Slightly higher activation in motor cortex at the presence of whole-body haptic feedback supports

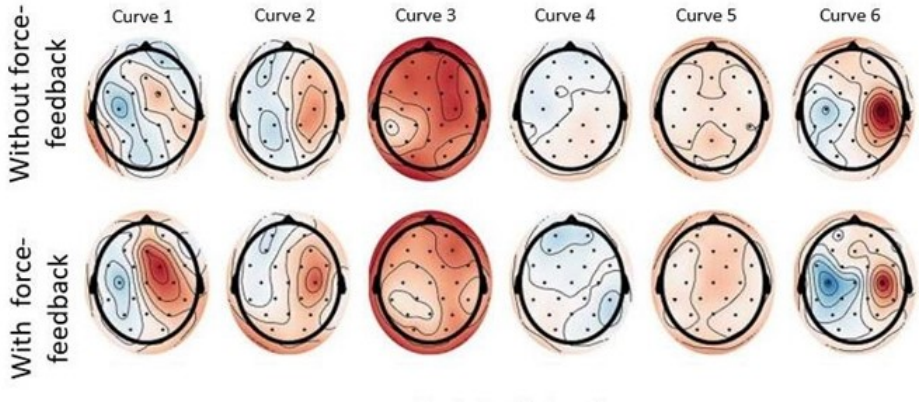

Figure 2.4: Topographical analysis of six simple curves. The first row represents the distribution of difference between baseline and scenarios at the absence of force feedback (the first row) and the at the presence of it (the second row)

an enhancement to drivers' engagement of required cognitive tasks of braking and steering control [Saha et al., 2017]. However, band frequency modulation based on ERP will be analyzed at the critical time intervals of curve negotiation. Our intent is to analyze the variability of frequency bands inside some temporal windows around 200 ms and 400 ms of latency.

2.4 Summary

The main differences between the two types of feedback found in this study is containing driver's visual responses. Fixed duration and pupil diameters found significantly higher while driving with haptic feedback in this preliminary study which could be due to the higher cognitive engagement. If it was the case, finding higher power in Theta band in frontal lobe is due to high vibration of system during haptic feedback activation and it is not relevant to the type of feedback. Therefore, the findings could be supported by the results that the high-fidelity driving simulator that can simulate various scenarios with high validity improvement of drivers' performance engages driver better [Groeger and Banks, 2007]. This preliminary study confirms

the possibility of EEG usage to alarm drivers properly within less than few seconds, once the system recognizes driver's cognitive stage and driving environment. We expect that the need for more number of channels for prediction of performance and drivers' cognitive state prior to hazard with other EEG measurements (e.g. ERP) would help us to develop a safer whole-body feedback to reduce cognitive workload and stress level of the driver, thereby enhance their control ability. In the future, we will design and analyze a haptic force feedback which could communicate with drivers through the seat and serves as an "intelligent messenger" that ensures human stays informed of the vehicle safety as well as driving environment which could play the role of "advance notification". In that regards, we will validate the preliminary findings with further analysis of power variation in each frequency within temporal duration as well as Event Related Potential (ERP). It could assist us to identify the perceptual operations of drivers on curves.

3 | Investigating Driver's Behavioral and Phsyiological Responses to Takeover Requests

Contemporary research on take over request has not fully transitioned from early stage, inclusive designs to those adhering to individualized levels of response. We found a paucity of research into transitions in conditionally automated vehicles, where drivers have different levels of situational awareness. Studies have shown that physiological measurements on individual drivers may provide better insight into the mental behavior and performance of each driver respectively. The aim of this preliminary study was to provide an initial step toward applying various physiological data sources in the limited take-over time budget, for two common take-over request (TOR) modalities used in conditionally automated vehicle.

3.1 Introduction

Conditionally automated vehicles (level 2 and 3 of automation) have been introduced with the aim of rapidly improving functionality, to the degree that highly automated driving will be introduced to the general public within the next few years [Cars, 2013]. Conditionally automated vehicles let drivers take their hands off the wheel and take their attention away from the primary task of driving. Currently, legal restraints dictate that a transition to driver take-over is made based on preset requirements or

limitations in the autonomous vehicle [SAE, 2014].In the transition, the automated system prompts the driver concerning the vehicle status and asks for a transition to manual driver control. This request for manual control is known as a take-over request (TOR). This requires drivers to take over from the automation system in a given, limited time budget (from the moment alarm goes off until a collision). However, shifting from an active controller role to that of a passive monitor causes drivers to stay out-of-the loop, which can in turn cause a loss of situation awareness [Endsley and Kiris, 1995] and driving skills in the long-term [De Winter et al., 2014]. This has been shown in recent investigations into deadly high-level automation accidents [Endsley, 2017, Banks et al., 2018], where the predominant cause of safety issues has been not providing adequate warnings to drivers to resume control [Griggs and Wakabayashi, 2018]. Hence, different types of hand-over strategies that actively monitor the most important human factors' constructs, which influence drivers' performances - such as drivers' situational awareness and mental workload [Paxion et al., 2014] - and keep drivers vigilant - in the loop - even when attention is on another task for a prolonged time, are missing. On the surface, any number of auditory and visual notifications might be adequate, but studies have shown that certain notification methods may in fact startle and stress the driver leaving the driver in a less capable state to make a life-saving decision as a result of affected situation awareness [Bliss and Acton, 2003]. In this study, we focus on the influence of workload, stress and the alarm type on takeover behavior with the help of physiological monitoring systems.

Autonomous vehicles have not been able to cope with all driving conditions, evidence by recent fatal crashes - in which autonomous systems failed to detect a pedestrian, poorly striped lane, or truck [Claybrook and Kildare, 2018, Levin and Woolf, 2016]. Some of these incidents could have been avoided with higher sensors sensitivity and by informing drivers about the abnormalities. As a result, despite the great advancement in the field of autonomous vehicles and the rapid growth in demand for them on the road, increasing the frequency of transitions from manual to autonomous and vice-versa could pose a significantly cognitively overwhelming experience for drivers. However, still most research in Human Factors has not profoundly considered human physiological aspects. On the one hand, considerable

research has focused on the time budget, conducted in a driving simulator [Damböck et al., 2013, Gold et al., 2013] and naturalistic settings [Eriksson et al., 2017]. On the other hand, mainly, the variation of TOR modality has been investigated [Melcher et al., 2015, Gold et al., 2017]. Therefore, there is still a need to develop a system to constantly consider a driver's physiological responses if one is to properly inform a driver.

Although there was a belief that automation could increase mental workload [Young and Stanton, 2002], a meta-analysis by [De Winter et al., 2014] showed controversial results. In fact, increasing the automation level reduced the mental workload. However, separate studies have shown the negative impact of driving with autonomous mode on mental workload, take-over performance, and reaction times [Strand et al., 2014, Zeeb et al., 2015, Bueno et al., 2016]. Physiological data has been used as an essential instrument for understanding and interpreting a driver's mental status. Applying neuropsychological and physiological measurements on drivers to investigate the relationship between mental behavior and performance while taking-over could provide us a profound understanding of what modalities provide useful TOR for autonomous vehicles. Whilst previously mentioned research has explored the best TOR modality, timing budget and driver behavior, relatively few studies have investigated the drivers' cognitive states at the time of transition. For the purpose of objectively obtaining the psychophysical state of the driver as accurate as possible, there is a need to shift from the simple questionnaires to a direct assessment of driver physiological responses and driver behavioral pattern. Therefore, this study has taken physiological data into account as the most reliable source of workload, stress and cognitive state analysis. Among all the physiological responses, we selected the following which produce more reliable measurements with a high temporal resolution necessary to detect vigilance difference in TOR: (1) electroencephalogram (EEG; electrical activity of the brain), (2) Eye-tracker, (3) photoplethysmography (PPG; electrical activity of the heart), and (4) Galvanic skin response (GSR; electrical activity of the skin). The primary purpose of the research in this paper is oriented to study psychophysical states of the driver by applying various physiological data streams to two common TOR modalities (visual-auditory

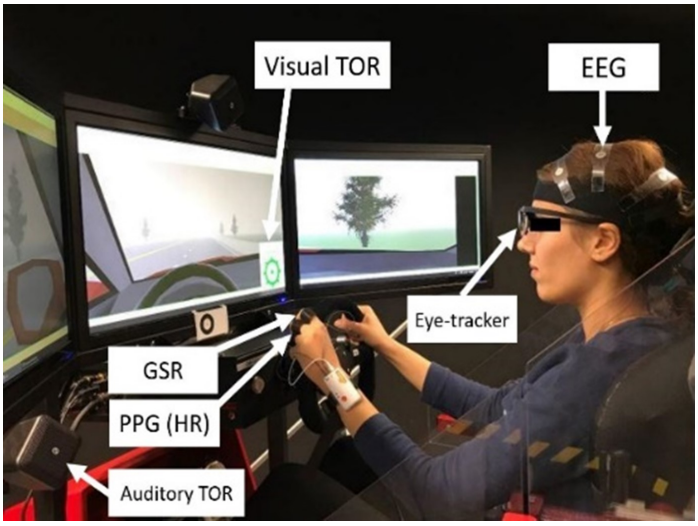

Figure 3.1: Experimental setup for recording EEG, HR, GSR and Eye movement.

and generic auditory), in the limited take-over time budget, investigating the modality cause more stress and workload for driver with a role of passive monitoring. This study provides elements of possible support for preventive of conditionally automated vehicles accidents.

3.2 Methodology

3.2.1 Experimental design

The experiment had a 2×2 repeated-measure factorial design with two within-subjects factors (warning modality and weather condition). Warning modality with two levels (visual-auditory and generic tones) and weather condition (rainy and sunny). It was hypothesized that visual-audio warning would promote proactive responses with improvement in physiological responses indicating greater cognitive engagement.

3.2.2 Participants

One graduate student (female, 29 yrs old with 5yrs of driving experience) participated in this preliminary study. However, a total of 35 participants ranging from 18-35 years old will be recruited - with the requirement that they have a valid driver's license, have at least a year of experience driving, normal vision without correction, and have no other health issues that may affect driving - will be recruited from the University of Virginia (in this preliminary study the results of this participant are reported). The Internal Review Board (IRB) at University of Virginia has approved the requirements and the study (IRB# 20606: Cognitive Trust in Human-Autonomous Vehicle Interaction).

3.2.3 Apparatus

Driving Simulator. The experiment was conducted in a high-fidelity driving simulator with the capability of 360° movements, and the ability to provide real-time feedback to the driver (401cr motion system by Force Dynamics, see Figure 3.1) The simulator was equipped with three 32" LCD screens with 1024×480 resolution, giving 120° horizontal field of view. The driving simulator was controlled by PreScan software. PreScan is widely used in many automotive OEMs and suppliers for concept studied, algorithm development to test advanced driver assistance systems (ADAS) and autonomous vehicles.

Eye-tracker. In highly automated driving, it is likely that drivers will engage in nondriving-related tasks which eye-tracker can let identifying whether the hazerdous objects in the visual frame were founded by drivers after receiving each of the TOR. It also helps to assess how long the driver focuses on those objects. One of the main components of eye-tracker is observation time which calculated as ratio between the fixation time and the time in which the object (e.g. pedestrian, truck, obstacle, etc.) appears in the visual frame. The smaller the ratio is, the less important (or not important at all) the object is to the driver or the driver failed to detect the object properly. Therefore, in order to capture the driver alertness to

capture the objects after reciving TOR, eye-tracker was use. A wearable pair of eye-tracker glasses (Tobii Pro-Glasses 2, Danderyd, Sweden; Tobii Pro-Glasses 2, 2017) with a sample rate of 60HZ were used. This device works wirelessly, which enabled us to capture exactly what the driver was attending to visually. Because the scenarios used consisted of many curves, and the driving simulator moves accordingly, the head-mounted eye tracker enabled us to measure the driver's gaze behavior accurately. In order to determine visual distraction, defined as not capturing the hazardous objects on the road, participants' gazes were manually coded with Tobii Pro Studio to the area of interest (AOI).

Electroencephalogram (EEG). In order to measure mental workload and engagement, EEG was recorded using a wireless B-Alert X24 system with 24 channels with the sample rate of 256 HZ. Wireless EEG signals were sent via Bluetooth to the data acquisition system. Also, in order to record the electoral activity of the brain, the sensor strip was placed according to the 10/20 extended standard and the channels were referenced using the mean of the mastoid processes. Analysis of event-related potentials (ERP) triggered by takeover alarm was conducted.

Heart Rate (HR). Fluctuation of heart rate in the time intervals between the nearby beats which occures as a result of emotional factors such as stress can be measured by variability of heart rate (HRV). In this study we focused on the time-domain indices of HRV by which we could quanify the amount of HRV after reciving TOR. These metrics used in this study include the standard deviation of normal beats (SDNN), root mean square of successive differences between normal heartbeats (RMSSD), number of adjacent NN intervals by more than 50 ms (NN50) and the proportion of NN50 to total number of NNs (pNN50).

Galvanic Skin Responses (GSR). Another physiological data considered in this study is the galvanic skin response (GSR), which indicates the conduction ability of the skin. As skin conductance is balanced by sweat secretion caused by sudomotor activity, any action accounting for muscle activation or automatic nervous system (ANS) like stress can can be objectified by GSR [Goshvarpour et al., 2017]. Therefore, in the case of emotional changes such as growing stress level, the magnitude

of the electrical resistance of the skin decreases, while the conductance increases. Along with HR parameters, GSR has been proven as one the most valid indicators of stress level. Thus, it was used in this study to indicate the stress level of drivers after receiving TOR. The skin conductance and Heart Rate were captured from the proximal phalanges of the index and the middle fingers of the non-dominant hand using Shimmer3 GSR.

3.2.4 Warning Modalities

In the visual-auditory conditions, the steering wheel color turned to one of three different colors(see Figure3.2). Green: when the system is on the autonomous mode and does not detect any hazardous object; red: as soon as the autonomous system detects a dangerous situation that might be out of the system limits, which alerts the participant to take over; and blue: if the participant has pressed both "on" buttons on the steering wheel, and switch to manual mode. Five sec after returning to a steady state and not detecting any dangerous situation, the visual steering wheel automatically turned to green to allow the participant to switch back to the autonomous mode. At the time of hazard detection, the auditory warning consisted of a single for the generic auditory warning, the sound matched that used for the visual-auditory warning, with the addition of a high frequency feedback tone (750Hz, duration: 75ms) presented at the time phase changing.

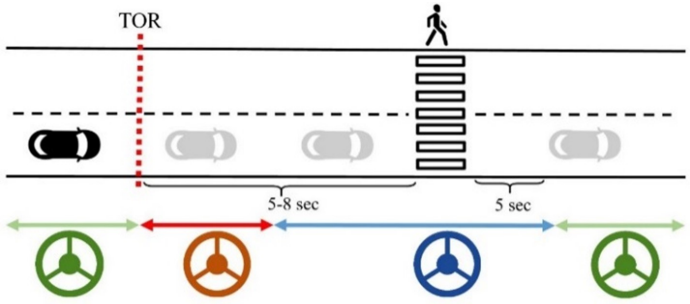

Figure 3.2: Experimental procedure for Visual takeover request (TOR)

3.2.5 Procedure

The driver was randomly assigned to the warning condition order. For both warning conditions, the participant was instructed on how and when to switch between the autonomous and manual modes. Participants were also told that warning condition indicated when the system detects a hazardous condition which exceeds its system limit. A 5-minute training session was carried out for the participant to become familiar with both the driving simulator and the driving task. Subsequently, the participant drove 16 experimental drives, which consisted of two weather conditions, divided into two blocks of eight drives. The participant was allowed a 5-minute break between each block. Each drive included a single weather condition and lasted for 3 minutes. Each trial composed of four hazardous objects or incidents, namely a pedestrian crossing the road, a cyclist and obstacle in the same lane as the vehicle, and a truck in the lane next to the vehicle. Within each trial, the incidents were randomly ordered. Therefore, as soon as the incident was detected and the system sent off the warning, the participant was given 5-8 seconds prior to the incident to take over [Eriksson and Stanton, 2017]. The participant was required to maneuver clear of the hazards by changing lanes or slowing down.

3.3 Data Processing

EEG

Data processing was performed using the EEGLAB and ERPLAB [Lopez-Calderon and Luck, 2014] toolboxes for MATLAB. First, a band pass filter (0.1 Hz – 30 Hz, 12dB) was applied to continuous data to account for linear trends in the data. Artifact removal was performed by eye for unusable data, and then independent component analysis was used to remove data from blinking and muscle movement. Data was then epoched into 1000 ms segments consisting of 200 ms before an alarm to 800 ms after an alarm for each TOR. Similar epochs were also created for alarm-free driving times. Further artifact rejection was performed using a moving window peak-to-peak

method, with a voltage threshold of 100 μV, a moving window width of 200 ms, and a window step of 50 ms. After this step, difference waves for visual-auditory TOR minus no alarm and auditory TOR minus no alarm were made. This was done to remove some of the noise in the signal that was generated simply from responding to the rich physical stimuli present in the simulator. For frontal, parietal, and central sites (F3, Fz, F4, C3, Cz, C4, P3, POz, P4), a negative peak amplitude (that is to say, the lowest point in a local voltage trough) within a window from 200 to 300 ms after stimulus onset was calculated to represent N2. The same was done for a positive peak amplitude (that is to say, the highest point in a local voltage peak) from 250 to 500 ms following stimulus onset to represent P3b. These ERP components were selected because increased amplitude in these components has been linked to better task engagement while driving [Lei et al., 2009]. For both measures, the 200 ms prior to stimulus onset was used as a baseline (to control for differences in voltage amplitude being based on an overall difference rather than a difference of reaction).

Eye-Tracker

Time spent-fixation, fixation duration, Time to First Fixation (TTFF-F) and pupil diameter were calculated using Tobii Pro Studio and iMotions. Analysis of AOI was performed to identify the moment participant's gaze is fixed on the hazardous objects.

Heart Rate (HR)

Heart rate data were transferred from Shimmer device to the data acquisition system (iMotions). Inter-beat interval (IBI) data were preserved at 128 Hz resolution as well as the equivalent beat per minute (bpm) heart rate values. For each TOR, HRV parameters were computed in the baseline point of time and 5sec interval while receiving warnings. Then, the collected data were fed into Kubios for processing of the HRV parameter. Kubios is a widely used software developed by the Biosignal Analysis and Medical Imaging Group of the University of Kuopio, Finland for analysis of HRV [Tarvainen et al., 2014]. This software allows the analysis of HRV and all the

heart rate time-domain analysis over discrete time.

Galvanic Skin Responses (GSR)

reprocessing GSR data were performed using MATLAB (The MathWorks, Natick, Massachusetts, USA). Data were filtered using a low-pass, third-order, zero-lag Butterworth type filter of 10-Hz cutoff frequency. Significant steering wheel events were detected when the derivative of steering wheel orientation signal was higher than the sum of mean signal plus two standard deviations (i.e.,sudden change of direction). Typical maximum values of EDR are 3s of latency between stimulus onset and EDR initiation, 3s between the EDR initiation and the variation peak, and 10 s between the variation peak and the point of recovery of 50% of the EDR amplitude [Dawson et al., 2007]. Thus, a 20s data window with 5s prestimulus and 10s poststimulus insured a data set large enough to perform computation of the amplitude of the event-related EDR.

3.4 Results

Performance before and after receiving warnings in the two weather conditions and the impact of the associated cognitive demand on heart rate, skin conductance, elecrocephlography signals, and driving performance were analyzed. Table 1 presents mean and standard deviation values for skin conductance and heart rate (commonly used HRV for each TOR modality). A two-way ANOVA (weather condition and warning modalities) with 0.5 level of significance was computed on each of the physiological data, and EEG was examined as an exploratory factor.

Eye-tracker

In order to obtain better insight about the eye movement that falls within the visual cue, areas of interest (AOI) were defined over the boundaries inside the display (see Figure 3.1). Analysis of paired t-test for the visual cue of the visual-auditory warning

	GSR	Heart Rate Variability (HRV)			
	Number of peaks	RMSSD	SDNN	PNN50	NN50
Weather					
Sunny	14.28 (0.45)	33.1(13.8)**	49.3(17.4)*	16.2(18.3)**	24.2(18.7)***
Rainy	15.6 (1.3)	23.2(9.8)	44.8(13.6)	8.3(10.8)	12.5(14.3)
Warning					
Visual-Audio	11.25 (2.88)**	28.6(14.5)	48.4(10.9)	13.0(15.7)	19.1(16.7)
Audio	9.18 (1.24)	29.4(10.1)	49.1(18.3)	13.85(11.2)*	20.3(15.3)

Table 3.1: standard deviation of normal beats (SDNN), root mean square of successive differences between normal heartbeats (RMSSD)) number of adjacent NN intervals by more than 50ms (NN50), the proportion of NN50 to total number of NNs (pNN50). $* = p < 0.05$, $** = p < 0.01$

on some of the eye-tracker features are as follow: (1) the time spent on the visual cue in two weather conditions, sunny (Mean= 895ms, SD= 491) and rainy (Mean= 1749 ms, SD=547) yielded a significant difference ($t = 3.29$, $df = 5$, $p = 0.001$), (2) the average fixation duration on sunny weather (Mean=177.0, SD=24.9) and rainy weather (Mean=421, SD=169) revealed a significant difference ($t = 2.94$, $df = 5$, $p = 0.021$).

EEG

Paired T-Tests were performed for the sites selected, comparing peak amplitude for P3b and N2. While there was no significant difference at N2 (mean difference = 0.04, $t = 0.13$, $df = 8$, $p = 0.548$), P3b was significantly greater across the electrodes selected for the visual-auditory TOR than for the auditory TOR (mean difference = 0.66, $t = 2.18$, $df = 8$, $p = 0.030$). Table 2 presents the mean values for each ERP component, and the results of the paired T-tests.

Figure. 3.4 displays the difference waves for visual-auditory TOR minus no alarm (in black) and auditory-only TOR minus no alarm (in red). P3b is indicated on the Figure 3.4 and Figure 3.5, and there is a clear pattern of higher voltage at this point across the electrodes of interest. Figure 3.5 also includes scalp maps for

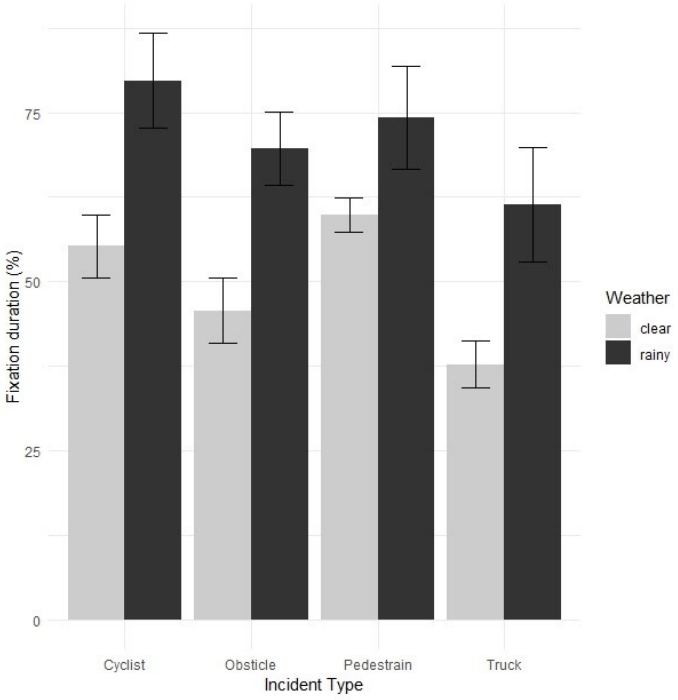

Figure 3.3: Effects of the visual-auditory Takeover-Request (TOR) on drivers' fixation duration w.r.t. the weather conditions

the difference waves at 256.706 ms (the average latency for the P3b peak). These show that the highest positive voltage activity across the scalp at these time points was centered on Fz and POz. These are also the sites with the greatest difference in activity between auditory-visual and auditory TOR. Across the scalp, visual-auditory TOR generated greater activity during at least one component of ERP related to better task engagement than auditory TOR. Further, these differences were greatest over the areas of the scalp with the most activity following TOR.

GSR

The number of peaks obtained from the GSR phasic data (frequency range:0.16HZ and above) as it linearly correlates to arousal which reflects both emotional and cognitive responses. The preliminary results are shown in Table 1.

TOR	Visual μV Mean (SD)	Audio μV Mean (SD)	Difference	t	df	p
P3b	3.58 (0.77)	2.92 (0.82)	0.66*	2.18	8.0	0.03
N2	1.31 (0.61)	1.27 (0.79)	0.04	0.13	8.0	0.548

Table 3.2: Mean ERP results across electrodes of interest

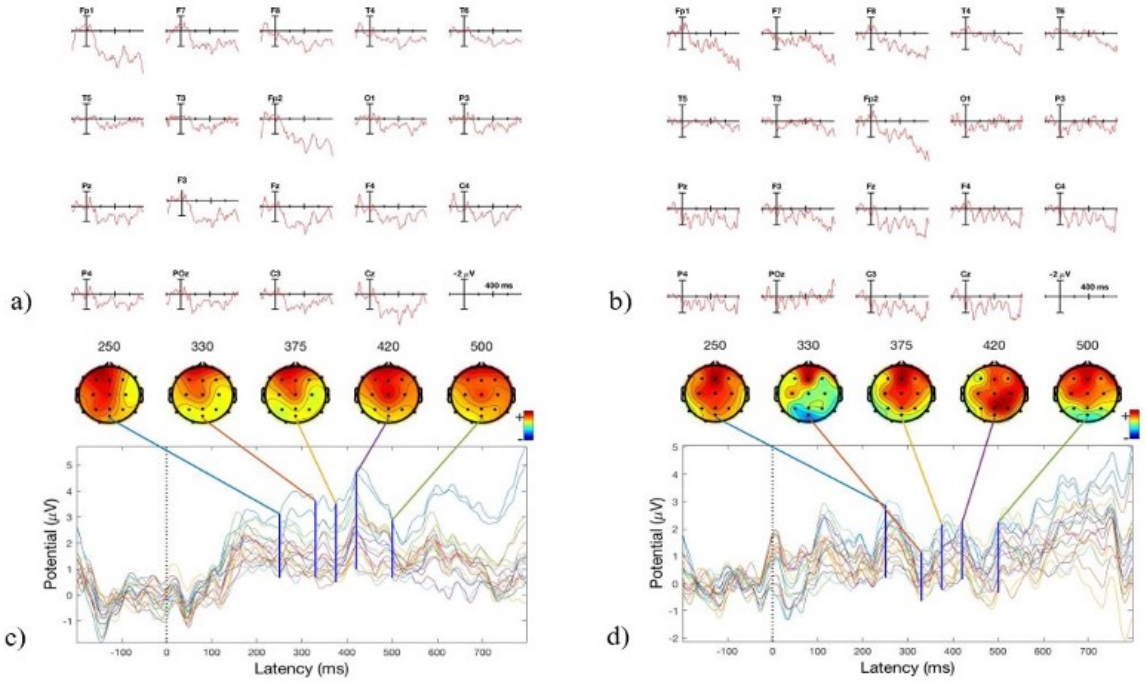

Figure 3.4: Results of ERP Analysis for twenty scalp's positions.

Heart Rate

Heart rate data were transferred from Shimmer device to the data acquisition system(iMotions) Inter-beat interval (IBI) data were preserved at 128 Hz resolution as well as the equivalent beat per minute (bpm) heart rate values. For each TOR, HRV parameters were computed in the baseline point of time and 5sec interval while receiving warnings.

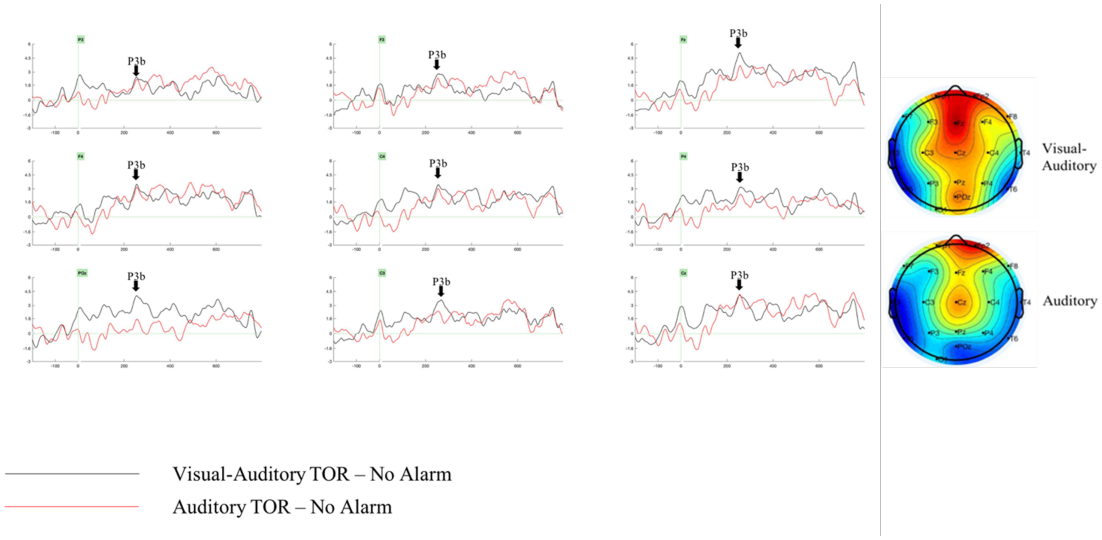

Figure 3.5: Results of ERP Analysis. Visual-audtiory TOR minus no alarm and Auditory TOR minus no alarm is shown for each electrode of interest, with the approximate latency of the P3b peak amplitude indicated for each. Scalp maps show the distribution of instantaneous amplitude at average P3b.

Reaction Time

In order to analyze the effect of TOR modalities on reaction times in the two weather conditions a two-way ANOVA was carried out. The reaction time was calculated as the time difference between warning sets and control switch. The analysis showed neither TOR modality ($F(1,15)=0.158$ $p = 0.699$) nor weather condition ($F(1,15)=0.81$, $p = 0.781$) had a significant effect on the reaction time of the participant. Fig.4 shows the reaction time on two different weather conditions for each modality.

3.5 Summary

This study examined a set of neurophysiological responses and driving performances over the warning cues of two different modalities, in a high-fidelity driving simulator with the random occurrences of roadway hazards under varying weather conditions.

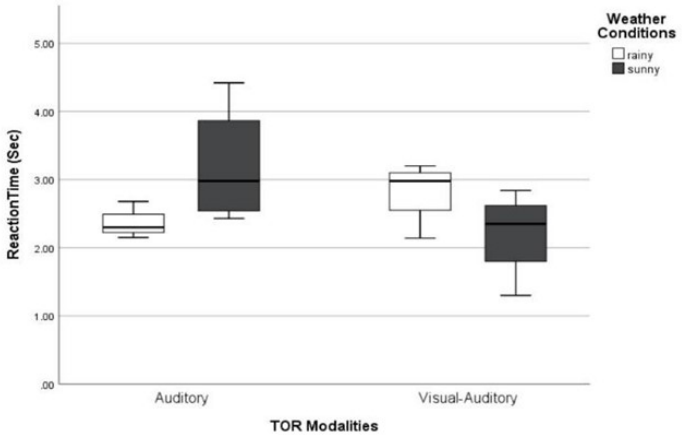

Figure 3.6: Reaction time of drivers after receiving two takeover modalities in rainy and sunny weather

It hypothesized color cues augmented with a single-tone hazard warning (i.e., visual-auditory modality) would be superior to auditory-only warning, possibly due to higher level of cognitive engagement and enhanced situation awareness. The results showed mixed results around this hypothesis; although no significant difference was observed in driving performance (mean reaction time), visual-auditory cues manifested enhanced physiological responses (in terms of GSR and HR-PNN50) as well as higher event-related potentials at close to 300 milli-seconds. This observation is not conclusive at this preliminary stage of data analysis. Despite no visible increase of driving performance, such modality-induced benefits in cognitive activation warrant further investigation. They can be exploit for the optimal design of warning in conditionally-automated vehicles.

III

4 | Predicting Visual Attention in Automated Vehicles

4.1 Background

Inspired by human visual attention, we propose a novel inverse reinforcement learning formulation using Maximum Entropy Deep Inverse Reinforcement Learning (MEDIRL) for predicting the visual attention of drivers in accident-prone situations. MEDIRL predicts fixation locations that lead to maximal rewards by learning a task-sensitive reward function from eye fixation patterns recorded from attentive drivers. Additionally, we introduce EyeCar, a new driver attention dataset in accident-prone situations. We conduct comprehensive experiments to evaluate our proposed model on three common benchmarks: (DR(eye)VE, BDD-A, DADA-2000), and our EyeCar dataset. Results indicate that MEDIRL outperforms existing models for predicting attention and achieves state-of-the-art performance. We present extensive ablation studies to provide more insights into different features of our proposed model

4.2 Introduction

Autonomous vehicles have witnessed significant advances in recent years. These vehicles promise better safety and freedom from the prolonged and monotonous task of driving. However, one of the remaining safety challenges of vision-based models integrated into these vehicles is how to quickly identify important visual cues

and understand risks involved in traffic environments at a time of urgency [Tawari and Kang, 2017]. Humans have an incredible visual attention ability to quickly detect the most relevant stimuli, to direct attention to potential hazards in complex situations [Pakdamanian et al., 2021], and to select only a relevant fraction of perceived information for more in-depth processing [Ungerleider and G, 2000]. Humans are able to guide their attention by a combination of bottom-up (*stimuli driven*, e.g., color and intensity) and top-down (*task driven*, e.g., current goals or intention) mechanisms [Deng et al., 2016, Katsuki and Constantinidis, 2014].

During *task-specific* activities, the *goal-directed* behavior of humans along with their underlying *target-based* selective attention, enables drivers to ignore objects and unnecessary details in their field of view that are irrelevant to their decisions [Chen et al., 2012, Chen et al., 2015]. For example, at one moment, a driver's goal might be to initiate an overtaking maneuver, thus a nearby vehicle becomes the target object. Later, the driver may need to stop abruptly to avoid an accident, thereby the brake light of the car in front becomes the target object. Despite recent progress in computer vision models for autonomous systems [Kim and Canny, 2017, Xu et al., 2017], they are still behind the foveal vision ability of humans [Ohn-Bar et al., 2020, Xia et al., 2020, Zelinsky et al., 2019].

Inverse reinforcement learning (IRL) algorithms are capable to address this problem by learning to imitate the efficient attention allocation produced by an expert, i.e., an attentive driver [Ng et al., 2000]. It is important that autonomous vehicles leverage human visual attention mechanisms to improve their performance, especially for better safety in critical situations where rare events can be encountered. In this paper, we introduce Maximum Entropy Deep Inverse Reinforcement Learning (MEDIRL) to learn *task-specific* visual attention policies to reliably predict attention in imminent rear-end collisions.

Prior efforts in bottom-up saliency models commonly prioritize pixel location (e.g., free-viewing fixation) [Kruthiventi et al., 2017, Pal et al., 2020, Stojić et al., 2020]. These models do not fully capture driver attention in goal-directed behavior [Einhäuser et al., 2020, Xia et al., 2020, Xia et al., 2020, Kummerer et al.,

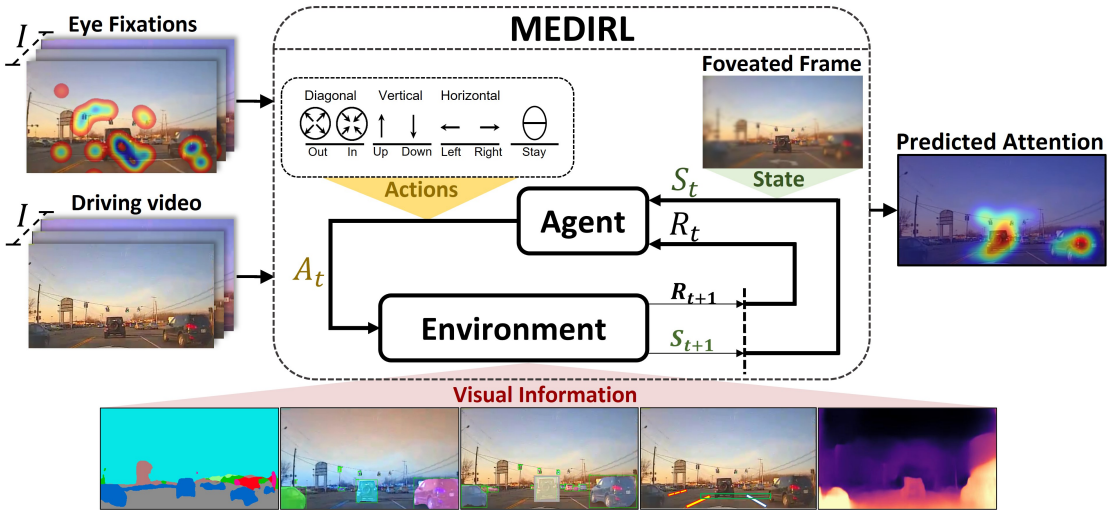

Figure 4.1: Given a driving video and corresponding eye fixations as inputs, MEDIRL learns to model the fixation selection as a sequence of states and actions (S_t, A_t). MEDIRL then predicts a maximally-rewarding fixation location by perceptually parsing a scene to extract rich visual information (environment) and accumulating a sequence of visual cues through fixations (state).

2017]. Moreover, video-based saliency models usually aggregate spatial features guided by saliency maps in each frame [Wang et al., 2018, Jiang et al., 2017, Hu et al., 2020, Yang et al., 2019]. However, most of these fixation prediction models utilized a particular source of information [Xia et al., 2020, Palazzi et al., 2018, Fang et al., 2019], and did not consider to jointly process spatial and temporal information [Wang et al., 2018, Hu et al., 2020]. In this work, we aim to predict eye fixation patterns made prior to critical situations, where these patterns can be either spatial (fixation map) or spatiotemporal (fixation sequences) features.

Inverse reinforcement learning (IRL) is an advanced form of imitation learning [Ziebart et al., 2008, Wulfmeier et al., 2015] that enables a learning agent to acquire skills from expert demonstrations [Tschiatschek et al., 2019]. Our proposed MEDIRL model learns a sequence of eye fixations by considering each fixation as a potential source of reward [Yang et al., 2020]. We leverage collective visual information that has been deemed relevant for video saliency in prior works [Min and Corso, 2019, Pal et al., 2020, Chen et al., 2021]. For example, if an autonomous

system tries to locate the salient regions of a driving scene before an imminent rear-end collision, the desired visual behavior can be demonstrated by studying the attention of a driver who effectively detects brake lights. In this way, the learning agent can infer a reward function explaining experts' behavior and optimize its own behavior accordingly. To this end, our proposed model predicts driver attention where a fixation pattern is represented as state-action pairs. Given a video frame input paired with eye fixations, MEDIRL predicts a maximally-rewarding fixation location (action) by perceptually parsing a scene to extract rich visual information (environment), and accumulating a sequence of visual cues through fixations (state) (see Figure 4.1).

Additionally, we introduce *EyeCar*, a new driver attention dataset in accident-prone situations. EyeCar is essential for training goal-directed attention models as it is the only dataset capturing attention before accidents in an environment with high traffic density. We exhaustively evaluate our proposed model on three common benchmarks (DR(eye)VE [Palazzi et al., 2018], BDD-A [Xia et al., 2018], DADA-2000 [Fang et al., 2019]) as well as our own EyeCar dataset. The experimental results show that MEDIRL outperforms state-of-the-art models on driver attention prediction. We also conduct extensive ablation studies to determine which input features are most important for driver attention prediction in critical situations. Our contributions can be summarized as follows:

- We propose MEDIRL, a novel IRL formulation for predicting driver visual attention in accident-prone situations. MEDIRL uses maximum entropy deep inverse reinforcement learning to predict maximally-rewarding fixation locations.

- We introduce EyeCar, a new driver attention dataset comprised of rear-end collisions videos for the goal-directed attention problem in critical driving situations.

- Extensive experimental evaluation on three driver attention benchmark datasets: DR(eye)VE [Palazzi et al., 2018], BDD-A [Xia et al., 2018], DADA-2000 [Fang et al., 2019], and EyeCar. Results show that MEDIRL

outperforms existing models for attention prediction and achieves state-of-the-art performance. Besides, we present ablation studies showing target (brake light), non-target (context), and driving tasks are important for predicting driver attention.

4.3 Related Work

Our work is broadly related to prior efforts on models for fixation prediction, using inverse reinforcement learning for visual tasks, and prior datasets for driving tasks.

Fixation Prediction. With increased access to large-scale annotated attention datasets and advanced data-driven machine learning techniques, prediction of human saliency has received significant interest in computer vision [Wang et al., 2019, Wang and Shen, 2017, Kruthiventi et al., 2017, Zhong et al., 2013, Cornia et al., 2018, Min and Corso, 2019]. A large number of previous studies explored bottom-up saliency models and visual search strategies over static stimuli [Fan et al., 2019, Li and Yu, 2015, Gong et al., 2015, Fu et al., 2015, Borji and Itti, 2015, Yun et al., 2013], and video [Zhong et al., 2013, Wang et al., 2015, Mathe and Sminchisescu, 2014, Min and Corso, 2019, Zahedian et al., 2019, Chen et al., 2021]. Generally, the output of these models is an attention map showing the probability of eye fixation distribution. In contrast to this approach, fewer works explored top-down attention models for explaining sequences of eye movements [Sprague and Ballard, 2004, Borji et al., 2012, Borji et al., 2010]. More recently, some works explored visual attention models in the context of driving [Guangyu Li et al., 2019, Xia et al., 2020, Gao et al., 2019]. Because task-specific instructions may change gaze distributions [Rothkopf et al., 2007], some models commonly detect salient regions of images or videos in a free-viewing task. Prior research also studied the pattern of eye movements associated with the task-specific activities [Mathe and Sminchisescu, 2014, Anderson et al., 2018]. Some of these works rely on the direct ties between eye movement and the demands of a task [Yang et al., 2020, Tatler et al., 2011, Sprague and Ballard, 2004]. These previously proposed attention models are trained mostly on static image-viewing

scenarios while human attention typically gets information in a sequential fashion. Further, recent video-saliency works have proposed joint bottom-up and top-down mechanisms for attention prediction using deep learning [Palazzi et al., 2018, Xia et al., 2018, Fang et al., 2019, Kim et al., 2020, Pal et al., 2020]. However, they did not consider to jointly process spatial and temporal information. We are interested in detecting the salient regions of a scene in a task-specific driving activity in which *estimating where the drivers are dynamically looking at*, and *reliably detecting the task-related objects (target objects)*.

Inverse Reinforcement learning. Our approach builds on works on modeling human visual attention with their fixation being a sequential decision process of the agent to detect salient regions [Mathe and Sminchisescu, 2013,Zelinsky et al., 2020,Liu et al., 2019]. The recently proposed work by Yang *et al.* [Yang et al., 2020] is the closest to our work as it proposes a model of visual attention in a visual search task of *static images*. We go further by addressing video saliency predictions in a dynamic and complex driving environment. Our model also does not require to predefine a set of targets but instead parses each driving video frame to extract rich scene context and candidate target objects. Next, it integrates visual cues with driver's eye fixations. It then recovers the intrinsic task-specific reward functions [Zheng et al., 2018] induced by visual attention allocation policies recorded from drivers in a driving environment. To do that, we propose to use maximum entropy deep IRL [Ziebart et al., 2008] which can handle raw image inputs and enables the model to handle the often sub-optimal and seemingly stochastic behaviors of drivers [Wulfmeier et al., 2015].

Driving Attention Datasets. Several driving behavior datasets have been proposed [Codevilla et al., 2019, Xu et al., 2017, Ramanishka et al., 2018]. However, only a few large-scale, publicly available, real-world video datasets with annotated visual attention exist in a driving context. DR(eye)VE [Palazzi et al., 2018] and BDD-A [Xia et al., 2018] are the most well-known large-scale annotated datasets in naturalistic and in-lab driving settings, respectively. Importantly, the recently-released annotated driving attention dataset with in-lab settings, DADA-2000 [Fang et al., 2019], is the only available dataset capturing scenes of collisions. This is because it is nearly impossible to collect enough driver attention data for collision

Dataset	collision	collision-POV	speed	GPS	#vehicles	#frames	#gaze
DR(eye)Ve	✗	✗	✓	✓	1.0	555k	8
BDD-A	✗	✗	✓	✗	4.4	318k	45
DADA-2000	✓	✗	✗	✗	2.1	658k	20
EyeCar	✓	✓	✓	✓	4.6	315k	20

Table 4.1: Compared to prior datasets, EyeCar is the only dataset that captured collisions from a point-of-view (POV) perspective and the host vehicle is involved in the collision. Previous datasets either did not capture attention from a collision point of view or had a less crowded scene.

or near-collision events. EyeCar further contributes to this area by having a more diverse array of driving events, beyond looking forward and lane-keeping. Unlike DADA-2000, EyeCar captures collisions from a collision point-of-view (POV) perspective (egocentric) where the ego-vehicle is involved in the accident. Table 4.1 compares EyeCar with similar datasets (more details in Sec. **??**).

4.4 Method

We propose MEDIRL for predicting drivers' visual attention in accident prone situations from driving videos paired with their eye fixations. MEDIRL learns a visual attention policy from demonstrated attention behavior. We formulate the problem as the learning of a policy function that models the eye fixations as a sequence of decisions made by an agent. Each fixation pattern is predicted given the present agent state and the current observed world configuration (i.e., a scene context).

4.4.1 Overview and Preliminaries

In this section, we introduce our notation and describe the features used in our proposed model.

Visual Information. During attention allocation in a dynamic and complex scene, relevant anchor objects–those with a spatial relationship to the target object–can guide attention to a faster reaction time, less scene coverage, and less time between fixating on the anchor and the target object [Võ et al., 2019, Helbing et al., 2020,

Beitner et al., 2021]. Therefore, we need to encode each frame of a given video to extract target and non-target features which an agent needs in order to effectively select the next fixation locations. Next, we describe in detail how this encoding is done (see Figure. 4.2). An overview of the visual encoder function is also outlined in Algorithm 1.

Given a family of driving video frame input, $I = \{I_t\}_{t=1}^{T}$, where T is the number of frames. We extract visual information in a discriminative way while keeping the relevant spatial information. Each frame has several fixation locations that are processed sequentially. At each step, we extracts visual features from the current input frame. To well represent a given video frame input to an agent, we consider both pixel- and instance-level representation (see Figure 4.1). The pixel-level representation determines the overall scene category by putting emphasis on understanding its global properties. The instance-level representation identifies the individual constituent parts of a whole scene as well as their interrelations on a more local instance-level.

For pixel-level representations, we extract features X_t from a given video frame (e.g., cars, trees). The feature extractor output is a tensor $X_t \in \mathcal{R}^{h \times w \times d}$, where h, w, and d are the height, width, and channel, respectively. At the instance-level, we represent the bounding box or instance-mask to reason explicitly over instances (e.g., lead-vehicle) rather than reasoning over all objects representation. We utilize a position-sensitive ROI average pooling layer [Yang et al., 2019] to extract region features Y_t for each box.

To extract features relevant to a driving task, we also consider the road lanes along with the lead vehicle features in our visual representation. The road lanes (G_t) are critical for the task-related visual attention of drivers since they are an important indicator of the type of maneuver [Do et al., 2017]. To amplify the predicted attention for pixels of the target objects, we detect the lead vehicle (M_t) which is important in rear-end collisions [Lyu et al., 2020]. The lead vehicle is a critical anchor object that can direct the driver attention to the target object, i.e. brake lights. We discretize each frame into an $n \times m$ grid where each patch matches the smallest (furthest)

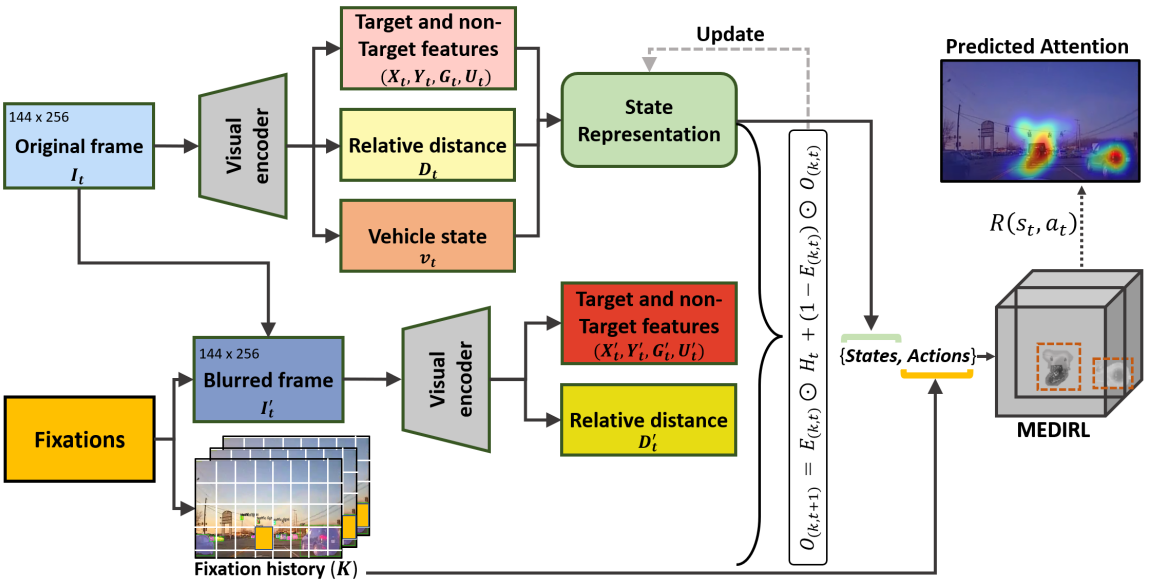

Figure 4.2: Overview of our state-representation. To simulate human fovea, the agent receives high-resolution information surrounding the attended location, and low-resolution information outside of this simulated fovea. At each fixation point, a new state is generated by applying Eq. 4.4.2.

size of the lead vehicle bounding box (see Figure. 4.2). In addition, we extract pixel locations of the brake lights by first converting each frame to the HSV color space, and then using a position-sensitive ROI max-pooling layer to extract region features for the lead vehicle box (U_t). The boxes and their respective features are treated as a set of objects.

Relative Distance. Drivers pay more attention to the objects which are relatively closer as opposed to those at a distance, since the chance of collision is significantly higher for the former case. Thus, relative distance between objects and the ego-vehicle is crucial for making optimal driving decisions [Pal et al., 2020]. To amplify nearby regions of a driving scene, we use dense depth map (D_t) and combine it with the general visual features (Y_t) by using the following formula:

$$Z_t = Y_t \oplus D_t = Y_t \odot \lambda * D_t + Y_t, \lambda = 1.2$$

where λ is an amplification factor

Driving Tasks. To discover which features of an observed environment are the

most driving task related, we need to determine the types (Q_t) of driving task. We observed three driving tasks ending to rear-end collisions across all videos: *lane-keeping*, *merging-in*, and *braking*. We use function f_{task} to define these driving tasks by two criteria: 1) ego-vehicle makes lane changing decision c and 2) the existence of a traffic signal I_{signal} in a given driving video.

$$\text{driving task} = \begin{cases} \text{lane-keeping}, & \text{if } c = 0,\ I_{signal} = 0 \\ \text{merging-in}, & \text{if } c = 1,\ I_{signal} = 0 \text{ or } 1 \\ \text{braking}, & \text{if } c = 0,\ I_{signal} = 1 \end{cases}$$

Vehicle State. We optionally concatenate the speed of the ego-vehicle v_t, which can influence the fixation selection [Yu et al., 2020, Palazzi et al., 2018, Pal et al., 2020], with the extracted visual representation, relative distance, and driving tasks.

4.4.2 MEDIRL

Attentive drivers predominantly attend to the task-related regions of the scene to filter out irrelevant information and ultimately make the optimal decisions. MEDIRL attempts to imitate this behavior by using the collective non-target and target features –extracted through parsing the driving scene– in the state representation. Subsequently, it integrates changes in the state representation with alterations in eye fixation point, to predict fixation. Therefore, the **state** of an agent is determined by a sequence of visual information that accumulates through fixations towards the target object (i.e., a brake light) which we call it a foveated frame, Figure 4.1 shows an example of a foveated frame. The **action** of an agent, the next fixation location, depends on the state at that time. The **goal** of an agent is to maximize internal **reward** by encapsulating the intended behavior of attentive drivers (experts) through changes in fixation locations. MEDIRL employs IRL to recover this reward function (R) from the set of demonstrations.

State Representation: MEDIRL considers the following components in the state representations: simulating the human visual system, collecting a context of spatial

cues, and modeling state dynamics. See Algorithm 1 for describing the overview of the state representation.

Human visual system (fovea): Human visual system accumulates information by attending to a specific location within the field of view. Consequently, humans selectively fixate on new locations to make optimal decisions. It means high-resolution visual information is available only at a central fixated location and the visual input outside of the attend location becomes progressively more blurred with distance away from the currently fixated location [Zelinsky et al., 2019]. We simulate human fovea by capturing high-resolution information about the current fixation location and a surrounding patch with a size 12×17 (about 1 visual angle), as well as low-resolution information outside of the simulated fovea [Zelinsky et al., 2019]. To effectively formulate this system, MEDIRL uses a local patch from the original frames of the video as the high-resolution foveal input and a blurred version of the frame to approximate low-resolution input L from peripheral vision [Zhang et al., 2018]. We obtain the blurred frames by applying a Gaussian smoothing with standard deviation $\sigma = 2 \times d$, which d is equal to Euclidean distance between the current fixation point $p_{k,t}$, where $k = 0, ..., \mathcal{K}$, and the size of the frame. Note that the number of fixations K varies from frame to frame.

Spatial cues: A driving task and the driving-relevant (anchor) objects of the scene can potentially direct drivers' attention to the primary target object. For example, drivers consider the distance to the lead vehicle when they brake. To approximate this guided selection of fixations, MEDIRL includes visual information in the state representation. This state representation collects the non-target and target features can create a context of spatial and temporal cues that might affect the selection of drivers' fixations.

Dynamics of state: To model the altering of the state representation followed by each fixation, we propose a dynamic state model. To begin with, the state is a low-resolution frame corresponding to peripheral visual input. After each fixation made by a driver, we update the state by replacing the portion of the low-resolution features with the corresponding high-resolution portion obtained at each new fixation

Algorithm 1 MEDIRL State Representation

function Visual Encoder(a video frame I)
  X := *HRnet*(I) ▷ global feature
  O := *mask-rcnn*(I) ▷ list of detected object
  Y := *ROI-average*(O, X) ▷ extract region features
  G, c := *VPG-net*(I) ▷ detect road lanes and lane changes
  M, I_{signal} := *mask-rcnn*(Y) ▷ detect lead-vehicle and traffic signal
  U := *ROI-max*(*HSV-color*(I),M) ▷ detect brake lights
  D := *MonoDepth2*(I) ▷ compute relative distance
  Z := Y ⊕ D ▷ amplify close objects
  Q := $f_{task}(c, I_{signal})$ ▷ compute driving task
  visual-cues = *concatenate*(G, M, U, Z) ▷ a context of spatial cues
  v := ego-vehicle speed ▷ vehicle state
  return *visual-cues, v, Q* ▷ return all extracted features
end function
function blur(frame I, fixation k)
  d = Euclidean(k,size(I))
  I' = *GaussianBlur*(I, σ), $\sigma = 2 \times d$ ▷ apply a Gaussian smoothing
  return *I'* ▷ return the low-resolution frame
end function
procedure State dynamics(frame I_t, fixations $\mathcal{K}$)
  for k ∈ K do **do**
    # collect context of spatial cues based on a simulated fovea movements
    H_t := VisualEncoder(I_t)
    $L_{k,t}$:= VisualEncoder(blur(I_t,k))
    # update the state that occurs following each fixation
    $O_{0,1} = L_{0,1}$ ▷ initialize frame corresponding to peripheral vision
    # $E_{k,t}$ is the circular mask generated from the k-th fixation
    $O_{k+1,t} = E_{k,t} \odot H_t + (1 - E_{k,t}) \odot O_{k,t}$
  end for
end procedure

location (see Figure. 4.2). At a given time step t, feature maps H for the original frame (high-resolution) and feature maps L for the blurred frame (low-resolution) are combined as follows:

$$O_{0,1} = L_{0,1}, O_{k+1,t} = E_{k,t} \odot H_t + (1 - E_{k,t}) \odot O_{k,t},$$

where $\odot$ is an element-wise product. $O_{k,t}$ is a context of spatial cues after k fixations. $E_{k,t}$ is the circular mask generated from the kth fixation (i.e., it is a binary map with 1 at current fixation location and 0 elsewhere in a discretize frame). To jointly aggregate all the temporal information, we update the next frame by considering all context of spatial cues in the previous frame as follows:

$$O_{k,t+1} = E_{k,t+1} \odot H_{t+1} + (1 - E_{k,t+1}) \odot O_{\mathcal{K},t},$$

where $O_{\mathcal{K},t}$ is visual information after all fixations $\mathcal{K}$ of time step t(previous frame).

Drivers have various visual behaviors while performing a driving tasks and many factors (e.g. speed) may affect the chosen fixation strategy [Yu et al., 2020, Palazzi et al., 2018, Pal et al., 2020]. To efficiently predict fixations for all drivers, we augment the state by aggregating it with a high-dimensional latent space that encodes the driving task Q_t. We then add another fully-connected layer to encode the current speed of the ego-vehicle v_t and concatenate the state with the speed vector. With the visual information and ego-vehicle state at each time step, we fuse all into a single state. The state of the agent is then complete in the sense that it contains all bottom-up, top-down, and historical information (more detail of these components can be found in the supplementary material).

Action Space: Herein we aim to predict the next eye fixation location of a driver. Therefore, the policy selects one out of $n * m$ patches in a given discretize frame. The center of the selected patch in the frame is a new fixation. Finally, the changes (Δ_x, Δ_y) of the current fixation and the selected fixation define the action space A_t: {left, right, up, down, focus-inward, focus-outward, stay}, as shown in Figure 4.1 which has three degrees of freedom (vertical, horizontal, diagonal).

Data	MethodTask	Merging-in			Lane-keeping			Braking		
		CC↑	s-AUC↑	KLD↓	CC↑	s-AUC↑	KLD↓	CC↑	s-AUC↑	KLD↓
DR(eye)VE	Multi-branch [Palazzi et al., 2018]	0.48	-	2.80	0.55	-	1.87	0.71	-	2.20
	HWS [Xia et al., 2018]	0.51	-	2.12	0.75	-	1.72	0.74	-	1.99
	SAM-ResNet [Cornia et al., 2018]	**0.78**	-	2.01	0.80	-	1.80	0.79	-	1.89
	SAM-VGG [Cornia et al., 2018]	**0.78**	-	2.05	0.82	-	1.84	0.80	-	1.81
	TASED-NET [Min and Corso, 2019]	0.68	-	1.89	0.73	-	1.71	0.70	-	1.89
	MEDIRL (ours)	**0.78**	-	**0.88**	**0.89**	-	**0.75**	**0.85**	-	**0.82**
BDD-A	Multi-branch [Palazzi et al., 2018]	0.58	0.51	2.08	0.75	0.72	2.00	0.69	0.77	2.04
	HWS [Xia et al., 2018]	0.53	0.59	1.95	0.67	0.89	1.52	0.69	0.81	1.59
	SAM-ResNet [Cornia et al., 2018]	0.74	0.61	2.00	0.89	0.79	1.83	0.85	0.88	1.89
	SAM-VGG [Cornia et al., 2018]	0.76	0.62	1.79	0.89	0.82	1.64	0.86	0.87	1.85
	TASED-NET [Min and Corso, 2019]	0.73	0.68	1.83	0.81	0.66	1.17	0.87	0.88	1.12
	MEDIRL (ours)	**0.82**	**0.79**	**0.91**	**0.94**	**0.91**	**0.85**	**0.93**	**0.92**	**0.89**
DADA-2000	Multi-branch [Palazzi et al., 2018]	0.44	0.53	3.65	0.69	0.54	2.85	0.67	0.64	2.91
	HWS [Xia et al., 2018]	0.49	0.59	3.02	0.72	0.53	2.65	0.69	0.77	2.80
	SAM-ResNet [Cornia et al., 2018]	0.65	0.61	2.39	0.78	0.64	2.32	0.75	0.81	2.34
	SAM-VGG [Cornia et al., 2018]	0.68	0.60	2.41	0.76	0.62	2.24	0.75	0.80	2.35
	TASED-NET [Min and Corso, 2019]	0.69	0.66	1.98	0.78	0.69	1.87	0.80	0.81	1.45
	MEDIRL (ours)	**0.70**	**0.68**	**1.31**	**0.89**	**0.71**	**0.92**	**0.81**	**0.88**	**0.99**

Table 4.2: Performance comparison of driver attention prediction on benchmarks.

Reward and Policy: To learn the reward function and policies of driver visual attention in rear-end collisions, we use a **maximum entropy** deep inverse reinforcement learning [Wulfmeier et al., 2015]. MEDIRL assumes the reward is a function of the state and the action, and this reward function can be jointly learned using the imitation policy.

The main goal of IRL is to recover the unknown reward function R from the set of demonstrations $\Xi = \{\xi_1, \xi_2, ..., \xi_q\}$, where $\xi_q = \{(s_1, a_1), ..., (s_\tau, a_\tau)\}$. We use maximum entropy deep IRL, which models trajectories as being distributed proportional to their exponentiated return:

$$p(\xi) = (1Z)exp(R(\xi)),$$

where Z is the partition function, $Z = \int_\xi exp(R(\xi))d\xi$. To approximate the reward function, we assume it can be represented as $R = \omega^T\phi$, where ω is a weight vector and ϕ is a feature vector. Such representation is constrained to be linear with respect to the input features ϕ. In order to learn a reward function with fewer constraints, we use deep learning techniques to determine $\Phi(\phi, \theta)$, a potentially higher dimensional feature space, and approximate the reward function as $R = \omega^T\Phi(\phi, \theta)(s, a)$. Note that the weight vectors of ω and the parameter vector θ are both associated with the network which is fine-tuned by jointly training the different category of driving tasks.

Loss Function: To learn the attention policies, MEDIRL maximizes the joint posterior distribution of fixation selection demonstrations Ξ, under a given reward structure and of the model parameter, θ. For a single frame and a given fixation sequence ξ with a length of $|\tau|$, the likelihood is:

$$\mathcal{L}_\theta = (1\Xi)\sum_{\xi^i \in \Xi} logP(\xi^i, \theta),$$

where $P(\xi^i, \theta)$ is the probability of the trajectory ξ^i in demonstration Ξ.

The algorithm tries to select a reward function that induces an attention policy with a maximum entropy distribution over all state-action trajectories and minimum empirical Kullback-Leibler divergence (KLD) from drivers state-action pairs. In each iteration (q) of maximum entropy deep IRL algorithm, we first evaluate the reward

value based on the state features and the current reward network parameters (θ_q). Then, we determine the current policy (π_q) based on the current approximation of reward (R_q), and transition matrix $\mathcal{T}$ (i.e., the outcome state-space of a taken action). We benefit from the maximum entropy paradigm, which enables the model to handle sub-optimal and stochastic visual behavior of drivers, by operating on the distribution over possible trajectories [Ziebart et al., 2008, Wulfmeier et al., 2015].

4.5 Dataset

Attentional lapses in normal situations (e.g., lane-following, empty road) do not cost the same as accident-prone situations (e.g., rear-end collision) where the cost of making an error is high. Nevertheless, collecting enough eye movements from drivers in accident-prone situations is nearly impossible because they are rather uncommon. In addition, driver attention data collected in-car has two main drawbacks [Xia et al., 2018, Xia et al., 2020]: 1) missing covert attention: eye-trackers can only record a single focus of drivers while a driver may be attending to multiple important objects, and 2) false positive gaze: drivers can be distracted to potential disturbances (e.g., side road advertisement) that are not relevant to the driving. Prior works [Xia et al., 2018, Xia et al., 2020] addressed these issues with in-lab data collection, collecting drivers' eye movements while performing simulated driving tasks.

Although in-lab driver attention collection is inevitably different from in-car driver attention, BDD-A in-lab experimental protocol showed that in-lab visual attention data reliably reveal where a driver should look at and identify the potential risks. Therefore, we follow their established and standardized experimental design protocol for collecting in-lab driver attention and create the EyeCar dataset exclusively for rear-end collisions. In order to incentivize users to pay attention and play the fall-back ready role in autonomous vehicles, we further modified the experimental design by sitting them in a low-fidelity driving simulator. The simulator consisting of a Logitech G29 steering wheel, accelerator, brake pedal, and eye-tracker (see supplementary materials for more details).

Data	Method\Task	Merging-in			Lane-keeping			Braking		
		CC↑	s-AUC↑	KLD↓	CC↑	s-AUC↑	KLD↓	CC↑	s-AUC↑	KLD↓
DR(eye)VE	Multi-branch [Palazzi et al., 2018]	0.36	0.37	6.46	0.51	0.49	4.80	0.69	0.49	3.38
	HWS [Xia et al., 2018]	0.38	0.34	4.38	0.71	0.51	4.44	0.72	0.61	3.30
	SAM-ResNet [Cornia et al., 2018]	0.49	0.48	4.29	0.73	0.55	3.90	0.74	0.66	3.27
	SAM-VGG [Cornia et al., 2018]	0.50	0.47	4.31	0.74	0.53	3.95	0.75	0.64	3.29
	TASED-NET [Min and Corso, 2019]	0.48	0.46	3.95	0.74	0.55	3.81	0.76	0.65	3.23
	MEDIRL (ours)	**0.51**	**0.51**	**2.32**	**0.76**	**0.57**	**3.11**	**0.79**	**0.69**	**3.07**
BDD-A	Multi-branch [Palazzi et al., 2018]	0.46	0.48	4.42	0.51	0.61	3.57	0.61	0.64	3.08
	HWS [Xia et al., 2018]	0.41	0.47	4.36	0.69	0.81	3.55	0.67	0.68	2.86
	SAM-ResNet [Cornia et al., 2018]	0.55	0.48	3.85	0.85	0.72	3.29	0.79	0.74	2.46
	SAM-VGG [Cornia et al., 2018]	0.53	**0.49**	3.92	0.84	0.70	3.22	0.77	0.70	2.49
	TASED-NET [Min and Corso, 2019]	0.55	**0.49**	3.78	0.84	0.71	3.12	0.77	0.76	2.47
	MEDIRL (ours)	**0.58**	**0.49**	**2.81**	**0.86**	**0.73**	**2.43**	**0.79**	**0.81**	**2.30**
DADA-2000	Multi-branch [Palazzi et al., 2018]	0.21	0.38	6.46	0.45	0.44	4.67	0.54	0.59	3.12
	HWS [Xia et al., 2018]	0.31	0.35	6.12	0.51	0.47	4.54	0.67	0.71	3.10
	SAM-ResNet [Cornia et al., 2018]	0.33	0.38	5.28	0.65	0.56	4.42	0.77	0.71	3.07
	SAM-VGG [Cornia et al., 2018]	0.30	0.39	5.35	0.69	0.57	4.31	0.74	0.69	3.10
	TASED-NET [Min and Corso, 2019]	0.32	0.38	4.76	0.68	0.57	3.99	0.73	0.74	3.01
	MEDIRL (ours)	**0.41**	**0.45**	**3.79**	**0.73**	**0.60**	**2.51**	**0.75**	**0.79**	**2.51**

Table 4.3: Performance comparison of driver attention prediction on EyeCar. The models trained on Dr(eye)VE, BDD-A, and DADA-2000 train sets and tested on EyeCar.

We recruited 20 participants (5 female and 15 male, ages 22-39) with at least three years of driving experience (Mean=9.7, SD=5.8). Participants watched all 21 selected dash-cam videos (each lasted approximately 30sec) to identify hazardous cues in rear-end collisions. The EyeCar dataset contains 3.5 hours of gaze behavior (aggregated and raw) captured from more than 315,000 rear-end collisions video frames. Each frame comprises 4.6 vehicles on average which makes EyeCar driving scenes more complex than other visual attention datasets (see Table 4.1). The extracted speed from each frame shows that 38% of vehicles were driving high ($65 \leq v$), 39% normal ($35 \leq v \leq 65$), and 23% low ($35 \geq v$). EyeCar also provides a rich set of annotations(e.g., scene tagging, object bounding, lane marking, etc.; more details in supplementary materials).

4.6 User Study

Training details. Driver attention is often strongly biased towards the vanishing point of the road and does not regularly change in a normal driving situation [Xia

et al., 2018,Pal et al., 2020]. However, attentive drivers regularly shift their attention from the center of the road to capture important cues in accident-prone situations. MEDIRL aims to predict driver attention in critical situations. Thus, to learn driving task-specific fixations and to avoid a strong center bias in our model two criteria were imposed when sampling training frames: 1) train on important frames, 2) exclude driving-irrelevant objects fixation sequence. Since a driver has to attend (fixate) to important visual cues which usually appear in critical situations, the important frames are defined as frames wherein the attention map greatly deviates from the average attention map. We use KLD to measure the difference between the attention over each video frame and the average attention map of the entire video. The average attention map of each frame is calculated by aggregating and smoothing the gaze patterns of all independent observers [Deng et al., 2019]. We then sample continuous sequences of six frames as the training frames where their KLD is at least 0.89. We also exclude fixation sequences with more than 40% focus on the irrelevant objects (e.g., trees, advertisement).

Datasets. We evaluate our model on three driver attention benchmark datasets: DR(eye)VE [Palazzi et al., 2018], BDD-A [Xia et al., 2018], DADA-2000 [Fang et al., 2019] and EyeCar. To predict driver attention related to rear-end collisions, we extract the full stopping events (resembling near-collisions) from DR(eye)VE and BDD-A, and rear-end collision events from DADA-2000. After applying the exclusion standard, we were left with 400, 1350, and 534 events in DR(eye)VE, BDD-A, and DADA-2000, respectively. Finally, within each type of driving task, we randomly split each of them into three sets of: 70% training, 10% validation, and 20% test.

4.6.1 Implementation Details

We resize each video frame input to 144×256. Then we normalize each frame by subtracting the global mean from the raw pixels and dividing by the global standard deviation. To encode visual information (see Sec. 4.4.2), we use several backbones: HRNetV2 [Wang et al., 2020]–pre-trained on Mapillary Vistas street-view scene [Neuhold et al., 2017], MaskTrack-RCNN [Yang et al., 2019]–pre-trained on

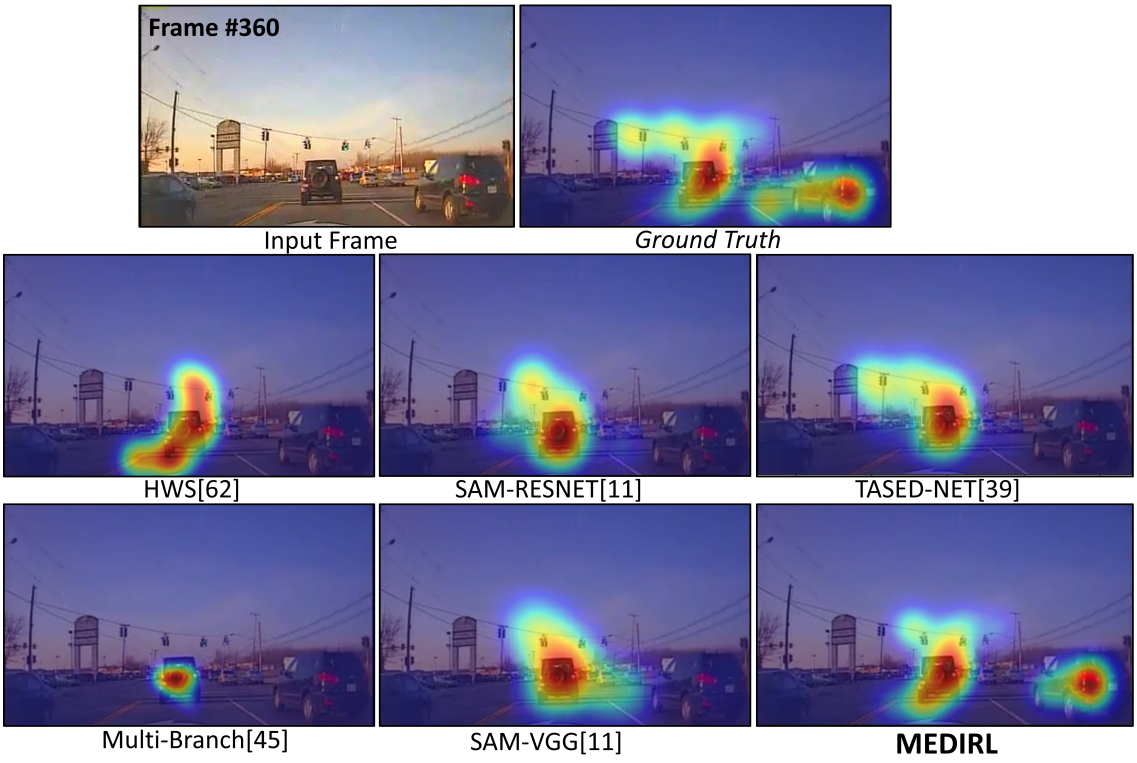

Figure 4.3: Predicted driver attention in a braking task for each compared model and MEDIRL. They all trained on BDD-A. MEDIRL can learn to detect most task-related salient stimuli (e.g., traffic light, brake light).

youtube-VIS, Monodepth2 [Godard et al., 2019]–pre-trained on KITTI 2015 [Geiger et al., 2012], and VPGNet [Lee et al., 2017]–pre-trained on VPGNet dataset.

MEDIRL consists of four hidden convolutional layers with 52, 34, 20, and 20 ReLu units, respectively; followed by seven softmax units to output a final probability map. We use batch normalization after ReLu activation and set the reward discount factor to 0.98. We also set the initial learning rate to 1.5×10^{-4}, and during the first 10 epochs, we linearly increase the learning rate to 5×10^{-4}. After epoch 11, we apply a learning rate decay strategy that multiplies the learning rate by 0.25 every three epochs. For training, we use Adam optimizer [Kingma and Ba, 2014] ($\beta_1 = .9, \beta_2 = .99$) and weight decay $= 0$. Overall, MEDIRL is trained on 36 epochs with a batch size of 20 sequences, and each sequence had six frames. The training time of MEDIRL is approximately 1.5 hours on a single NVIDIA Tesla V100 GPU and it

	[165mm]Ablated versions of MEDIRLDataset	EyeCar			BDD-A		
		CC↑	KLD↓	F_β ↑	CC↑	KLD↓	F_β ↑
1	global image + IRL	0.18	4.21	0.10	0.22	4.38	0.12
2	non target + IRL	0.19	4.15	0.12	0.20	4.29	0.12
3	target+non target + IRL	0.29	3.51	0.18	0.36	3.85	0.25
4	target+non target+distance + IRL	0.30	3.62	0.19	0.38	3.77	0.27
5	lead vehicle+lane + IRL	0.30	3.57	0.23	0.29	3.51	0.28
6	target+non target + lane+lead vehicle + IRL	0.36	3.53	0.21	0.41	3.47	0.32
7	target+non target+distance + lane+lead vehicle + IRL	0.33	3.43	0.26	0.35	3.07	0.34
8	target+non target+distance + lane+driving task + IRL	0.51	3.41	0.31	0.57	2.18	0.59
9	target+non target+distance + lead vehicle+driving task + IRL	0.66	2.91	0.49	0.73	1.07	0.66
10	target+non target+distance+lane+lead vehicle+driving task + IRL	0.70	2.78	0.60	0.87	**0.87**	0.75
11	**MEDIRL**: target+non target+distance+lane+lead vehicle+driving task + speed + IRL	**0.74**	**2.51**	**0.61**	**0.89**	**0.88**	**0.78**

Table 4.4: Quantitative evaluation of the ablated versions of MEDIRL and full MEDIRL. All models trained on BDD-A train set and tested on EyeCar and BDD-A test sets.

takes about 0.08 seconds to process each frame.

Evaluation Metrics. To evaluate attention prediction, we use location-based and distribution-based saliency metrics: KLD, shuffled Area under the ROC curve (s-AUC), and Correlation Coefficient (CC) [Bylinskii et al., 2018]. We report s-AUC since it penalizes models with more central prediction [Borji et al., 2012, Bylinskii et al., 2018, Gao et al., 2019].

4.7 Results

Table 4.2 provides the **quantitative evaluation** results of MEDIRL and five baseline attention prediction models including Multi-branch [Palazzi et al., 2018], HWS [Xia et al., 2018], SAM-ResNet [Cornia et al., 2018], SAM-VGG [Cornia et al., 2018], TASED-NET [Min and Corso, 2019]. For fair comparisons, we directly report available results released by the authors or reproduce experimental results via publicly available source codes. In this evaluation, we trained models on BDD-A and tested on each benchmark. The results highlight that MEDIRL surpasses almost all models under all evaluation metrics. Most significantly, our approach can effectively predict driver attention while performing various driving tasks. Although we are unable to calculate s-AUC for Dr(eye)VE as the original fixation were not reported, the results in Table 4.2 also indicates that the MEDIRL's superiority is not limited to a dataset.

Further, we evaluate MEDIRL along with other attention models on EyeCar dataset, reported in Table 4.3. In this experiment, we **trained models on each benchmark** (i.e., BDD-A, DR(eye)VE, DADA) and **tested on EyeCar**. MEDIRL performs favorably against other counterparts. However, there is a big performance gap between Table 4.2 and 4.3, which may indicate EyeCar has different distributions. To investigate this matter, we **trained models on EyeCar** and **tested on each benchmark**. We obtained the following results; $(CC:$ 0.89, $KLD:$ 0.80$)$, $(CC:$ 0.94, $s\text{-}AUC:$ 0.91, $KLD:$ 0.85$)$, $(CC:$ 0.85, $s\text{-}AUC:$ 0.77, $KLD:$ 0.99$)$ on DR(eye)VE, BDD-A, and DADA-2000, respectively, that are average values for all types of driving tasks. These results show the effectiveness of EyeCar on representing salient regions in

critical situations and also show that EyeCar attention distribution prior to accident-prone situations is more informative than benchmarks.

Figure 4.3 shows **qualitative comparison** of MEDIRL against other models. MEDIRL can reliably capture the important visual cues in a braking task in the case of a complex frame. In contrast, nearly all other models partially capture the spatial cues and predict attention mainly towards the center of the frame, thereby ignoring the target and non-target objects (i.e., spatial cues). Please refer to the supplementary material for more examples.

4.7.1 Ablations Studies

To investigate how different features in our model affect its performance, we compare several ablated versions of our model against two testing sets (i.e., EyeCar and BDD-A), using F_β ($\beta^2 = 1$ [Pal et al., 2020]), CC, and KLD. All ablated versions of our model are trained on BDD-A.

The results show that crucial features in the model include the context of spatial cues related to target and non-target (L3), driving-specific objects (Line 8, 10), followed by driving task (L9) features. MEDIRL without target (L2) and non-target (L5) shows a significant performance drop. From the results in Table 4.4, we can observe that compared with the ablated versions, our full model achieves better performance, which demonstrates the necessity of each feature in our proposed model.

4.8 Summary

We proposed MEDIRL, a novel inverse reinforcement learning formulation for predicting driver attention in accident-prone situations. MEDIRL effectively learns to model the fixation selection as a sequence of states and actions. MEDIRL predicts a maximally-rewarding fixation location by perceptually parsing a scene and accumulating a sequence of visual cues through fixations. To facilitate our study, we provide a new driver attention dataset comprised of rear-end collision

videos with richly annotated eye information. We investigate the effectiveness of attention prediction model by experimental evaluation on three benchmarks and EyeCar. Results show that MEDIRL outperforms existing models for attention prediction and achieves state-of-the-art performance.

5 | Predicting Drivers' Takeover Performance

Automated vehicles promise a future where drivers can engage in non-driving tasks without hands on the steering wheels for a prolonged period. Nevertheless, automated vehicles may still need to occasionally hand the control back to drivers due to technology limitations and legal requirements. While some systems determine the need for driver takeover using driver context and road condition to initiate a takeover request, studies show that the driver may not react to it. We present DeepTake, a novel deep neural network-based framework that predicts multiple aspects of takeover behavior to ensure that the driver is able to safely take over the control when engaged in non-driving tasks. Using features from vehicle data, driver biometrics, and subjective measurements, DeepTake predicts the driver's intention, time, and quality of takeover. We evaluate DeepTake performance using multiple evaluation metrics. Results show that DeepTake reliably predicts the takeover intention, time, and quality, with an accuracy of 96%, 93%, and 83%, respectively. Results also indicate that DeepTake outperforms previous state-of-the-art methods on predicting driver takeover time and quality. Our findings have implications for the algorithm development of driver monitoring and state detection.

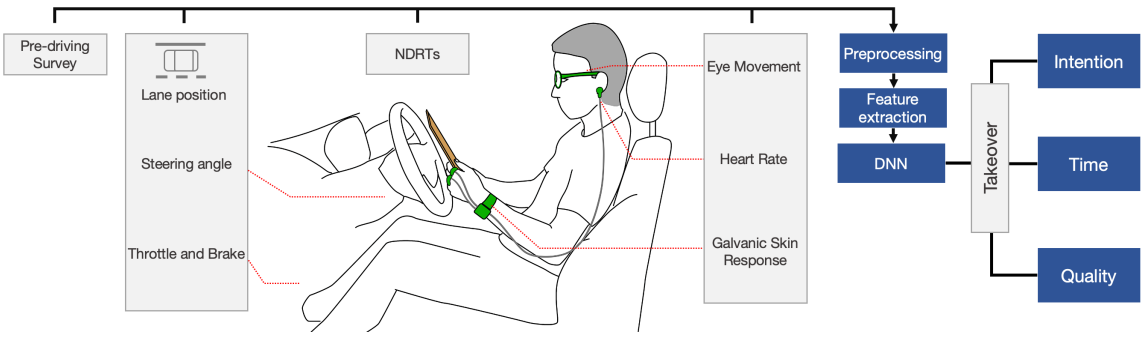

Figure 5.1: DeepTake uses data from multiple sources (pre-driving survey, vehicle data, non-driving related tasks (NDRTs) information, and driver biometrics) and feeds the preprocessed extracted features into deep neural network models for the prediction of takeover intention, time and quality.

5.1 Introduction

The rapid development of autonomous driving technologies promis-es a future where drivers can take their hands off the steering wheels and instead engage in non-driving related tasks (NDRTs) such as reading or using mobile devices. Incorporating cameras, sensors, global positioning systems (GPS), adaptive cruise control, light detection and ranging, and advanced driver assistance systems, automated vehicles (AVs) can navigate automatically. In Level 3 of autonomy (i.e., conditionally automated driving), as defined by the Society of Automotive Engineers (SAE international [SAE, 2018]), the driver does not need to continuously monitor the driving environment. Nevertheless, due to current technology limitations and legal restrictions, AVs may still need to handover the control back to drivers occasionally (e.g., under challenging driving conditions beyond the automated systems' capabilities) [McCall et al., 2019]. In such cases, AVs would initiate takeover requests (TORs) and alert drivers via auditory, visual, or vibrotactile modalities [Naujoks et al., 2014, Wan and Wu, 2018, Pakdamanian et al., 2018] so that the drivers can resume manual driving in a timely manner. However, there are challenges in making drivers safely take over control. Drivers may need a longer time to shift their attention back to driving in some situations, such as when they have been

involved in NDRTs for a prolonged time [Zeeb et al., 2017] or when they are stressed or tired [Feldhütter et al., 2018]. Even if TORs are initiated with enough time for a driver to react, it does not guarantee that the driver will safely take over [McDonald et al., 2019]. Besides, frequent alarms could startle and increase drivers' stress levels leading to detrimental user experience in AVs [Pakdamanian et al., 2020, Körber et al., 2018, Lee and Yang, 2020]. These challenges denote the need for AVs to constantly monitor and predict driver behavior and adapt the systems accordingly to ensure a safe takeover.

The vast majority of prior work on driver takeover behavior has focused on the empirical analysis of high-level relationships between the factors influencing takeover time and quality (e.g., [Mok et al., 2017, Zhang et al., 2019a, Du et al., 2020c, Ebnali et al., 2019]). More recently, the prediction of driver takeover behavior using machine learning approaches has been drawing increasing attention. However, only a few studies have focused on the prediction of either takeover time [Lotz and Weissenberger, 2018, Berghöfer et al., 2018] or takeover quality [Braunagel et al., 2017, Deo and Trivedi, 2019, Du et al., 2020b, Du et al., 2020d]; and their obtained accuracy results (ranging from 61% to 79%) are insufficient for the practical implementation of real-world applications. This is partly due to the fact that takeover prediction involves a wide variety of factors (e.g., drivers' cognitive and physical states, vehicle states, and the contextual environment) that could influence drivers' takeover behavior [Zeeb et al., 2015].

In this paper on the other hand, we present a novel approach, named **DeepTake**, to address these challenges by providing reliable predictions of multiple aspects of takeover behavior. **DeepTake** is a unified framework for the prediction of driver takeover behavior in three aspects: (1) *takeover intention* – whether the driver would respond to a TOR; (2) *takeover time* – how long it takes for the driver to resume manual driving after a TOR; and (3) *takeover quality* – the quality of driver intervention after resuming manual control. As illustrated in Figure 5.1, DeepTake considers multimodal data from various sources, including driver's pre-driving survey response (e.g., gender, baseline of cognitive workload and stress levels), vehicle data (e.g., lane position, steering wheel angle, throttle/brake pedal angles), engagement

in NDRTs, and driver biometrics (e.g., eye movement for detecting visual attention, heart rate and galvanic skin responses for the continuous monitoring of workload and stress levels). This data can easily be collected in AVs' driving environment. For instance, all of the driver biometrics utilized in DeepTake can be captured by wearable smartwatches and deployed eye-tracking systems. The multitude of sensing modalities and data sources offer complementary information for the accurate and highly reliable prediction of driver takeover behavior. DeepTake extracts meaningful features from the preprocessed multimodal data, and feeds them into deep neural network (DNN) models with mini-batch stochastic gradient descent. We built and trained different DNN models (which have the same input and hidden layers, but different output layers and weights) for the prediction of takeover behavior: intention, time and quality. We validate DeepTake framework feasibility using data collected from a driving simulator study. Finally, we evaluate the performance of our DNN-based framework with six machine learning-based models on prediction of driver takeover behavior. The results show that DeepTake models significantly outperform six machine learning-based models in all predictions of takeover intention, time and quality. Specifically, DeepTake achieves an accuracy of 96% for the binary classification of takeover intention, 93%, and 83% accuracy for multi-class classification of takeover time and quality, respectively. These accuracy results also outperform results reported in the existing work.

The main contribution of this work is the development of DeepTake framework that predicts driver takeover intention, time and quality using vehicle data, driver biometrics and subjective measurements[1]. The intersection between ubiquitous computing, sensing and emerging technologies offers promising avenues for DeepTake to integrate modalities into a novel human-centered framework to increase the robustness of drivers' takeover behavior prediction. We envision that DeepTake can be integrated into future AVs, such that the automated systems can make optimal decisions based on the predicted driver takeover behavior. For example, if the predicted takeover time exceeds the duration that the vehicle can detect situations

[1]DeepTake framework configurations, implementation details and code are available at https://github.com/erfpak7/DeepTake

requiring TORs, or the predicted takeover quality is too low to respond to TORs, the automated systems can warn the driver to engage less with the NDRT. In other words, DeepTake facilitates drivers to be distracted as long as they can properly respond and safely maneuver the vehicle. The reliable prediction of driver takeover behavior provided by DeepTake framework would not only improve the safety of AVs, but also improve drivers' user experience and productivity in AVs (e.g., drivers can focus on NDRTs without worrying about missing any TORs and potential tragic circumstances). We believe that our work makes a step towards enabling NDRTs in automated driving, and helps HCI researchers and designers to create user interfaces and systems for AVs that adapt to the drivers' context.

5.2 Related work

We discuss prior work on the analysis of takeover time and quality, and position our work in the context of state-of-the-art takeover behavior prediction research.

5.2.1 Takeover time

In this paper, we consider the *takeover time* as the period of time from the initiation of TOR to the exact moment of the driver resuming manual control (see Figure 5.3), following the ISO standard definition in [ISO 21959:2020, 2020]. Note that the same concept has also sometimes been named as takeover reaction time or response time in the literature (e.g., [Johns et al., 2016, Kim and Yang, 2017, Petermeijer et al., 2017a, Eriksson and Stanton, 2017]). The empirical literature defines a large variety of takeover time from a mean of 0.87s to brake [Winter et al., 2016], to an average of 19.8s to response to a countdown TOR [Politis et al., 2018] and 40s to stabilize the vehicle [Merat et al., 2014]. This range is derived from influential factors impacting perception, cognitive processing, decision-making and resuming readiness [Gold et al., 2016, Zeeb et al., 2015]. A meta-analysis of 129 studies by Zhang et al. [Zhang et al., 2019a] found that a shorter takeover time is associated with the following factors: a higher urgency of the driving situation, the driver not performing a non-

driving related task (NDRT) such as using a handheld device, the driver receiving an auditory or vibrotactile TOR rather than no TOR or a visual-only TOR. Recent studies by Mok et al. [Mok et al., 2017] and Eriksson et al. [Eriksson and Stanton, 2017] both confirmed that drivers occupied by NDRTs have higher responses to TORs. Similarly, [Feldhütter et al., 2017] found a significant increase in reaction time induced by NDRTs. It is further concluded that the visual distraction causes higher reaction time when it is loaded with cognitive tasks [Tang et al., 2020]. Studies have also revealed several driving environments, TOR modalities [van der Heiden et al., 2017, Tang et al., 2020], driving expectancy [Ruscio et al., 2015], age [Walch et al., 2017] and gender [Warshawsky-Livne and Shinar, 2002] associated with takeover time. The present study extend previous findings by considering various NDRTs, gender, and objective and subjective measurements of mental workload into the DeepTake framework.

5.2.2 Takeover quality

In addition to takeover time, it is essential to assess the *takeover quality*, which is defined as the quality of driver intervention after resuming manual control [ISO 21959:2020, 2020]. There are a variety of takeover quality measures, depending on different takeover situations (e.g., collision avoidance, lane-keeping), including objective measures (e.g., mean lateral position deviation, steering wheel angle deviation, metrics of distance to other vehicles or objects, minimum time to collision, frequency of emergency braking) and subjective measures (e.g., expert-based assessment, self-reported experience). Prior work has found that takeover quality can be influenced by factors such as drivers' cognitive load [Du et al., 2020a, Zeeb et al., 2016], emotions and trust [Dillen et al., 2020, Du et al., 2020c, Hergeth et al., 2017], and distraction of secondary NDRTs [Martelaro et al., 2019, Dogan et al., 2019]. Takeover time to an obstacle [Zeeb et al., 2016] has been used widely studies as an indicator of takeover performance [Eriksson and Stanton, 2017]. However, a study by Louw et al. [Louw et al., 2017] showed that takeover time and quality appear to be independent. This lack of consensus could be due to the fact that studies apply

various time budget for takeover control.

5.2.3 Takeover prediction

While existing literature mostly focus on the empirical analysis of drivers' takeover time and quality, there are a few recent efforts on the predication of drivers' takeover behavior using machine learning (ML) approaches. Lotz and Weissenberger [Lotz and Weissenberger, 2018] applied a linear support vector machine (SVM) method to classify takeover time with four classes, using driver data collected with a remote eye-tracker and body posture camera; the results achieve an accuracy of 61%. Braunagel et al. [Braunagel et al., 2017] developed an automated system that can classify the driver's takeover readiness into two levels of low and high (labeled by objective driving parameters related to the takeover quality); their best results reached an overall accuracy of 79% based on a linear SVM classifier, using features including the traffic situation complexity, the driver's gazes on the road and NDRT involvement. Deo and Trivedi [Deo and Trivedi, 2019] proposed a Long Short Term Memory (LSTM) model for continuous estimation of the driver's takeover readiness index (defined by subjective ratings of human observers viewing the feed from in-vehicle vision sensors), using features representing the driver's states (e.g., gaze, hand, pose, foot activity); their best results achieve a mean absolute error (MAE) of 0.449 on a 5 point scale of the takeover readiness index. Du et al. [Du et al., 2020b, Du et al., 2020d] developed random forest models for classifying drivers' takeover quality into two categories of good and bad (given by subjective self-reported ratings), using drivers' physiological data and environment parameters; their best model achieves an accuracy of 70%.

In summary, the existing works only focus on the prediction of either takeover time or takeover quality. By contrast, DeepTake provides a unified framework for the prediction of all three aspects of takeover behavior: intention, time and quality together. Furthermore, DeepTake achieves better accuracy results: 96% for takeover intention (binary classification), 93% for takeover time (three classes), and 83% for takeover quality (three classes).

5.3 DeepTake: A New Approach for Takeover Behavior Prediction

In this section, we present a novel deep neural network (DNN)-based approach, DeepTake, for the prediction of a driver's takeover behavior (i.e., intention, time, quality). Figure 5.1 illustrates an overvi-ew of DeepTake. First, we collect multimodal data such as driver biometrics, pre-driving survey, types of engagement in non-driving related tasks (NDRTs), and vehicle data. The multitude of sensing modalities and data streams offers various and complementary means to collect data that will help to obtain a more accurate and robust prediction of drivers' takeover behavior. Second, the collected multimodal data are preprocessed followed by segmentation and feature extraction. The extracted features are then labeled based on the belonging takeover behavior class. In our framework, we define each aspect of takeover behavior as a classification problem (i.e., takeover intention as a binary classes whereas takeover time and quality as three multi-classes). Finally, we build DNN-based predictive models for each aspect of takeover behavior. DeepTake takeover predictions can potentially enable the vehicle autonomy to adjust the timely initiation of TORs to match drivers' needs and ultimately improve safety. We describe the details of each step as follows.

5.3.1 Multimodal Data Sources

Driver Biometrics

The prevalence of wearable devices has made it easy to collect various biometrics for measuring drivers' cognitive and physiological states. Specifically, we consider the following three types of driver biometrics in DeepTake.

Eye movement. Drivers are likely to engage in non-driving tasks when the vehicle is in the automated driving mode [Borojeni et al., 2018, Wintersberger et al., 2018, Pakdamanian et al., 2020]. Therefore, it is important to assess the drivers' visual attention and takeover readiness before the initiation of TORs. There is a

proven high correlation between a driver's visual attention and eye movement [Zeeb et al., 2015, Wu et al., 2019, Alsaid et al., 2019]. DeepTake uses eye movement data (e.g., gaze position, fixation duration on areas of interest) measured by eye-tracker devices. We utilize a pair of eye-tracking glasses in our user study (see Section 6.3). But the aforementioned eye movement data can be captured with any eye-tracking device.

Heart rate. Studies have found that *heart rate variability* (HRV), fluctuation of heart rate in the time intervals between the nearby beats, is a key factor associated with drivers' workload [Paxion et al., 2014], stre-ss [Dillen et al., 2020], and drowsiness [Vicente et al., 2011]. DeepTake uses features extracted from HRV analysis for monitoring drivers' situational awareness and readiness to respond to TORs. Heart rate can be measured in many different ways, such as checking the pulse or monitoring physiological signals. DeepTake employes photoplethysmographic (PPG) signal, which can be collected continuously via PPG sensors commonly embedded in smartwatches. PPG sensors monitor heart rate by the emission of infrared light into the body and measure the reflection back to estimate the blood flow. Unlike some heart rate monitoring devices that rely on the placement of metal electrodes on the chest, PPG sensors provide accurate heart rate measures without requiring intrusive body contact. Therefore, a PPG signal is preferred for monitoring drivers' heart rate.

Galvanic skin response (GSR). Along with HRV, GSR has been identified as another significant indicator of drivers' stress and workload [Dillen et al., 2020, Foy and Chapman, 2018, Mehler et al., 2012, Radlmayr et al., 2014]. A GSR signal measures the skin conduction ability. Drivers' emotional arousal (e.g., stress) can trigger sweating on the hand, which can be detected through distinctive GSR patterns. DeepTake incorporates features extracted from the GSR signal for monitoring drivers' stress levels. GSR sensors are also embedded in many wearable devices, including smartwatches.

Pre-Driving Survey

In addition to the objective measurements of driver biometrics, DeepTake exploits subjective pre-driving survey responses, because drivers' prior experience and background may influence their takeover behavior [Zhang et al., 2019a]. However, any subjective rating of factors affecting a driver's cognitive and physical ability as well as driving experience prepare a complete specification of objective metrics, potentially enhancing the distinctive attributes of an algorithm. DeepTake framework exerts demographic information, NASA-Task Load Index (NASA-TLX) [Hart and Staveland, 1988], and the 10-item Perceived Stress Scale (PSS-10) [Cohen et al., 1983] to measure drivers' perceived workload and psychological stress. In our user study (see Section 6.3), we asked participants to fill in questionnaires at the beginning of each trial.

Non-Driving Related Tasks (NDRTs)

As described in Section 6.2, prior studies have found that engaging in NDRTs can undermine drivers' takeover performance. Diverse NDRTs require different levels of visual, cognitive and physical demands; thus, the influence varies when drivers are asked to interrupt the secondary task and resume manual control of the vehicle. DeepTake accounts for the impact of different NDRTs on the prediction of drivers' takeover behavior. In our user study, we considered four NDRTs in which drivers are very likely to engage in automated vehicles: (1) *having a conversation with passengers*, (2) *using a cellphone*, (3) *reading*, and (4) *solving problems* such as simple arithmetic questions (more details in Section 5.4.3). We chose these NDRTs because they are commonly used in driving studies [Gerber et al., 2020, Dogan et al., 2019], and they follow the framework of difficulty levels in the flow theory [Csikszentmihalyi and Csikzentmihaly, 1990]. We further designed reading and arithmetic problem solving with two difficulty levels (easy and medium adapted from [Nourbakhsh et al., 2012], which reported a strong correlation between the questions and the physiological responses). Nevertheless, DeepTake framework can be easily adjusted to any NDRTs.

Vehicle Data

DeepTake also considers a wide range of data streams captured from the automated vehicles, including lane position, distance to hazards, angles of the steering wheel, throttle and brake pedal angles, and the vehicle velocity. Such vehicle data can help to determine the driving condition, the urgency of a takeover situation, and the impact of drivers' takeover behavior.

5.3.2 Data Preparation

Feature Extraction and Multimodal Data Fusion

The goal of DeepTake is to provide a procedure to reliably predict drivers' takeover behavior (i.e., intention, time and quality) before a TOR initiation. Hence, the taken procedure for data preparation depends on the driving setting, collected data and the context. Herein, we incorporate data of drivers' objective and subjective measurements, as well as vehicle dynamic data. We initially apply data preprocessing techniques including outliers elimination, missing value imputation using mean substitutions, and smoothing to reduce artifacts presented in raw data. It is worth mentioning that we exclude any data stream providing insights about the unknown future (e.g., type of alarm) or containing more than 50% missing value. The preprocessed time series data are then segmented into 10-second fixed time windows *prior to the occurrences of TORs*. In other words, if TOR happened at time t, we only used data captured in the fixed time window of [t-10s, t] and did not include any data later than t. We started with time window values of 2s and 18s, suggested in the literature [Du et al., 2020d, Braunagel et al., 2017, Zhang et al., 2019a], and experimentally settled on 10s, as real-world applications require a shorter time window with better prediction. We then aggregated the values of all multimodal data over this time interval, resulting in 256 (max sampling rate)$\times 10sec = 2560$ observations *per takeover* event. However, depending on specific applications and contextual requirements, the selected time window length could vary. Subsequently, the segmented windows from modalities are processed to extract meaningful features

Table 5.1: List of extracted features used in DeepTake

Data Source	Feature	Type	Values
Eye movement	Gaze position	float	(1920×1080)
	Pupil size	float	(0-7)
	Time to first fixation	int	(1-90)
	Fixation duration	float	(100-1500ms)
	Fixation sequence	int	(1-2500)
Heart rate (PPG signal)	SDNN	float	(45-75ms)
	RMSSD	float	(25-43ms)
	pNN50	float	(18-28%)
GSR signal	Number of peaks	int	(1-6)
	Amplitude of peaks	float	(0.01- 1.58µs)
Pre-driving survey	Gender	binary	(M-W)
	NASA-TLX	categorical	(1-21)
	PSS-10	categorical	(0-4)
Secondary tasks	NDRTs	categorical	(C,U,R,S)[1]
Vehicle data	Right lane distance	float	(0.73-2.4m)
	Left lane distance	float	(1.02-2.8m)
	Distance to hazard	float	(98-131m)
	Steering wheel angle	float	(-180-114°)
	Throttle pedal angle	float	(15-21°)
	Brake pedal angle	float	(0-17°)
	Velocity	float	(0-55mph)

1: C; Conversation, U; Using cellphone, R; Reading articles on tablet, and S: Solving arithmetic questions

describing the attributes impacting takeover behavior.

For the eye movement, we acquire interpolated features extracted from raw data through iMotion software [iMotions, 2015a]. The extracted eye movement attributes include gaze position, pupil diameters of each eye, time to first fixation, and fixation duration/sequence on the detected area of interest (i.e., cellphone, tablet and monitor).

To compute the heart rate features, we first apply a min-max normalization on the raw PPG signal, and then filter the normalized PPG signal by applying a 2order Butterworth high pass filter with a cut-off of 0.5Hz followed by a 1order Butterworth low pass filter with a cut-off frequency of 6Hz. We use an open-source

toolkit HeartPy [van Gent et al., 2019] to filter the PPG signals and extract the following features from heart rate variability (HRV) analysis: the standard deviation of normal beats (SDNN), root mean square of successive differences between normal heartbeats (RMSSD), and the proportion of pairs of successive beats that differ by more than 50ms (pNN50). These metrics are to correlate with driver's cognitive workload and stress [Peruzzini et al., 2019].

Furthermore, we obtain two common and important GSR features: the number and amplitude of peaks [Manawadu et al., 2018, Nourbakhsh et al., 2012]. A peak occurs when there is a quick burst of raised conductance level. The peak amplitude measures how far above the baseline the peak occurred. Thus, peaks are valuable indicator of stress and mental workload.

While the variety of a driver's subjective and objective measurements along with vehicle dynamic data provide complementary information to draw better insights into drivers' takeover behavior, we need to finally fuse these multimodal data into a joint representation as input to the DNN model. Beforehand, however, we employ the Z-score normalization for most of the features except extracted PPG features to accentuate key data and binding relationships within the same range. To normalize the features associated with PPG, we use the min-max normalization, as explained above. For any remaining features still containing missing values, their missing values are imputed by using their means. Table 5.1 summarizes the list of data sources and extracted features used in DeepTake. Finally, the generated features from each modality concatenated to create a rich vector representing driver takeover attributes. The joint representations of all feature vectors with the provision of their associated labels are eventually fed into DNN models for training. Below, the labeling procedure of these feature vectors is explained.

Data Labeling

The target labels greatly depend on the context in which the labels are presented. Herein, we define the ground truth labeling for an attribute set denoting the feature vector. Each label indicates the classification outcome of takeover intention, time, and

quality that is more representative of our user study and the three takeover behavior aspects.

Takeover intention. DeepTake classifies a driver's takeover intention into the binary outcomes, indicating whether or not the driver would resume manual control of the vehicle. In our user study, if a participant initiated the takeover action by pressing the two buttons mounted on the steering wheel (see Figure 5.2) upon receiving a TOR, we label the feature vector as "TK", showing the takeover intention; if no takeover action was initiated between the moment of TOR initiation and the incident (e.g., obstacle avoidance), we use a "NTK" label displaying the absence of intention.

Takeover time. Recall from Section 6.2 that takeover time is defined as the time period between a TOR and the exact moment of a driver resuming manual control. Prior works have considered the starting time of manual control as the first contact with the steering wheel/pedals [Zeeb et al., 2015] or the takeover buttons [Kim and Yang, 2017]. In our user study, we timed the takeover moment once a participant pressed the two takeover buttons on the steering wheel simultaneously (see Figure 5.2). We categorize takeover time into three classes, using threshold values consistent with the pre-defined i^{th} percentile of takeover time in prior driving studies [Coley et al., 2009]. Let T denote the takeover time, thus the labels are defined as "low" when $T < 2.6s$, "medium" when $2.6s \leq T \leq 6.1s$, or "high" when $T > 6.1s$.

Takeover quality. As we alluded to earlier in Section 6.2, there are a wide range of metrics [ISO 21959:2020, 2020] for measuring takeover quality, depending on the needs of various takeover scenarios. In our user study (see Section 6.3), we consider a motivating scenario where the driver needs to take over control of the vehicle and swerve away from an obstacle blocking the same lane; meanwhile, the vehicle should not deviate too much from the current lane, risking crashing into nearby traffic. Therefore, we measure the takeover quality using the lateral deviation from the current lane, denoted by P. In our study, we design a 4-lane rural highway with a lane width of $3.5m$. Therefore, we label the feature vectors into three classes of

takeover quality: "low" or staying in a lane when $P < 3.5m$, "medium" or maneuver the obstacle but too much deviations when $7m < P \leq 10m$, or "high" or maneuver safely and one lane deviates when $3.5 \leq P \leq 7m$.

5.3.3 DNN Models for Takeover Behavior Prediction

DeepTake utilizes a feed-forward deep neural network (DNN) with a mini-batch stochastic gradient descent. The DNN model architecture begins with an input layer to match the input features, and each layer receives the input values from the prior layer and outputs to the next one. There are three hidden layers with 23, 14, and 8 ReLu units, respectively. The output layer can be customized for the multi-class classification of takeover intention, takeover time and takeover quality. For example, for the classification of takeover quality, the output layer consists of three Softmax units representing three classes (low-, medium-, and high-) of takeover quality. DeepTake framework uses Softmax cross-entropy loss with an Adam optimizer with a learning rate of 0.001 to update the parameters and train the DNN models over 400 epochs. In each iteration, DeepTake randomly samples a batch of data in order to compute the gradients with a batch size of 30. Once the gradients are computed, the initiated parameters get updated. The early stopping method set to 400 epochs prevents overfitting. In addition, DeepTakes randomly divides the given labeled data into 70% for training (necessary for learning the weights for each node), 15% for validation (required to stop learning and overtraining), and 15% for testing (the final phase for evaluating the proposed model's robustness to work on unseen data). Finally, in order to address imbalanced data issues where the number of observations per class is not equally distributed, DeepTake utilizes Synthetic Minority Oversampling Technique (SMOTE) [Chawla et al., 2002] which uses the nearest neighbor's algorithm to generate new and synthetic data.

In summary, our DeepTake framework employs different DNN models to predict takeover intention, takeover time and takeover quality. All of the DNN models in DeepTake have the same number of inputs and hidden layers, yet different output layers and associated weights.

5.4 User Study

To test the feasibility of our proposed DeepTake framework, we conducted a user study with 20 participants featuring takeover behavior using a driving simulator[2]. The following section describes the experimental setup and design of our user study as follows.

5.4.1 Participants

In this study, 20 subjects (11 female, 9 male) aged 18-30 (mean= 23.5, SD= 3.1) were recruited. All participants were hired through the university and were required to have normal or corrected-to-normal vision, to not be susceptible to simulator sickness, and to have at least one year of driving experience to be eligible for participation in this study. Before the experiment, participants were questioned as to their age and driving experience. None of them had prior experience of interaction with AVs. They were reminded of their right to abort their trial at any point with no question asked. Three participants' data were later excluded from the analysis, due to biometric data loss and a large amount of missing values. Participants received $20 to compensate for the time they spent in this study.

5.4.2 Apparatus

Figure 5.2 shows our low fidelity driving simulator setup, which consists of a Logitech G29 steering wheel, accelerator, brake pedal and paddle shifters. The simulator records driver control actions and vehicle states with a sampling frequency of 20Hz and sent the captured data through a custom API using iMotions software [iMotions, 2015a]. The simulated driving environments along with the tasks were created using PreScan Simulation Platform. The driving environment was displayed on a 30-inch monitor. The distance between the center of the Logitech G29 steering wheel and the monitor was set at 91cm. A set of stereo speakers was used to generate the

[2]This study complies with the American Psychological Association Code of Ethics and was approved by the Institutional Review Board at University of Virginia.

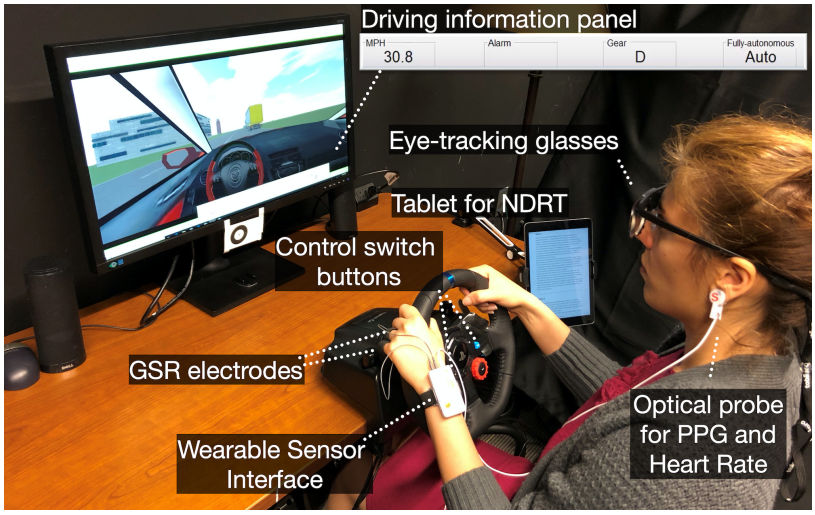

Figure 5.2: User study setup. This custom driving simulator consists of a 30-inch monitor, a Logitech G29 steering wheel, and 10.5-inch Apple iPad Air on which the non-driving tasks are displayed. For switching between the automated and manual control of the vehicle, the participant needs to press the two blue buttons on the steering wheel simultaneously. The participant wears a pair of eye-tracking glasses, and a wearable device with GSR and PPG sensors for the biometrics acquisition.

driving environment sounds along with the auditory alarm of TORs (more details in Section 5.4.3). An Apple iPad Air (10.5-inch) was positioned to the right side of the driver and steering wheel to mimic the infotainment system and displayed an article for NDRT.

We used Tobii Pro-Glasses 2 with the sample rate of 60Hz to collect the eye movement data, and a Shimmer3+ wearable device with a sampling rate of 256Hz to measure PPG and GSR signals. To maintain consistency across all participants, we positioned the Shimmer3+ to the left of all subjects. This consistency helps reduce the motion artifact where the subjects needed to frequently interact with the tablet on the right-hand side. Although we designed our scenarios in a way to minimize the inevitable motion artifacts, we performed necessary signal processing on the PPG and GSR signals to remove potentially corrupted data, as discussed in Section 5.3.1.

Table 5.2: Non-driving related tasks (NDRTs) used in our study

Task Type	Definition
Conversation with passenger	Interacting with the experimenter who sits close to the participants
Using cellphone	Interacting with their cellphones for texting and browsing
Reading articles	Reading three types of articles (i.e.easy,mid,hard) on the tablet
Solving questions	Answering 2-level arithmetic questions (i.e. easy and medium)

5.4.3 Experimental design

A within-subjects design with independent variables of stress and cognitive load manipulated by NDRTs and the TOR types was conducted with three trials in a controlled environment as shown in Figure 5.2. We designed driving scenarios in which the simulated vehicle has enough functionality similar to AVs, such that the full attention of the driver was not required at all times.

Non-Driving Related Tasks. We used four common NDRTs with various difficulty levels and cognitive demand as shown in Table 5.2. Participants used the tablet to read the designated articles and answer the arithmetic questions. Additionally, they were asked to use their own hand-held phones, needed for the browsing tasks. Each participant performed all NDRTs with the frequency of four times in each trial (except for solving the arithmetic questions which occurred three times;15×3 in total). The conditions and the three driving scenarios were counterbalanced among all participants to reduce order and learning effects. To have natural behavior to the greatest extent possible, participants were allowed to depart from NDRTs to resume control of the vehicle at any given time. During manual driving, participants controlled all aspects of the vehicle, including lateral and longitudinal velocity control.

Driving Scenarios. The driving scenarios comprised a 4-lane rural highway, with various trees and houses placed alongside the roadway. We designed five representative situations where the AVs may need to prompt a TOR to the driver, including novel and unfamiliar incidents that appear on the same lane. Figure 5.3 shows an example of a takeover situation used in our study. The designed unplanned takeovers let participants react more naturally to what they would normally do in

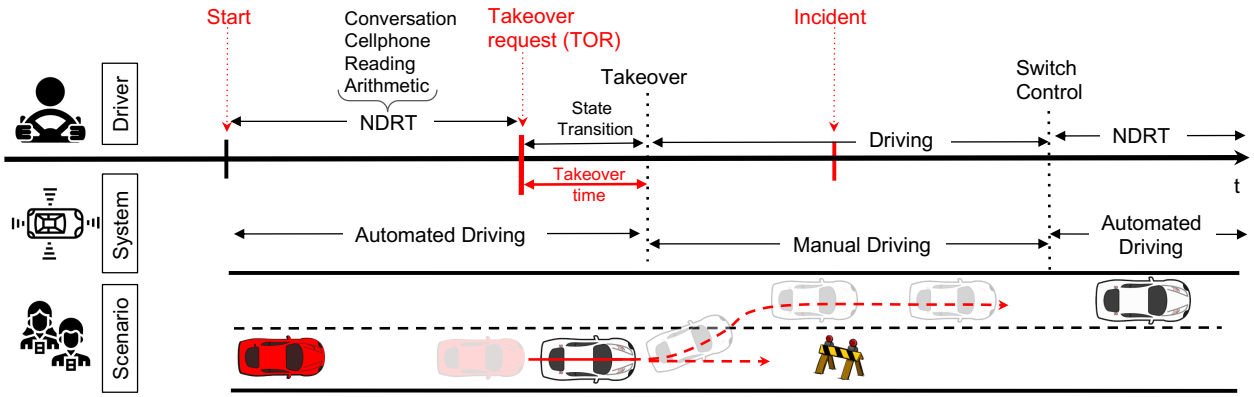

Figure 5.3: A schematic view of an example of a takeover situation used in our study, consisting of: 1) takeover timeline associated with participants' course of action; 2) system status; and 3) takeover situation. The vehicle was driven in the automated mode to the point after the TOR initiation and transitioning preparation period. The ego vehicle is shown in red and the lead car is white. When the Ego vehicle reaches its limits, the system may initiate (true alarm) or fail (no alarm) to initiate the TOR, and the driver takes the control back from the automated system.

AVs [McCall et al., 2019] or as introduced by Kim and Yang [Kim and Yang, 2017], participants' reaction times are in detectable categories. In other words, participants have no previous knowledge of incident appearance, which might happen among other incidents requiring situational awareness and decision-making.

Takeover Requests. In order to incorporate DeepTake in the design of adaptive in-vehicle alert systems in a way that not only monitors driver capability of takeover, but also to enhance takeover performance of automated driving, various types of TOR were required. An auditory alarm was used to inform participants about an upcoming hazard that required takeover from the automated system. The warning was a single auditory tone (350Hz, duration: 75ms) presented at the time of hazard detection (≈140m or ≈13sec before the incidents, depending the speed of the vehicle). In a precarious world, AVs should be expected to fail to always provide correct TORs. Herein, the scenarios were constructed conservatively to include flawed TORs by which subjects would not over-trust the system's ability. In other words, the scenario demands that the participant be partially attentive and frequently perceive

the environment. In order to cover the scenarios that one might encounter while driving an AV, we designed multiple critical types of TORs, including an explicit alarm (true alarm), silent failure (no alarm), and nuisance alarm (false alarm). True alarm indicates the situation in which the system correctly detects the hazard and triggers a TOR, no alarm represents the system's failure to identify the existing hazard, and false alarm presents misclassification of a non-hazardous situation as an on-road danger requiring takeover. We randomized the 15 TOR occurrences in each trial (45 in total for each participant) with 6, 3, 6 repetitions for true alarm, no alarm, false alarm, respectively. In addition, we also designed an information panel where the participants could see the status of the vehicle along with the cause of TOR (see Figure 5.2).

5.4.4 Procedure

Upon arrival in the lab, participants were asked to sign a consent form and fill out a short demographic and driving history questionnaires. Subsequently, they were briefed on how the automated system functions, how to enable the system by simultaneously pressing two blue buttons on the steering wheel, and what they would experience during NDRTs. They were further instructed that if the system detected a situation beyond its own capabilities to handle, it would ask (true alarm) or fail to ask (no alarm) to take over control. Afterward, participants completed a short training drive along a highway for a minimum of 5 minutes to get familiar with the driving and assure a common level of familiarity with the setup, NDRTs, and auditory signals pitch.

Once the subjects felt comfortable with the driving tasks and NDRTs, they proceeded to the main driving scenario. Prior to beginning the main experiment, we calibrated the eye-tracking glasses (repeated at the beginning of each trial) and set participants up with the Shimmer3+ wearable device. Then, participants were required to complete the baseline NASA-TLX questionnaire followed by the PSS-10 questionnaire. The participants were also instructed to follow the lead car, stay on the current route, and follow traffic rules as they normally do. The participants

were cautioned that they were responsible for the safety of the vehicle regardless of its mode (manual or automated). Therefore, they were required to be attentive and to safely resume control of the vehicle in case of failures and TORs. Since the scenarios were designed to have three types of TORs, they needed to adhere to the given instruction whenever they felt the necessity. The given instruction enabled the drivers to respond meticulously whenever it was required and to reinforce the idea that they were in charge of the safe operation of the vehicle. Due to the system's limitations, participants were told to maintain the speed within the acceptable range ($<$ 47mph). The experiment was conducted utilizing scenarios consisting of sunny weather conditions without considering the ambient traffic. The order of NDRT engagement was balanced for participants (see Figure 5.3).

The remainder of the experiment consisted of three trials, each containing 15 TORs, followed by a 5-minute break between trials. At the end of each trial, participants were requested to fill out the NASA-TLX. After completion of the last trial, participants filled out the last NASA-TLX followed by a debrief and a $20 compensation. The experiment took about one hour for each participant.

5.5 Performance Evaluation

We evaluate the performance of DeepTake framework using the multimodal data collected from our user study. We describe the baseline methods, metrics, results, and analysis as follows.

5.5.1 Baseline Methods

Overall, we obtained about 2 million observations to train, test, and validate DeepTake with; 2560 observations per TOR $\times$ 15 TORs per trial $\times$ 3 trials $\times$ 17 subjects. We evaluate the performance of DeepTake DNN-based models with six other ML-based predictive models, including Logistic Regression, Gradient Boosting, Random Forest, Bayesian Network, Adaptive Boosting (Adaboost), and Regularized Greedy Forest (RGF). Our process of choosing the ML models is an exploratory

task with trials and tests of multiple off-the-shelf algorithms and choosing those that perform the best. To evaluate the prediction performance of DeepTake framework with other ML models, we were obligated to utilize some feature importance techniques. The reasons to apply feature importance techniques for an ML algorithm are: to train the predictive model faster, reduce the complexity and increase the interpretability and accuracy of the model. In order to do so, after splitting the labeled data into training, testing, and validation sets (see Section 5.3.3), we employ the following feature importance methods on each training set: Absolute Shrinkage and Selection Operator (LASSO), and random forest. LASSO helps us with not only selecting a stable subset of features that are nearly independent and relevant to the drivers' takeover behavior, but also with dimensionality reduction. The random forest method, on the other hand, ranks all of the features based on their importance levels with the drivers' takeover behavior. The overlapped features chosen by the two methods were used to train the ML-based classification models of takeover behavior.

5.5.2 Metrics

We apply 10-fold cross-validation on training data to evaluate the performance of selected features in the prediction of driver takeover intention, time and quality. Cross-validation provides an overall performance of the classification and presents how a classifier algorithm may perform once the distribution of training data gets changed in each iteration. In cross-validation, we utilize the training fold to tune model hyper-parameters (e.g., regularization strength, learning rate, and the number of estimators), which maximizes prediction performance. Therefore, we train predictive models with the best hyper-parameters. Cross-validation randomly partitions the training data into n subsets without considering the distribution of data from a subject in each set. A possible scenario is that data from one subject could be unevenly distributed in some subsets, causing overestimation of the prediction performance of a model. To avoid this situation, we check the subjects' identifiers in both the training and testing sets to ensure that they belong to just one group. We achieve this by forcing the subject to be in one group. To determine the *accuracy* of the

binary classification of takeover intention performed by predictive models, accuracy was defined as

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \tag{5.1}$$

where TP, TN, FP, and FN represent True Positive, True Negative, False Positive, and False Negative, respectively.

For the multi-class classification of takeover time and quality, we used the average accuracy per class. We also used the metric of *weighted F1 scores* given by

$$WF_1 = \sum_{n=1}^{l} \frac{2 \times Pr_i \times Rc_i}{Pr_i + Rc_i} \times W_i, \tag{5.2}$$

where

$$Pr_i = \frac{\sum_{i=1}^{l} \frac{TP_i}{TP_i + FP_i}}{l} \tag{5.3}$$

$$Rc_i = \frac{\sum_{i=1}^{l} \frac{TP_i}{TP_i + FN_i}}{l} \tag{5.4}$$

are the precision and the recall, respectively. In addition, W_i is the weight of the i^{th} class depending on the number of positive examples in that class. It is worth mentioning that to deal with our imbalanced data, where the number of observations per class is not equally distributed, DeepTake framework along with ML-based predictive models use SMOTE to have a well-balanced distribution within class (see Section 5.3.3).

Given multiple classifiers, we use the *Receiver Operating Characteristic* (ROC) curve to compare the performance of DeepTake alongside other ML-based models. The ROC curve is a widely-accepted method that mainly shows the trade-off between TP and FP rates. A steep slope at the beginning of the curve shows a higher true positive (correct) classification of the algorithm, whereas increasing the FP rate causes the curve to flatten. The ROC curve provides an effective way to summarize the overall performance of classification algorithms by its only metric, AUC. The AUC values provided in Figure 5.4 can be interpreted as the probability of correctly

classifying the driver takeover behavior into the candidate category compared to a random selection (black line in Figure 5.4). In addition, we use the *confusion matrix* to further illustrate the summary of DeepTake's performance on the distinction of takeover intention, time, and quality per class.

5.5.3 Results and Analysis

Multiple classification algorithms were employed to compare the performance of DeepTake on obtaining a reliable discriminator of driving takeover behavior, including intention, time, and quality. As the prediction of driver takeover time and quality are contingent upon the driver's intention to take over from the autonomous systems after receiving TOR, the classification algorithms were initially carried out on this first stage of driver takeover prediction, followed by takeover time and quality.

Takeover intention. Analysis of the binary classification of drivers' takeover intention is shown in Table 5.3. The results show that DeepTake outperforms other ML-based models. However, among the ML-based algorithms, RGF attains the highest accuracy and weighted F1 score (92% and 89%) followed by AdaBoost (88% and 88%) and Logistic Regression (77% and 88%). Moreover, ROC was applied in order to better evaluate each of the classifiers. Figure 5.4.a shows ROC curves and AUC values for all six ML models along with DeepTake to infer the binary classification of takeover intention. Although DeepTake shows outperformance on correctly classifying a driver's intention (AUC=0.96) using the multimodal features, RGF shows promising performance with an AUC of 0.94. Similar to the accuracy level, AdaBoost had a slightly lower performance with an AUC= 0.91. Furthermore, we obtained the confusion matrix for takeover intention (Figure 5.6.a) showing that the percentage of misclassifications is insignificant. Table 5.3, together with the results obtained from the AUC in Figure 5.4.a and the confusion matrix in Figure 5.6.a, ensure that our multimodal features with the right DNN classifier surpass the takeover intention prediction.

Takeover time. DeepTake's promising performance in takeover intention estimation leads us to a more challenging multi-class prediction of driver takeover

time. As some of the ML-based models attained reasonably high accuracy in the binary classification of takeover, their performances, along with our DeepTake DNN based in classifying multi-class classification of takeover time could assess the robustness of the DeepTake.

Figure5.4.b shows a comparison amongst the models explored in this paper along with DeepTake for prediction of takeover time. It displays that DeepTake produces the best overall result with an AUC value of 0.96 ± 0.02 for each takeover low-, mid-, and high- time. We next consider the accuracy comparison of our DeepTake model with other classifier algorithms, reported in Table 5.3. It is evident that DeepTake outperforms all of the classic algorithms. In the three-class classification of takeover time (low, mid, high), DeepTake achieves a weighted-F1

Table 5.3: Classification performance comparison.

Target value	**Classifier**	**Accuracy**	**W-F1[1] score**
Takeover Intention	Logistic Regression	0.77	0.81
	Gradient Boosting	0.76	0.75
	RF[2]	0.75	0.72
	Naive Bayes	0.71	0.66
	Ada Boost	0.88	0.87
	RGF[3]	0.92	0.89
	DeepTake	**0.96**	**0.93**
Takeover Time	Logistic Regression	0.47	0.45
	Gradient Boosting	0.47	0.46
	RF	0.44	0.45
	Naive Bayes	0.36	0.38
	Ada Boost	0.64	0.58
	RGF	0.73	0.71
	DeepTake	**0.93**	**0.87**
Takeover Quality	Logistic Regression	0.65	0.63
	Gradient Boosting	0.60	0.59
	RF	0.53	0.52
	Naive Bayes	0.41	0.39
	Ada Boost	0.42	0.39
	RGF	0.82	0.77
	DeepTake	**0.83**	**0.78**

1: *Weighted F1-score*; 2:*Random Forest*; 3:*Regularized Greedy Forests*

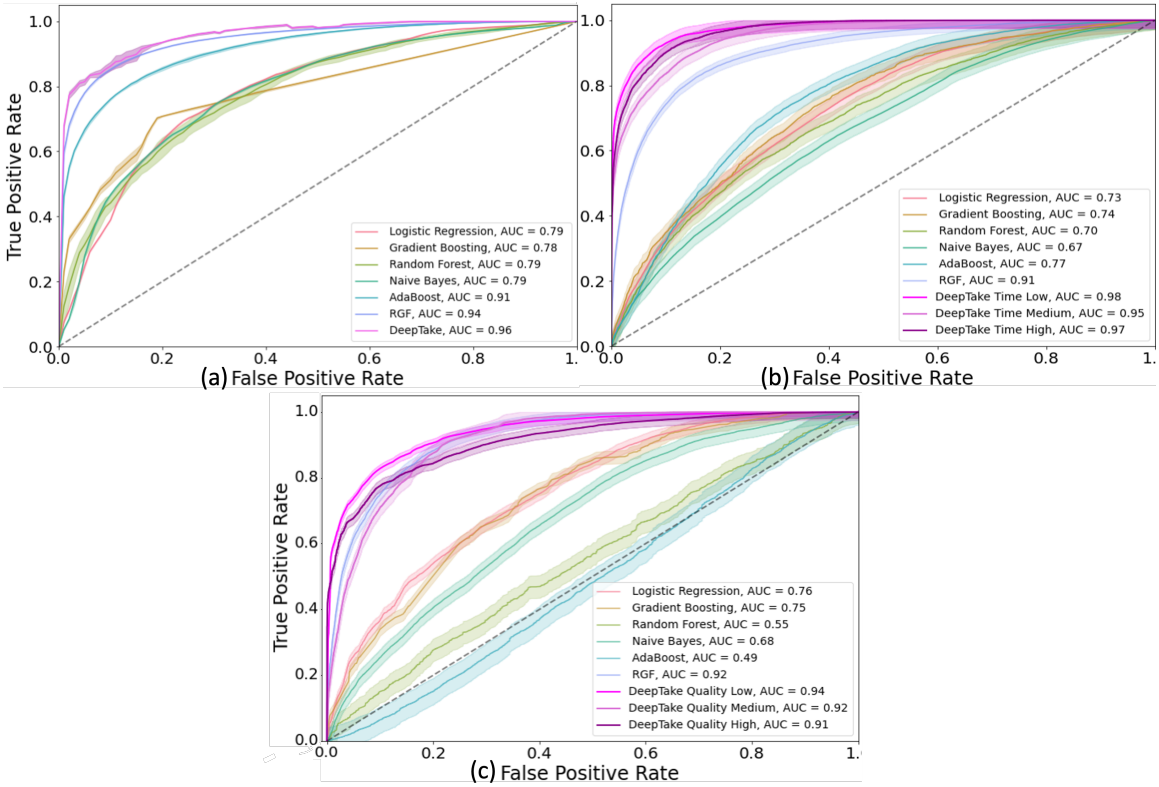

Figure 5.4: The ROC curve comparison of our DeepTake and six ML classification algorithms for classification of takeover behavior: (a) takeover intention, (b) takeover time, and (c) takeover quality. The ROC curve shows the average performance of each classifier and the shadowed areas represent the 95% confidence interval. The macro AUC associated with each classifier is shown where AUC value of 0.5 refers to a chance.[Best viewed in color]

score of 0.87, thereby achieving the best performance on this task by a substantially better accuracy result of 92.8%. Among the classifiers, RGF and AdaBoost still performed better (73.4% and 64.1%). As shown in Figure 5.5, DeepTake gained a high accuracy for both the training and testing sets. However, the model did not significantly improve and stayed at around 92% accuracy after the epoch 250.

To capture a better view of the performance of DeepTake on the prediction of each class of takeover time, we also computed the confusion matrix. Figure 5.6 displays the performance of DeepTake DNN model as the best classifier of three-class takeover time. As the diagonal values represent the percentage of elements for which

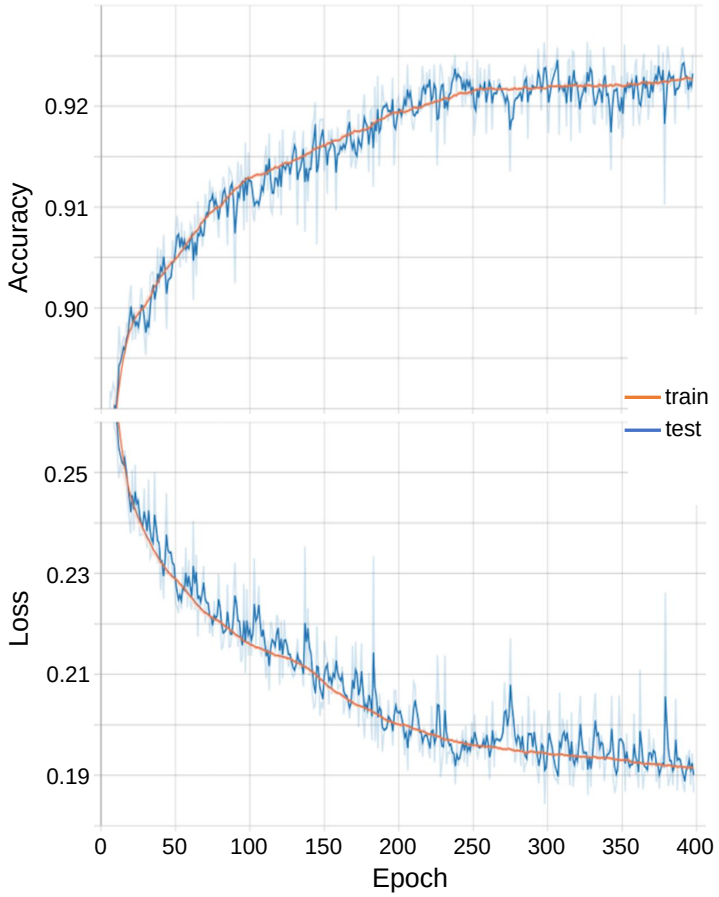

Figure 5.5: The top graph shows the prediction accuracy of training and test sets for 400 epochs, whereas the bottom graph indicates the loss for DeepTake on prediction of three classes of low-, mid-, and high- takeover time.

the predicted label is equal to the true label, it can be seen that the misclassification in medium takeover time is the highest. Also, marginal misclassifications are found in the 2%-5% of the high and low takeover time classes, respectively. Overall, all three evaluation metrics of AUC, accuracy, and confusion matrix indicate that DeepTake robustness and promising performances in correctly classifying the three-class takeover time.

Takeover quality. The test accuracy results of the 3-class classification of all

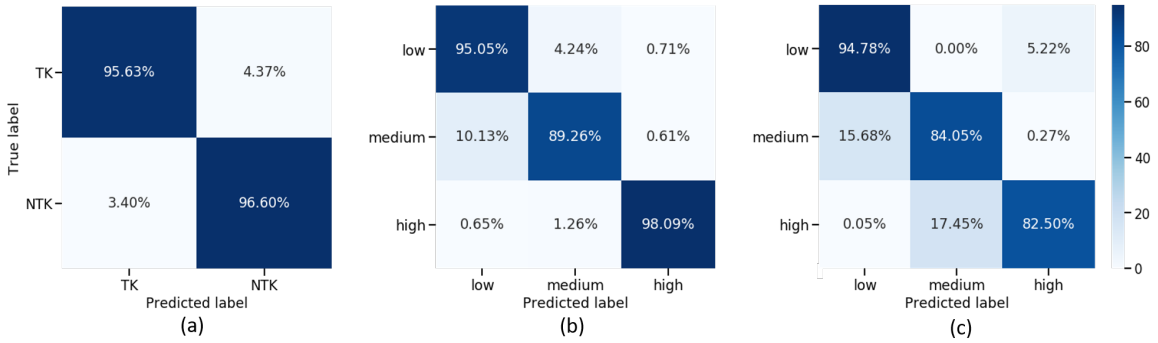

Figure 5.6: Confusion matrix for the prediction of takeover behavior.The results are averaged over 10 fold cross validation splits. (a) Binary class takeover intention takeover(TK) vs. Not Takeover(NTK), (b) 3-Class classification results of takeover time, (c) 3-class classification of takeover quality.

classifiers are presented in Table 5.3. DeepTake achieves the highest accuracy with an average takeover quality of 83.4%. While the value of RGF was close to DeepTake, the rest of the algorithms were not reliable enough to discriminate each class of takeover. However, we should note that RGF training time is very slow and it takes about two times longer than DeepTake to perform prediction.

In addition, Figure 5.4.c presents the multi-class classification of takeover quality. Analysis of the discriminatory properties of DeepTake achieve the highest AUC of 0.92 $\pm$ 0.01 scores among the other models for each individual class. RGF model yields an impressive average macro AUC of 0.91. Such a model indicates a high-performance achievement with informative features.

We further investigated DeepTake robustness in correctly classifying each class of takeover quality and the results achieved by the method are shown in Figure 5.6.c. For the 3-class quality estimation, DeepTake achieved an average accuracy of 87.2%.

5.6 Discussion & Summary

5.6.1 Summary of major findings

In the current design of takeover requests, AVs do not account for human cognitive and physical variability, as well as their possibly frequent state changes. In addition, most previous studies emphasize the high-level relationships between certain factors and their impacts on takeover time or quality. However, a safe takeover behavior consists of a driver's willingness and readiness together. The focus of this paper is to utilize multimodal data into a robust framework to reliably predict the three main aspects of drivers' takeover behavior: takeover intention, time and quality. To the best of our knowledge, the DeepTake framework is the first method for the estimation of all three components of safe takeover behavior together within the context of AVs and it has also achieved the highest accuracy compared to previous studies predicting each aspect individually. To ensure the reliability of DeepTake's performance, we applied multiple evaluation metrics and compared the results with six well-known classifiers. Despite the promising accuracy of some of the classifiers, namely the RGF classifier, the accuracy of DeepTake surpassed in its prediction of takeover behavior. In general, our model performed better in classifying driver takeover intention, time and quality with an average accuracy of 96%, 93%, and 83%, respectively.

In order to further assess the robustness of DeepTake, we increase the number of classes to the more challenging five-class classification of takeover time where the classes defined as "lowest" when $T < 1.5s$, "low" when $1.5s \leq T < 2.6s$, "medium" when $2.6s \leq T < 4.7s$, "high" when $4.7s \leq T \leq 6.1s$, or "highest" when $T > 6.1s$. Figure 5.7 represents the performance of DeepTake on classifying the five-class takeover time. Although DeepTake was not as distinctive in five-class classification as in the three-class, it still achieved promising results. Lowest, high, and medium takeover times are the top three pairs that were the most frequently misclassified by the DNN model. The reason might be that the selected features do not have the required distinctive characteristics to perfectly divide the low and medium takeover time. In each class, it could still distinguish between five other classes with an average

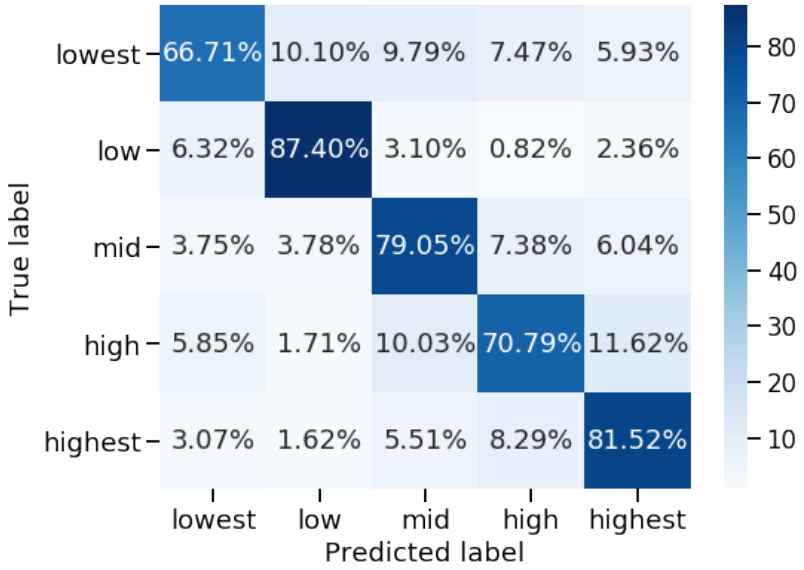

Figure 5.7: Confusion matrix for the prediction of five classes of driver takeover time.

accuracy of 77%. With a future larger amount of data collection satisfying each class need, DeepTake could further improve its distinctive aspect of each feature for more precise classification.

5.6.2 Descriptive analysis of takeover time and quality

Although DeepTake takes advantage of a DNN-based model integrated into its framework, understanding the reasons behind its predictions is still a black-box and a challenging problem which will be tackled in our future works. However, to comprehend the effects of multimodal variables on takeover time and quality, a repeated measure Generalized Linear Mixed (GLM) model with a significance level of $\alpha = 0.05$ to assess the correlation of suboptimal features was used to predict takeover time and quality. The analysis of the results shows the significant main effect of NDRTs on takeover time and quality ($F_{3,28} = 13.58$, $p < 0.001$) followed by fixation sequence ($F_{1,28} = 35.87$, $p < 0.001$) and vehicle velocity ($F_{1,28} = 13.06$, $p < 0.001$). Post-hoc tests using Bonferroni demonstrated a higher impact of interaction with the tablet and reading articles ($p < 0.001$) as opposed to a conversation with

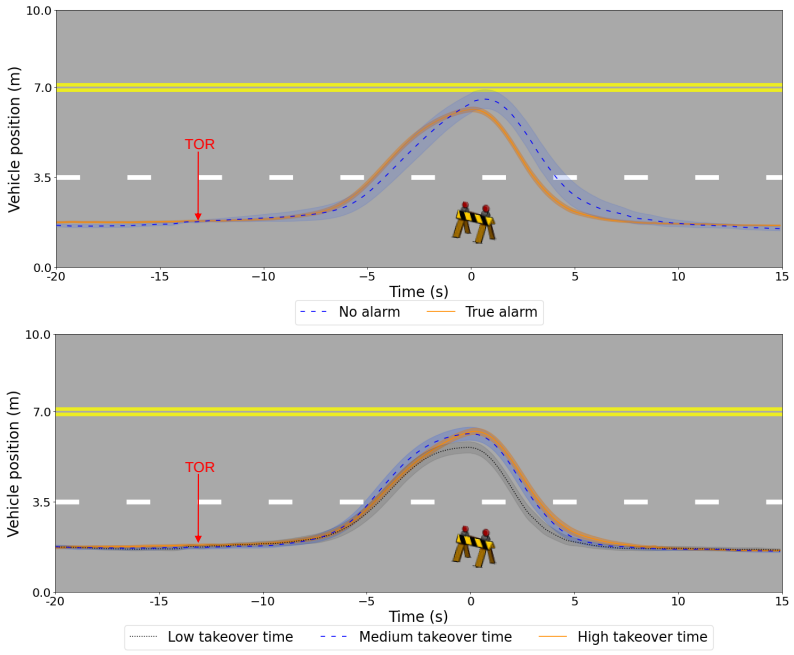

Figure 5.8: Average trajectories when drivers took over control from automated system after receiving TORs. Top graph shows the lateral position of the vehicle with respect to no alarm (silent failure) and true alarm (explicit alarm). Bottom graph shows the lateral position of the vehicle for three categories of takeover time (low, mid, and high). The light shaded area representing standard deviation at each time point.

passengers. This result could be based on the amount of time spent and the level of cognitive load on the takeover task. This finding is aligned with the previous results of [Feldhütter et al., 2017, Eriksson and Stanton, 2017]. Additionally, there was no significant effect of brake and throttle pedal angle on the takeover time($F_{1,28} = 3.05$, $p = 0.085$) and quality ($F_{1,28} = 1.27$ $p = 0.256$). This could be because our scenarios did not take place on crowded roads and participants were not forced to adopt a specific behavior after the TOR. Therefore, they could maneuver the vehicle without significant adjustment to either pedal.

On the other hand, takeover quality tied into drivers' lane keeping control and was impacted by the alarm type and the category of takeover time shown in Figure 5.8. Although we did not consider the type of alarm and category of takeover time for prediction of takeover behavior as they could simply manipulate DeepTake outcomes

by providing insights about the future, it is worth additional investigation of their impacts on the takeover quality. Since participants' takeover times and the speed of the vehicle on the manual driving were different, Figure 5.8 shows the average time of TOR. The top graph in Figure 5.8 depicts the average lateral position of the vehicle with respect to no alarm and true alarm. These two types of the alarm were considered due to the necessity of taking over. Under the impact of the true alarm, the vehicle deviates less than when there is no alarm, yet not significantly ($F_{2,28} = 7.07$, $p = 0.78$). Moreover, the drivers performed more abrupt steering wheel maneuvers to change lanes on true alarm. Similarly, the bottom graph in Figure 5.8 shows the lateral position with respect to different takeover times (low, mid, and high). It can be seen that the longer the takeover time is, the farther the vehicle deviates from the departure lane. Differences in takeover time were also analyzed to investigate the takeover quality. The main effect of the type of takeover time was not significant ($F_{2,19} = 0.44$). Although prior research has revealed various timing efforts to fully stabilize the vehicle [Merat et al., 2014], our observations are comparable to [Naujoks et al., 2019] and [Bueno et al., 2016].

5.6.3 Feature Selection

Neural networks are essentially black-box models, which generate a prediction based on input features and some learned weights. In critical applications, it is imperative to understand how and why the model gives the predictions, by identifying the *important features* that have the highest impact on the model predictions. We therefore designed and evaluated a framework to determine feature importance as viewed by the model.

We examined off-the-shelf state-of-the-art methods such as SHAP [Lundberg and Lee, 2017], LIME [Ribeiro et al., 2016] and Integrated Gradients (IG) [Sundararajan et al., 2017]. SHAP and Integrated Gradients are white-box techniques whereas LIME is a black-box method for attribution analysis. Given a set of input samples, we generate an importance vector of size, $1 \times$ n_features per sample. We randomly selected 3000 samples and created a $3000 \times$ n_features importance matrix. From the importance matrix, we computed the number of times a feature

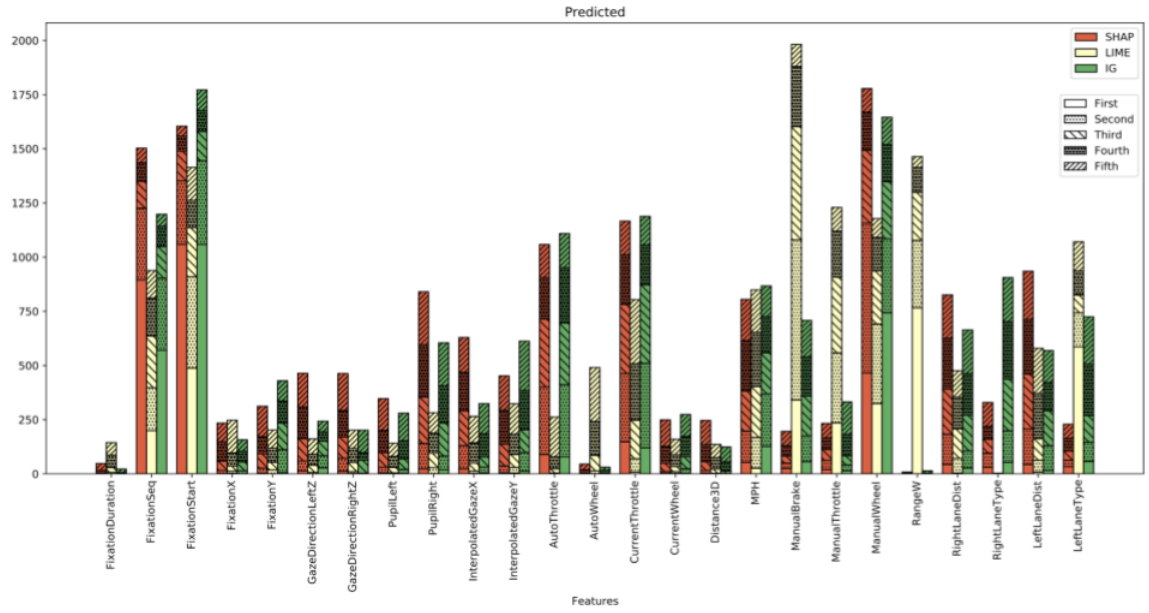

Figure 5.9: Distribution showing the number of times a feature was regarded top-5 important

was regarded as top-k important feature (with the respective method) and created a dictionary where for each feature there are k values and each value signifies the number of times that feature was regarded as k^{th} important feature. For validation, we dropped the features that were found of less importance, re-trained network using the same architecture, and evaluated the resulting accuracy.

In Figure 5.9, for each feature, the measured importance values are plotted for the three evaluated methods (SHAP, LIME and IG). Each bar has 5 parts, demonstrating the top-5 importance values. It can be observed that *FixationSeq, FixationStart* and *Manualwheel* are given high importance values by the three methods, whereas *ManualBrake* is given a high importance value only by LIME. Some features, such as *FixationDuration, AutoWheel, RightLaneType, RangeW* appear to have low importance values. Figure 5.10 depicts the model accuracy after dropping low-importance features. It also validates the importance values as measured by the three methods. For example, *ManualWheel* and *FixationSeq* are important features hence dropping those results in lower accuracy. Dropping *FixationDuration, RangeW*

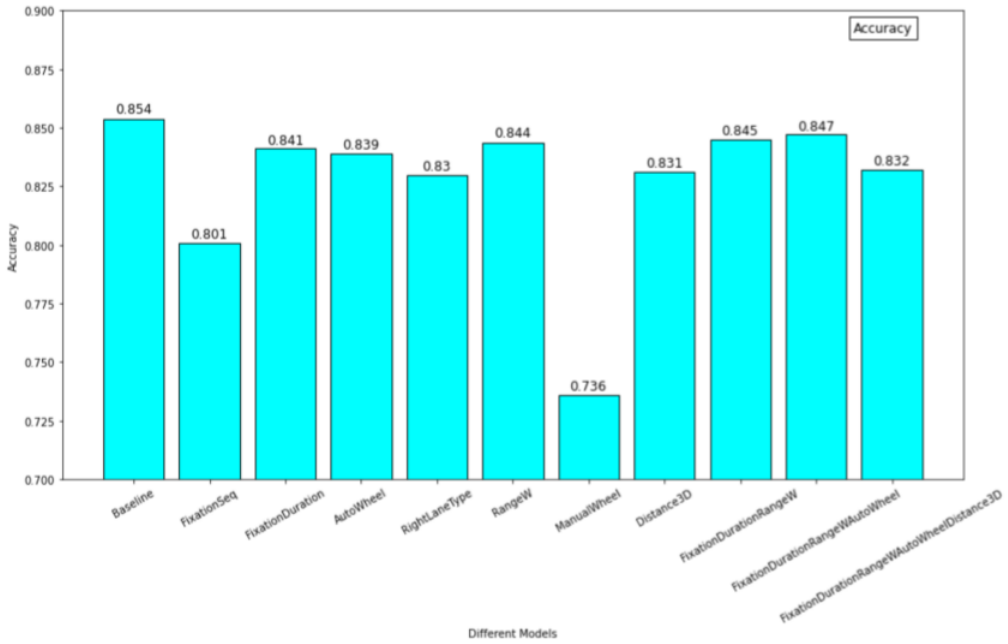

Figure 5.10: Bar plot depicting the accuracy of models trained with dropped features

and *AutoWheel* results in a model with comparable accuracy, demonstrating that they are indeed of low importance. The results indicate that SHAP and IG have similar performance, with LIME giving some outliers. The experiments indicate that existing attribution techniques can indeed be used to understand the model behaviour and furthermore can be used to optimize the model (by dropping some features that have little influence over model predictions).

5.6.4 Implications on the design of future interactive systems

We believe that our human-centered framework makes a step towards enabling a longer interaction with NDRTs for automated driving. DeepTake helps the system to constantly monitor and predict the driver's mental and physical status by which the automated system can make optimal decisions and improve the safety and user experience in AVs. Specifically, by integrating the DeepTake framework into the monitoring systems of AVs, the automated system infers when the driver has the intention to takeover through multiple sensor streams. Once the system confirms a

strong possibility of takeover intention, it can adapt its driving behavior to match the driver's needs for acceptable and safe takeover time and quality. Therefore, a receiver of TOR can be ascertained as having the capability to take over properly, otherwise, the system would have allowed the continued engagement in NDRT or warned about it. Thus, integration of DeepTake into the future design of AVs facilitates the human and system interaction to be more natural, efficient and safe. Since DeepTake should be used in safety-critical applications, we further validated it to ensure that it meets important safety requirements [Grese et al., 2021]. We analyzed DeepTake sensitivity and robustness with several techniques using the Marabou verification tool. The sensitivity analysis provides insight into the importance of input features, in addition to providing formal guarantees with respect to the regions in the input space where the DeepTake behaves as expected.

DeepTake framework provides a promising new direction for modeling driver takeover behavior to lessen the effect of the general and fixed design of TORs which generally considers homogeneous takeover time for all drivers. This is grounded in the design of higher user acceptance of AVs and dynamic feedback [Seppelt and Lee, 2019, Ekman et al., 2017]. The information obtained by DeepTake can be conveyed to passengers as well as other vehicles letting their movement decisions have a higher degree of situational awareness. We envision that DeepTake would help HCI researchers and designers to create user interfaces and systems for AVs that adapt to the drivers' state.

5.6.5 Limitations and future work

The following limitations should be taken into consideration for future research and development of DeepTake.

First, it is acknowledged that the DeepTake dataset is vulnerable to the low fidelity driving simulator used for data collection. It is possible that the takeover behavior of subjects were influenced by the simplicity of driving setup and activities. To apply DeepTake on the road, we will need more emphasis on various user's activities and safety, and exclude subjective surveys causing biases. Second, while we increased

the number of classes, future development of DeepTake should predict takeover time numerically. For this purpose, a larger dataset will be needed which accounts for a high variation of individual takeover time and probabilistic nature of DNNs by which the DeepTake framework can still learn and reliably predicts takeover time.

Third, although we tried to avoid overfitting, it is possible that DeepTake emphasized more on few features that frequently appeared in TORs, and the performance may not be the same if more scenarios are being tested. Thus, DeepTake decision boundaries need to be experimented with different adversarial training techniques. Forth, DeepTake lacks using real-world data which often significantly different and could potentially impact the results of DeepTake framework. Testing the framework on real-world data helps users to gain confidence in DeepTake's performance. DeepTake was developed and assessed offline using a driving simulator in a controlled environment. Future work should explore the deployment of DeepTake online and in the wild for real-world applications in future AVs. We plan to integrate the DeepTake and its verification results [Grese et al., 2021] into the safety controller, which will be then evaluated using the on-road vehicle. In our future work we also plan to try to reduce the number of features in the model by using the results from the sensitivity analysis along with feature importance analysis techniques (i.e. LIME and SHAP) to discover features that may be able to be dropped from the model.

IV

6 | Designing Context-Aware In-vehicle Alert System

In conditionally automated driving, drivers decoupled from driving while immersed in non-driving-related tasks (NDRTs) could potentially either miss the system-initiated takeover request (TOR) or a sudden TOR may startle them. To better prepare drivers for a safer takeover in an emergency, we propose novel context-aware advisory warnings (CAWA) for automated driving to gently inform drivers. This will help them stay vigilant while engaging in NDRTs. The key innovation is that CAWA adapts warning modalities according to the context of NDRTs. We conducted a user study to investigate the effectiveness of CAWA. The study results show that CAWA has statistically significant effects on safer takeover behavior, improved driver situational awareness, less attention demand, and more positive user feedback, compared with uniformly distributed speech-based warnings across all NDRTs.

6.1 Introduction

The rapid development of autonomous driving technologies promises a future where drivers can take their hands off the steering wheels, foot off the pedals, and instead engage in non-driving related tasks (NDRTs) such as reading or using mobile devices. While full self-driving vehicles are not yet commercially available, we are at the stage that conditionally automated driving (level 3 of autonomy, defined by the Society of Automotive Engineers (SAE) [SAE, 2018]) provides various forms of driver

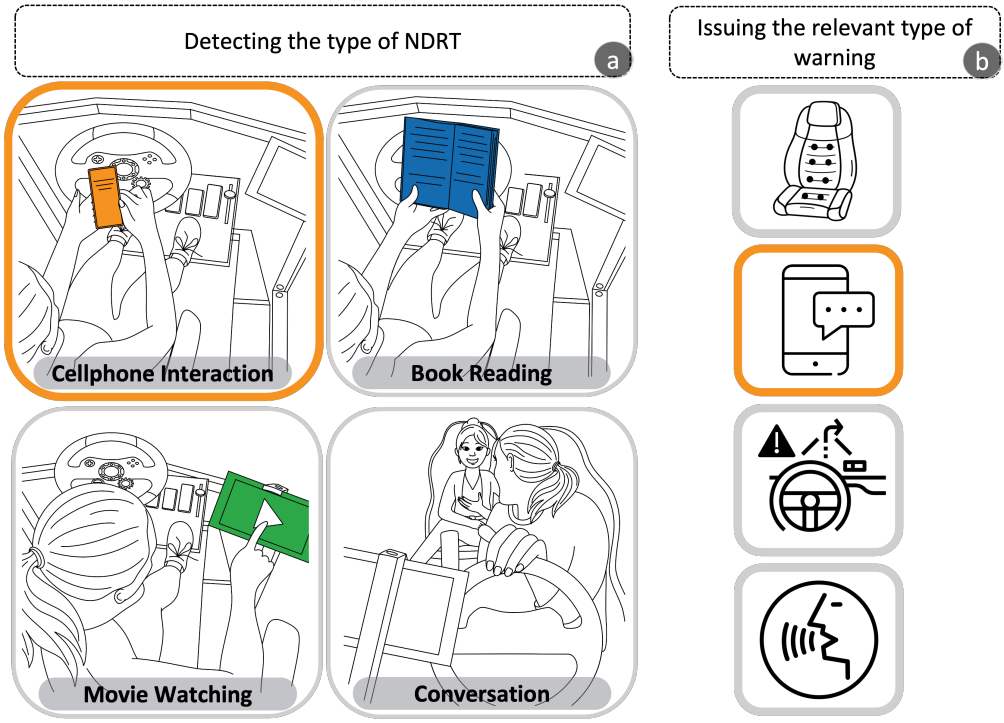

Figure 6.1: The study's proposed context-aware advisory warning method, *CAWA*. a) Detection of the NDRT in which the driver is engaged, b) Selecting the type of modality according to detected activity.

assistance, advanced monitoring systems, and control of the longitudinal and lateral vehicle kinematics on a sustained basis. Although in conditionally automated driving, drivers do not need to continuously monitor the driving environment, due to current technology limitations and legal restrictions, the automated system still needs to relinquish the control back and ask the human driver to resume the control in case of system failures, anticipated dangerous situation, or exceeding its operational limit via a so-called take-over request (TOR) [Bazilinskyy et al., 2018, Gold et al., 2013].

A growing body of research shows that being immersed in NDRTs for an extended period of time causes the level of situation awareness to fall below a comfortable point to safely recover manual control, mainly in urgent situations [Weaver and DeLucia, 2020, Du et al., 2020b, Pakdamanian et al., 2021, Marberger et al., 2017]. Importantly, the control transition process and taking control back cause longer reconfiguration of cognitive and motoric states for drivers to

react properly [Marberger et al., 2017, Kerschbaum et al., 2015]. Thus, human factors researchers argue while most vehicles are not completely self-driving, safety hurdles arise in automated vehicles. Recent fatal crashes indicate drivers' failures to promptly and properly respond to a TOR due to the loss of situation awareness [Board, 2020]. Hence, a key challenge is how to maintain driver readiness for a safe takeover while enabling an enjoyable user experience of engaging in NDRTs. Most existing works focus on the design of TORs, such as its timing [Yoon et al., 2021, Du et al., 2020a] and modalities [Yoon et al., 2019, Salminen et al., 2019]. On the one hand, limitations on current vehicle sensing technologies pose constraints on how early hazardous road incidents can be detected for initiating TORs. The takeover time-budget between the TOR initiation and the incident occurrence is typically 5-7 seconds [Zhang et al., 2019a], which may not be long enough for drivers immersed in NDRTs to regain situational awareness and resume manual driving in a timely and safe fashion. On the other hand, current incorporated unimodal or multimodal TOR may suddenly inform drivers about an upcoming hazard [Zhang et al., 2019a], which may in fact startle and stress the driver and leaving the driver in a less capable state to execute a life-saving maneuver.

To address the aforementioned limitations, we propose context-aware advisory warnings (CAWA) for automated driving to gently and adaptively inform drivers (see Figure 6.1), helping them stay vigilant while engaging in NDRTs. Previous studies on advisory warnings mainly regard manual driving system settings that alert drivers prior an upcoming hazard [Seeliger et al., 2014, Maag et al., 2015]. In contrast, we consider advisory warnings for automated driving system to let drivers know that they are entering the incipient phase of error creation. Then, the key contributions of CAWA are two-fold: (1) CAWA adapts warning modalities according to the NDRT context in which a driver is immersed, for reducing the likelihood that a warning will go unnoticed. (2) CAWA provides gentle warnings in contrast with sudden and startling TORs. For example, if a driver is playing a game on her mobile phone and is wearing headphones, CAWA sends a text message warning to the phone to grab the driver's attention, while auditory or visual warnings may be missed.

In this study each participant experienced two driving scenarios, CAWA and

baseline. In the CAWA trial, advisory warnings were issued depending on the context of NDRTs (e.g., text message warning when the driver is playing a game on her mobile phone, visual warning when the driver is having a conversation) (see Figure 6.1). In the baseline, however, auditory warnings were given uniformly for all NDRTs. We compared CAWA with auditory warning as these are omnidirectional and have already widely applied by auto-manufacturers. The user study demonstrated promising results. Compared with the baseline, CAWA has statistically significant effects on safer takeover behavior, improved driver situational awareness, less attention demand for workload, and more positive driver perceptions.

To the best of our knowledge, this is the first study on context-aware advisory warnings for automated driving. We believe that our work has the potential to provoke future HCI research on integrating advisory warnings into the design of automated vehicles, taking a step toward improving the safety and user experience of automated driving.

6.2 Related Work

Takeover performance can be explained by both reaction time and post-takeover control [McDonald et al., 2019]. Despite many factors have been identified contributing to better reaction time and takeover control such as traffic density [Gold et al., 2016] and driver cognitive state [Sadeghian Borojeni et al., 2018, Van der Heiden et al., 2021] or emotion [Sanghavi et al., 2020], the impact of time budget ("lead time") [Eriksson and Stanton, 2017] and TOR modality [Borojeni et al., 2017] have been widely studied by researchers. For example, studies show that additional second of time budget lead to increase of reaction time by on average 0.27second [Zhang et al., 2019a, McDonald et al., 2019]. If drivers are given more time to gain sufficient situation awareness, they could prepare for the upcoming transition of control. Gold et al. [Gold et al., 2013] has shown that shorter takeover times lead to faster responses but worse maneuvers. On the other hand, a study by Merat et al. [Merat et al., 2014] suggests 20-40second of time budget for a safe takeover to fully stabilised the vehicle

after reclaiming control. As supplying such time budget may not be technologically feasible at the moment, researchers are required to study alternative approaches to enable drivers gaining enough situation awareness as a function of available time [Lu et al., 2017].

To improve takeover time and quality, many warning modalities have been studied such as audio [Politis et al., 2015b], visual [Kim et al., 2017], vibrotactile [Bazilinskyy et al., 2018] and combination of these warning modalities [Baldwin et al., 2012]. Prior studies explored priming drivers before asking them to resume vehicle control. In the study by van der Heiden [van der Heiden et al., 2017], participants received audio warnings 20 seconds prior to TORs, which caused them to disengage from the NDRT earlier and look at the road more closely. In another study [Holländer and Pfleging, 2018], participants received visual warnings indicating the remaining driving time or distance until a TOR would be issued. Compared with these existing works, our study employed a richer set of warning modalities including speech-based cues, visual head-up-displays, text messages, and vibrotactile cues.

Previous research has extensively studied different modalities for in-vehicle alerts, in particular TORs. One of the most prevalent modalities is auditory cues, which can be divided into two categories: nonspeech- and speech-based. Compared with nonspeech-based auditory tones, speech-based messages offer more information and are more favorable to drivers [Wu and Boyle, 2021]. Various representations of visual cues have been designed and utilized, such as a head-up-display [Gerber et al., 2020], augment reality [Lorenz et al., 2014], and LED lights [Borojeni et al., 2018]. Studies also found that vibrotactile and haptic cues can effectively alert drivers [Dass Jr et al., 2013, Telpaz et al., 2015, Morrell and Wasilewski, 2010]. Recent efforts have been increasingly focusing on multi-modal alerts where multiple modalities are triggered simultaneously [Petermeijer et al., 2017a, Bazilinskyy et al., 2018, Sanghavi et al., 2021]. While multi-modal alerts were found to be more effective (e.g., leading to shorter takeover reaction time), they were perceived as more urgent and annoying [Politis et al., 2015a]. Our study takes a different approach from these existing works by incorporating advisory warnings instead of TORs. Moreover, in

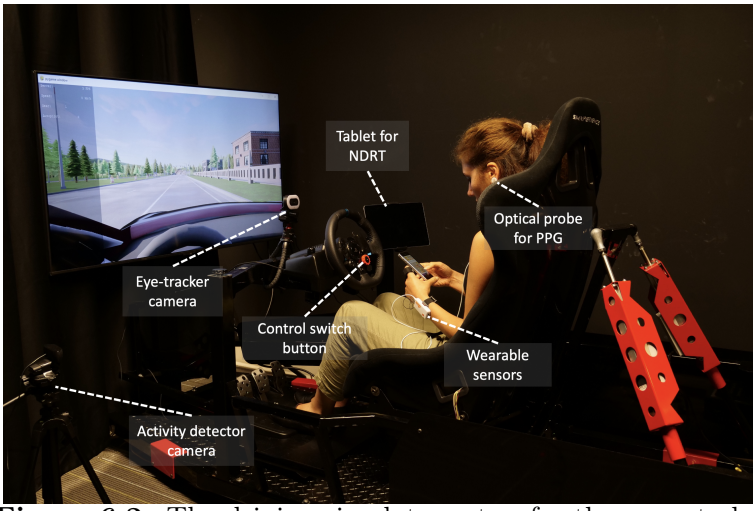

Figure 6.2: The driving simulator setup for the user study.

order to avoid prevalence alert fatigue, CAWA chooses a proper advisory warning from multiple modalities according to the context of NDRTs, rather than triggering all modalities simultaneously.

6.3 Methodology

In this section, we describe the experimental setup, design and procedure. The study protocol was approved by the Institutional Review Board (IRB) at the anonymous university (# anonymous protocol number).

6.3.1 Participants

We recruited a total of 20 participants (14 males; 6 females) with the age range of 18-32 years old (mean= 22.65years; SD= 4.01years). All eligible participants had normal or corrected-to-normal vision, as well as a valid driver's license (mean= 2.8 years, SD = 3.1 years). None of the participants had previous experience with automated driving or prior knowledge about the user study. We used 19 participants' data for the result analysis, excluding one participant due to largely missing biometric data.

6.3.2 Experimental Apparatus

Driving simulator. The study was conducted in a fixed-based driving simulator from SimXperience (Stage 5 Full Motion Racing simulator, fig:Sim). The setup consists of a 55-inch display (1280 × 720 pixel resolution) placed within a horizontal field-of-view and approximately 63-inch away from the driving seat, a racing car seat, a Logitech G29 steering wheel, and sport pedals. No gearshift was required and participants could switch between automated and manual driving modes by pressing a designated button on steering wheel (see Figure 6.2 for details). An Apple iPad Pro with a 9.7-inch display was mounted on the right side of the driving seat for watching movies. Tablet was mounted in common height of the infotainment systems in a landscape format. A 2.0 channel sound bar speaker was placed behind the driver seat for the auditory warnings. The virtual driving environment was created using CARLA [Dosovitskiy et al., 2017], an open-source driving simulation environment built on top of the Unreal Engine. The vehicle was programmed to simulate an SAE Level 3 automation, which handled the longitudinal and lateral vehicle kinematics, and responded to traffic elements.

Biometrics. In this study, we collected drivers' psychophysiological, vehicle-related metrics, workload, and perceived safety. We used a Shimmer3+ wearable device to measure the driver's heart rate (PPG) and galvanic skin response (GSR) signals with a sampling rate of 256 Hz. Heart rate variability (the time elapsed between two successive R-waves) from PPG and maximum and mean phasic components were calculated as the objective metrics reflecting cognitive load variation and stress, respectively.

Face and activity cameras. We installed one high resolution camera (NexiGo N930E 1080p webcam with ring light) above the steering wheel to monitor the driver's eye and head movements. Since CAWA required real-time detection of gaze behavior, we employed state-of-the art pupil and iris localization models [Park et al., 2018b, Xiong et al., 2019] and modified it to fit our needs by integrating deep pictorial gaze estimation [Park et al., 2018a]. Thus, we were able to reliably estimate position and direction of gaze in real-time. Figure 6.3 shows an example of the face video

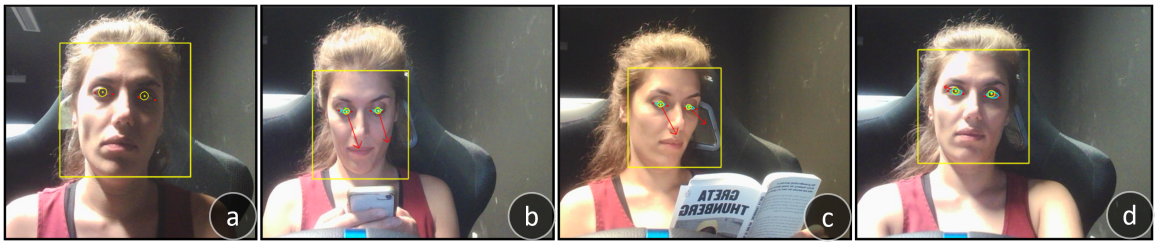

Figure 6.3: Examples of estimated eye region landmarks around the iris and eyelid edges along with gaze direction while performing NDRTs and after a takeover control. a) four main landmarks of eyes and pupil detection, b) gaze direction while looking at the phone, c) gaze direction while reading a book, d) looking at the road after takeover control resumption.

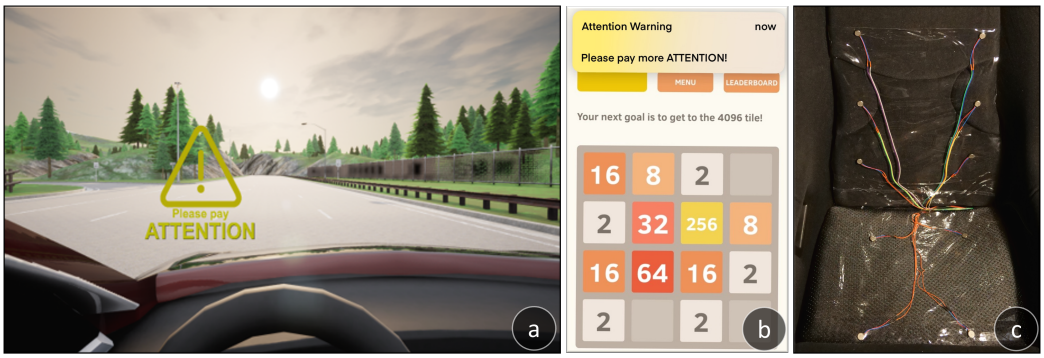

Figure 6.4: Advisory warning modalities: (a) visual warning from the ego's vehicle view , (b) text message, (c) vibrotactile.

examined to capture drivers' eye movements and gaze directions. These videos helped to monitor and to identify when a driver detected a threat or when took her eyes off the driving scene. Furthermore, a high resolution camera (Logitech Ultra HD 1080p) was used to extract participant's driving and engagement activities. Finally, we developed multiple APIs to forward all stream of data to iMotions biometric platform for the real-time aggregation and synchronization.

6.3.3 Experimental Design

We used a within-subject design with driver's cognitive load, and the modality of advisory warnings as independent variables (see Section 6.3.4). The cognitive load was manipulated via the difficulty of the NDRTs (low: watching movie; mid: reading and having an informal conversation; high: playing 2048 game) (see

Table 6.1: CAWA adapts advisory warning modalities based on the context of NDRTs

Non-Driving Related Tasks (NDRTs)	Advisory Warning Modalities
Playing the 2048 game on the mobile phone	Text message
Watching a movie displayed on the tablet	Vibrotactile
Reading a book	Speech-based
Having a conversation with the passenger	Visual

Table6.1). These four activities were selected as the common activities drivers will most likely engage with in L3 [Louw et al., 2019, Naujoks et al., 2016]. Based on prior literature [Petermeijer et al., 2017a, Körber et al., 2018], four takeover events were designed in urban areas with typical roadway features (see Figure 6.5). The difficulty of the scenarios was designed to be approximately the same. Each participant executed two sessions (CAWA and baseline) and the order of sessions was counterbalanced across participants. Per session, the participant experienced 16 possible takeover events (4 TORs per NDRT). In order to avoid predictably and over-trusting of the automated system, we randomly assigned 4 more TORs in each trial to be false alarms, where no hazardous incident was actually detected but a TOR was issued. Although participants interacted with all NDRTs, the given advisory warnings were different in each session. In the CAWA session, the modality of advisory warnings adapted to the context of NDRTs, whereas in baseline, all advisory warnings across different NDRTs use the same auditory modality. In both experimental sessions, the simulated vehicle was equipped with SAE Level 3 automation which could issue TORs (350 Hz acoustic tone with 75 ms duration) to ask the driver to resume the control once it detected an unfamiliar situation out of its capabilities. In the manual driving mode, participants could control the vehicle via the steering wheel and pedals (see details in Sec. 6.3.5).

Figure 6.5: Examples of the TOR four takeover situations, a) Fallen trees. b) Working zone. c) Police set up roadblocks. d) Breakdown cars.

6.3.4 Independent Variables

Modalities

Text message. We developed a Python API that can automatically send a text message containing an advisory warning of "Please pay ATTENTION!" to the driver's mobile phone (see Figure 6.4). The developed attention warning message was displayed at the top of the screen. While drivers are immersed with playing a game on phone, they may potentially miss the auditory and visual cues. In such situation, a notification that grabs users' attention with a quick-to-the-point warning could abruptly direct their attention to the driving scene.

Vibrotactile. We attached 10 vibrotactile actuators (Tatoko 10 mm x 3 mm vibration motor, 3V, 12000 rpm) to the driver's seat as shown in Figure6.4(c), and used an Arduino Uno microcontroller and L9910 motor drivers to drive the vibrotacile actuators. The generated vibrotactile feedback pattern involves two 200 ms long vibrations at maximum amplitude, separated by a 200 ms delay between them.

Speech-based. Previous research has shown that semantics and emotional

tone leads to higher perceived urgency [Baldwin, 2011, Politis et al., 2015a, Ljungberg et al., 2012]. So, it is important to consider whether the message is comprehensible and pleasant for a driver to react upon in a timely manner. We created a gentle warning message "Please pay attention" with a female voice and an American accent.

Visual. Head-up-displays are increasingly used for effective visual communication with drivers [Doshi et al., 2008]. We designed the visual advisory warning as a windshield projected head-up-display shown in Figure6.4(a), which includes a warning sign icon accompanying the text "Please pay ATTENTION".

Please note that we implemented a unimodal advisory warning in CAWA to be effective for each NDRT and to avoid resource sharing conflicts defined by Wickens' multiple resource theory [Wickens, 2002].

Non-driving activities

Participants were asked to perform four NDRTs with three cognitive difficulty levels (i.e. Low: watching movies; Mid: reading and informal conversation; High: playing a mentally demanding game) while setting the vehicle in an automated driving mode. They were also informed that they needed to take control of the vehicle in case a TOR is issued. Studies have shown that engaging with a NDRT for more than three minutes could lead significant decline in situation awareness [De Winter et al., 2014]. Thus, in this study, each NDRT lasted about 219 seconds (SD=15s) before the system initiated a TOR. Participants interacted with each NDRT for about 657 seconds in each block of experiment. Both blocks of experiment consisted of the following NDRTs:

Watching. We selected two movies in the same Action/Thriller genre to prevent potential effects from one specific genre. Participants were given two Netflix movies to choose from, "Extraction" by Sam Hargrave or "Ava" by Tate Taylor.

Reading. "No One Is Too Small to Make a Difference" by Greta Thunberg was selected for the users to read. Participants were also instructed to read out loud to make sure they are surely reading the book.

Conversing. The subjects were asked to have a conversation with the

experimenter sitting behind them to simulate conversation with another passenger regarding everyday topics (e.g., plans for summer vacation).

Gaming. The participant played a 2048 smartphone game, a single-player sliding block puzzle game, whose objective is to slide and combine numbers on a grid with the purpose of achieving a sum of 2048. This game challenge physical and visual demands for receiving an emergency alert.

6.3.5 Procedure

Upon arrival, the participants were briefed about the study. Participants then signed an informed consent form and completed a demographics questionnaire, followed by a 5-minute practice drive to get familiar with the driving simulator and NDRTs. We fitted the participant with the Shimmer3+ wearable device and calibrated the eye-tracker algorithm (which was re-calibrated at the beginning of each trial). Participants were informed that there was no need to actively monitor the driving environments or resume the control of the vehicle unless a TOR was issued. However, they were instructed to resume the vehicle control as soon as a TOR was issued, then switch back to the automated driving once the incident had passed and continue the engagement with a NDRT.

At the beginning of the drive, the participants were asked to activate the automated mode and perform a NDRT based on the experimenter's instructions, followed by three more NDRTs (see Table 6.1). Previous research finds that participants engaging with a NDRT for more than 180 seconds could lead to a significant decline in situation awareness [De Winter et al., 2014]. In this study, immersion to a NDRT lasted 200 seconds on average (SD=15s) before being interrupted by a TOR, which was programmed to be triggered automatically about 111 meters (≈5s) before detection of a dangerous incident. The advisory warnings were also triggered 38-45 seconds (M=40.3s, SD=1.6s) prior TOR to make drivers vigilant of vehicle's state. Overall, participants engaged with each NDRT per trial for 12-15min. The chosen time window is twice as long as in previous studies [van der Heiden et al., 2017,Borojeni et al., 2018] in order to evaluate CAWA's impact on driver

takeover readiness. At the end of each trial, the questionnaire on workload (DALI) and perceived safety and urgency were administered.

After the participant completed all of the driving trials, the experimenter conducted a semi-structured interview to seek the participant's general feedback about the study. The interview guideline was prepared following a prior study [Trösterer et al., 2017]. The entire study took about 100-130 minutes, and the participant received a $30 gift card for completing the study.

6.3.6 Dependent Variables

To investigate the proposed research questions, we used the following objective measurements and subjective feedback as dependent variables.

RQ1 questions driver takeover behavior. We measured the driver's reaction time (i.e., the time difference between the TOR initiation and the exact moment of the driver pressing the button on the steering wheel to resume manual control), and the lateral vehicle control (i.e., deviation from the lane during the takeover).

RQ2 asks about driver situational awareness. As gaze behavior shown to be a reliable indicator of situation awareness [Bhavsar et al., 2017, Recarte and Nunes, 2000, Li et al., 2012], we applied the state-of-the-art computer vision techniques [Park et al., 2018b, Park et al., 2018a] to estimate the gaze behavior of drivers in real-time. We calculated two metrics: (i) percentage of drivers looking at the road; and (ii) fixation duration of when a driver's eyes are on/off the road.

RQ3 evaluates driver stress and cognitive workload. We used the biometric data to calculate metrics including heart rate variability and the number of GSR signal peaks, showing mental workload and stress respectively.pNN50 was calculated as the number of two consecutive intervals (called NN) in which the change in consecutive normal sinus intervals exceeds 50 milliseconds divided by the total number of NN intervals measured. Furthermore, we report the number GSR peaks from the time of advisory warning receipt to moment of takeover control. We also asked participants to complete the Driving Activity Load Index (DALI) [Pauzié, 2008], which customizes NASA-TLX for the automotive domain.

RQ4 inquires about driver perceptions. We asked participants to rate their perceived safety, disruptiveness, and the urgency of advisory warnings on a 5-point Likert-type scale ranging from 1 (strongly disagree) to 5 (strongly agree), which was adapted from the rating questionnaire used in the prior study by Iqbal et al. [Iqbal et al., 2011]. At the end of the study, we interviewed the participants about their preferences for the different advisory warnings and solicited their rationales for the order of preference and usefulness.

6.4 Results

We analyzed the data collected from the user study for the proposed research questions. We set the statistical significance level as $\alpha = 0.05$.

6.4.1 Quantitative Measurements

Effects on Driver Takeover Behavior (RQ1)

We observed in the study that participants were able to take over the vehicle control following TORs with a high success rate. Out of the 456 TORs (19 participants $\times$ 2 trials $\times$ 12 true TORs per trial), only 4 takeovers were failed (e.g., the driver was playing a game on the mobile phone and failed to take over in a timely manner, causing the vehicle to collide with an obstacle). We conducted statistical analysis using the data of 452 successful takeovers to investigate drivers' takeover behavior.

Takeover Quality. We plotted the vehicle trajectories in Figure 6.6. To calculate the lateral RMSE after the issue of the TOR, we estimated an optimal lane change path using heuristic methods. We then compared the position of the vehicle and the path to obtain the lateral error during each time frame .

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=0}^{N} \|\Delta y_i\|^2} \tag{6.1}$$

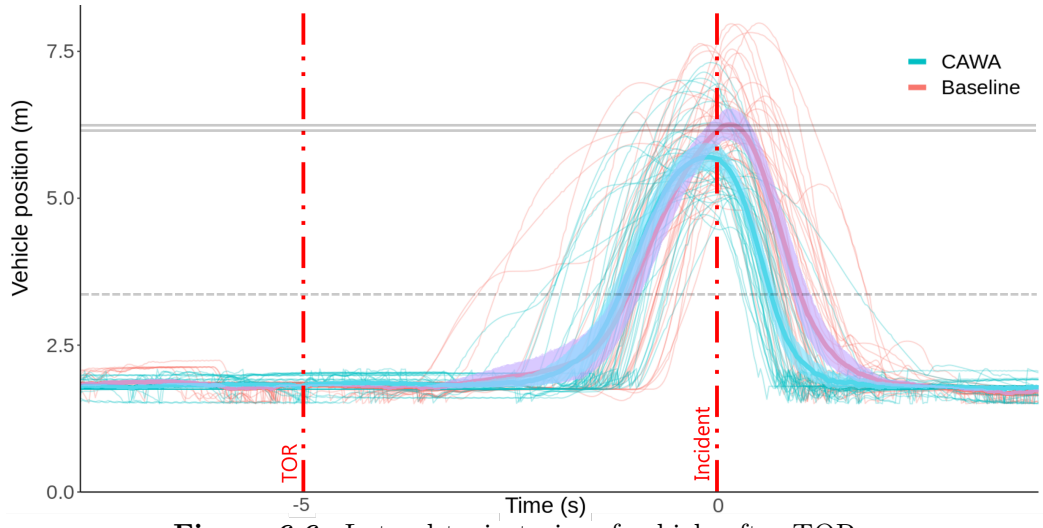

Figure 6.6: Lateral trajectories of vehicle after TORs.

It shows substantial variation in control strategies and higher takeover control after receiving CAWA, as opposed to the baseline, indicating better takeover quality.

A two-way repeated-measures ANOVA also found statistically significant effects on the lateral vehicle control ($F(1, 443) = 13.46$, $p < 0.01$, $\eta^2 = 0.15$) by comparing CAWA and the baseline. Post-hoc showed that the *visual* warning resulted in lower lateral deviation compared to all other modalities ($p < 0.01$). This means that the drivers who were looking at the road while holding a conversation had better control of the car as opposed to other modalities.

Reaction Time. A two-way repeated-measures analysis of variance (ANOVA) analysis found a significant main effect of type of NDRTs ($F(3, 443) = 2.39$, $p < 0.05$, $\eta^2 = 0.049$) and type of advisory warnings ($F(1, 443) = 185.53$, $p < 0.001$, $\eta^2 = 0.47$) on reaction time, showing CAWA can lead to a faster reaction time than the baseline. For types of NDRTs, post-hoc analyses with Bonferroni revealed that there was a significant difference between gaming on the phone an conversation with the experimenter ($p < 0.01$) and between gaming and watching a movie on tablet ($p < 0.01$), indicating that conversing with passengers and watching movie leads to quicker reaction time than gaming (see Figure 6.7 (ii)).

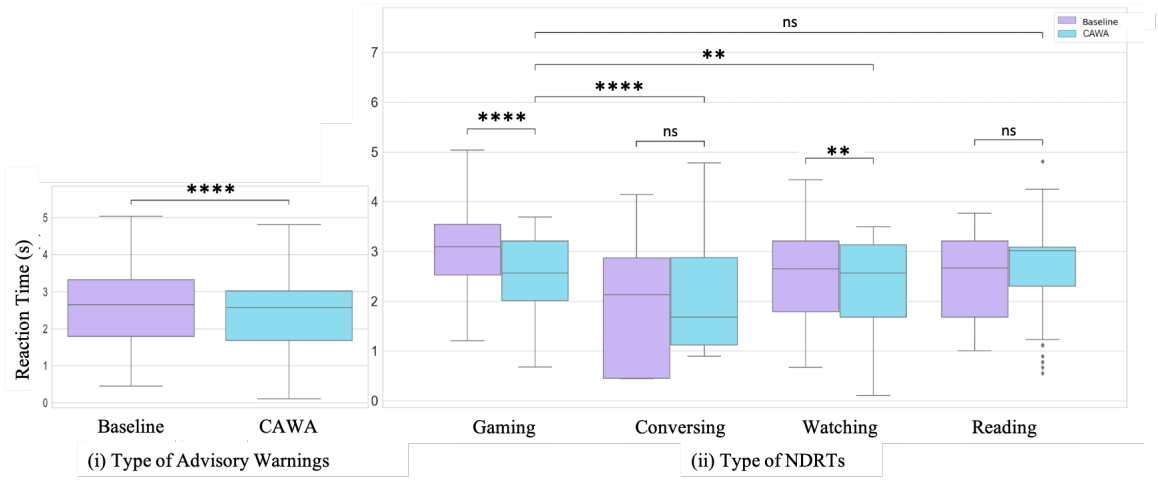

Figure 6.7: Comparisons between the participants' takeover reaction time in relation to the type of advisory warning and the imposed modality. **: $p < 0.01$, ****:$p < 0.0001$

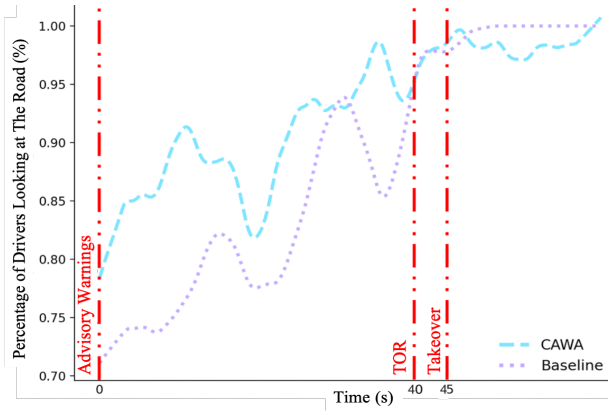

Figure 6.8: Results on the percentage of drivers looking at the road.TOR: issue of TOR; Takeover: the longest time of takeover

Effects on Driver Situational Awareness (RQ2)

Figure 6.8 displays the percentage of drivers looking at the road from the time they received the advisory warning to 20 seconds after resuming vehicle control (i.e., the number of drivers looking at the road at a given time divides the total number of participants). On average, 87.6% of the drivers look at the road from the time of receiving an advisory warning, to the time of actual takeover of control, showing an enhancement on driver's situation awareness. Shortly after the TOR, more than 95%

of drivers shifted their visual attention to the screen. However, more participants stayed vigilant in baseline after taking the vehicle control. Furthermore, analyzing the eye-gaze vector for investigating the fixation time on/off the road, shows the standard deviations across the mean of participants. We ran ANOVA and found significant main effect of type of advisory warnings ($F(1, 443) = 39.47$, $p < 0.05$, $\eta^2 = 0.23$) on the fixation time. Although conversing resulted in higher time fixation on the road, there was no significant difference was observed between the type of NDRTs.

Effects on Driver Stress and Cognitive Workload (RQ3)

We investigated the effect of CAWA and baseline on stress (i.e. GSR) and cognitive load(i.e. heart rate variability(HRV)). The results show no significant effect of NDRT type ($F(3, 443) = 0.95$, $p = 0.42$, $\eta^2 = 0.007$) and type of advisory warnings ($F(1, 443) = 2.23$, $p = 0.14$, $\eta^2 = 0.006$) on HRV (i.e., pNN50). Besides, the statistical analysis showed that the number of GSR peaks from the time of receiving advisory warnings to moment of takeover was significantly impacted by type of NDRT ($F(3, 443) = 0.95$, $p = 0.42$, $\eta^2 = 0.007$), no significant effect of the type of advisory warnings was found ($F(1, 443) = 2.23$, $p = 0.14$, $\eta^2 = 0.006$). Post-hoc test with Bonferroni on the number of GSR peaks indicated a statistically significant difference between watching a movie with conversing ($p < 0.05$) and reading ($p < 0.05$).

We also analyzed the participants' subjective ratings on DALI, which includes six dimensions of workload as shown in Figure 6.9. ANOVA analysis found significant effects on attention demand ($F(2, 54) = 3.70$, $p < 0.05$, $\eta^2 = 0.12$). Post-hoc testing with Bonferroni on attention demand also indicated a significant difference between CAWA and the baseline ($p = 0.029$), which means that the attention required by the baseline was much more demanding than CAWA. However, no statistically significant effects were found in other workload dimensions.

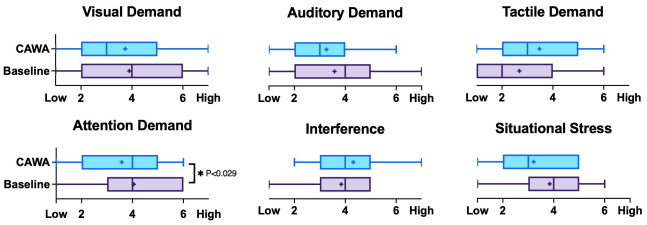

Figure 6.9: Results on DALI ratings about workload.

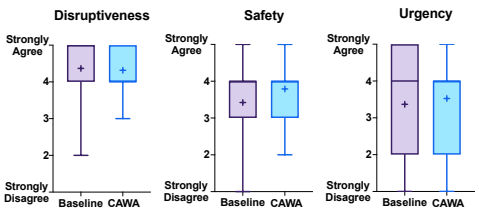

Figure 6.10: Results on driver perceived safety, disruptiveness, and urgency of advisory warnings.

Driver Perceptions (RQ4)

Figure 6.10 shows the survey results on drivers' perceived safety, disruptiveness and urgency of advisory warnings. The results of safety ($F(2, 54) = 0.799$, $p = 0.377$, $\eta^2 = 0.021$), disruptiveness ($F(2, 54) = 0.0.498$, $p = 0.485R$, $\eta^2 = 0.014$) and urgency ($F(2, 54) = 2.866$, $p = 0.099$, $\eta^2 = 0.074$), did not show a significant main effect on type of advisory warnings. Even though more participants rated CAWA to be safer with higher urgency than the baseline, yet they found it more disruptive.

6.4.2 Qualitative Measurements

Preferences and Challenges

For the qualitative evaluation, the details of the interviews for each subject were recorded verbatim. We transcribed the audio recordings from the post-session semi-structured interviews into text and arranged the texts according to the condition. Then, based on the participants' statements on each condition describing their observation, we compared the similarities and differences. Overall, seventeen participants rated the CAWA as more gentle than baseline warnings. Two participants perceived baseline as more gentle mainly due to the "shocking" of the *Vibrotactile* modality. CAWA was referred to as "safer" alternative by fifteen participants.

They described a feeling of a need for learning why they received the warnings in order to adapt to the situation. However, four participants found CAWA

"disruptive", "too pressuring", and "urgent". Although they stated that only a beep is "not enough" or there is "not enough information", they still preferred being less disturbed and let continuing engagement in NDRTs.

Types of Modalities

Participants were asked to express their perception about pros and cons of each implemented type of warning modalities, and their preferences were varied. Participants' preferences of the most suitable types of warning modality for increasing situation awareness and takeover readiness were ranked as *Text messages* (N = 7), *speech-based* (N = 6) and *Vibrotactile* (N = 4), *Visual* (N = 1). Only one participant mentioned that he doesn't need any warnings at all. Four participants expressed the main reasons for preferring the *vibrotactile* modality over the others was as "*it directly connected to my body and waked me up*" and they preferred feeling the cues rather than being interrupted via visual or auditory alarms (P15). For example, a participant stated "*... I guess if you are in the car and your have your music up really loud and watching TV really loud then vibrotactile warnings would be really helpful*".

However, five participants did not favor the *vibrotactile* modality found it difficult to know where their attention should be directed to, for example, P2 stated - "*it did not vibrate anywhere I need to pay attention or I was close to accident. I don't know in which condition did it vibrate or what I should do*". "*The only condition, may it helps would be when I was sleeping, otherwise I didn't find it useful*" Several participants who dislike the textile feels that the textile is not application-specific, only when they are feeling drowsy, for example sleeping, for example, One participant found the textile feedback confusing as it directs his attention to nowhere, as P1 stated "For other context-aware warnings, I know where to look at " The majority (two-thirds) of participants found the textile notification as the most gentle and potentially useful way as an input of notification,

Seven subject found the text messages notification "very useful", "creative", "attention-grabbing" while engaging with the games on the cellphone. For example, P5 stated - *"The game was the hardest. With the game I was using my hands and*

my eyes, and then when the computer says takeover, I need to redirect my eyes to the screen, and put down the phone and then hit the button, with the book I can quickly put it down, and then put it back". However, five of participants who opposing the text messages mentioned two main reasons. They found it "disruptive" as it could block the other urgent text notifications during everyday life. In addition, it was expressed by one that the workload that they need to not only pay attention but also read the text message.

Participants had mixed feelings over the *Speech-based* modality, as they perceived it as most interruptive and "jarring" of the four, yet effective; Six participants valued it as "*it stands out from everything else, and immediately brought me back.* Contrarily, over half of participants perceived the speech-based modality as "robotic". Several participants mentioned that hearing the robotic voice perceived as jarring that they may paying more attention and felt more urgent. "*A warning can fade into the background when you were doing the task, but you are always able to hear the voice, all of the games, the voice was way louder than any others, it stands out from everything else, and immediately brought me back.*" However, participants disliked the voice notifications felt the voice interruptive and sounds similar to takeover request that they sometimes over-responded to it "*the voice was shocking and startling, i was nearly jumped out when i was playing, and also it sounded very similar to the takeover request.*"

Three participants favored the *Visual* modality, as the most "practical" type of warning. These participants backed their choice as it required "less attention" and it was found "less annoying". For instance, P15 stated that *Visuals* is "*easy to understand compared to the text messages that I still need to read the words.*" Three participants opposed visual warnings as it potentially "*occluded the vision of the situation*" (P1, P5, P7) and could be "distracting" (P7). For example, P5 commented - "*it can blend into the background*".

Moreover, two participants found the Visuals being too gentle that they sometimes ignores it, thus less useful than other three warnings, especially when they were playing the game on the mobile phone.

The only condition, may it helps would be when I was sleeping, otherwise I didn't find it useful". Another common complaint was that the *vibrotactile* feeling made them feel uncomfortable. Potential changes can be providing numbers of textile responses showing the urgency of advisory warnings as the prediction of the driver's current engagements. Baseline warnings were perceived as most interruptive of the four and being "too loud". Participant who liked baseline modality in the way that it can accompany them, e.g., P4 mentioned - "*When I am watching the movie as I felt that the car can accompany me and I would not feel alone.*" Participants have mixed feelings about the robotic voice perceived as "jarring" but more effect because they paid more attention and felt more urgent to react to it. "*other warnings can fade into the background when you were doing the task, but you are always able to hear the voice, all of the games, the voice was way louder than any others, it stands out from everything else, and immediately brought me back.*" (P9). However, participants disliked the baseline warnings perceived the voice being too interruptive as an advisory warning - "*the voice was shocking and startling, I was nearly jumped out when I was playing*" (P11) and also it "sounded similar to takeover request" (P17) so that they over-responded to it.

6.5 Discussion

This study aimed to investigate the effects of context-aware advisory warnings on takeover readiness and performance. In order to do so, we proposed a novel context-aware advisory warning system (CAWA). CAWA adapts its warning modalities based on the context a driver is immersed in. In contrast to pre-alert systems [van der Heiden et al., 2017] that startle and stress the driver to take an immediate action, advisory warnings are non-assertive. Although a large body of literature has investigated the influence of various warnings on takeover time [Lu et al., 2017, Eriksson and Stanton, 2017, Petermeijer et al., 2017a] and quality [Du et al., 2020a, Weaver and DeLucia, 2020], to the best of our knowledge, it is the first study to employ multiple modalities for "advising" drivers of automated vehicles to pay attention to the driving scene and

to be more conscious of the automated driving status, specifically via text messages.

Some studies on transitioning the vehicle control in automated driving have suggested that cognitive state of drivers are more important than motor readiness when designing an effective TOR, while other studies have suggested that the physical attributes that are captivated by NDRTs are essential. Therefore, we considered both aspects in our detailed assessments.

Results revealed the importance of advisory warnings for driver's readiness and situation awareness. Results showed a significant differences between CAWA and baseline in controlling the vehicle at moment of takeover. There were also significant differences in reaction time due to increases in fundamental frequency. In terms of situation awareness, while more drivers tend to look at the screen longer after receiving CAWA, there were no significant difference. DALI also revealed a significant difference on attention demand. The overall results show that divers received CAWA have higher takeover quality and lower reaction time when compared to drivers received auditory warning.

6.5.1 Takeover Behavior

Takeover reaction times and quality were measured and analyzed to compare differences due to perceived CAWA and auditory warning. In line with previous studies that found auditory warning leads to significantly higher reaction time [Eriksson and Stanton, 2017, Politis et al., 2015a], we observed significantly higher reaction times with baseline as opposed to CAWA. Further, the results showed that conversing yielded the lowest reaction time, but the results may reflect the fact that the conversation with the experimenter did not need shifting visual attention. The most cognitively and visually demanding task, playing 2048 game, showed higher reaction time. Although the react times were varied, CAWA helped drivers to resume the control faster. The range of reaction time obtained in our study slightly differ from previous studies [Eriksson and Stanton, 2017, Zhang et al., 2019a], showing that participants were somewhat prepared to take the control or anticipated a takeover after receiving an advisory warning. Despite the research of [Gold et al., 2018]

indicating that the complexity of NDRTs is not a significant variable for reaction time, our experiment's findings indicated that takeover time was significantly impacted by physical and cognitive loads needed for performing NDRTs.

We also observed that CAWA assisted drivers to depart earlier and helped less deviation from the center of the lane (see Figure 6.6). This finding of vehicle control after receiving TOR is in line with our expectations based on previous studies [Politis et al., 2015a, Manawadu et al., 2018] showing non-auditory warnings provides relatively better control of the vehicle. Our findings also suggest that a safer takeover is a composite of multiple factors (e.g. type of NDRT and its level of complexity, type of modalities, etc.) and they may have a greater effect on readiness and takeover.

6.5.2 Situation Awareness

We observed higher rates of monitoring of the road after receiving CAWA compared to the baseline. More specifically, after the vehicle approached to advisory warning time, 14% more of driver looked back at the road and stayed more visually attentive. In general, our results shows that receiving advisory warnings increases 26% likelihood of looking at the road as opposed to the results reported in [van der Heiden et al., 2017].

6.5.3 User Experience

Concerning the usability aspect of proposed method, the users perceptions towards advisory warnings' safety and disturbance were analyzed along with their subjective workload using DALI survey. Participants' rating of their perceived safety, disruptiveness and urgency, favord CAWA, but did not differ significantly between the two conditions. Post-study interviews revealed that users believed that CAWA could avoid being missed, but it leads to higher annoyance. Even though we extended the timing of advisory warning suggested by literature to 200s on average, we acknowledge that a better experimental design with less frequent interruption could have increased

CAWA's usability. In addition to driver's perceptions, only the significant difference in the attention demand subscale of the DALI supported the hypotheses. Despite slightly better score in visual and auditory demand of CAWA, participants' subjective workload rating did not differ significantly between the conditions. It is possible that the similar time budget to takeover between the two conditions was perceived as alike workload. Another possibility for the absence of significance in the subscales of DALI could be due to the within-subject design where we only collected one data point to compare the conditions.

6.6 Limitations

We applied unimodal advisory warning rather than multimodal modalities. While multimodal modalities were found to improve reaction time [Petermeijer et al., 2017a] and quality of takeover [Naujoks et al., 2014], prior studies reported them as urgent [Kutchek and Jeon, 2019] and annoying [Politis et al., 2015a]. We utilized unimodal modalities (1) to avoid resource sharing conflicts according to Wickens' multiple resource theory [Wickens, 2002], (2)to investigate the impact of non-assertive advisory warnings on takeover behavior. However, we acknowledge that a more exhaustive picture would have been available if we combined multiple modalities to urge drivers to pay attention to the driving scene.

Another limitation is using a driving simulator. While driving simulator studies are very common due to advantages in creating standardized situations for experimental control, they come with limited external validity. Participants may react differently in the lab than they do naturally while driving in the wild. Despite randomizing the time interval for advisory warnings, participant could still expect to encounter a TOR.

Despite these limitations, this study takes the first steps toward enabling CAWA for automated driving, which can provoke many exciting future research directions. In this study warnings were triggered for a fixed period (about 40 seconds) before TORs in the study. Future work could leverage recent advances in predicting

driver takeover behavior and readiness [Yoon et al., 2021, Pakdamanian et al., 2021], and develop agent-based systems to intelligently decide when and how to trigger warnings based on driver state predictions.

V

7 | Conclusions & Future Work

Previous studies suggest that drivers are incapable of responding effectively to critical situations resulting from limitations or failures of highly automated driving since they are 'out-of-the-loop'. According to these studies, drivers are generally less aware of their surroundings when automated driving is engaged, thus leading to crashes, and response time to critical incidents is slower, sometimes resulting in crashes. This could be due to the fact that drivers take some time after disengaging automation to reorient their attention to the driving scene after it is deactivated. The objective of this dissertation was to address this challange by:

- Investigating the effects of drivers' mental workload and type of TORs on their takeover performance (i.e. takeover time and takeover quality) and psychophysiological responses (gaze behavior, heart rate activities, GSR, and EEG).
- Developing, to our knowledge, the first neural network model to predict drivers' takeover performance (i.e. takeover intention, takeover time, and takeover quality) by utilizing drivers' physiological data and driving environment.
- Developing the first end-to-end adaptive alert system that informs drivers about the loss of SA using a context-aware warning system. To evaluate the system's practicality, we conduct a preliminary proof-of-concept human-subject experiments to study their takeover performance, perceived safety, acceptance, and preparedness in multiple traffic scenarios.

We initially evaluated the effects of cognitive load, traffic density, and types

of TOR on driving behaviors (takeover timeliness and quality) and psychological responses (eye movements, galvanic skin response, heart rate activity). After understanding each feature importance and its practicality for real-world applications, we applied advanced machine learning algorithms to develop computational models that predict drivers' takeover performance based on their physiological information and driving environment information. Finally, predicting the driver's reaction time and takeover behavior should lead to safely bringing human back to-the-loop. As a result, we developed an end-to-end adaptive alert system that warns driver about the loss of SA based on the type of immersion.

7.1 Future Work

There are several interesting future directions where this body of work can be extended.

First, despite the fact that objective measures of mental workload and situation awareness using physiological assessments are useful in the research phase, these measures are of relatively minor importance in an automated driving system intended for public use. Thus, future research should aim to develop an objective method of assessing a driver's state and behavior that is non-invasive and non-intrusive. Such non-invasive technologies would allow for a precise determination of the driver's state, thereby allowing certain elements of the driver-vehicle interaction to be tailored to meet the needs of each driver.

Secondly, in this thesis, several frequency-domain based features were identified through feature analysis, which enhanced our proposed machine learning models. However, it's essential to apply advanced signal processing methods to further enhance their values. The interpretation of different frequencies of heart rate and pupil diameter and their possible links with mental workload or driver states will require further research.

Another key limitation that cuts across all our studies especially for affect prediction is the relatively small amount of data across few participants. There is an

opportunity to further validate, and evaluate the generalizability of our approaches by collecting a larger sample with more diverse participants.

To conclude, it is hoped that the research presented in this thesis will encourage designers of contemporary and future automated vehicles to take into account human factors principles to create a safer and more accessable human-centered automated vehicle.

References

[Almahasneh et al., 2014] Almahasneh, H., Chooi, W.-T., Kamel, N., and Malik, A. S. (2014). Deep in thought while driving: An eeg study on drivers' cognitive distraction. *Transportation research part F: traffic psychology and behaviour*, 26:218–226.

[Alsaid et al., 2019] Alsaid, A., Lee, J. D., and Price, M. (2019). Moving into the loop: An investigation of drivers' steering behavior in highly automated vehicles. *Human factors*, page 0018720819850283.

[Anderson et al., 2014] Anderson, J. M., Nidhi, K., Stanley, K. D., Sorensen, P., Samaras, C., and Oluwatola, O. A. (2014). *Autonomous vehicle technology: A guide for policymakers*. Rand Corporation.

[Anderson et al., 2018] Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., and Zhang, L. (2018). Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086.

[Automotive, 2016] Automotive, I. (2016). Autonomous vehicle sales forecast to reach 21 mil. globally in 2035.(aug. 2016). *Retrieved August*, 18:2017.

[Baldwin, 2011] Baldwin, C. L. (2011). Verbal collision avoidance messages during simulated driving: perceived urgency, alerting effectiveness and annoyance. *Ergonomics*, 54(4):328–337.

[Baldwin et al., 2012] Baldwin, C. L., Eisert, J. L., Garcia, A., Lewis, B., Pratt, S. M., and Gonzalez, C. (2012). Multimodal urgency coding: auditory, visual, and tactile parameters and their impact on perceived urgency. *Work*, 41(Supplement 1):3586–3591.

[Banks et al., 2018] Banks, V. A., Plant, K. L., and Stanton, N. A. (2018). Driver error or designer error: Using the perceptual cycle model to explore the circumstances surrounding the fatal tesla crash on 7th may 2016. *Safety science*, 108:278–285.

[Banks and Stanton, 2016] Banks, V. A. and Stanton, N. A. (2016). Keep the driver in control: Automating automobiles of the future. *Applied ergonomics*, 53:389–395.

[Bazilinskyy et al., 2018] Bazilinskyy, P., Petermeijer, S. M., Petrovych, V., Dodou, D., and de Winter, J. C. (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.

[Beitner et al., 2021] Beitner, J., Helbing, J., Draschkow, D., and Võ, M. L.-H. (2021). Get your guidance going: Investigating the activation of spatial priors for efficient search in virtual reality. *Brain Sciences*, 11(1):44.

[Berghöfer et al., 2018] Berghöfer, F. L., Purucker, C., Naujoks, F., Wiedemann, K., and Marberger, C. (2018). Prediction of take-over time demand in conditionally automated driving-results of a real world driving study. In *Proceedings of the Human Factors and Ergonomics Society Europe Chapter 2018 Annual Conference*, pages 69–81.

[Bhavsar et al., 2017] Bhavsar, P., Srinivasan, B., and Srinivasan, R. (2017). Quantifying situation awareness of control room operators using eye-gaze behavior. *Computers & chemical engineering*, 106:191–201.

[Bliss and Acton, 2003] Bliss, J. P. and Acton, S. A. (2003). Alarm mistrust in automobiles: how collision alarm reliability affects driving. *Applied ergonomics*, 34(6):499–509.

[Board, 2020] Board, N. (2020). Collision between a sport utility vehicle operating with partial driving automation and a crash attenuator mountain view, california. *accessed October*, 30.

[Boggs et al., 2020] Boggs, A. M., Arvin, R., and Khattak, A. J. (2020). Exploring the who, what, when, where, and why of automated vehicle disengagements. *Accident Analysis & Prevention*, 136:105406.

[Borji et al., 2010] Borji, A., Ahmadabadi, M. N., Araabi, B. N., and Hamidi, M. (2010). Online learning of task-driven object-based visual attention control. *Image and Vision Computing*, 28(7):1130–1145.

[Borji and Itti, 2015] Borji, A. and Itti, L. (2015). Cat2000: A large scale fixation dataset for boosting saliency research. *arXiv preprint arXiv:1505.03581*.

[Borji et al., 2012] Borji, A., Sihite, D. N., and Itti, L. (2012). Probabilistic learning of task-specific visual attention. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 470–477. IEEE.

[Borojeni et al., 2017] Borojeni, S. S., Wallbaum, T., Heuten, W., and Boll, S. (2017). Comparing shape-changing and vibro-tactile steering wheels for take-over requests in highly automated driving. In *Proceedings of the 9th international conference on automotive user interfaces and interactive vehicular applications*, pages 221–225.

[Borojeni et al., 2018] Borojeni, S. S., Weber, L., Heuten, W., and Boll, S. (2018). From reading to driving: priming mobile users for take-over situations in highly automated driving. In *Proceedings of the 20th International Conference on Human-Computer Interaction with Mobile Devices and Services*, pages 1–12.

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

[Braunagel et al., 2017] Braunagel, C., Rosenstiel, W., and Kasneci, E. (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.

[Bueno et al., 2016] Bueno, M., Dogan, E., Selem, F. H., Monacelli, E., Boverie, S., and Guillaume, A. (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)*, pages 2040–2045. IEEE.

[Burgess et al., 2009] Burgess, P. W., Alderman, N., Volle, E., Benoit, R. G., and Gilbert, S. J. (2009). Mesulam's frontal lobe mystery re-examined. *Restorative neurology and neuroscience*, 27(5):493–506.

[Bylinskii et al., 2018] Bylinskii, Z., Judd, T., Oliva, A., Torralba, A., and Durand, F. (2018). What do different evaluation metrics tell us about saliency models? *IEEE transactions on pattern analysis and machine intelligence*, 41(3):740–757.

[Cars, 2013] Cars, V. (2013). Volvo car group initiates world unique swedish pilot project with self-driving cars on public roads. *See: https://www. media. volvocars. com/global/engb/media/pressreleases/136182/volvo-car-group-initiates-world-unique-swedish-pilotproject-with-self-driving-cars-on-public-roads*.

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

[Chen et al., 2012] Chen, A. J.-W., Britton, M., Turner, G. R., Vytlacil, J., Thompson, T. W., and D'Esposito, M. (2012). Goal-directed attention alters the tuning of object-based representations in extrastriate cortex. *Frontiers in human neuroscience*, 6:187.

[Chen et al., 2015] Chen, C., Seff, A., Kornhauser, A., and Xiao, J. (2015). Deepdriving: Learning affordance for direct perception in autonomous driving. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2722–2730.

[Chen et al., 2021] Chen, J., Song, H., Zhang, K., Liu, B., and Liu, Q. (2021). Video saliency prediction using enhanced spatiotemporal alignment network. *Pattern Recognition*, 109:107615.

[Claybrook and Kildare, 2018] Claybrook, J. and Kildare, S. (2018). Autonomous vehicles: No driver... no regulation? *Science*, 361(6397):36–37.

[Codevilla et al., 2019] Codevilla, F., Santana, E., López, A. M., and Gaidon, A. (2019). Exploring the limitations of behavior cloning for autonomous driving. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 9329–9338.

[Cohen et al., 1983] Cohen, S., Kamarck, T., and Mermelstein, R. (1983). A global measure of perceived stress. *Journal of health and social behavior*, pages 385–396.

[Coley et al., 2009] Coley, G., Wesley, A., Reed, N., and Parry, I. (2009). Driver reaction times to familiar, but unexpected events. *TRL Published Project Report*.

[Cornia et al., 2018] Cornia, M., Baraldi, L., Serra, G., and Cucchiara, R. (2018). Predicting human eye fixations via an lstm-based saliency attentive model. *IEEE Transactions on Image Processing*, 27(10):5142–5154.

[Csikszentmihalyi and Csikzentmihaly, 1990] Csikszentmihalyi, M. and Csikzentmihaly, M. (1990). *Flow: The psychology of optimal experience*, volume 1990. Harper & Row New York.

[Damböck et al., 2013] Damböck, D., Weißgerber, T., Kienle, M., and Bengler, K. (2013). Requirements for cooperative vehicle guidance. In *16th international IEEE conference on intelligent transportation systems (ITSC 2013)*, pages 1656–1661. IEEE.

[Dass Jr et al., 2013] Dass Jr, D. E., Uyttendaele, A., and Terken, J. (2013). Haptic in-seat feedback for lane departure warning. In *Proceedings of the 5th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, pages 258–261.

[Dawson et al., 2007] Dawson, M. E., Schell, A. M., Filion, D. L., Cacioppo, J. T., Tassinary, L. G., and Berntson, G. (2007). Handbook of psychophysiology. *The Electrodermal System*, pages 159–181.

[De Waard and Brookhuis, 1996] De Waard, D. and Brookhuis, K. (1996). The measurement of drivers' mental workload.

[De Winter et al., 2014] De Winter, J. C., Happee, R., Martens, M. H., and Stanton, N. A. (2014). Effects of adaptive cruise control and highly automated driving on workload and situation awareness: A review of the empirical evidence. *Transportation research part F: traffic psychology and behaviour*, 27:196–217.

[Deng et al., 2019] Deng, T., Yan, H., Qin, L., Ngo, T., and Manjunath, B. (2019). How do drivers allocate their potential attention? driving fixation prediction via convolutional neural networks. *IEEE Transactions on Intelligent Transportation Systems*, 21(5):2146–2154.

[Deng et al., 2016] Deng, T., Yang, K., Li, Y., and Yan, H. (2016). Where does the driver look? top-down-based saliency detection in a traffic driving environment. *IEEE Transactions on Intelligent Transportation Systems*, 17(7):2051–2062.

[Deo and Trivedi, 2019] Deo, N. and Trivedi, M. M. (2019). Looking at the driver/rider in autonomous vehicles to predict take-over readiness. *IEEE Transactions on Intelligent Vehicles*.

[Dillen et al., 2020] Dillen, N., Ilievski, M., Law, E., Nacke, L. E., Czarnecki, K., and Schneider, O. (2020). Keep calm and ride along: Passenger comfort and anxiety as physiological responses to autonomous driving styles. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–13.

[Do et al., 2017] Do, Q. H., Tehrani, H., Mita, S., Egawa, M., Muto, K., and Yoneda, K. (2017). Human drivers based active-passive model for automated lane change. *IEEE Intelligent Transportation Systems Magazine*, 9(1):42–56.

[Dogan et al., 2019] Dogan, E., Honnêt, V., Masfrand, S., and Guillaume, A. (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.

[Doshi et al., 2008] Doshi, A., Cheng, S. Y., and Trivedi, M. M. (2008). A novel active heads-up display for driver assistance. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 39(1):85–93.

[Dosovitskiy et al., 2017] Dosovitskiy, A., Ros, G., Codevilla, F., Lopez, A., and Koltun, V. (2017). Carla: An open urban driving simulator. In *Conference on robot learning*, pages 1–16. PMLR.

[Du et al., 2020a] Du, N., Kim, J., Zhou, F., Pulver, E., Tilbury, D., Robert, L., Pradhan, A., Yang, X. J., et al. (2020a). Evaluating effects of cognitive load, takeover request lead time, and traffic density on drivers' takeover performance in conditionally automated driving. *AutomotiveUI*.

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

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

[Du et al., 2020d] Du, N., Zhou, F., Pulver, E. M., Tilbury, D. M., Robert, L. P., Pradhan, A. K., and Yang, X. J. (2020d). Predicting driver takeover performance in conditionally automated driving. *Accident Analysis & Prevention*, 148:105748.

[Ebnali et al., 2019] Ebnali, M., Hulme, K., Ebnali-Heidari, A., and Mazloumi, A. (2019). How does training effect users' attitudes and skills needed for highly automated driving? *Transportation research part F: traffic psychology and behaviour*, 66:184–195.

[Einhäuser et al., 2020] Einhäuser, W., Atzert, C., and Nuthmann, A. (2020). Fixation durations in natural scene viewing are guided by peripheral scene content. *Journal of vision*, 20(4):15–15.

[Ekman et al., 2017] Ekman, F., Johansson, M., and Sochor, J. (2017). Creating appropriate trust in automated vehicle systems: A framework for hmi design. *IEEE Transactions on Human-Machine Systems*, 48(1):95–101.

[Endsley, 2017] Endsley, M. R. (2017). Autonomous driving systems: A preliminary naturalistic study of the tesla model s. *Journal of Cognitive Engineering and Decision Making*, 11(3):225–238.

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

[Eriksson et al., 2017] Eriksson, A., Banks, V., and Stanton, N. (2017). Transition to manual: Comparing simulator with on-road control transitions. *Accident Analysis & Prevention*, 102:227–234.

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

[Fan et al., 2019] Fan, D.-P., Wang, W., Cheng, M.-M., and Shen, J. (2019). Shifting more attention to video salient object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8554–8564.

[Fang et al., 2019] Fang, J., Yan, D., Qiao, J., Xue, J., Wang, H., and Li, S. (2019). Dada-2000: Can driving accident be predicted by driver attention analyzed by a benchmark. In *2019 IEEE Intelligent Transportation Systems Conference (ITSC)*, pages 4303–4309. IEEE.

[Feldhütter et al., 2017] Feldhütter, A., Gold, C., Schneider, S., and Bengler, K. (2017). How the duration of automated driving influences take-over performance and gaze behavior. In *Advances in ergonomic design of systems, products and processes*, pages 309–318. Springer.

[Feldhütter et al., 2018] Feldhütter, A., Kroll, D., and Bengler, K. (2018). Wake up and take over! the effect of fatigue on the take-over performance in conditionally automated driving. In *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, pages 2080–2085. IEEE.

[Flemisch et al., 2012] Flemisch, F., Heesen, M., Hesse, T., Kelsch, J., Schieben, A., and Beller, J. (2012). Towards a dynamic balance between humans and automation: authority, ability, responsibility and control in shared and cooperative control situations. *Cognition, Technology & Work*, 14(1):3–18.

[Flemisch et al., 2008] Flemisch, F., Schieben, A., Kelsch, J., and Löper, C. (2008). Automation spectrum, inner/outer compatibility and other potentially useful human factors concepts for assistance and automation. *Human Factors for assistance and automation*.

[Foy and Chapman, 2018] Foy, H. J. and Chapman, P. (2018). Mental workload is reflected in driver behaviour, physiology, eye movements and prefrontal cortex activation. *Applied ergonomics*, 73:90–99.

[Fu et al., 2015] Fu, K., Gong, C., Gu, I. Y.-H., and Yang, J. (2015). Normalized cut-based saliency detection by adaptive multi-level region merging. *IEEE Transactions on Image Processing*, 24(12):5671–5683.

[Gao et al., 2019] Gao, M., Tawari, A., and Martin, S. (2019). Goal-oriented object importance estimation in on-road driving videos. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 5509–5515. IEEE.

[Geiger et al., 2012] Geiger, A., Lenz, P., and Urtasun, R. (2012). Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3354–3361. IEEE.

[Gerber et al., 2020] Gerber, M. A., Schroeter, R., Xiaomeng, L., and Elhenawy, M. (2020). Self-interruptions of non-driving related tasks in automated vehicles: Mobile vs head-up display. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–9.

[Gheorghe, 2017] Gheorghe, L. A. (2017). Detecting eeg correlates during preparation of complex driving maneuvers. Technical report, EPFL.

[Godard et al., 2019] Godard, C., Mac Aodha, O., Firman, M., and Brostow, G. J. (2019). Digging into self-supervised monocular depth estimation. In *Proceedings of the IEEE international conference on computer vision*, pages 3828–3838.

[Gold et al., 2013] Gold, C., Damböck, D., Lorenz, L., and Bengler, K. (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*, volume 57, pages 1938–1942. Sage Publications Sage CA: Los Angeles, CA.

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

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

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

[Gong et al., 2015] Gong, C., Tao, D., Liu, W., Maybank, S. J., Fang, M., Fu, K., and Yang, J. (2015). Saliency propagation from simple to difficult. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2531–2539.

[Goshvarpour et al., 2017] Goshvarpour, A., Abbasi, A., and Goshvarpour, A. (2017). An accurate emotion recognition system using ecg and gsr signals and matching pursuit method. *Biomedical journal*, 40(6):355–368.

[Grese et al., 2021] Grese, J. M., Pasareanu, C., and Pakdamanian, E. (2021). Formal analysis of a neural network predictor inshared-control autonomous driving. In *AIAA Scitech 2021 Forum*, page 1580.

[Griggs and Wakabayashi, 2018] Griggs, T. and Wakabayashi, D. (2018). How a self-driving uber killed a pedestrian in arizona. *The New York Times*, 13.

[Groeger and Banks, 2007] Groeger, J. A. and Banks, A. (2007). Anticipating the content and circumstances of skill transfer: Unrealistic expectations of driver training and graduated licensing? *Ergonomics*, 50(8):1250–1263.

[Guangyu Li et al., 2019] Guangyu Li, M., Jiang, B., Che, Z., Shi, X., Liu, M., Meng, Y., Ye, J., and Liu, Y. (2019). Dbus: Human driving behavior understanding system. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 0–0.

[Hart and Staveland, 1988] Hart, S. G. and Staveland, L. E. (1988). Development of nasa-tlx (task load index): Results of empirical and theoretical research. In *Advances in psychology*, volume 52, pages 139–183. Elsevier.

[Helbing et al., 2020] Helbing, J., Draschkow, D., and Võ, M. L.-H. (2020). Search superiority: Goal-directed attentional allocation creates more reliable incidental identity and location memory than explicit encoding in naturalistic virtual environments. *Cognition*, 196:104147.

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

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

[Holländer and Pfleging, 2018] Holländer, K. and Pfleging, B. (2018). Preparing drivers for planned control transitions in automated cars. In *Proceedings of the 17th International Conference on Mobile and Ubiquitous Multimedia*, pages 83–92.

[Hu et al., 2020] Hu, Z., Li, S., Zhang, C., Yi, K., Wang, G., and Manocha, D. (2020). Dgaze: Cnn-based gaze prediction in dynamic scenes. *IEEE transactions on visualization and computer graphics*, 26(5):1902–1911.

[iMotions, 2015a] iMotions (2015a). Affectiva imotions biometric research platform.

[iMotions, 2015b] iMotions, A. (2015b). Affectiva imotions biometric research platform.

[Iqbal et al., 2011] Iqbal, S. T., Horvitz, E., Ju, Y.-C., and Mathews, E. (2011). Hang on a sec! effects of proactive mediation of phone conversations while driving. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 463–472.

[ISO 21959:2020, 2020] ISO 21959:2020 (2020). Road vehicles — Human performance and state in the context of automated driving. Standard, International Organization for Standardization.

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

[Jiang et al., 2017] Jiang, L., Xu, M., and Wang, Z. (2017). Predicting video saliency with object-to-motion cnn and two-layer convolutional lstm. *arXiv preprint arXiv:1709.06316*.

[Johns et al., 2016] Johns, M., Mok, B., Sirkin, D., Gowda, N., Smith, C., Talamonti, W., and Ju, W. (2016). Exploring shared control in automated driving. In *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 91–98. IEEE.

[Katsuki and Constantinidis, 2014] Katsuki, F. and Constantinidis, C. (2014). Bottom-up and top-down attention: different processes and overlapping neural systems. *The Neuroscientist*, 20(5):509–521.

[Kerschbaum et al., 2015] Kerschbaum, P., Lorenz, L., and Bengler, K. (2015). A transforming steering wheel for highly automated cars. In *2015 ieee intelligent vehicles symposium (iv)*, pages 1287–1292. IEEE.

[Kim and Yang, 2017] Kim, H. J. and Yang, J. H. (2017). Takeover requests in simulated partially autonomous vehicles considering human factors. *IEEE Transactions on Human-Machine Systems*, 47(5):735–740.

[Kim and Canny, 2017] Kim, J. and Canny, J. (2017). Interpretable learning for self-driving cars by visualizing causal attention. In *Proceedings of the IEEE international conference on computer vision*, pages 2942–2950.

[Kim et al., 2020] Kim, J., Moon, S., Rohrbach, A., Darrell, T., and Canny, J. (2020). Advisable learning for self-driving vehicles by internalizing observation-to-action rules. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9661–9670.

[Kim et al., 2017] Kim, N., Jeong, K., Yang, M., Oh, Y., and Kim, J. (2017). " are you ready to take-over?" an exploratory study on visual assistance to enhance driver vigilance. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, pages 1771–1778.

[Kingma and Ba, 2014] Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

[Körber et al., 2018] Körber, M., Prasch, L., and Bengler, K. (2018). Why do i have to drive now? post hoc explanations of takeover requests. *Human factors*, 60(3):305–323.

[Kruthiventi et al., 2017] Kruthiventi, S. S., Ayush, K., and Babu, R. V. (2017). Deepfix: A fully convolutional neural network for predicting human eye fixations. *IEEE Transactions on Image Processing*, 26(9):4446–4456.

[Kummerer et al., 2017] Kummerer, M., Wallis, T. S., Gatys, L. A., and Bethge, M. (2017). Understanding low-and high-level contributions to fixation prediction. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4789–4798.

[Kutchek and Jeon, 2019] Kutchek, K. and Jeon, M. (2019). Takeover and handover requests using non-speech auditory displays in semi-automated vehicles. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–6.

[Lee and Yang, 2020] Lee, J. and Yang, J. H. (2020). Analysis of driver's eeg given take-over alarm in sae level 3 automated driving in a simulated environment. *International journal of automotive technology*, 21(3):719–728.

[Lee et al., 2017] Lee, S., Kim, J., Shin Yoon, J., Shin, S., Bailo, O., Kim, N., Lee, T.-H., Seok Hong, H., Han, S.-H., and So Kweon, I. (2017). Vpgnet: Vanishing point guided network for lane and road marking detection and recognition. In *Proceedings of the IEEE international conference on computer vision*, pages 1947–1955.

[Lei et al., 2009] Lei, S., Welke, S., and Roetting, M. (2009). Driver's mental workload assessment using eeg data in a dual task paradigm. In *In Proceedings of 21 st International Technical Conference on the Enhanced Safety of Vehicle*. Citeseer.

[Levin and Woolf, 2016] Levin, S. and Woolf, N. (2016). Tesla driver killed while using autopilot was watching harry potter, witness says. *The Guardian*, 1.

[Li and Yu, 2015] Li, G. and Yu, Y. (2015). Visual saliency based on multiscale deep features. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5455–5463.

[Li et al., 2012] Li, W.-C., Chiu, F.-C., and Wu, K.-J. (2012). The evaluation of pilots performance and mental workload by eye movement.

[Liu et al., 2019] Liu, C., Chen, Y., Tai, L., Ye, H., Liu, M., and Shi, B. E. (2019). A gaze model improves autonomous driving. In *Proceedings of the 11th ACM Symposium on Eye Tracking Research & Applications*, pages 1–5.

[Ljungberg et al., 2012] Ljungberg, J. K., Parmentier, F. B., Hughes, R. W., Macken, W. J., and Jones, D. M. (2012). Listen out! behavioural and subjective responses to verbal warnings. *Applied Cognitive Psychology*, 26(3):451–461.

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

[Lopez-Calderon and Luck, 2014] Lopez-Calderon, J. and Luck, S. J. (2014). Erplab: an open-source toolbox for the analysis of event-related potentials. *Frontiers in human neuroscience*, 8:213.

[Lorenz et al., 2014] Lorenz, L., Kerschbaum, P., and Schumann, J. (2014). Designing take over scenarios for automated driving: How does augmented reality support the driver to get back into the loop? In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 58, pages 1681–1685. SAGE Publications Sage CA: Los Angeles, CA.

[Lotz and Weissenberger, 2018] Lotz, A. and Weissenberger, S. (2018). Predicting take-over times of truck drivers in conditional autonomous driving. In *International Conference on Applied Human Factors and Ergonomics*, pages 329–338. Springer.

[Louw et al., 2015] Louw, T., Kountouriotis, G., Carsten, O., and Merat, N. (2015). Driver inattention during vehicle automation: how does driver engagement affect resumption of control? In *4th International Conference on Driver Distraction and Inattention (DDI2015), Sydney: proceedings*. ARRB Group.

[Louw et al., 2019] Louw, T., Kuo, J., Romano, R., Radhakrishnan, V., Lenné, M. G., and Merat, N. (2019). Engaging in ndrts affects drivers' responses and glance patterns after silent automation failures. *Transportation research part F: traffic psychology and behaviour*, 62:870–882.

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

[Lu et al., 2017] Lu, Z., Coster, X., and De Winter, J. (2017). How much time do drivers need to obtain situation awareness? a laboratory-based study of automated driving. *Applied ergonomics*, 60:293–304.

[Lundberg and Lee, 2017] Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

[Luo et al., 2019] Luo, R., Wang, Y., Weng, Y., Paul, V., Brudnak, M. J., Jayakumar, P., Reed, M., Stein, J. L., Ersal, T., and Yang, X. J. (2019). Toward real-time assessment of workload: A bayesian inference approach. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 63, pages 196–200. SAGE Publications Sage CA: Los Angeles, CA.

[Lyu et al., 2020] Lyu, F., Cheng, N., Zhu, H., Zhou, H., Xu, W., Li, M., and Shen, X. (2020). Towards rear-end collision avoidance: adaptive beaconing for connected vehicles. *IEEE Transactions on Intelligent Transportation Systems*.

[Maag et al., 2015] Maag, C., Schneider, N., Lübbeke, T., Weisswange, T. H., and Goerick, C. (2015). Car gestures–advisory warning using additional steering wheel angles. *Accident Analysis & Prevention*, 83:143–153.

[Manawadu et al., 2018] Manawadu, U. E., Hayashi, H., Ema, T., Kawano, T., Kamezaki, M., and Sugano, S. (2018). Tactical-level input with multimodal feedback for unscheduled takeover situations in human-centered automated vehicles. In *2018 IEEE/ASME International Conference on Advanced Intelligent Mechatronics (AIM)*, pages 634–639. IEEE.

[Marberger et al., 2017] Marberger, C., Mielenz, H., Naujoks, F., Radlmayr, J., Bengler, K., and Wandtner, B. (2017). Understanding and applying the concept of "driver availability" in automated driving. In *international conference on applied human factors and ergonomics*, pages 595–605. Springer.

[Martelaro et al., 2019] Martelaro, N., Teevan, J., and Iqbal, S. T. (2019). An exploration of speech-based productivity support in the car. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–12.

[Mathe and Sminchisescu, 2013] Mathe, S. and Sminchisescu, C. (2013). Action from still image dataset and inverse optimal control to learn task specific visual scanpaths. In *Advances in neural information processing systems*, pages 1923–1931.

[Mathe and Sminchisescu, 2014] Mathe, S. and Sminchisescu, C. (2014). Actions in the eye: Dynamic gaze datasets and learnt saliency models for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 37(7):1408–1424.

[McCall et al., 2019] McCall, R., McGee, F., Mirnig, A., Meschtscherjakov, A., Louveton, N., Engel, T., and Tscheligi, M. (2019). A taxonomy of autonomous vehicle handover situations. *Transportation research part A: policy and practice*, 124:507–522.

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

[Mehler et al., 2012] Mehler, B., Reimer, B., and Coughlin, J. F. (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.

[Melcher et al., 2015] Melcher, V., Rauh, S., Diederichs, F., Widlroither, H., and Bauer, W. (2015). Take-over requests for automated driving. *Procedia Manufacturing*, 3:2867–2873.

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

[Merat et al., 2014] Merat, N., Jamson, A. H., Lai, F. C., Daly, M., and Carsten, O. M. (2014). Transition to manual: Driver behaviour when resuming control from a highly automated vehicle. *Transportation research part F: traffic psychology and behaviour*, 27:274–282.

[Min and Corso, 2019] Min, K. and Corso, J. J. (2019). Tased-net: Temporally-aggregating spatial encoder-decoder network for video saliency detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2394–2403.

[Mok et al., 2017] Mok, B., Johns, M., Miller, D., and Ju, W. (2017). Tunneled in: Drivers with active secondary tasks need more time to transition from automation. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 2840–2844.

[Montgomery, 2018] Montgomery, W. D. (2018). Public and private benefits of autonomous vehicles.

[Morrell and Wasilewski, 2010] Morrell, J. and Wasilewski, K. (2010). Design and evaluation of a vibrotactile seat to improve spatial awareness while driving. In *2010 IEEE Haptics Symposium*, pages 281–288. IEEE.

[Naujoks et al., 2017] Naujoks, F., Befelein, D., Wiedemann, K., and Neukum, A. (2017). A review of non-driving-related tasks used in studies on automated driving. In *International Conference on Applied Human Factors and Ergonomics*, pages 525–537. Springer.

[Naujoks et al., 2014] Naujoks, F., Mai, C., and Neukum, A. (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 et al., 2016] Naujoks, F., Purucker, C., and Neukum, A. (2016). Secondary task engagement and vehicle automation–comparing the effects of different automation levels in an on-road experiment. *Transportation research part F: traffic psychology and behaviour*, 38:67–82.

[Naujoks et al., 2019] Naujoks, F., Purucker, C., Wiedemann, K., and Marberger, C. (2019). Noncritical state transitions during conditionally automated driving on german freeways: Effects of non–driving related tasks on takeover time and takeover quality. *Human factors*, 61(4):596–613.

[Neuhold et al., 2017] Neuhold, G., Ollmann, T., Rota Bulo, S., and Kontschieder, P. (2017). The mapillary vistas dataset for semantic understanding of street scenes. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4990–4999.

[Ng et al., 2000] Ng, A. Y., Russell, S. J., et al. (2000). Algorithms for inverse reinforcement learning. In *Icml*, volume 1, page 2.

[Nourbakhsh et al., 2012] Nourbakhsh, N., Wang, Y., Chen, F., and Calvo, R. A. (2012). Using galvanic skin response for cognitive load measurement in arithmetic and reading tasks. In *Proceedings of the 24th Australian Computer-Human Interaction Conference*, pages 420–423.

[Noy et al., 2018] Noy, I. Y., Shinar, D., and Horrey, W. J. (2018). Automated driving: Safety blind spots. *Safety science*, 102:68–78.

[Ntousakis et al., 2015] Ntousakis, I. A., Nikolos, I. K., and Papageorgiou, M. (2015). On microscopic modelling of adaptive cruise control systems. *Transportation Research Procedia*, 6:111–127.

[Ohn-Bar et al., 2020] Ohn-Bar, E., Prakash, A., Behl, A., Chitta, K., and Geiger, A. (2020). Learning situational driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11296–11305.

[Olaverri-Monreal, 2020] Olaverri-Monreal, C. (2020). Promoting trust in self-driving vehicles. *Nature Electronics*, 3(6):292–294.

[Olaverri-Monreal and Jizba, 2016] Olaverri-Monreal, C. and Jizba, T. (2016). Human factors in the design of human–machine interaction: An overview emphasizing v2x communication. *IEEE Transactions on Intelligent Vehicles*, 1(4):302–313.

[Pakdamanian et al., 2018] Pakdamanian, E., Feng, L., and Kim, I. (2018). The effect of whole-body haptic feedback on driver's perception in negotiating a curve. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, pages 19–23. SAGE Publications Sage CA: Los Angeles, CA.

[Pakdamanian et al., 2020] Pakdamanian, E., Namaky, N., Sheng, S., Kim, I., Coan, J. A., and Feng, L. (2020). Toward minimum startle after take-over request: A preliminary study of physiological data. In *12th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, pages 27–29.

[Pakdamanian et al., 2021] Pakdamanian, E., Sheng, S., Baee, S., Heo, S., Kraus, S., and Feng, L. (2021). Deeptake: Prediction of driver takeover behavior using multimodal data. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–14.

[Pal et al., 2020] Pal, A., Mondal, S., and Christensen, H. I. (2020). "looking at the right stuff"-guided semantic-gaze for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11883–11892.

[Palazzi et al., 2018] Palazzi, A., Abati, D., Solera, F., Cucchiara, R., et al. (2018). Predicting the driver's focus of attention: the dr (eye) ve project. *IEEE transactions on pattern analysis and machine intelligence*, 41(7):1720–1733.

[Parasuraman and Wickens, 2008] Parasuraman, R. and Wickens, C. D. (2008). Humans: Still vital after all these years of automation. *Human factors*, 50(3):511–520.

[Park et al., 2018a] Park, S., Spurr, A., and Hilliges, O. (2018a). Deep pictorial gaze estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 721–738.

[Park et al., 2018b] Park, S., Zhang, X., Bulling, A., and Hilliges, O. (2018b). Learning to find eye region landmarks for remote gaze estimation in unconstrained settings. In *Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications*, pages 1–10.

[Pauzié, 2008] Pauzié, A. (2008). A method to assess the driver mental workload: The driving activity load index (dali). *IET Intelligent Transport Systems*, 2(4):315–322.

[Paxion et al., 2014] Paxion, J., Galy, E., and Berthelon, C. (2014). Mental workload and driving frontiers in psychology.

[Peruzzini et al., 2019] Peruzzini, M., Tonietti, M., and Iani, C. (2019). Transdisciplinary design approach based on driver's workload monitoring. *Journal of Industrial Information Integration*, 15:91–102.

[Petermeijer et al., 2017a] Petermeijer, S., Doubek, F., and de Winter, J. (2017a). Driver response times to auditory, visual, and tactile take-over requests: A simulator study with 101 participants. In *2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 1505–1510. IEEE.

[Petermeijer et al., 2017b] Petermeijer, S. M., Cieler, S., and De Winter, J. C. (2017b). Comparing spatially static and dynamic vibrotactile take-over requests in the driver seat. *Accident analysis & prevention*, 99:218–227.

[Politis et al., 2015a] Politis, I., Brewster, S., and Pollick, F. (2015a). Language-based multimodal displays for the handover of control in autonomous cars. In *Proceedings of the 7th international conference on automotive user interfaces and interactive vehicular applications*, pages 3–10.

[Politis et al., 2015b] Politis, I., Brewster, S., and Pollick, F. (2015b). To beep or not to beep? comparing abstract versus language-based multimodal driver displays. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems*, pages 3971–3980.

[Politis et al., 2018] Politis, I., Langdon, P., Adebayo, D., Bradley, M., Clarkson, P. J., Skrypchuk, L., Mouzakitis, A., Eriksson, A., Brown, J. W., Revell, K., et al. (2018). An evaluation of inclusive dialogue-based interfaces for the takeover of control in autonomous cars. In *23rd International Conference on Intelligent User Interfaces*, pages 601–606.

[Prewett et al., 2011] Prewett, M. S., Elliott, L. R., Walvoord, A. G., and Coovert, M. D. (2011). A meta-analysis of vibrotactile and visual information displays for improving task performance. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 42(1):123–132.

[Radlmayr et al., 2014] Radlmayr, J., Gold, C., Lorenz, L., Farid, M., and Bengler, K. (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*, volume 58, pages 2063–2067. Sage Publications Sage CA: Los Angeles, CA.

[Rahman et al., 2019] Rahman, M. M., Deb, S., Strawderman, L., Burch, R., and Smith, B. (2019). How the older population perceives self-driving vehicles. *Transportation research part F: traffic psychology and behaviour*, 65:242–257.

[Ramanishka et al., 2018] Ramanishka, V., Chen, Y.-T., Misu, T., and Saenko, K. (2018). Toward driving scene understanding: A dataset for learning driver behavior and causal reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[Recarte and Nunes, 2000] Recarte, M. A. and Nunes, L. M. (2000). Effects of verbal and spatial-imagery tasks on eye fixations while driving. *Journal of experimental psychology: Applied*, 6(1):31.

[Reimer et al., 2011] Reimer, B., Mehler, B., Coughlin, J. F., Roy, N., and Dusek, J. A. (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.

[Ribeiro et al., 2016] Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. *CoRR*, abs/1602.04938.

[Rothkopf et al., 2007] Rothkopf, C. A., Ballard, D. H., and Hayhoe, M. M. (2007). Task and context determine where you look. *Journal of vision*, 7(14):16–16.

[Ruscio et al., 2015] Ruscio, D., Ciceri, M. R., and Biassoni, F. (2015). How does a collision warning system shape driver's brake response time? the influence of expectancy and automation complacency on real-life emergency braking. *Accident Analysis & Prevention*, 77:72–81.

[Sadeghian Borojeni et al., 2018] Sadeghian Borojeni, S., Boll, S. C., Heuten, W., Bülthoff, H. H., and Chuang, L. (2018). Feel the movement: Real motion influences responses to take-over requests in highly automated vehicles. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–13.

[SAE, 2014] SAE (2014). International: on-road automated vehicle standards committee. *Taxonomy and definitions for terms related to on-road motor vehicle automated driving systems, information report*.

[SAE, 2018] SAE (2018). Taxonomy and definitions for terms related to driving automation systems for on-road motor vehicles. *SAE International: Warrendale, PA, USA*.

[Saffarian et al., 2012] Saffarian, M., De Winter, J. C., and Happee, R. (2012). Automated driving: human-factors issues and design solutions. In *Proceedings of the human factors and ergonomics society annual meeting*, volume 56, pages 2296–2300. Sage Publications Sage CA: Los Angeles, CA.

[Saha et al., 2017] Saha, A., Konar, A., and Nagar, A. K. (2017). Eeg analysis for cognitive failure detection in driving using type-2 fuzzy classifiers. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 1(6):437–453.

[Salminen et al., 2019] Salminen, K., Farooq, A., Rantala, J., Surakka, V., and Raisamo, R. (2019). Unimodal and multimodal signals to support control transitions in semiautonomous vehicles. In *Proceedings of the 11th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, pages 308–318.

[Sanghavi et al., 2021] Sanghavi, H., Jeon, M., Nadri, C., Ko, S., Sodnik, J., and Stojmenova, K. (2021). Multimodal takeover request displays for semi-automated vehicles: Focused on spatiality and lead time. In *International Conference on Human-Computer Interaction*, pages 315–334. Springer.

[Sanghavi et al., 2020] Sanghavi, H., Zhang, Y., and Jeon, M. (2020). Effects of anger and display urgency on takeover performance in semi-automated vehicles. In *12th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, pages 48–56.

[Schwall et al., 2020] Schwall, M., Daniel, T., Victor, T., Favaro, F., and Hohnhold, H. (2020). Waymo public road safety performance data. *arXiv preprint arXiv:2011.00038*.

[Seeliger et al., 2014] Seeliger, F., Weidl, G., Petrich, D., Naujoks, F., Breuel, G., Neukum, A., and Dietmayer, K. (2014). Advisory warnings based on cooperative perception. In *2014 IEEE Intelligent Vehicles Symposium Proceedings*, pages 246–252. IEEE.

[Seppelt and Lee, 2019] Seppelt, B. D. and Lee, J. D. (2019). Keeping the driver in the loop: Dynamic feedback to support appropriate use of imperfect vehicle control automation. *International Journal of Human-Computer Studies*, 125:66–80.

[Shah et al., 2015] Shah, S. J., Bliss, J. P., Chancey, E. T., and Brill, J. C. (2015). Effects of alarm modality and alarm reliability on workload, trust, and driving performance. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 59, pages 1535–1539. SAGE Publications Sage CA: Los Angeles, CA.

[Shanker et al., 2013] Shanker, R., Jonas, A., Devitt, S., Huberty, K., Flannery, S., Greene, W., Swinburne, B., Locraft, G., Wood, A., Weiss, K., et al. (2013). Autonomous cars: Self-driving the new auto industry paradigm. *Morgan Stanley blue paper*, pages 1–109.

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

[Sprague and Ballard, 2004] Sprague, N. and Ballard, D. (2004). Eye movements for reward maximization. In *Advances in neural information processing systems*, pages 1467–1474.

[Stanton, 2015] Stanton, N. A. (2015). Responses to autonomous vehicles. *Ingenia*, 62(9):221–233.

[Stanton et al., 2011] Stanton, N. A., Dunoyer, A., and Leatherland, A. (2011). Detection of new in-path targets by drivers using stop & go adaptive cruise control. *Applied ergonomics*, 42(4):592–601.

[Stojić et al., 2020] Stojić, H., Orquin, J. L., Dayan, P., Dolan, R. J., and Speekenbrink, M. (2020). Uncertainty in learning, choice, and visual fixation. *Proceedings of the National Academy of Sciences*, 117(6):3291–3300.

[Strand et al., 2014] Strand, N., Nilsson, J., Karlsson, I. M., and Nilsson, L. (2014). Semi-automated versus highly automated driving in critical situations caused by automation failures. *Transportation research part F: traffic psychology and behaviour*, 27:218–228.

[Sundararajan et al., 2017] Sundararajan, M., Taly, A., and Yan, Q. (2017). Axiomatic attribution for deep networks. *CoRR*, abs/1703.01365.

[Tang et al., 2020] Tang, Q., Guo, G., Zhang, Z., Zhang, B., and Wu, Y. (2020). Olfactory facilitation of takeover performance in highly automated driving. *Human Factors*, page 0018720819893137.

[Tarvainen et al., 2014] Tarvainen, M. P., Niskanen, J.-P., Lipponen, J. A., Ranta-Aho, P. O., and Karjalainen, P. A. (2014). Kubios hrv–heart rate variability analysis software. *Computer methods and programs in biomedicine*, 113(1):210–220.

[Tatler et al., 2011] Tatler, B. W., Hayhoe, M. M., Land, M. F., and Ballard, D. H. (2011). Eye guidance in natural vision: Reinterpreting salience. *Journal of vision*, 11(5):5–5.

[Tawari and Kang, 2017] Tawari, A. and Kang, B. (2017). A computational framework for driver's visual attention using a fully convolutional architecture. In *2017 IEEE Intelligent Vehicles Symposium (IV)*, pages 887–894. IEEE.

[Telpaz et al., 2015] Telpaz, A., Rhindress, B., Zelman, I., and Tsimhoni, O. (2015). Haptic seat for automated driving: preparing the driver to take control effectively. In *Proceedings of the 7th international conference on automotive user interfaces and interactive vehicular applications*, pages 23–30.

[Trösterer et al., 2017] Trösterer, S., Meschtscherjakov, A., Mirnig, A. G., Lupp, A., Gärtner, M., McGee, F., McCall, R., Tscheligi, M., and Engel, T. (2017). What we can learn from pilots for handovers and (de) skilling in semi-autonomous driving: An interview study. In *Proceedings of the 9th international conference on automotive user interfaces and interactive vehicular applications*, pages 173–182.

[Tschiatschek et al., 2019] Tschiatschek, S., Ghosh, A., Haug, L., Devidze, R., and Singla, A. (2019). Learner-aware teaching: Inverse reinforcement learning with preferences and constraints. In *Advances in Neural Information Processing Systems*, pages 4145–4155.

[Ungerleider and G, 2000] Ungerleider, S. K. and G, L. (2000). Mechanisms of visual attention in the human cortex. *Annual review of neuroscience*, 23(1):315–341.

[Urmson, 2015] Urmson, C. (2015). The view from the front seat of the google self-driving car. *Backchannel. Retrived from https://backchannel. com/the-view-from-the-front-seat-of-the-google-self-driving-car-46fc9f3e6088*.

[van der Heiden et al., 2017] van der Heiden, R. M., Iqbal, S. T., and Janssen, C. P. (2017). Priming drivers before handover in semi-autonomous cars. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 392–404.

[Van der Heiden et al., 2021] Van der Heiden, R. M., Kenemans, J. L., Donker, S. F., and Janssen, C. P. (2021). The effect of cognitive load on auditory susceptibility during automated driving. *Human factors*, page 0018720821998850.

[van Gent et al., 2019] van Gent, P., Farah, H., van Nes, N., and van Arem, B. (2019). Heartpy: A novel heart rate algorithm for the analysis of noisy signals. *Transportation research part F: traffic psychology and behaviour*, 66:368–378.

[Vicente et al., 2011] Vicente, J., Laguna, P., Bartra, A., and Bailón, R. (2011). Detection of driver's drowsiness by means of hrv analysis. In *2011 Computing in Cardiology*, pages 89–92. IEEE.

[Võ et al., 2019] Võ, M. L.-H., Boettcher, S. E., and Draschkow, D. (2019). Reading scenes: how scene grammar guides attention and aids perception in real-world environments. *Current opinion in psychology*, 29:205–210.

[Vogelpohl et al., 2018] Vogelpohl, T., Kühn, M., Hummel, T., Gehlert, T., and Vollrath, M. (2018). Transitioning to manual driving requires additional time after automation deactivation. *Transportation research part F: traffic psychology and behaviour*, 55:464–482.

[Walch et al., 2017] Walch, M., Mühl, K., Kraus, J., Stoll, T., Baumann, M., and Weber, M. (2017). From car-driver-handovers to cooperative interfaces: Visions for driver–vehicle interaction in automated driving. In *Automotive user interfaces*, pages 273–294. Springer.

[Wan and Wu, 2018] Wan, J. and Wu, C. (2018). The effects of vibration patterns of take-over request and non-driving tasks on taking-over control of automated vehicles. *International Journal of Human–Computer Interaction*, 34(11):987–998.

[Wang et al., 2020] Wang, J., Sun, K., Cheng, T., Jiang, B., Deng, C., Zhao, Y., Liu, D., Mu, Y., Tan, M., Wang, X., et al. (2020). Deep high-resolution representation learning for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*.

[Wang and Shen, 2017] Wang, W. and Shen, J. (2017). Deep visual attention prediction. *IEEE Transactions on Image Processing*, 27(5):2368–2378.

[Wang et al., 2018] Wang, W., Shen, J., Guo, F., Cheng, M.-M., and Borji, A. (2018). Revisiting video saliency: A large-scale benchmark and a new model. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4894–4903.

[Wang et al., 2015] Wang, W., Shen, J., and Shao, L. (2015). Consistent video saliency using local gradient flow optimization and global refinement. *IEEE Transactions on Image Processing*, 24(11):4185–4196.

[Wang et al., 2019] Wang, W., Song, H., Zhao, S., Shen, J., Zhao, S., Hoi, S. C., and Ling, H. (2019). Learning unsupervised video object segmentation through visual attention. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3064–3074.

[Wang et al., 2014] Wang, Y., Reimer, B., Dobres, J., and Mehler, B. (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.

[Warshawsky-Livne and Shinar, 2002] Warshawsky-Livne, L. and Shinar, D. (2002). Effects of uncertainty, transmission type, driver age and gender on brake reaction and movement time. *Journal of safety research*, 33(1):117–128.

[Weaver and DeLucia, 2020] Weaver, B. W. and DeLucia, P. R. (2020). A systematic review and meta-analysis of takeover performance during conditionally automated driving. *Human factors*, page 0018720820976476.

[Wickens, 2002] Wickens, C. D. (2002). Multiple resources and performance prediction. *Theoretical issues in ergonomics science*, 3(2):159–177.

[Winter et al., 2016] Winter, J. D., Stanton, N. A., Price, J. S., and Mistry, H. (2016). The effects of driving with different levels of unreliable automation on self-reported workload and secondary task performance. *International journal of vehicle design*, 70(4):297–324.

[Wintersberger et al., 2018] Wintersberger, P., Riener, A., Schartmüller, C., Frison, A.-K., and Weigl, K. (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*, pages 53–65.

[Wright et al., 2017] Wright, T. J., Svancara, A., and Horrey, W. J. (2017). An initial review of the instructional and operational variability among automated systems in passenger vehicles. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 61, pages 1979–1979. SAGE Publications Sage CA: Los Angeles, CA.

[Wu and Boyle, 2021] Wu, X. and Boyle, L. N. (2021). Auditory messages for intersection movement assist (ima) systems: effects of speech-and nonspeech-based cues. *Human factors*, 63(2):336–347.

[Wu et al., 2019] Wu, Y., Kihara, K., Takeda, Y., Sato, T., Akamatsu, M., and Kitazaki, S. (2019). Assessing the mental states of fallback-ready drivers in automated driving by electrooculography. In *2019 IEEE Intelligent Transportation Systems Conference (ITSC)*, pages 4018–4023. IEEE.

[Wulfmeier et al., 2015] Wulfmeier, M., Ondruska, P., and Posner, I. (2015). Maximum entropy deep inverse reinforcement learning. *arXiv preprint arXiv:1507.04888*.

[Xia et al., 2020] Xia, Y., Kim, J., Canny, J., Zipser, K., Canas-Bajo, T., and Whitney, D. (2020). Periphery-fovea multi-resolution driving model guided by human attention. In *The IEEE Winter Conference on Applications of Computer Vision*, pages 1767–1775.

[Xia et al., 2018] Xia, Y., Zhang, D., Kim, J., Nakayama, K., Zipser, K., and Whitney, D. (2018). Predicting driver attention in critical situations. In *Asian conference on computer vision*, pages 658–674. Springer.

[Xiong et al., 2019] Xiong, Y., Kim, H. J., and Singh, V. (2019). Mixed effects neural networks (menets) with applications to gaze estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[Xu et al., 2017] Xu, H., Gao, Y., Yu, F., and Darrell, T. (2017). End-to-end learning of driving models from large-scale video datasets. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2174–2182.

[Yang and Coughlin, 2014] Yang, J. and Coughlin, J. F. (2014). In-vehicle technology for self-driving cars: Advantages and challenges for aging drivers. *International Journal of Automotive Technology*, 15(2):333–340.

[Yang et al., 2019] Yang, L., Fan, Y., and Xu, N. (2019). Video instance segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5188–5197.

[Yang et al., 2020] Yang, Z., Huang, L., Chen, Y., Wei, Z., Ahn, S., Zelinsky, G., Samaras, D., and Hoai, M. (2020). Predicting goal-directed human attention using inverse reinforcement learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 193–202.

[Yoon et al., 2019] Yoon, S. H., Kim, Y. W., and Ji, Y. G. (2019). The effects of takeover request modalities on highly automated car control transitions. *Accident Analysis & Prevention*, 123:150–158.

[Yoon et al., 2021] Yoon, S. H., Lee, S. C., and Ji, Y. G. (2021). Modeling takeover time based on non-driving-related task attributes in highly automated driving. *Applied ergonomics*, 92:103343.

[Young and Stanton, 2002] Young, M. S. and Stanton, N. A. (2002). Attention and automation: new perspectives on mental underload and performance. *Theoretical issues in ergonomics science*, 3(2):178–194.

[Yu et al., 2020] Yu, F., Chen, H., Wang, X., Xian, W., Chen, Y., Liu, F., Madhavan, V., and Darrell, T. (2020). Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2636–2645.

[Yun et al., 2013] Yun, K., Peng, Y., Samaras, D., Zelinsky, G. J., and Berg, T. L. (2013). Studying relationships between human gaze, description, and computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 739–746.

[Zahedian et al., 2019] Zahedian, S., Sadabadi, K. F., and Nohekhan, A. (2019). Localization of autonomous vehicles: Proof of concept for a computer vision approach. In *2019 ITS America Annual Meeting*. ITSWC.

[Zeeb et al., 2015] Zeeb, K., Buchner, A., and Schrauf, M. (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 et al., 2016] Zeeb, K., Buchner, A., and Schrauf, M. (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 et al., 2017] Zeeb, K., Härtel, M., Buchner, A., and Schrauf, M. (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.

[Zelinsky et al., 2019] Zelinsky, G., Yang, Z., Huang, L., Chen, Y., Ahn, S., Wei, Z., Adeli, H., Samaras, D., and Hoai, M. (2019). Benchmarking gaze prediction for categorical visual search. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0.

[Zelinsky et al., 2020] Zelinsky, G. J., Chen, Y., Ahn, S., Adeli, H., Yang, Z., Huang, L., Samaras, D., and Hoai, M. (2020). Predicting goal-directed attention control using inverse-reinforcement learning. *arXiv preprint arXiv:2001.11921*.

[Zhang et al., 2019a] Zhang, B., de Winter, J., Varotto, S., Happee, R., and Martens, M. (2019a). 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.

[Zhang et al., 2019b] Zhang, B., Wilschut, E. S., Willemsen, D. M., and Martens, M. H. (2019b). Transitions to manual control from highly automated driving in non-critical truck platooning scenarios. *Transportation research part F: traffic psychology and behaviour*, 64:84–97.

[Zhang et al., 2018] Zhang, R., Liu, Z., Zhang, L., Whritner, J. A., Muller, K. S., Hayhoe, M. M., and Ballard, D. H. (2018). Agil: Learning attention from human for visuomotor tasks. In *Proceedings of the european conference on computer vision (eccv)*, pages 663–679.

[Zheng et al., 2018] Zheng, Z., Oh, J., and Singh, S. (2018). On learning intrinsic rewards for policy gradient methods. In *Advances in Neural Information Processing Systems*, pages 4644–4654.

[Zhong et al., 2013] Zhong, S.-h., Liu, Y., Ren, F., Zhang, J., and Ren, T. (2013). Video saliency detection via dynamic consistent spatio-temporal attention modelling. In *Twenty-seventh AAAI Conference on Artificial Intelligence*.

[Ziebart et al., 2008] Ziebart, B. D., Maas, A. L., Bagnell, J. A., and Dey, A. K. (2008). Maximum entropy inverse reinforcement learning. In *Aaai*, volume 8, pages 1433–1438. Chicago, IL, USA.