

# 博士論文

Predicting Readiness and Performance of a  
Driver for Safe and Personalized Transitions  
from Automated to Manual Driving

(安全でパーソナライズされた自動から手動運転への遷移  
を行うためのドライバのレディネスと運転能力の予測)

黄 超



THE UNIVERSITY OF TOKYO

DOCTORAL THESIS

---

**Predicting Readiness and Performance of a Driver  
for Safe and Personalized Transitions from  
Automated to Manual Driving**

---

*Author:*

Chao HUANG

*Supervisor:*

Prof. Kimihiko NAKANO

*A thesis submitted in fulfillment of the requirements  
for the degree of Doctor of Philosophy  
in the*

Department of Mechanical Engineering

School of Engineering

February 12, 2024





# Abstract

Automated driving has been believed to be a promising technology in reducing road traffic crashes and releasing drivers from the heavy burden of driving tasks. However, due to various technical, legal and societal requirements, automated vehicles may still need to hand over control to drivers occasionally, known as takeover. This is a challenging task in that it is difficult to decide the sufficient time for a driver to safely take over the vehicle. This denotes the necessity for automated driving systems to constantly monitor the driver and predict drivers' takeover behaviors and adapt the system to ensure safe takeovers.

To build such a prediction model, a variety of factors need to be taken into consideration. Therefore, in this research, a thorough literature review was conducted firstly by classifying the influencing factors into system-, scenario-, and human-related factors, along with a complete architecture for evaluation and prediction of drivers' takeover behaviors. Based on this, two new factors impacting drivers' takeover performance that have not been studied were identified, including the time interval before the TOR (denoted as duration of monitoring) and drivers' personality traits as obtained from the big-five personality tests. Results showed that their effects were indeed statistically significant. Curious were that effects of duration of monitoring seemed to be nonlinear, and effects of different aspects of personality seemed to affect drivers' takeover performance in different ways.

Based on the above results, 38 features were identified from the dataset, and drivers' takeover behaviors were modeled in three aspects. Firstly, since takeover time is a continuous variable, it was modeled as a regression problem, and the best model was found out to be XGBoost regressor, with a mean absolute error less than 0.5 s. Secondly, since takeover readiness is a categorical variable, it was modeled as a classification problem, and again, the best model was found out to be XGBoost classifier, with both accuracy and recall over 95%. Finally, since takeover maneuvers were time series data, takeover style was modeled as a clustering problem, and dynamic-time-warping-based clustering was utilized to classify drivers' takeover maneuvers into three distinctive styles. Besides, for the regression and classification problems, SHAP explainer was utilized to better understand the effects of the most important features in predicting takeover time and readiness, giving us direct guidance on selecting the most important features in building the prediction model in an efficient way.

Finally, crashes and near-crashes data during the experiments were analyzed, and five patterns of unsafe behaviors were identified that may have caused those crashes and near-crashes. Correspondingly, several suggestions were put forward, which we believe could be utilized for more advanced HMI design and to mitigate these patterns of crashes.

# Acknowledgements

At the first place, I would like to express my sincere gratitude towards my supervisor, Prof. Kimihiko Nakano. I am deeply indebted to Prof. Nakano for his invaluable guidance, unwavering support, and scholarly insights throughout the entire research process. His mentorship has been instrumental in shaping the direction and quality of this research.

Secondly, I would like to extend my gratitude to the members of my thesis committee, Prof. Yoshihiro Suda, Prof. Yuji Yamakawa, Prof. Takanori Fukao and Dr. Yuki Asano from the University of Tokyo for their insightful feedback, scholarly guidance, and commitment to the development of this thesis.

Thirdly, I would like to thank Dr. Bo Yang from the University of Tokyo, who has given me much advice about my research works as the co-author of my journal articles. He has given me not only support on academic research but also suggestion on career development.

Fourthly, I am grateful to my fellow doctoral candidates and colleagues for their camaraderie, intellectual exchange, and shared experiences, which have been a source of motivation and support. I would also like to extend my appreciation to the participants who generously contributed their time to this study. Their involvement was crucial in gathering meaningful data and perspectives.

Special thanks go to my family for their unwavering encouragement, love, and understanding throughout this challenging academic journey. I am also grateful to my girlfriend, Chengyue Li, who provided both academic and emotional support. I am deeply grateful.

Finally, I acknowledge the financial support provided by the National Institute of Information and Communications Technology (NICT), Japan. Their support has enabled me to dedicate time and resources to this research, contributing significantly to its successful completion. And I would also like to show my gratitude to <https://www.jikken-baito.com> for recruitment of participants.



# Contents

<b>Abstract</b>	<b>i</b>
<b>Acknowledgements</b>	<b>iii</b>
<b>List of Figures</b>	<b>xv</b>
<b>List of Tables</b>	<b>xvi</b>
<b>List of Abbreviations</b>	<b>xvii</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Background . . . . .	1
1.2 Objectives and Significance . . . . .	4
1.3 Outline . . . . .	7
<b>2 Takeover Process and Literature Review</b>	<b>9</b>
2.1 Takeover Process . . . . .	9
2.2 Architectures for Safe Takeovers . . . . .	10
2.3 Factors Influencing Takeover Behaviors . . . . .	13
2.3.1 System-Related Factors . . . . .	13
<i>Level of Automation</i> . . . . .	13
<i>Non-Driving-Related Tasks</i> . . . . .	14
<i>HMI and TOR</i> . . . . .	16
2.3.2 Scenario-Related Factors . . . . .	17
<i>Traffic Density</i> . . . . .	17
<i>Vehicle Speeds</i> . . . . .	18
<i>Time Budget</i> . . . . .	19
<i>Scenarios Configurations</i> . . . . .	20
2.3.3 Human-Related Factors . . . . .	21
<i>Age and Experience</i> . . . . .	21
<i>Gaze Behaviors</i> . . . . .	23
<i>Duration of Driving</i> . . . . .	25

	<i>Reaction Speed</i> . . . . .	27
2.4	Modeling of Takeover Behaviors . . . . .	27
2.5	Performance Evaluation Metrics . . . . .	31
2.5.1	Time-Related Measures . . . . .	31
2.5.2	Quality-Related Measures . . . . .	32
2.6	Future Challenges and Discussions . . . . .	33
2.7	Summary of Chapter 2 . . . . .	35
<b>3</b>	<b>Takeover Experiments and Data Analysis</b>	<b>37</b>
3.1	Purpose of the Experiment . . . . .	37
3.2	General Experimental Design . . . . .	38
3.2.1	Participants and Equipment . . . . .	38
3.2.2	Non-Driving Related Task . . . . .	39
3.2.3	Takeover Scenarios . . . . .	40
3.2.4	Experiment Design . . . . .	42
3.2.5	Personality Tests . . . . .	43
3.2.6	Experimental Procedures . . . . .	43
3.3	Analysis 1: Impact of Duration of Monitoring . . . . .	44
3.3.1	Research Questions of Analysis 1 . . . . .	44
3.3.2	Data Processing and Performance Metrics . . . . .	46
	<i>Takeover Time</i> . . . . .	46
	<i>Percentage of Area of Interest</i> . . . . .	46
	<i>Heat Map and Gaze Entropy</i> . . . . .	47
	<i>Pupil Dilation and Eyelid Opening</i> . . . . .	47
3.3.3	Results and Analysis . . . . .	48
	<i>Analyzing Methods</i> . . . . .	48
	<i>Effects of DoM and Scenarios (Q1)</i> . . . . .	48
	<i>Gaze Behavior and Eye Movements (Q2)</i> . . . . .	51
3.3.4	General Discussions . . . . .	57
	<i>Effects of Duration of Monitoring (Q3)</i> . . . . .	57
	<i>Implications of Eye Tacking Data (Q4)</i> . . . . .	59
	<i>Impacts of Takeover Scenarios (Q5)</i> . . . . .	63
3.3.5	Summary of Analysis 1 . . . . .	63
3.4	Analysis 2: Impact of Personality . . . . .	64
3.4.1	Hypotheses of Analysis 2 . . . . .	64
3.4.2	Data Processing and Performance Metrics . . . . .	66
3.4.3	Results and Analysis . . . . .	66
	<i>ANOVA Results</i> . . . . .	67

	<i>Takeover Time</i>	70
	<i>Longitudinal Performance</i>	71
	<i>Lateral Performance</i>	73
	<i>Turn Signal Missing Rates</i>	75
	<i>Summary</i>	75
3.4.4	General Discussions	76
	<i>Role of DoM (H4)</i>	76
	<i>Effect of Neuroticism (H1-1)</i>	76
	<i>Effect of Agreeableness (H1-2)</i>	77
	<i>Effect of Conscientiousness (H1-3)</i>	77
	<i>Effect of Extraversion and Openness (H2 &amp; H3)</i>	77
3.4.5	Summary of Analysis 2	78
3.5	Summary of Chapter 3	79
<b>4</b>	<b>Modeling Takeover Behaviors</b>	<b>80</b>
4.1	Dataset Preparation	80
4.1.1	Raw Data	80
4.1.2	Independent Variables	80
	<i>Driver Attributes</i>	80
	<i>System Attributes</i>	82
	<i>Scenario Attributes</i>	83
4.1.3	Dependent Variables	83
4.1.4	Data Cleaning	83
4.2	Modeling Takeover Time	85
4.2.1	Problem Statement and Objectives	85
4.2.2	Basics of XGBoost and SHAP	87
	<i>Objective Function</i>	87
	<i>Gradient Tree Boosting</i>	88
	<i>Shapley Value</i>	89
4.2.3	Results and Discussion	90
	<i>Training Parameters and Metrics</i>	90
	<i>XGBoost Regressor Performance</i>	91
	<i>SHAP Explanation–Global</i>	93
	<i>SHAP Explanation–Local</i>	98
4.2.4	Summary	103
4.3	Modeling Takeover Readiness	103
4.3.1	Problem Statement and Objectives	103
4.3.2	Oversampling and Undersampling Technique	104

	<i>Purpose of Oversampling and Undersampling</i>	104
	<i>Synthetic Minority Oversampling Technique</i>	105
	<i>Random Undersampling Technique</i>	105
4.3.3	Results and Discussion	105
	<i>Training Parameters and Metrics</i>	105
	<i>XGBoost Classifier Performance</i>	107
	<i>SHAP Explanation–Global</i>	109
	<i>SHAP Explanation–Local</i>	116
4.3.4	Summary	116
4.4	Modeling Takeover Style	119
4.4.1	Problem Statement and Objectives	119
4.4.2	DTW-Based Clustering	120
	<i>Data Processing and Feature Selection</i>	120
	<i>DTW-Based K-Means Clustering</i>	120
4.4.3	Results and Discussions	124
	<i>Effects of DoM on Patterns of Evasive Maneuvers</i>	124
	<i>Effects of Scenarios on Patterns of Evasive Maneuvers</i>	126
4.4.4	Summary	128
4.5	Summary of Chapter 4	135
<b>5</b>	<b>Main Results and Unsafe Takeover Behaviors</b>	<b>136</b>
5.1	Main Results Discussions	136
5.2	Failure Cases Discussions	138
5.2.1	Data Processing and Annotation	138
	<i>Phase 1: Identifying Crashes and Near-Crashes</i>	138
	<i>Phase 2: Categorizing Crashes and Near-Crashes</i>	139
5.2.2	Analysis of Safe and Unsafe Takeover Behaviors	141
	<i>Safe Takeover Behaviors</i>	141
	<i>C1—Neglect of Mirror</i>	142
	<i>C2—Use of Mirror at Improper Time</i>	143
	<i>C3—Improper Judgment</i>	144
	<i>C4—Improper Attention Allocation</i>	145
	<i>C5—Improper Vehicle Speed</i>	146
5.2.3	Implication for Safe Takeovers	147
<b>6</b>	<b>Conclusions</b>	<b>150</b>
6.1	Key Findings	150
6.2	Main Contributions	151
6.3	Future Works	153



<b>A</b>	<b>Experiment Design</b>	<b>155</b>
A.1	Description of Takeover Scenarios . . . . .	155
A.2	Experimental Arrangements . . . . .	157
A.3	Pre-Experiment Questionnaire . . . . .	158
A.4	Modified NASA-TLX After Each Scenario . . . . .	160
<b>B</b>	<b>Data Processing</b>	<b>163</b>
B.1	SuRT Data Processing . . . . .	163
B.2	NASA-TLX Data Processing . . . . .	165
B.3	Eye Gazes Data Processing . . . . .	166



# List of Figures

1.1	Structure of this thesis. . . . .	8
2.1	System-initiated takeover process [1]. . . . .	9
2.2	Simple architecture for evaluating takeover readiness [2]. . . . .	11
2.3	Complete architecture for evaluating takeover readiness. . . . .	12
3.1	Driving simulator. . . . .	39
3.2	Surrogate Reference Task (SuRT). . . . .	40
3.3	Illustration of procedures of each takeover scenario. . . . .	44
3.4	Left: Takeover time given different DoM. Right: Takeover time in scenarios with different criticality. ns means non-significant. . . . .	49
3.5	Percentage of eye gazes directed to the defined AOIs, where RightMirror and RightScreen are combined as RightSide, and LeftMirror, LeftScreen and Distractor are combined as LeftSide. Ⓢ and Ⓣ represent end of SuRT and TOR respectively. From top to bottom are short, medium, and long DoM respectively. . . . .	52
3.6	Stacked scatter plots of gaze points of drivers before and after the TOR given different DoM. From top to bottom are short, medium, and long DoM respectively. . . . .	53
3.7	Heat maps of eye gazes of drivers before and after the TOR given different DoM. From top to bottom are short, medium, and long DoM respectively. . . . .	54
3.8	Box and scatter plots of gaze entropy given different DoM. . . . .	55
3.9	Eyelid opening and pupil dilation before and after the TOR given different DoM. Dashed blue lines mark the end of the SuRT, solid red lines mark the TOR and dotted black lines mark the takeovers. Peaks after the TOR are marked in red circles. . . . .	56
3.10	Regression model of takeover time over DoM in high- and medium-critical scenarios respectively. . . . .	57
3.11	Regression model of takeover time over DoM in medium- and high-critical scenarios, where the solid and the dashed lines are results of linear and polynomial regression, respectively. . . . .	58

3.12	Eyelid opening, pupil dilation and eye gazes directed to CenterScreen before and after TOR given different DoM. Dashed blue lines mark the end of the SuRT, solid red lines mark the TOR and dotted black lines mark the takeovers. . . . .	62
3.13	Above: Takeover time of different levels of extraversion. Below: Takeover time of different levels of extraversion given different DoM. . . . .	70
3.14	Above: Takeover time of different levels of openness. Below: Takeover time of different levels of openness given different DoM. . . . .	71
3.15	Above: Mean speed of different levels of neuroticism. Below: Mean speed of different levels of neuroticism given different DoM. . . . .	72
3.16	Above: STD of speed of different levels of neuroticism. Below: STD of speed of different levels of neuroticism given different DoM. . . . .	72
3.17	Speed profile of different levels of neuroticism given DoM <sub>L</sub> in each takeover scenario. . . . .	73
3.18	Above: STD of steering angle of different levels of agreeableness. Below: STD of steering angle of different levels of agreeableness given different DoM. . . . .	74
3.19	Steering profile of different levels of agreeableness given DoM <sub>M</sub> in each takeover scenario. . . . .	74
3.20	Turn signal missing rates for each personality trait. . . . .	75
4.1	Histogram of the takeover time. . . . .	87
4.2	SHAP summary bar plot for takeover time. . . . .	95
4.3	SHAP summary beeswarm plot for takeover time. . . . .	96
4.4	SHAP main effects scatter plots for takeover time. . . . .	97
4.5	SHAP interaction effects scatter plots for takeover time. . . . .	99
4.6	SHAP individual explanation for takeover time prediction 1.223 s. . . .	100
4.7	SHAP individual explanation for takeover time prediction 2.637 s. . . .	101
4.8	SHAP individual explanation for takeover time prediction 3.833 s. . . .	102
4.9	SHAP summary bar plot for takeover readiness. . . . .	110
4.10	SHAP summary beeswarm plot for takeover readiness. . . . .	111
4.11	SHAP main effects scatter plots for takeover readiness. . . . .	114
4.12	SHAP interaction effects scatter plots for takeover readiness. . . . .	115
4.13	SHAP individual explanation for takeover readiness prediction (high). .	117
4.14	SHAP individual explanation for takeover readiness prediction (low). .	118
4.15	Patterns of evasive maneuvers in TOR1 with different DoMs. Dotted, dashed, and solid lines represent Type I, Type II and Type III patterns of evasive maneuvers respectively. . . . .	125

4.16	Patterns of evasive maneuvers in scenarios TOR1, TOR2 and TOR3. Dotted, dashed, and solid lines represent Type I, Type II and Type III patterns of evasive maneuvers respectively. . . . .	129
4.17	Patterns of evasive maneuvers in scenarios TOR4, TOR5 and TOR6. Dotted, dashed, and solid lines represent Type I, Type II and Type III patterns of evasive maneuvers respectively. . . . .	130
4.18	Patterns of evasive maneuvers in scenarios TOR7, TOR8 and TOR9. Dotted, dashed, and solid lines represent Type I, Type II and Type III patterns of evasive maneuvers respectively. . . . .	131
4.19	Patterns of evasive maneuvers in scenarios TOR10, TOR11 and TOR12. Dotted, dashed, and solid lines represent Type I, Type II and Type III patterns of evasive maneuvers respectively. . . . .	132
5.1	Left: Eye gazes of drivers in scenario TOR1 under each DoM. From above to below to right were DoM <sub>S</sub> , DoM <sub>M</sub> and DoM <sub>L</sub> respectively. Different colors represent different fixations, and the bigger circles represent the center of the fixations. Right: Subprocesses of eye gazes of drivers with safe takeover behaviors. From above to below are eye gazes before, during and after lane change. . . . .	142
5.2	Typical eye gazes of drivers with C1 unsafe behaviors. . . . .	143
5.3	Left: Typical eye gazes of drivers with C2 unsafe behaviors. Right: Subprocesses of eye gazes of drivers with C2 unsafe behaviors. From above to below are eye gazes before, during and after lane change. . . . .	144
5.4	Left: Typical eye gazes of drivers with C3 unsafe behaviors. Right: Subprocesses of eye gazes of drivers with C3 unsafe behaviors. From above to below are eye gazes before, during and after lane change. . . . .	145
5.5	Left (Case 1): Subprocesses of eye gazes of drivers with C4 unsafe behaviors before and after colliding with the vehicle. Right (Case 2): Subprocesses of eye gazes of drivers with C4 unsafe behaviors before and after colliding with the obstacles. . . . .	146
5.6	Subprocesses of eye gazes of drivers with C5 unsafe behaviors before and after colliding with the vehicle ahead. . . . .	147



# List of Tables

2.1	Template to specify a test scenario [3]. . . . .	20
2.2	Classification of psychological dimension of scenarios [4]. . . . .	21
2.3	Summary of modeling of takeover behaviors. . . . .	29
2.4	Summary of factors influencing takeover behaviors. . . . .	34
2.5	Conclusions regarding factors impacting takeover behaviors. . . .	36
3.1	Distributions of ages and driving experience of the participants. .	38
3.2	Summary of the takeover scenarios. . . . .	41
3.3	Test order of the takeover scenarios. . . . .	43
3.4	Takeover time (mean and standard deviation) in scenarios with different criticality given different DoM. . . . .	49
3.5	Two-way ANOVA results performed on DoM and scenarios for takeover time in high- and medium critical scenarios. . . . .	50
3.6	ANOVA results and takeover time (mean and standard deviation) in high- and medium-critical scenarios given different DoM. . . .	51
3.7	Friedman test results performed on DoM for eyelid opening and pupil dilation 5 s, 10 s and 15 s after the TOR. . . . .	56
3.8	Summary of the performance metrics (effects of personality). . . .	66
3.9	Grouping criterion and grouping results of the participants per values of the personality traits (each participated in 6 scenarios). .	67
3.10	Summary of ANOVA results (effects of personality). . . . .	69
3.11	Summary of the analysis results (effects of personality). . . . .	75
4.1	Independent variables for modeling takeover behaviors. . . . .	81
4.2	Coding results of driver attributes 1. . . . .	82
4.3	Dependent variables for modeling takeover behaviors. . . . .	84
4.4	Summary of drivers' reaction speed and correction rates. . . . .	85
4.5	Summary of driver-related features. . . . .	86
4.6	Drivers' ratings on scenario attributes. . . . .	86
4.7	Descriptive statistics of takeover time. . . . .	87
4.8	XGBoost regressor performance given different scenarios and win- dow size. . . . .	91

4.9	XGBoost regressor vs. baseline regression models. . . . .	93
4.10	Best parameters for regression models. . . . .	93
4.11	Truth table for binary classification. . . . .	107
4.12	XGBoost classifier performance given ratios of oversampling and window size. . . . .	108
4.13	XGBoost classifier vs. baseline classification models. . . . .	112
4.14	Best parameters for regression models. . . . .	112
4.15	Descriptive statistics of each feature for each driving style in TOR1 given different DoMs. . . . .	126
4.16	Clustering results of the takeover style. . . . .	127
4.17	Descriptive statistics of each feature for each driving style in sce- narios TOR1–TOR6. . . . .	133
4.18	Descriptive statistics of each feature for each driving style in sce- narios TOR7–TOR12. . . . .	134
A.1	Description of the takeover scenarios. . . . .	156
A.2	Summary of the experimental arrangements. . . . .	157
B.1	An excerpt of SuRT raw data. . . . .	163
B.2	An excerpt of NASA-TLX raw data. . . . .	165



# List of Abbreviations

<b>ACC</b>	<b>Adaptive Cruise Control</b>
<b>ADS</b>	<b>Automated Driving Systems</b>
<b>AEB</b>	<b>Automatic Emergency Braking</b>
<b>AOI</b>	<b>Area Of Interest</b>
<b>ANOVA</b>	<b>ANalyses Of VAriance</b>
<b>ANOSIM</b>	<b>ANalysis Of SIMilarities</b>
<b>AV</b>	<b>Automated Vehicles</b>
<b>CAD</b>	<b>Conditionally Automated Driving</b>
<b>CAS</b>	<b>Collision Avoidance System</b>
<b>CNN</b>	<b>Convolutional Neural Networks</b>
<b>DA</b>	<b>Discriminant Analysis</b>
<b>DCS</b>	<b>Driver Controllability Set</b>
<b>DDT</b>	<b>Dynamic Driving Task</b>
<b>DNN</b>	<b>Deep Neural Networks</b>
<b>DoM</b>	<b>Duration of Monitoring</b>
<b>DTW</b>	<b>Dynamic Time Warping</b>
<b>EoR(G)</b>	<b>Eyes-on-Road (Gazes)</b>
<b>GB</b>	<b>Gradient Boosting</b>
<b>HARA</b>	<b>Hazard Analysis and Risk Assessment</b>
<b>HMI</b>	<b>Human-Machine Interface</b>
<b>kNN</b>	<b>k-Nearest Neighbor</b>
<b>LCA</b>	<b>Lane Change Assistance</b>
<b>LD</b>	<b>Linear Discriminant</b>
<b>LeadT</b>	<b>Lead Time</b>
<b>LKA</b>	<b>Lane Keeping Assistance</b>
<b>LOA</b>	<b>Level Of Automation</b>
<b>LR</b>	<b>Logistic Regression</b>
<b>LSTM</b>	<b>Long Short Term Memory</b>
<b>MAE</b>	<b>Mean Absolute Error</b>
<b>MLR</b>	<b>Multiple Linear Regression</b>

<b>MRC</b>	<b>Minimum Risk Conditions</b>
<b>MRM</b>	<b>Minimum Risk Maneuvers</b>
<b>MDTW</b>	<b>Multivariate Dynamic Time Warping</b>
<b>NB</b>	<b>Naive Bayes</b>
<b>NDRT</b>	<b>Non-Driving Related Tasks</b>
<b>ODD</b>	<b>Operational Design Domain</b>
<b>ORI</b>	<b>Observable Readiness Index</b>
<b>RBF</b>	<b>Radial Basis Function</b>
<b>RF</b>	<b>Random Forest</b>
<b>RGF</b>	<b>Regularized Greedy Forest</b>
<b>RtI</b>	<b>Request to Intervene</b>
<b>SHAP</b>	<b>SHapley Additive exPlanations</b>
<b>SD/STD</b>	<b>Standard/STandard Deviation</b>
<b>SEM</b>	<b>Structural Equation Model</b>
<b>SuRT</b>	<b>Surrogate Reference Task</b>
<b>SVM</b>	<b>Support Vector Machine</b>
<b>TLC</b>	<b>Time to Lane Crossing</b>
<b>TOR</b>	<b>Take-Over Request</b>
<b>TQT</b>	<b>Twenty Questions Task</b>
<b>TTC</b>	<b>Time To Collision</b>

# Chapter 1

## Introduction

### 1.1 Background

Automated driving has been believed to be a promising technology in reducing road traffic crashes, which have been found to be mostly related with driver carelessness, e.g., recognition error [5]. However, after several fatal crashes that involved automated driving systems (ADS), e.g., the fatal crash caused by an Uber automated vehicle in 2018 [6], the safety and reliability of ADS itself has become a great concern in recent years. From functional safety (ISO 26262:2018 [7]) to safety of the intended functionality (ISO/PAS 21448:2019 [8]) and cybersecurity (ISO/SAE 21434:2021, [9]), and to the suggesting of a minimum set of assumptions in the development of safety-related models for ADS (IEEE 2846:2022 [10]), aspects regarding development and deployment of ADS have been increasingly considered. To ensure safety of automated vehicles (AV), the most straightforward way might be to always keep drivers in the loop. However, it will make the technology less meaningful considering that one of the advantages of it is that it renders drivers the possibility to engage in non-driving related tasks (NDRTs). In fact, even during manual driving, it is almost unavoidable for drivers to get out of the control loop, especially for younger drivers [11]. Although highly and fully automated vehicles are desirable in the future, due to various technical, legal and societal problems [12], conditionally automated driving (CAD) might be a good compromise at present.

Per definition of ISO/SAE 22736:2021 [13], in conditionally automated vehicles, the fallback-ready user (mostly the driver) should always be “receptive to ADS-issued requests to intervene, as well as to DDT (dynamic driving task) performance-relevant system failures in other vehicle systems, and will respond appropriately”. Hence, one of the peculiar aspects of CAD is that transitions of control authority between human drivers and ADS would be inevitable, including driver-initiated and system-initiated transitions [1]. Of especial concern are

the latter where drivers are out of control loop due to various NDRTs, making it challenging to make a safe takeover. Within operational design domain (ODD), ADS should be capable of driving autonomously. Otherwise, the fallback-ready user should be able to take over. For the takeover to be timely and successful, however, three basic questions are essential—“When should the takeover occur?”, “How do we know the driver is ready to take over?” and “How can we safely and successfully carry out the transitions based on the evaluation results?”.

Regarding the first question, some typical cases when transitions were necessary were summarized in [14] and [15]. As general discussions, both driver-initiated and system-initiated transitions were included, however, the principles they based on were a bit different. In [14], based on “Is the transition required?”, “Who initiates the transition?”, and “Who is in control after the transition?”, six types of transitions between the driver and the ADS were defined. And using a flowchart, five scenarios were discussed in detail in [15], namely, “when the driver takes over the vehicle actively in normal driving conditions”, “when the driver intervenes in the vehicle actively in situations where the risks are not urgent”, “when the driver is required to intervene passively facing urgent risks the ADS cannot handle”, “when the driver fails to take over facing urgent risks but realizes it before it is too late”, and “when the driver fails to take over facing urgent risks and realizes it only after it is too late”. In these studies, the scenarios were defined quite abstractly. To put it into practice and to issue takeover requests (TOR) “within sufficient time” [13], hazard analysis and risk assessment (HARA) would be necessary. In [16], it was mentioned that the average takeover lead time (time from the takeover request to collision) used in the studies was  $6.37 \pm 5.36$  s, which would be sufficient to guarantee safety of automated vehicles in general. However, it is to be noted that the word “sufficient” is more of a relative term than an absolute one. A sufficient time for a vigilant driver may not be sufficient for a sleepy driver anymore. Besides, there may be the situation when a takeover needs to be issued, but no takeover request is provided due to system failures; there may also be the situation where the system has issued a takeover request, yet the driver fails to take over control of the vehicle. Sufficient time would be quite different in accordance with states of the driver. This leads us to the second question—“How do we know the driver is ready to take over?”.

For us human beings, to decide whether the driver is ready or not is subjective and relatively easy. However, if we want a machine to undertake this duty, the evaluation measures have to be quantified. For this to be true, many factors need to be taken into considerations, which can roughly be classified

into three categories, namely, system-related factors, scenario-related factors and human-related factors [17]. Factors such as types of human-machine interfaces (HMI), level of automation and vehicle states, etc., that have vehicles involved fall into the first category. Of special importance are HMIs, as indicated in [18], “advanced user interfaces can enhance the safety and acceptance of automated driving” by using auditory instead of visual output modalities [19–21], and by displaying takeover guiding information [22, 23], etc. As for scenario-related factors, per defined in [24], “a scenario describes the temporal development between several scenes in a sequence of scenes”, and since “a scene describes a snapshot of the environment including the scenery and dynamic elements” [24], environmental conditions, type of roads, obstacles, lead time, traffic densities and vehicle speeds can all be put into the second category. A typical research regarding effect of traffic densities was shown in [25], where it was indicated that the presence of traffic led to longer takeover time and worsened the takeover quality. And to explore the effect of lead time, a wide range of lead time was investigated in [26] (3 s, 6 s, 10 s, 15 s, 30 s and 60 s). The results showed that the optimal takeover performance was recorded when the lead time was equal or longer than 10 s, which was higher than the usual recommendation of 5–6 s. This may be related to the types of the NDRTs used and urgency of the scenarios during the experiments. As humans can hardly be controlled while doing experiments, human-related factors may be the most complicated among the three. Ages, genders, NDRTs, driving experience, duration of driving, sleeping habits, and personality, etc., have all been found to have certain influence on situation awareness and takeover performance of drivers, and therefore, fall into this category. Considering these, a deep understanding of human beings is essential to better evaluate and predict takeover behaviors of drivers. As a reference of the usually used NDRTs, [27] provided a sound summary, where the NDRTs were classified into two categories—everyday and standardized NDRTs. The first category includes those that are common in our everyday driving, such as operating the in-vehicle information system, reading, eating, listening to music, texting or talking on the phone, and watching movies, etc. While the second category refers to specially designed tasks and is relatively complicated. Such examples include tracking tasks, where one is required to keep track objects within tolerance limits; 20-questions-task (TQT), where one is asked to answer 20 questions to guess an item; rotated figures task, where one is asked to decide whether two differently rotated figures match or not; and addition task, where one is required to add numbers shown on a display; etc. With each study emphasizing on different aspects of the influencing factors, there is the necessity to

gather all these results together to better evaluate takeover behaviors of drivers. This is the main purpose of this thesis.

Based on the evaluation results, it then becomes possible to answer the third question—“How can we safely and successfully carry out the transitions?”. To answer this question, at least three aspects of research need to be addressed. Foremost, fallback strategies when the driver is not ready or fails to take over within required time at the time that the driver needs to take over the vehicle. Different strategies may be necessary facing different situations. For example, a fallback strategy was proposed for the avoidance or mitigation of rear-end collisions when the driver’s ability to respond to ADS-issued requests was compromised in [28]. And in [29], a fallback strategy containing 3 degraded levels with 7 fallback scenarios was proposed, so that the vehicle can be brought to minimal risk condition when different functional failures occurred. When the driver is ready to take over, then comes the second aspect—driver intervention identification. As indicated in [30], for the safe transitions between the driver and the ADS, timely and precise identification of intervention of the driver was necessary. Similar discussion can also be found in [31]. Finally, the switching process when everything is ready. The main principle is that this process has to be smooth and stable enough, so that the stability of the vehicle and comfort of the driver can be guaranteed.

Although these are all important questions that are worthy of discussing, as evaluation of takeover behaviors of the driver forms the basis for the third question, and also to limit the scope of the study, this thesis will focus on the second one. To be specific, it will concentrate on takeover process and general architectures for evaluating and modeling of takeover behaviors of the driver (Chapter 2), factors that might influence takeover behaviors of the driver (Chapters 2 and 3), and finally, methods for modeling takeover behaviors of the driver (Chapters 2 and 4).

## 1.2 Objectives and Significance

A vast majority of studies on driver takeover behaviors have focused on empirical analysis of the relationship between the various factors and drivers’ takeover behaviors. Although some review results have been provided [32–35] regarding the influence of various factors on takeover performance, new insights still seem to be lacking in those results, especially with respect to scenario-related factors like scenario predictability and driver-related factors like personality. Of course, they have also been missing in the modeling of drivers’ takeover

behaviors. Considering these, the **main goal** of this research is to model drivers' takeover behaviors by integrating a variety of factors, so that drivers' takeover behaviors could be predicted in advance and systems could adapt their strategies accordingly to ensure safe takeovers. To achieve this goal, the following four objectives were proposed in this thesis.

1. Give a thorough analysis of the factors influencing takeover behaviors of drivers during automated driving, along with a framework for evaluation and prediction of takeover behaviors (**Objective 1**).
2. Based on the results of Objective 1, design experiments to explore new factors that have not been studied in the previous research through statistical analysis, with insight into eye-tracking data (**Objective 2**).
3. Model drivers' takeover behaviors in three aspects—takeover time, takeover readiness and takeover style, as regression, classification and clustering problems, respectively (**Objective 3**). This is completed using machine learning models based on the factors extracted in Objective 1 and 2.
4. Analyze failure cases during the experiments through video recordings and dynamic eye gazes of drivers, and give some suggestions on design of advanced human-machine interface to ensure safe takeovers (**Objective 4**).

The results of this research are meant to be applied in passenger cars equipped with advanced driver-assistance systems or automated driving systems, so that drivers' states and intention could be better understood, which helps to improve safety of the system and experience of the driver. Industrially, this research holds significance in the following ways.

1. **Enhance Safety:** This study addresses the critical issue of takeover events in automated driving, where control is handed back to the driver. By identifying factors influencing takeover behaviors and developing predictive models, the research contributes to enhancing the safety of automated driving systems. This is of paramount importance in reducing road traffic crashes and improving overall road safety.
2. **HMI Design:** The findings provide insights into the factors affecting takeover performance. This information is valuable for designing more effective Human-Machine Interfaces (HMIs) that consider human factors, ensuring a seamless transition of control between the automated system and the driver.

3. **Modeling Takeover Behaviors:** The research contributes to modeling takeover behaviors in various aspects, such as regression for takeover time, classification for takeover readiness, and clustering for takeover maneuvers. The use of advanced techniques demonstrates the application of state-of-the-art methods to address complex problems in automated driving systems.

In summary, the research has industrial significance by advancing our understanding of takeover events in automated driving, providing actionable insights for HMI design, and contributing to the development of more robust and safe automated driving systems. The findings are relevant for manufacturers, designers, and policymakers involved in the development and regulation of autonomous vehicles.

Moreover, this research also holds significant academic importance in the following ways.

1. This research addresses a critical aspect of automated driving system—the takeover process. The identification of influencing factors, including some novel considerations, contributes to a comprehensive understanding of the complex dynamics involved in the human-automation interaction.
2. The identification of previously unstudied factors demonstrates the research's ability to explore new dimensions in predicting drivers' takeover behaviors. The statistical significance of these factors reinforces their importance in the overall prediction model.
3. The application of machine learning models to predict takeover time and readiness adds practical value to the research, and modeling takeover styles recognizes the temporal dynamics of the takeover process, contributing to a more nuanced understanding and classification of drivers' takeover maneuvers.
4. The analysis of crashes and near-crashes data and the identification of patterns of unsafe behaviors contribute to the real-world safety implications of automated driving systems. The suggested improvements in HMI design based on these patterns offer practical guidance for enhancing the safety of automated driving systems.

In summary, this research significantly advances the knowledge in the field of automated driving systems by addressing critical challenges in predicting and understanding drivers' takeover behaviors. The incorporation of novel factors,



machine learning models, and real-world safety implications enhances its academic significance and practical relevance for the development and deployment of automated driving technologies.

## 1.3 Outline

Chapter 2–Chapter 5 in this thesis aim to fulfill Objective 1–Objective 4, respectively. Their relationships are graphically shown in Fig. 1.1, and the main contents in each chapter are summarized below:

- **Chapter 1:** Background of this research was introduced in Section 1.1, from which, we gained motivations, and the objectives were summarized in Section 1.2. Finally, the outline of this research in this section.
- **Chapter 2:** Takeover process was briefly analyzed in Section 2.1, followed by architecture for evaluation and prediction of takeover behaviors in Section 2.2. A thorough analysis of the factors influencing takeover behaviors was conducted in Section 2.3, including system-, scenario- and human-related factors. Then, methods for modeling takeover behaviors were discussed in Section 2.4, along with the metrics used for evaluating takeover behaviors in Section 2.5. Finally, some challenges were pointed out in Section 2.6.
- **Chapter 3:** On basis of the analysis in Chapter 2, two new factors were identified that have not been studied in the previous research, including duration of monitoring (a period of time before the TOR after the NDRT has ended, Fig. 3.3) and drivers' personality. They were discussed in Section 3.3 and Section 3.4, respectively. For completeness, the purpose of the experiments was briefly discussed in Section 3.1, followed by details of the experimental design in Section 3.2.
- **Chapter 4:** On basis of results in Chapter 2 and Chapter 3, takeover behaviors were modeled using the data collected in the previous experiments. Firstly, data processing and preparation were described in Section 4.1. Then, takeover behaviors were modeled in three aspects:
  - Takeover time (Section 4.2): Since takeover time is a continuous variable, it was modeled as a regression problem.
  - Takeover readiness (Section 4.3): Since takeover readiness is a categorical variable, it was modeled as a classification problem.

- Takeover style (Section 4.4): Since takeover maneuver is time series data, it was modeled as a clustering problem.
- **Chapter 5:** Main results from Chapter 2–Chapter 4 were briefly discussed in Section 5.1. Then, some failure cases during the experiments discussed in Section 5.2, along with some suggestions for HMI design to ensure safe takeovers.
- **Chapter 6:** In Section 6.1, main conclusions in each chapter were summarized in correspond to the objectives proposed in Chapter 1. Then, the contributions of this research were summarized in Section 6.2. Finally, limitations of this research and future works following this research.

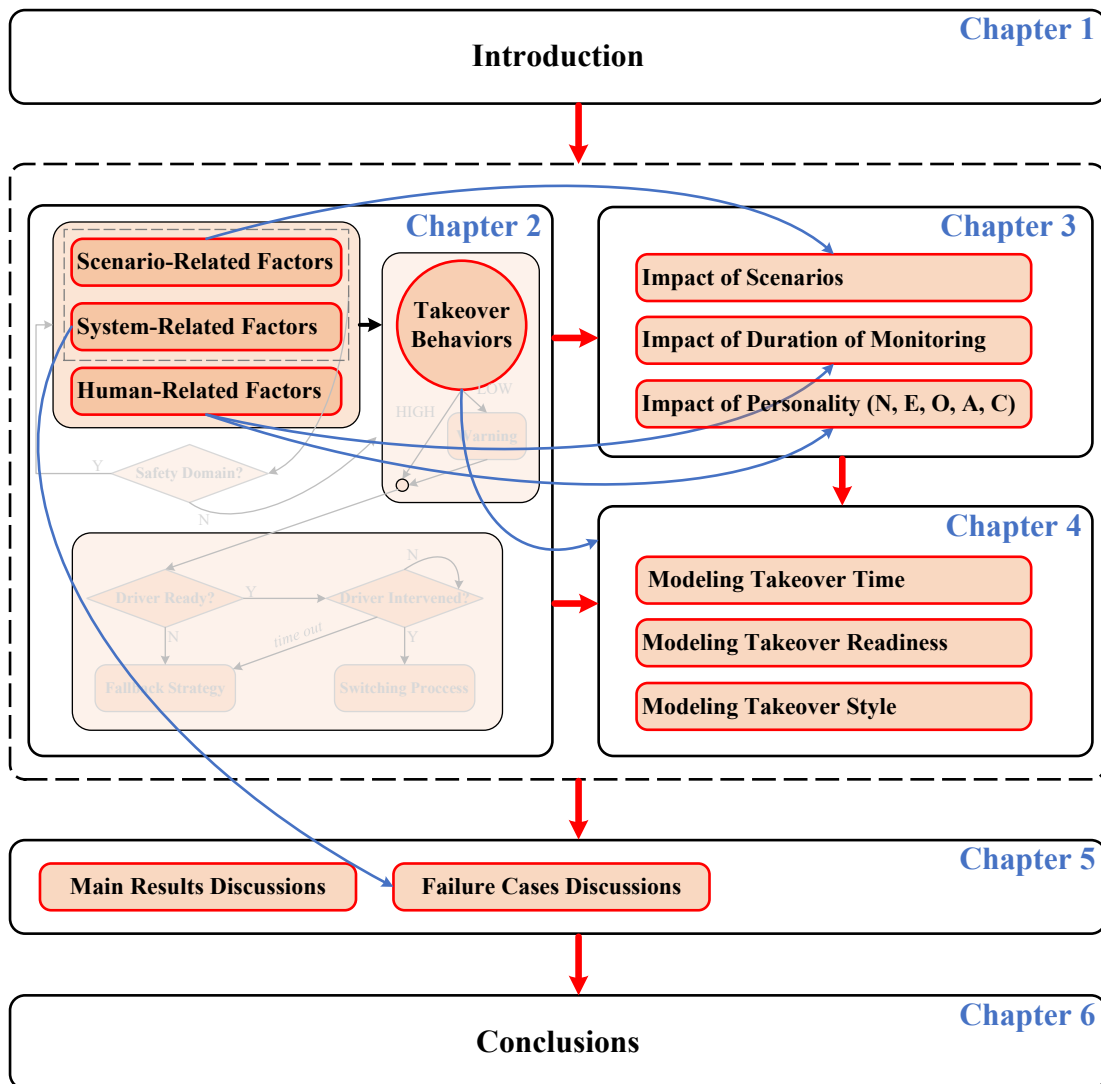


Figure 1.1: Structure of this thesis.

## Chapter 2

# Takeover Process and Literature Review

### 2.1 Takeover Process

Transitions of control occur both from vehicle to driver and driver to vehicle. Transition from driver to vehicle depends mostly on driver's willingness to transfer control and use the ADS. However, transition from vehicle to driver, also known as "takeover", depends on many factors. As indicated in Chapter 1, they can roughly be classified as system-related, scenario-related and human-related factors. Although takeovers can also be driver initiated, of particular interest in this thesis are takeovers initiated by the system.

Fig. 2.1 shows the process model for a system-initiated transition from vehicle to the driver [1, 3]. For complete explanation, one can reference to [1]. Below are the definitions of some of the key concepts for convenience of reference:

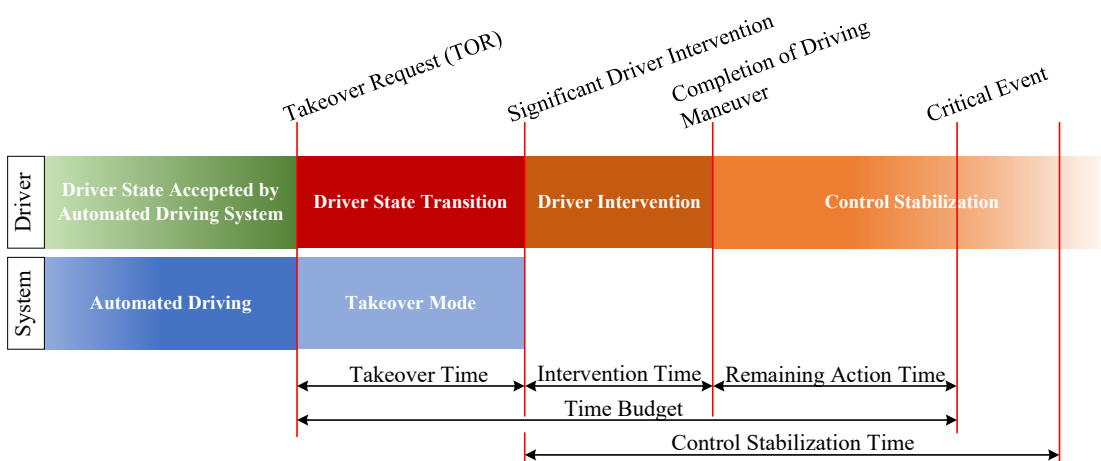


Figure 2.1: System-initiated takeover process [1].

- **Takeover Request:** Notification of the ADS to a driver/fallback-ready user indicating that s/he should take over vehicle control (Comment 2.1.1).

- **Critical Event:** Situation that can be specified in time and space that the ADS cannot handle safely and that will occur in case the driver does not intervene.
- **Takeover Mode:** System behavior after a TOR has been issued, which depends on the automation level of the ADS (Comment 2.1.2).
- **Driver State Transition:** Process of transforming the actual driver state to a target driver state suitable to effectively take over manual control (Comment 2.1.3).
- **Significant Driver Intervention:** Action initiated by the driver to request manual control of the vehicle (e.g., buttons, switches, pedals, or steering wheel).
- **Post Transition Control:** A time window to analyze the quality of manual control driving after a TOR has been issued.

**Comment 2.1.1** The system may issue a TOR: (1) When it finds a DDT performance relevant system failure or an object/event which cannot be handled by the system; (2) When existing the ODD for which it was designed.

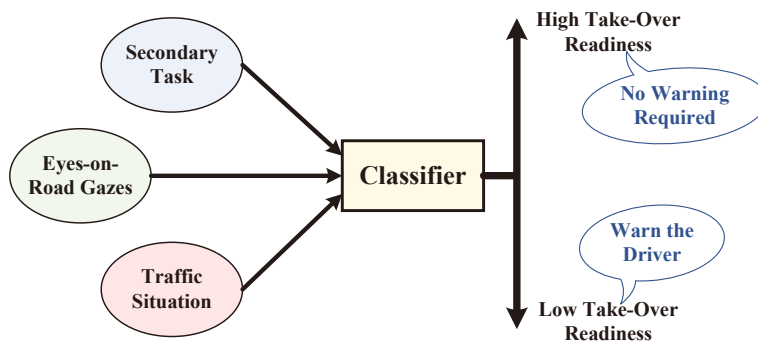
**Comment 2.1.2** The system may terminate immediately after issuing a TOR for level 2 while it shifts to the takeover mode following a TOR before termination for levels 3 and 4. When the driver does not initiate intervention within the takeover mode, the system may shift to the minimum risk maneuver (MRM) to stop the vehicle safely for level 3 and level 4 ADS.

**Comment 2.1.3** This process can be analyzed on a sensory (e.g., taking one's eyes off the NDRT and fixing eye gazes on the road), motoric (e.g., freeing one's hands and taking them back on the steering wheel) and cognitive (e.g., processing information gathered from the environments and making decisions) level. Detailed analysis of this process can be referenced to [36] and [37].

## 2.2 Architectures for Safe Takeovers

Factors that might influence takeover behaviors of drivers have gained extensive attention over the years. Nevertheless, to thoroughly evaluate takeover behaviors, integration of several sources of information is necessary. In view of this, Braunagel et al. [2] proposed the architecture as shown in Fig. 2.2. In this architecture, the classifier took traffic situations, NDRTs and eyes-on-road gazes (EoRG) of drivers as inputs, and outputted the takeover readiness. When the

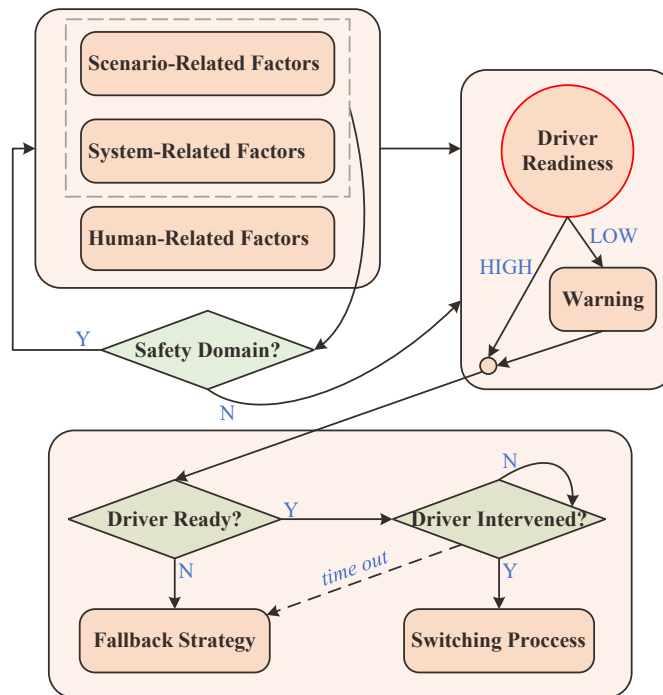
readiness was evaluated as high, no further measures were needed; whereas when the readiness was low, a warning should be issued to remind drivers of the situation. Eighty-one participants were involved in a driving simulator study, and it was shown that the proposed system could detect takeover readiness of drivers with an accuracy of 79%. Although the result was not quite satisfying, it was the one of the earliest approaches trying to integrate the former findings together and build an automated classification system of takeover readiness of drivers.



**Figure 2.2:** Simple architecture for evaluating takeover readiness [2].

However, this architecture is kind of simplified, and a more complete one is shown in Fig. 2.3. It can be observed that the upper part of this architecture is similar to that in Fig. 2.2, but with a few differences: 1) Traffic situation is generalized to scenario-related factors; 2) NDRTs and EoRG are incorporated as parts of the human-related factors; 3) System-related factors are also considered. As traffic situation is part of scenarios, and secondary tasks and EoRG are parts of human-related factors, the first two differences come easily. To better understand the third one, it is important to first comprehend the concept of safety domain [13], which refers to the situation where the automated vehicle is capable of running safely without intervention of human drivers. On one hand, HMI [18, 19], vehicle states [38], etc., have been found to influence situation awareness of drivers; on the other hand, by taking the vehicle states into consideration, together with scenario-related factors, it is possible to activate the system only when necessary. In this way, it could not only enhance accuracy of evaluation of takeover readiness of drivers, but also improve efficiency and lengthen lifespan of the system.

Apart from the above differences, the biggest difference lies in the lower part of this architecture, which takes effect in the transition stage when the driver is required to take over the vehicle. During this period of time, the driver still needs to be constantly monitored and evaluated, until the takeover process has



**Figure 2.3:** Complete architecture for evaluating takeover readiness.

been successfully completed. Overall, the operating process can be summarized as below:

1. Scenario-related factors, e.g., external environment, system-related factors, e.g., vehicle states, and the driver are closely monitored. If the vehicle is about to drive or has driven out of safety domain → Step 2.
2. Readiness of the driver shall be evaluated, if the takeover readiness of the driver is low, a warning shall be raised. Either way → Step 3.
3. The system shall decide whether the driver is ready, if the driver is not ready yet, fallback strategies shall take effect. Otherwise → Step 4.
4. The system shall detect whether the driver has intervened, if the driver fails to take over within required time, fallback strategies shall take effect. Otherwise → Step 5.
5. The switching process shall begin.
6. Return to Step 1 after reengagement of the ADS.

In either architecture, evaluation of takeover readiness of drivers plays an essential role in the interaction between the vehicle and the human driver. In an era where the automated driving system is not so autonomous, it is critical to the safety of the vehicle as well as the driver. And it is to be noted that the same architecture can also be used to predict takeover time and takeover quality.

## 2.3 Factors Influencing Takeover Behaviors

### 2.3.1 System-Related Factors

System-related factors are factors that have vehicles involved, e.g., level of automation, HMI, NDRT, TOR, and vehicle speeds, etc. However, since vehicle speed is more related with traffic scenarios, it will be discussed in section 2.3.2. This section will focus on level of automation, NDRTs, HMIs and TORs. Since TOR is also part of HMI, it will be discussed together with HMI.

#### Level of Automation

Definition of level of automation (LOA) of autonomous systems have been discussed thoroughly over the past few decades. In automotive industry, the most widely accepted definitions might be that given in [13], where the LOA was classified into 5 levels (except for purely manual driving). Basically, systems that can help the driver maintain either longitudinal or lateral control is referred to as “driver assistance system” or “level 1 automated driving system”, such as adaptive cruise control (ACC), lane keeping assistance (LKA), etc. When the system can help the driver maintain both longitudinal and lateral control but requires the driver to continuously monitor the surroundings, it is referred to as “partially automated driving system” or “level 2 automated driving system”. Further, when the driver is not required to monitor the surroundings continuously, the system can be referred to as “conditionally/highly/fully automated driving system” or “level 3/4/5 automated driving system” according to the situations it can deal with.

Many studies in the automotive domain were based on the above definitions. Basically, it is indicated that high levels of vehicle automation usually results in reduced situation awareness. For example, it was pointed out in [39] that, with high LOA, drivers were more inclined to in-vehicle entertainment tasks and also showed increasing symptoms of fatigue, which were all signs of reduced situation awareness. Further taken system failures into consideration, to what extent the driving performance was influenced by LOA during critical situations caused by automation failures was investigated in [38]. A driving simulator study revealed that in such critical situations, the semi-automated driving was safer than highly automated driving, and when the extent of failure increased, the situation became more critical. Similar results can also be found in [40]. This suggested that silent failures in ADS can be rather dangerous, especially when drivers are underload or distracted by various NDRTs. Hence, it is important

that an optimum level of automation is designed that neither underloads or overloads the driver.

It should be noticed that, increase of LOA is usually accompanied with more frequent engagement in secondary tasks, as pointed out in [39, 41, 42]. However, this is not necessarily true, it may also depend on the driver's familiarity with the automated driving system. In the on-road experiment conducted in [43], experimental results collected from 32 participants of different age groups and with diverse experience with ACC showed that the frequency of engagement in secondary tasks increased only with participants with previous experience with vehicle automation. And curiously, a decrease in primary task performance was not found due to increase of secondary task engagement, which might be attributed to the type of the secondary tasks involved as well as situation-adaptive behavior of the driver.

### **Non-Driving-Related Tasks**

Effects of NDRTs on driving and takeover performance have been broadly discussed over the years, and based on different criterion, NDRTs can also be classified differently.

In [44], according to level of demands on drivers, NDRTs were grouped into three categories—simple secondary tasks, moderate secondary tasks and complex secondary tasks. In this sense, adjusting radio, talking to passenger in adjacent seat, drinking, smoking, etc., belong to simple secondary tasks; talking/listening to hand-held devices, inserting or retrieving CD/cassette, eating, etc., belong to moderate secondary tasks; and dialing a hand-held device, operating/viewing a PDA, reading, etc., belong to complex secondary tasks. These secondary tasks were summarized from the 100-car naturalistic driving study data [45, 46], and the complete list can be referenced to Appendix A in [44].

In [27], based on characteristics of the tasks, NDRTs were separated into two categories—everyday NDRTs and standardized NDRTs. The first category includes those that are common in our everyday driving, such as operating the in-vehicle information system, reading, eating, listening to music, texting or talking on the phone, and watching movies etc. While the second category refers to specially designed tasks and is relatively complicated. Such examples include tracking tasks, where one is required to keep track objects within tolerance limits; 20-questions-task (TQT), where one is asked to answer 20 questions to guess an item; rotated figures task, where one is asked to decide whether two differently rotated figures match or not; and addition task, where one is required to add



numbers shown on a display; etc. A relatively complete list of the everyday and standardized NDRTs can be referenced to Table 1 and Table 2 in [27], respectively.

The merit of the standardized tasks is that they are designed for evaluating certain aspects of human performance and are therefore convenient for comparison between different studies. A typical example involving the TQT was given in [47], where interactions among three factors—drive (manual, automated), secondary task (TQT, no TQT) and scenario (critical incident, no critical incident), were discussed in detail. Fifty participants were involved in the study. The results showed that in the absence of TQT, performance of drivers during automated driving was similar to that during manual driving. However, when the driver was required to perform the TQT, drivers' reaction slowed down facing critical incident, and it was also found that drivers made less lane changes in response to the variable message sign. Overall, the performance was worst when the driver was required to takeover the vehicle while performing demanding secondary tasks facing critical incidents. In [48], the surrogate reference task (SuRT), where one is required to detect a slightly larger circle within a field of smaller circles and respond by indicating the position of the circle via a touch pad, was used as indicator of level of situation awareness of the driver. The results suggested that “the uncertainty group solved fewer secondary tasks in critical situations and more secondary tasks in noncritical situations”. This task was also utilized in [49], whereas it was used to make the driver out of the loop, so that the steering wheel concepts put forward could be better evaluated. Further in [50], the SuRT was compared with n-Back task, where one is asked to observe a serial presentation of targets and distractors and identify when the current stimulus matches the target  $n$  steps earlier. The results showed that the two tasks showed similar effects on driver behavior during takeover. Other standardized tasks such as addition [51], peripheral detection tasks [52], oddball tasks [53], etc., are not used frequently, and one can also design new tasks when the existing ones can not meet one's own needs.

Since standardized NDRTs are artificial in nature, they sometimes can not reflect the real driving situations, therefore, daily NDRTs have also been researched in depth. In [54], utilizing reading as the secondary task, the impact of manual and cognitive tasks on driver's ability to take over vehicle was examined carefully. Over one hundred participants (95 were valid) were recruited for the experiment, where they were asked to read (low cognitive load) or proofread (high cognitive load) a text on a tablet which was either handheld (high manual load) or mounted in the vehicle (low manual load). Results showed that manual task load prolonged reaction times and deteriorated takeover quality irrespective

of the type of interventions. However, cognitive tasks load was found to affect the lateral and longitudinal maneuvers differently. Although it was obvious that the increase of cognitive load impaired takeover quality in the lateral maneuver, the opposite was true in the longitudinal maneuver, which shall be examined in future studies. In [55], a variety of tasks were offered, including playing a game, reading a magazine, eating a snack, watching a DVD, doing a puzzle, listening to the radio, etc., to test the driver's willingness to engage in different types of NDRTs. The results showed that the drivers were most likely to listen to the radio or watch a DVD as the level of automation increased. In most research, NDRTs are not studied independently, rather as a tool to distract or burden the driver. For example, in [56], reading a magazine or playing a game was used as a distraction to the driver to investigate the most suitable patterns of vibrations that can assist the driver during takeover. In [36], "texting" and "internet search" tasks were served as visual distractions. In [57], the phone call tasks were used to induce cognitive distraction to test the driver's reactions under high cognitive load. In [58], movie watching task was utilized to place the driver out of the loop and prevent him/her from obtaining information about the road. More detailed discussions could be found in other sections of this chapter.

## HMI and TOR

To facilitate control transitions between ADS and the driver, HMIs would be indispensable, which aims to help to increase situation awareness of drivers with proper designs, so that takeovers can be timely, safe and comfortable. This includes not only interfaces that are used to provide information to drivers, but also TOR design that is used to warn the driver to take over the vehicle efficiently.

Almost all the previous research on takeover studies have provided certain auditory and visual signals to convey a TOR to the driver. Overall, it seems that auditory and vibrotactile TORs have been shown to elicit faster reaction time than purely visual ones [18–21]. This is understandable in that during automated driving, drivers are probably distracted or busy engaging in various NDRTs, information provided by a visual TOR may be easily overlooked by drivers. This may be one of the reasons why auditory TORs are most frequently used in the literature [32].

Moreover, as discussed in Section 2.1, takeover is a complicated process, involving sensory, motoric, and cognitive states of drivers. Hence, in addition to providing simple warning messages to drivers, HMIs may also be designed to provide information that supports drivers in making decisions, which may

possibly support the driver to make decisions faster and more correctly. In view of this, Eriksson et al. [22] designed 4 HMIs that supported drivers in different degrees, e.g. using an augmented sphere to highlight the slow-moving vehicle ahead, using an augmented overlay to inform drivers whether there was a gap in the left lane, using augmented reality arrows to indicate whether the driver could change lane or brake. Results suggested that such kind of HMIs did not significantly affect drivers' initial reaction to the TOR. However, they indeed showed an improvement in drivers' correct decisions. Exemplifying the usefulness of such kind of advanced HMIs.

### 2.3.2 Scenario-Related Factors

Taxonomy regarding traffic scenarios has been well defined in [24]. As these definitions are relatively abstract, in practical applications, what attributes and value facets to choose may depend on the specific research questions. An example of how was the taxonomy defined can be referenced to [59], which was used for assessment of communication process between automated and other vehicles. Among the many factors concerned, this section will mainly discuss traffic density, vehicle speeds, time budget, and scenario configurations.

#### Traffic Density

Traffic density is defined as the number of vehicles located on a road segment within a given time interval. Generally, with low traffic density, drivers are inclined to NDRTs and show signs of fatigue easily. As the density of the traffic increased, it was shown in [39] that alertness of drivers also increased, where the driver focused more attention to the roadway and became receptive to the changing demands imposed by heavy traffic. However, rise of alertness of drivers does not necessarily guarantee better takeover performance. Besides, these results were based on only two traffic densities, namely, low and high traffic densities, with 500 and 1500 vehicles per hour per lane respectively (This is not a good unit for expressing traffic density, since it would be dependent on vehicle speeds. For example, when the vehicle runs at a speed of 60 km/h, 500 and 1500 vehicles per hour correspond to 8 and 25 vehicles per kilometer).

Given this situation, a more detailed discussion was given in [25], where three traffic densities were defined—0, 10, 20 vehicles per kilometer. Seventy-two participants were involved in a driving simulator experiment. Based on six dependent variables—hand-on time, takeover time, maximum longitudinal acceleration, maximum lateral acceleration, minimal time to collision (TTC) and

horizontal gaze dispersion, it was demonstrated that, whilst hand-on time was found to be independent of traffic densities, traffic density actually had a distinct negative impact on takeover performance of drivers. Specifically, a higher traffic density: 1) delayed the maneuver and led to longer takeover time; 2) resulted in higher accelerations and lower TTCs; and 3) brought about a higher crash probability. Compared with the former research, more indexes were considered in the latter one, making it more plausible. Nevertheless, since only 3 traffic densities were considered, more research might be helpful in considering traffic densities lower than 10 and higher than 20 vehicles per kilometer. Similar results can also be observed based on two traffic densities in [60].

Further in [50], besides traffic density, the lanes in which the obstacles occurred and the states of the adjacent lanes were also taken into consideration, where four situations were defined on a three lane highway: 1) obstacles in the middle, adjacent lanes blocked; 2) obstacles in the right, none in the left; 3) obstacles in the left, none in the right; and 4) obstacles in the middle, none in the adjacent lanes. The results indicated that situation 1 with high traffic density induced the most critical behavior during takeover, which was kind of in correspond to the conclusions in [25] and [60].

### Vehicle Speeds

Vehicle speed is defined as the mean speed of all vehicles passing a specific location within a given time interval. The effect of vehicle speed on takeover readiness of the driver was analyzed at length based on statistical data in [61]. For the analysis, the dataset was divided into 7 categories: being in standstill, driving below 10 km/h, driving in urban areas, driving on rural roads, driving on highways below 110 km/h, driving on highways between 110 and 160 km/h, and driving on highways above 160 km/h. The results indicated that highly demanding manual-visual were preferably avoided at high speeds on the highway, and the type of roads seemed to have no impact. Nevertheless, the latter might be due to the fact that the road categories were too broadly categorized. Therefore, further research involving more road types are necessary.

It is also worth noting that takeover performance of drivers might be affected by the driver's perception of vehicle speeds, which can be influenced by many factors like actual speed, road type, driving experience and gender, etc. As indicated in [62], drivers with driving experience estimated the vehicle speed more accurately than those without. Besides, speeds were most accurately estimated within the range 25–35 mph (40–60 km/h approximately). When the speed was

below this range, it tended to be overestimated; and when the speed was above this range, it tended to be underestimated. Hence, while considering effect of vehicle speeds on takeover performance of drivers, this should also be taken into consideration.

### Time Budget

Time budget, also known as lead time, typically refers to the TTC or TLC at the time of TOR. Although results were a bit different across different studies, time budget has been found to affect takeover time and takeover quality greatly. Generally, shorter time budgets lead to shorter takeover time, and deteriorate takeover quality, which is associated with greater maximum accelerations, shorter  $TTC_{min}$ , higher crash rates, and greater standard deviation of lane positions and steering wheel angle, etc. In a meta-analysis [63], Gold et al. indicated that 1 s increase in time budget would result in 0.329 s increase in takeover time through a non-linear regression. Besides, McDonald et al. [32] conducted another meta-analysis, and found this value to be 0.27 s, proving the rationality of the results.

However, it is hard to define a certain value of time budget to be suitable or optimal, as it largely depends on the takeover scenarios. For example, in [16], the suitable time budget was recommended to be 5–6 s, whereas Wan et al. [26] claimed that the optimal takeover performance was recorded when the time budget was over 10 s. This may be attributed to the types of the NDRTs and urgency of the scenarios involved. In non-critical situations like exiting from highways, [64] even suggested a time budget of 16–30 s. Hence, while carrying on such kind of research, criticality of the scenarios must also be taken into consideration. In [16], it was concluded that the mean takeover time budget was  $6.37 \pm 5.36$  s with a mean reaction time of  $2.96 \pm 1.96$  s in the literature. Since 3 s would be too short [65] for safe takeovers, in practical experiments, 5–6 s would be a bit critical whereas enough to avoid accidents for most drivers, and 7–8 s would be less critical.

Time budget refers to the time interval after the TOR. Sometimes, a warning before the TOR may also exist, namely, the two-stage warning system. The benefit of a warning before the TOR has been indisputable [65–68], what has become a concern is the timing to issue the two warnings. In the study by [65], the two warnings were issued 9 s and 3 s prior to a construction site and the results showed that takeover before the second step indeed yielded better performance and fewer crashes. However, many takeovers before the TOR was actually issued

also implied that 3 s left for stage two warning might be too short to make a safe and comfort takeover. Considering that, [66] and [67] lengthened the time budget for the second stage to 5 s and 7 s respectively, and correspondingly, the time budget for the first stage was provided 12 s and 10 s before colliding with the takeover events. [68] made a finer research and issued the first warning 3 s, 5 s, 7 s and 9 s before the TOR, and then left the driver with 7 s to complete the takeover. The results indicated that both the 5 s and 7 s time intervals yielded more rapid takeovers and were rated more appropriate than the 3 s and 9 s time intervals.

### Scenarios Configurations

In order to describe a test scenario in a standardized way, classification schemes can be utilized to make sure that all the relevant elements have been covered, including physical and psychological dimensions. And a template [3] can be referenced to Table 2.1.

**Table 2.1:** Template to specify a test scenario [3].

Test Scenario Parameters		Parameter Values
Physical dimension	Preceding the test scenario	
	Start scene	Traffic infrastructure
		Environmental conditions
		Traffic constellation
	Evolution of test scenario	
	End of test scenario	
Psychological dimension	Visual sketch	
	Urgency	
	Predictability	
	Perceived criticality	
	Complexity	

Meaning of the psychological aspects can be referenced to [4], where a brief literature was given regarding testing scenarios for human factors research in level 3 automated vehicles and the scenarios were categorized according to urgency, predictability, criticality, and complexity of the driver response. The classification criterion were summarized in Table 2.2. And an example of classification based on this criterion can be referenced to Table 3 in [4].

- The **urgency** of a testing scenario indicates how fast a driver takeover reaction is required.

- The **predictability** of a testing scenario refers to available knowledge about the existence and location of a system limit.
- The **criticality** of a testing scenario refers to the cost of failing to take over vehicle control in time.
- The **complexity** of the required drivers' response refers to the human action needed to resolve a transition demand.

**Table 2.2:** Classification of psychological dimension of scenarios [4].

Factors	Low	Medium	High
Urgency	High time budget	Medium time budget	Small time budget
Predictability	Near-term detection of the system limit	Predictable, but occurrence dependent on situation conditions	Known from map, backend, V2V, etc.
Criticality	Low safety risk	Increased safety risk	High safety risk
Complexity	Low complexity (e.g. stabilizing)	Medium complexity (e.g. steering)	High Complexity (e.g. lane change)

### 2.3.3 Human-Related Factors

Humans are complicated objects that are difficult to study, whereas it is still possible to gain some insights into effects of human-related factors through statistical methods. Besides demographics (age and gender), the most frequently researched factors, including experience, eye movements and gaze behaviors, duration of automated driving and reaction speed, will be covered in this section.

#### Age and Experience

It is commonly believed that one's cognitive ability decreases with ages. Hence, it is also anticipated that one's driving performance also degrades gradually as one gets older. Although there might be variations among individuals [69], the overall tendency can not be negated.

For manual driving, the impact of this decline has been studied thoroughly. As summarized in [70], cognitive reduction, sensory impairment, and physical limitations were three of the factors that influenced the old driver's capacity to drive safely. Obviously, visual impairment, hearing loss, grip strength reduction and flexibility decrease, etc., were commonly perceived conditions within old people and could lead to increased crashes. Effects of cognitive ability, however, were two-sided. On one hand, cognitive reduction can possibly lead to attention



deficit, which was associated with crash risks. On the other hand, accumulation of experience and change of cognitive ability may also influence the old driver's beliefs about his own capacity to drive safely. Therefore, it was highly possible that an old driver with cognitive reduction can change his driving behavior accordingly, making the driving safer.

For automated driving, some of the results obtained above still apply, whereas with some peculiarities. Facing the growing portion of older drivers in the population, takeover performance of two age groups—older drivers ( $\geq 60$  years) and younger drivers ( $\leq 28$  years), was investigated and compared in [57], with each consisting of 36 objects. In the experiment, the participants were asked to drive either with or without a NDRT (between-subject factor), and with either no, medium or high traffic density (within-subject factor, and is the same as [25]). Contrary to the common belief, the results showed that the older drivers responded as fast as the younger drivers, although they had different patterns of driving. Besides, learning effects were also observed in both age groups. As the drivers became familiar with the system, the takeover time was shorter, the TTC was longer and the maximum lateral acceleration was smaller, which are indications of safer driving. A similar study was given in [71], but the conclusion was a little different. Thirty-seven younger drivers (20-35 years) and thirty-nine older (60-81 years) were involved in this study. Results showed that the older drivers took more time to finish the takeover process than the younger ones. However, except some operational differences, both the old and young drivers can finish the takeover process successfully and efficiently.

A finer study was conducted in [72], where the drivers were separated into 4 age groups. The objective was to investigate the ability of drivers within different age groups to inhibit visual and auditory distractions. Eighty-nine subjects were involved in a driving simulator study, including 24 young drivers (19-25 years), 17 middle-ages drivers (35-45 years), 24 old drivers (56-65 years) and 24 older drivers (70-80 years). While driving on a simulator, the participants were asked to respond to critical events, including the distractor stimuli and the brake lights of the car ahead, with their responses recorded. Results showed that visual distractors had a stronger impact than acoustic ones, and in critical traffic situations, the young and oldest age groups suffered more from inhibition deficit than the other age groups, which should be kept in mind when developing driver assistance systems. In this sense, it seems that there is not a linear relationship between performance and age of drivers, which kind of explained the results in [57].

In addition, as mentioned above, learning effect, or experience gained over



the years might also affect driving behaviors as well as takeover performance of drivers greatly. This may partly explain why the results obtained in different studies are quite different. To verify this relationship, a driving simulator study was conducted in [73], where the specific purpose was to verify the relevance of age and experience to psychomotor skills and its impact on adaptation to automation. Ninety-six participants were involved in this study. For the effect of age, 3 age groups were involved, and the results once again confirmed the conclusion that reaction time of drivers increased with age. The difference was that reaction times were more carefully researched that included extra metrics like mean reaction times, mean motor times, dispersion reaction times, etc. For the effect of experience, the participants were again made into 3 groups, and it was confirmed that response time and the method of taking over the vehicle were highly related to experience of drivers. Besides, it was also found that the older the subjects were, the lower was the frequency of involving in dangerous behaviors. This implies that aging is not all without benefits, and when the drivers are properly trained and educated, safety of drivers and vehicles can also be greatly enhanced.

Overall, it can be concluded that older drivers respond more slowly than the younger ones [74, 75] and also behaves differently [76, 77]. Specifically, in [74], it was found that “younger drivers disengaged significantly quicker than the older cohorts”; in [75], it was concluded that “compared to younger participants, older adults showed longer response times”; in [76], it was shown that younger and older drivers have “distinct preferences on the type of activity for each age group”; and in [77], it was pointed out that younger drivers engaging in both single and multiple NDRT engagements perform better than that without NDRT engagement, whereas older drivers were not. Nevertheless, it does not necessarily indicate that takeover performance of older drivers will be worse than younger ones. Many factors need to be taken into consideration before the final conclusion can be made. At least, experience of older drivers can in a sense counterbalance the negative effect of late response due to aging. As summarized in [73], “deficiencies in older drivers, as well as some elements related to the lack of experience in younger drivers, can have a significant impact on the behavior of drivers when driving an automated vehicle.”

### **Gaze Behaviors**

Since most information is perceived visually while we are driving, vision distraction might be the most safety-critical type of distraction [36]. As Horrey

and Wickens indicated in [78], as glance duration of drivers on NDRTs rose, the response time to hazard events also increased, as well as the likelihood of getting in a collision. A more direct argument came from [79], where it was shown that during demanding visual task, the viewing time spent on the road center reduced from 80% in baseline driving to 29%. Since road center region was hypothesized to carry the most important information for driving, the reduced glance from road center was therefore associated with reduced path guidance information and increased reaction time to changes. As the task became more difficult, the effect also became more obvious [80]. S.G. Klauer et al. conducted a more thorough research in [44], and the results from the 100-car naturalistic driving study data indicated that “drivers engaging in visually and/or manually complex tasks have a three-times higher near-crash/crash risk than drivers who are attentive”, whereas “short and brief glances away from the forward roadway for the purpose of scanning the driving environment are safe and actually decrease near-crash/crash risk”.

However, these results mostly came from manual driving, to explore the relationship between the drivers’ gaze behavior and takeover performance during automated driving, K. Zeeb et al. proposed a model in [36] that was representative of the processes underlying takeover from automated driving while doing secondary tasks. Data collected from 107 participants revealed that compared with manual driving, driver’s single glance at the secondary tasks was 11 times higher (12.6 s v.s. 1.1 s), while the glance time on road seemed to be unaffected (0.8 s v.s. 0.6 s). The latter might be due to the fact that “the glances at the road have to be of a certain duration in order to adequately perceive the driving environment and to update the mental model of the driving situation” [36]. Based on the number of glances on secondary task and the maximum duration of off-road glances, drivers were classified into low, medium and high-risk drivers, where the high-risk drivers showed less frequent glances at the central display and longer maximum eyes-off-road time than the others. It was demonstrated that facing an emergent takeover situation, high-risk drivers started braking later than the other two groups of drivers and also showed more collisions with the surrounding traffic. This is in good correspondence to the result in [78].

Further in [81], gaze behavior of truck drivers during critical takeover conditions was studied. From results of nine takeover situations, it was found that the drivers were less likely to gaze into the rear mirrors within one second after the takeover request, and the oncoming traffic on the left lane was only considered between two and five seconds. Hence, it was concluded that, “due to temporal criticality, the drivers had no enough time to safely manage the situation even if

reaction times were very quick” and “although drivers overtaken critical obstacles with low TTCs, the gaze behavior analysis offered indications that the takeovers were not of sufficient quality.” When evaluating the takeover performance of the driver, gaze behavior of drivers is therefore necessary. In fact, driver’s gaze behavior can in a sense reflect driver’s trust in the automated system, as suggested in [82], “participants reporting a higher trust level spent less time looking at the road or instrument cluster and more time looking at the NDRT”. Similar results can also be observed in [83]. Therefore, automated vehicles should be designed to achieve optimal trust [84], so that the drivers can keep appropriate attention on the road conditions.

### Duration of Driving

The negative impact of long time driving on human performance is obvious, as suggested in [85], “human drivers can have major takeover issues if they are driving conditionally automated for longer periods of time” and “takeover performance decreases with increasing time of automated driving”. However, exactly how long the duration of automated driving could be is difficult to decide. Results obtained from different studies can also be very different, even with contradictory results.

For comparison of the influence of duration of automated driving on takeover performance, 25 and 50 min of automated driving were implemented in [85]. Nevertheless, these were not necessarily the most optimal time period. As indicated in [86], the drivers reported a drowsiness only after 15 min of automated driving, which was followed by a decline of takeover performance. However, in [87], a slightly different result was reported. Similar to [85], this study aimed to investigate the effect of different duration of automation on takeover performance as well as gaze behavior during automated driving, with different time periods (5 v.s. 20 min). Results showed that whilst reaction time increased significantly after 20 min of automated driving in the group performing SuRT as NDRT, the other parameters concerning takeover performance (TTC, longitudinal and lateral acceleration, etc.) did not show much difference. Hence, it was concluded that no significant difference between 5 and 20 min of automated driving on takeover performance could be found. This maybe due to insufficient length of the chosen intervals or because of the training effects prior to the study to get the participants familiarized with the driving simulator. As for the effects on eye gazes of the driver, it was indicated that “the participants showed more self-initiated distraction after longer duration of automated driving by averting

the eyes from the driving scene and letting the gaze wander due to monotony and boredom". This was in correspond to the first conclusion.

A more distinctive result was presented in [53], where vigilance decrement (assessed by an auditory secondary task) and passive fatigue (measured based on eye tracking results) of the driver were evaluated in a driving simulator experiment for 42.5 min of driving with partially automated systems. Twenty objects were involved in the experiment. Eye tracking results did indicate that there was an increment in fatigue during the 42.5 min of automated driving. Nevertheless, no significant increase of the reaction time was found, possibly because of the tasks involved.

In view of the results presented, it can be seen that it's very difficult to achieve a uniform conclusion, as the influencing factors are various. A similar concern was expressed in [88], which said "the study did not show that a longer engagement led to a worsening performance" and "additional tests are needed to further investigate the dependency on duration of engagement". In this study, the duration of automated driving was randomized within one of the three intervals—0-10 min, 10-20 min, and 20-30 min. Results from a driving simulator manifested that although the duration of automated driving indeed had certain effect on maximum lane offset (the maximum distance the vehicle drifts from the center line after takeover), it did not seem to affect the integral offset ratio (integral drift over an S-turn curve compared to a baseline obtained in manual driving). Besides, contrary to the common belief that the longest duration of automated driving worsens the performance most, it was indicated that duration between 10-20 min appeared to be worse than that within 20-30min. More research might be needed to verify this result.

Furthermore, considering the fact that the driving time of above research is relatively short, impact of long periods of driving was inspected in [58]. In this study, the driving scenario was designed to include three phases of automated driving, lasting for 10, 60, and 10 min respectively. Each phase was followed by a takeover request (TOR), and after each TOR, the drowsiness of the driver was verbally evaluated using a five-level Likert scale. The results revealed that after 60 min of driving, the driver takes on average 0.5 s longer to begin the lane change operation facing an accident. In addition, the avoidance maneuver and quality of control also were found to be worse after 60 min of automated driving. Moreover, it was claimed that sequencing long and short periods of automated driving can enhance safety during takeover, which was said to be able to help the drivers maintain a good level of vigilance over time. The last statement has also been verified in [89], which asserted that "when disengagement was more

predictable and system-based, drivers' attention towards the road center was higher and more stable". This could possibly be taken into consideration when designing human-machine interfaces.

### Reaction Speed

Takeover is essentially a task-switching process, where drivers switch from the out-of-the-loop-state to in-the-loop-state to take up longitudinal and lateral control of the vehicle. Hence, it is anticipated that drivers' ability to take over vehicle control when being engaged in NDRTs is probably related to drivers' multitasking ability and reaction speed [90]. In view of this, Körber et al. [91] conducted an experiment aiming to predict drivers' takeover time by results of two psychometric tests, including an multitask test and an reaction time test. Results showed that effect of reaction time was significant. However, a stable difference could found between the worst and the best multitaskers, although the effect diminished gradually in the course of the experiment, possibly due to training effects.

Yoon et al. [37] took one step further, and studied drivers' reaction speed in two aspects—motor reaction (reflexive physical actions to prepare the body for driving) and mental reaction (cognitive process where drivers process information gathered from the surrounding environment and make decisions). Through a conceptual framework, it was found out that physical attributes of NDRTs have a significant influence on motor readiness of drivers (e.g. which hand in use, mounted or handheld), whereas visual and cognitive attributes of NDRTs have significant influence on mental preparation time. Compared with real takeover scenarios, takeover time obtained through simple experiments in [37] was shorter. However, it could still be used as basis to determine the minimum time required for drivers to take over the vehicle. In this study, 2 s was suggested.

## 2.4 Modeling of Takeover Behaviors

When an automated vehicle has driven out of safety domain and needs to hand over control over the vehicle to the driver, it is important to make sure that the driver is ready to take over the vehicle. According to the level of readiness of the driver, different strategies can be taken to ensure safety of the driver as well as the vehicle to the largest extent. Compared with research on influencing factors, literature regarding evaluation and classification of takeover readiness of drivers

working with ADS is relatively few. To enhance precision for classification and also to raise safety level of the systems, more research in this area is necessary.

In order to characterize vehicle states where drivers were capable of safely operating the vehicle, a driver controllability set (DCS) was introduced in [92], which was defined as a subset of the vehicle states and could be updated online for real-time applications. The basic idea was, as long as the vehicle states did not leave the DCS during the time that the driver needed to take over control, a transition to manual driving could be considered as safe. However, how to properly define the boundaries of the DCS is not an easy task. In [92], an example was given based on two sets of real world data. One was for identification of the DCS, collected from manual driving, and another was for assessing safety of the takeover, collected during activated adaptive cruise control (ACC). Results showed that the method could distinguish between less and more critical situations given the DCS, which could give an indication of whether the transition was safe or not. As this method has not taken drivers' states into consideration, the problem is that it is relatively conservative and might be prone to false alarms.

Rather than vehicle states, a method for evaluating takeover readiness of drivers via recognizing driving posture was proposed in [93]. This method is in essential a pattern-matching method. The basic idea was to create a dataset that includes all the normal models. In real-time evaluation, driving posture of the driver was constantly monitored using an on-board depth sensor, and when the recognized posture deviated from the normal models too much as measured by certain distance metric, the posture was treated as abnormal. Nevertheless, since certain postures can hardly be uniquely corresponded to states of the driver, the method might not be so robust. Hence, it is better to take other factors into consideration. Moreover, a data-driven approach was proposed in [94] based purely on in-vehicle vision sensors. In rating takeover readiness, subjective ratings were collected first, provided by human observers viewing videos from the driving simulator. Using the normalized and interpolated ratings as the ground truth, termed as Observable Readiness Index (ORI), a machine learning algorithm was trained and evaluated to estimate the level of takeover readiness. In real-time applications, the features were extracted framewise based on CNNs (Convolutional Neural Networks), and then, an LSTM (Long Short Term Memory) was used for learning their temporal dependencies and giving the estimated ORI. The results showed that the best model could achieve a mean absolute error (MAE) between the predicted and assigned ORI values of 0.449 on a 5 point rating scale. With vanilla LSTMs, the results could be better.



However, as the essential part of this method, the ORI is prone to influence of the drivers' individual inclinations, how to guarantee the objectiveness of the ORI is a problem that must be solved.

**Table 2.3:** Summary of modeling of takeover behaviors.

Research	IVs	DVs	Methods
[63]	Age, Repetition, TB, TD, NDRT, Lane	Takeover Time, Time to Collision, Crash, Brake Application	Nonlinear Regression, Logistic Regression
[37]	Physical Features, Cognitive Features, Gaze Features	Takeover Time	MLR
[17]	Driver Features, Subjective Trust, Monitoring Strategy, System Features, Environment Features	Takeover Time, Takeover Quality	SEM
[2]	NDRT, EoRG, Traffic Situation	Takeover Quality	kNN, NB, RBF SVM, Linear SVM, LD
[95]	HR Indices, GSR Indices, Gaze Behaviors, TD, TB, Scenario	Takeover Quality	SVM, NB, RF, kNN, DA, LR
[96]	Age, LAD, SIM, TOR Features, NDRT Features, TB, Urgency, etc.	Takeover Time	XGBoost, Linear Regression, SVM, Fine Tree, RF
[97]	Eye Movement, Heart Rate, GSR Signal, Questionnaires, NDRTs, Vehicle Data	Takeover Intention, Takeover Time, Takeover Quality	DNN, LR, GB, RF, NB, Ada Boost, RGF

TB: Time Budget; TD: Traffic Density; HR: Heart Rate; GSR: Galvanic Skin Response.  
LAD: Level of Automated Driving; SIM: Fidelity of a SIMulator.  
MLR: Multiple Linear Regression; SEM: Structural Equation Model.  
GB: Gradient Boosting; RGF: Regularized Greedy Forest.

To improve accuracy of classification, instead of relying on single source of information, several kinds of factors were combined in some models proposed in [2, 17, 63, 94, 95, 97], etc., which were summarized in Table 2.3. With all those factors in hand, problems then arise as to how to obtain an appropriate classifier and how to choose a relatively small number of features. In [63], regression models were built for prediction of takeover time, time to collision, crash and brake probabilities. As the models were expressed explicitly as time-budget,

traffic density, NDRT, repetition, the current lane and driver's age, etc., they were quite understandable and had high interpretability, allowing detailed insights into the influencing factors that affected takeover performance in critical takeover scenarios. In [17], a structural equation model (SEM) was proposed based on the ACT-R cognitive architecture, which took driver's personal characteristics, subjective trust, monitoring strategy, system characteristics, and environmental characteristics into consideration, and used drivers' takeover time and takeover quality as endogenous variables. The results revealed that time budget was the most critical factor in promoting the safety and stability of the takeover process. Together with traffic density and driver characteristics, drivers' takeover quality could be determined directly. Besides, it was also found out that drivers' trust in the ADS, as an intermediate variable, could also explain a major portion of the variance in takeover time, suggesting again importance of a certain level of trust on safety of the ADS.

In [2], in order to get an appropriate classifier, five classifiers—k-Nearest Neighbor (kNN), Support Vector Machine (SVM) with a Radial Basis Function (RBF) kernel, SVM with a linear kernel, Naive Bayes (NB) and Linear Discriminant (LD), were evaluated and compared, and the results showed that compared with the other four classifiers, SVM with a linear kernel obtained the highest accuracy. In addition, it also had a modest execution time. Hence, it was chosen as the classifier in selecting an appropriate number of features. By applying the chosen classifier on different combinations of the features, it was shown that the EoRG features had a relatively lower influence on the classification performance than the other features. This provide some insights on the selection of features for balancing between execution time and accuracy. Similar research can also be found in [95], where as inputs were driver's physiological data (heart rate, galvanic skin response, eye-tracking data, etc.) and external environmental data (scenario type, traffic density, TOR lead time, etc.), as output was the driver's takeover performance, which was categorized as good or bad according to the driving behaviors during takeover. Based on data collected from 102 objects, six classifiers were trained and compared, including SVM, NB, Random Forest (RF), kNN, Discriminant Analysis (DA) and Logistic Regression (LR). The results showed that RF yielded the best result, with an accuracy of 84.3% and an F1-score of 64%.

The classifiers tested in [2] and [95] are mostly traditional machine learning methods, with the enhancement of computing power of electronic control units and development of more advanced algorithms, deep learning methods have also been more and more introduced into automotive applications, such as



CNN and LSTM applied in [94]. As a unified framework for predicting takeover performance, DeepTake was proposed in [97], which aimed to predict takeover performance in three respects—takeover intention, takeover time and takeover quality. For this purpose, three different DNN (deep neural networks) were built and trained that had the same input and hidden layers whereas different output layers. As input, various sources were considered, including driver's pre-driving survey response, vehicle data, engagement in NDRTs, and driver biometrics, etc. Results collected from driving simulator experiments involving 20 participants manifested that DeepTake can achieve an accuracy of 96% for the binary classification of driver intention, 93% and 83% for the multi-class classification of takeover time and quality, which outperformed other six machine learning-based methods (Logistic Regression, Gradient Boosting, Random Forest, Bayes Network, AdaBoost, Regularized Greedy Forest) in all aspects. Considering that the sample is relatively small (20 objects), more tests and experiments are necessary to develop robust and applicable algorithms.

## 2.5 Performance Evaluation Metrics

For a complete overview of takeover performance measures, a good summary was provided in Table 3 of [3]. Roughly, they can be classified into time-related and quality-related measures, which were summarized in [1]. For convenience of reference, some frequently used concepts in the literature are listed below.

### 2.5.1 Time-Related Measures

#### 1. Takeover Time

Time interval between onset of TOR (or RtI, Request to Intervene) and driver-initiated intervention or deactivation of the ADS. The transition process can be further decomposed into further sub-processes:

- time to first driver reaction (e.g. interruption of NDRTs);
- time to start of visual re-orientation;
- time to visually fixate RtI message (if visual HMI is involved);
- time to visually fixate road center (or other aspects of the scenery);
- time to start to move hand to wheel/feet to pedals;
- time to grasp wheel/touch pedals;
- time to start to operate relevant vehicle controls;

- time to onset to override or deactivate the ADS.

## 2. Intervention time

Time interval required by the driver to handle the imminent take-over situation by performing an appropriate driving maneuver.

## 3. Stabilization time

Time duration it takes for an individual user to reach a similar or comparable quality level of manual driving performance as in ordinary manual driving by an average driver.

## 2.5.2 Quality-Related Measures

### 1. Safety-oriented, objective measures

- ability to avoid collision/prevent run off-road events;
- omission of visual checks (e.g. mirror use);
- operation errors (especially related to system deactivation);
- maximum longitudinal/lateral acceleration;
- frequency of strong or emergency braking;
- minimum time to collision ( $TTC_{min}$ );
- minimum time to lane crossing ( $TLC_{min}$ );
- frequency of “near misses”; etc.

### 2. Sensitivity-oriented, objective measures

As opposed to the safety-oriented measures that are relevant to the critical situation, other measures can be used to show potential mid- and long-term detrimental effects of having to regain manual control after an extended period of automated driving. These measures include:

- Standard deviation (SD) of lateral position/steering wheel angle;
- yaw rate error and SD of yaw rate error;
- metrics of distance to other vehicles/objects;
- metrics for longitudinal control quality, e.g. speed behavior.

### 3. Subjective measures

- by the driver
- by an external observer

## 2.6 Future Challenges and Discussions

Although the influencing factors have been discussed separately in most literature, it should be noticed that, among them, interactive influence also exists. In [72], for example, instead of studying effect of driver's age only, 3 more factors were considered—task instructions given to the driver, stimulus and distraction types. To explore the relationship between every two of the factors, a method called analyses of variance (ANOVA) was utilized. Then, homogeneity of the variance was assessed using Levene's test. Based on the results, conclusions can be extracted that have been briefly discussed above. Similarly, studies where multiple factors were taken into consideration can also be referenced to [25, 26, 38, 43, 57, 98], etc. A brief summary and comparison of the factors considered in some of the studies is given in Table 2.4. As more factors are considered, the complexities will also go up. Therefore, one of the challenges is to find a method to integrate all these factors together. To realize this, a thorough experiment involving as many factors as possible is necessary. Based on the experimental results, classical machine learning methods or modern deep learning methods can be applied, as that in [60, 97, 99], etc. The problem with these methods is that they did not build the interactive influences between various factors into their model, which can add more interpretability to their models. This is one of the directions in this area of research. Only based on the experimental results is it possible to choose the most influencing ones, possibly using machine learning-based methods.

Another challenge is the perception problem. As most research have been conducted on driving simulators, we tend to make the assumption that all the factors needed could be provided in advance and exactly. However, this is not necessarily true in real scenarios. During driving, traffic situations, NDRTs, EoRG, or vehicle states, etc., are all dynamically changing factors that can not be known ahead of time. This poses high demand on the algorithms and computing unit used for evaluation of these factors. Vision-based methods have been widely studied and are very promising, especially with the fast development of artificial intelligence. However, some problems still remain, particularly with the robustness and speed of the algorithms, which are highly valued in real-time applications. Some tasks that are trivial to human beings can sometimes be quite confusing to machines, such as change of illumination, eyeglasses worn by the driver etc. Hence, it is quite essential to develop more robust, more precise and faster algorithms for driver monitoring as well as traffic situations estimation. This is the basis for evaluation of takeover readiness of drivers.

**Table 2.4:** Summary of factors influencing takeover behaviors.

Related Work	Factors	Levels of the Factor
[38]	level of automation extent of failure	semi- or highly automation; complete, severe or moderate
[25]	traffic density secondary tasks	0, 10, or 20 vehicles per km; with/without a verbal TQT
[72]	ages secondary tasks stimulus types task instructions	young, middle-aged, old or older; visual or acoustic distractor; brake light or brake light in combination with Go-/NoGo-distractor; perception, detection or discrimination
[39]	level of automation traffic density	manual or highly automated driving; light or heavy
[43]	level of automation familiarity with ACC ages	non-assisted, driver assistance or partial automation; prior experience with ACC; > 50 years vs. ≤ 50 years
[47]	level of automation scenario secondary tasks	manual or highly automated driving; critical or non-critical situations; with or without a verbal TQT
[55]	level of automation secondary tasks	manual, semi- or highly automation; 6 different NDRTs
[57]	ages traffic density secondary tasks	≥ 60 years or ≤ 28 years; no, medium or high traffic density; with or without a verbal TQT
[98]	secondary tasks driver related factors	3 different NDRTs; prior experience with ACC and learning effects
[81]	traffic situations duration of driving secondary tasks	9 takeover scenarios in 3 blocks; 30 s, 120 s or 240 s; video or game
[88]	ages vehicles speeds duration of driving	18–35 years, 35–55 years or 55+ years; low (55 mph) or high (65 mph); < 10 min, 10–20 min or 20–30 min
[71]	ages road types level of automation	younger (20–35 years) or older (60–81 years); city road and motor way; monitoring driving or disengagement from driving
[73]	ages years of driving distractor	18–26 years, 27–39 years ,or 40–65 years; Mean = 3 years, 10 years or 25 years; no distractor, fog or SuRT
[26]	lead time secondary tasks	short, medium or long; 5 different NDRTs or monitor

Finally, when we talk about evaluation of takeover readiness of drivers, we usually are evaluating current takeover readiness of drivers based on information in the past, including takeover readiness (whether the driver is ready to take over the vehicle), takeover time (how long will it take before the driver successfully takes over the vehicle) and takeover quality (quality of the takeover after resuming control of the vehicle), etc. This is important, yet not enough. During automated driving, the drivers are constantly monitored and evaluated, so that different strategies can be taken with different levels of attentiveness and readiness. In situations the ADS can not handle, if the driver has a high level of readiness, timely and smooth switching from ADS to the driver shall be able to be guaranteed; in case the driver is not ready, however, the ADS shall also be capable of compensating and transferring the system to a safe state with acceptable risks. Then arise the questions: Will it be better if the driver's activities can be learned, tracked and predicted? In view of this, research regarding real-time driving tracking, prediction and behavior learning becomes necessary. By evaluating and quantifying driver readiness, different decisions can hence be taken, ensuring safety of the vehicle to the greatest extent.

## 2.7 Summary of Chapter 2

Chapter 2 aims to fulfill Objective 1, and the main results and conclusions are:

- Factors influencing takeover behaviors of drivers during automated driving were discussed in detail, which can roughly be classified into system-related, scenario-related, and human-related factors.
- To gather all these influencing factors together, a complete framework for evaluation of takeover behaviors of drivers was proposed. This architecture takes scenario characteristics, system characteristics as well as driver characteristics as inputs, and outputs the takeover behaviors.
- As the final step for evaluating drivers' takeover behaviors, several modeling methods were also discussed. Although information utilized in each study is different, the tendency to use machine-learning-based methods in the final classification is the same.
- Gaps and future challenges were also pointed out.

Overall, by gathering all these ingredients together, this chapter aims to answer the question "How do we know the driver is ready to take over?", so that optimal measures can be taken by the system, and the takeover process can be

**Table 2.5:** Conclusions regarding factors impacting takeover behaviors.

Factors		Main Conclusions
System-Related	Level of Automation	Higher LOA leads to reduced SA and more engagement in NDRTs.
	NDRT	Demanding NDRTs, especially manual NDRTs increase reaction time, and decrease takeover quality.
	HMI & TOR	Auditory and vibrotactile TORs better than visual ones, and more advanced HMIs help drivers make decisions.
Scenario-Related	Traffic Density	Higher density increases alertness of drivers, takeover time and crash rates.
	Vehicle Speed	Higher speeds reduce engagement in demanding NDRTs and accuracy of speed estimation.
	Time Budget	Shorter time budgets lead to shorter takeover time, and deteriorate takeover quality.
Human-Related	Age and Experience	Older drivers are believed to react more slowly, whereas this can be compensated by their driving experience.
	Gaze Behaviors	ADS leads to more glance at NDRTs, associated with longer reaction time and higher collision rates.
	Duration of Driving	Longer duration of driving usually leads to decreased performance, however, optimal duration of driving may exist.
	Reaction Speed	Drivers' reaction time is related with their multitask ability and reaction speed, including motor reaction and mental reaction.

taken timely, safely and smoothly. Based on the results, in Chapter 3, experiments will be designed and conducted to explore effects of some factors that have not been studied in the existing research, which we believe will contribute to the existing literature.

## Chapter 3

# Takeover Experiments and Data Analysis

### 3.1 Purpose of the Experiment

In most of the studies in Chapter 2, TORs were issued during NDRTs when drivers were visually distracted. Issuing the TOR under such a condition would be like interrupting the driver during one task with another, which has been found to increase one's anxiety and time to complete the new task than when the same task was presented between the tasks [100]. Based on this idea, [101] proposed an attention aware system to issue TORs at emerging task boundaries and found that compared with issuing TORs within tasks, issuing TORs at task boundaries could decrease stress and lead to better takeover performance. The results were very meaningful, whereas kind of limited in that they did not tell us about the situations when drivers were left a few seconds before the TOR to observe the surroundings after the NDRT has ended. That is similar to a two-stage warning system, except without an explicit warning before the TOR.

Since it was discovered that “requiring drivers to constantly attend to the forward roadway did not necessarily benefit drivers' situation awareness compared to more naturalistic driving without specific task requirements” [102], instead of delivering the first warning explicitly and remind the driver to pay attention like that in a two-stage warning system, we would like to simply issue the TOR after the NDRT has ended for a few seconds and give the driver some time to monitor the surroundings, just like that by [101]. As demonstrated by [89] and [103], while transferring to manual driving from automated driving, drivers' eyes-on-road gazes (or percentage road center) increased slowly in the first 15 seconds, and after reaching a peak, they would lower a bit [89] and remained steady [103] afterwards, which may guide us in the experiment design.

Moreover, although comprehensive models for modeling takeover behavior

involving a lot of factors have also been proposed in [17, 63, 95–97], etc, the effect of personality on takeover performance has received little attention, which have been considered to be highly related to driving behaviors, especially the risk-taking attitudes of drivers [104]. Hence, one of the purposes of the experiment is to explore impacts of duration of monitoring (DoM, defined as the time left for drivers to divert their attention away from NDRTs and monitor the surroundings before the TOR) and drivers' personality traits on takeover behaviors. Finally, together with other factors, including drivers' age and experience, reaction speed, time budget, and NDRT, etc, to build a prediction model to predict drivers takeover behaviors before the TOR, so that optimal decisions can be made by the ADS based on the predicted results. In addition, using the data collected, drivers patterns of takeover behaviors will also be analyzed, along with some suggestions for safe takeovers.

## 3.2 General Experimental Design

### 3.2.1 Participants and Equipment

This experiment was approved by the Ethics Review Committee of the University of Tokyo (NO. 17-14). Forty-eight subjects (male: 38, female: 10) participated in the experiment, most of whom were students from the universities in Tokyo, Japan. All of them have a valid Japanese driver's license and have highway driving experience. And the distributions of ages and driving experience of the participants were summarized in Table 3.1. As for their experience with driving assistance systems, 6, 7, 3 and 7 participants have experience with Automatic Emergency Braking (AEB), Adaptive Cruise Control (ACC), Lane Change Assistance (LCA) and Collision Avoidance System (CAS) respectively. Besides, 9 of them have prior experience with driving simulators.

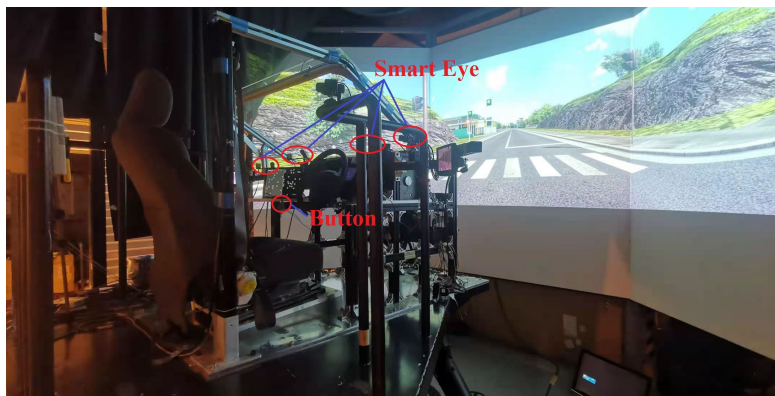
**Table 3.1:** Distributions of ages and driving experience of the participants.

Ages	Num of Participants	Driving Experience (Years)	Num of Participants
18–23	26	$\leq 1$	12
24–30	16	1–3	14
> 30	6	3–5	11
-	-	5–10	6
-	-	> 10	5

A high-fidelity driving simulator with a Stewart motion platform of six degrees of freedom was used for the experiment (Fig. 3.1). The position and



height of the seat were adjustable, allowing each participant to adjust the seat to a comfortable position. The front field vision was  $163.8^\circ$ , which was comparable to realistic driving scenarios. Two screens of 7 inches were used as left- and right-view mirrors on both sides of the driving simulator respectively, and another screen of 7 inches was used as speedometer installed behind the steering wheel, the same as that in a real vehicle. Finally, as the driver's seat in Japan is on the right side, a Surface Pro 7 was installed on the left-front side of the driver to simulate the infotainment system.



**Figure 3.1:** Driving simulator.

To activate the ADS, a button was installed on the left hand side of the driver. To deactivate the ADS, the driver could use either steering wheel, brake or accelerator pedals, where the thresholds for the steering wheel and the pedals were  $0.8 \text{ N}\cdot\text{m}$  and 20% respectively. These thresholds were set for this driving simulator to consider both the sensitivity and false alarm rates and may not be suitable for other driving simulators.

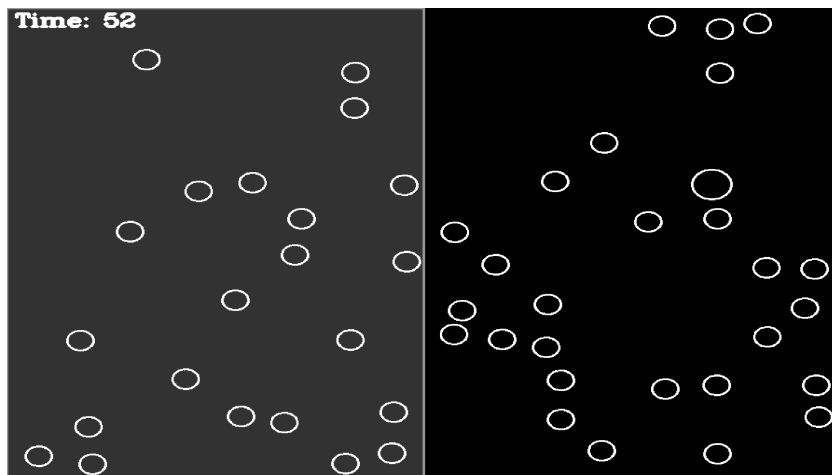
To detect eye movement and track eye gazes of the participants, the Smart Eye Pro eye tracking system with a four camera configuration was adopted. Of the four cameras, two were installed around the left and right A-pillar respectively, one was put right in front of the driver, and the last one was positioned left-front of the driver (Fig. 3.1). With this configuration, an accurate gaze output was available for head rotations up to about  $\pm 75^\circ$ . Besides, an additional high-speed camera was installed left-behind the driver's seat in the latter half of the experiments (the last 27 participants), so that dynamic eye gazes of drivers could be recorded and analysed using Smart Recorder and MAPPS.

### 3.2.2 Non-Driving Related Task

To distract the driver, the Surrogate Reference Task (SuRT) [105] was chosen as the NDRT in this experiment (Fig. 3.2). In this task, the screen was divided into two

parts, with one part including a circle bigger than the others (50 smaller circles). The task of the participant was to locate the part of the screen with the bigger circle using “left” and “right” keys, and then press “enter” to decide the answer. For this, a small keyboard with only direction and enter keys was fixed beside the tablet (Fig. 3.1). The duration of the SuRT was set as 60 s, with a countdown displayed on the upper-left corner. All the operations of the participants were recorded in a .csv file.

While choosing NDRT, we have referenced standard ISO/TS 14198:2019, where three standardized tasks are suggested for incurring attentional demands. And SuRT is suitable in this experiment in that: (1) SuRT is a standardized task that can be used to produce a range of statistically stable, repeatable and comparable secondary task demands for participants in an experiment; (2) In our experiments, drivers need to be visually distracted. Due to SuRT’s nature as a visual-manual task with proper cognitive demand, it is superior than the n-back task which induces a high level of cognitive demand whereas no visual or manual demands; (3) Although critical tracking task is also a visual-manual task, since it is too manually demanding compared with SuRT, SuRT would be the preferable choice; (4) The difficulty level of SuRT is adjustable, making it convenient for the later experiments that take the task demand into consideration.



**Figure 3.2:** Surrogate Reference Task (SuRT).

### 3.2.3 Takeover Scenarios

In consideration of duration of monitoring (DoM, defined as the time left for drivers to divert their attention away from NDRTs and monitor the surroundings before the TOR, Fig. 3.3), lead time (LeadT, defined as the time-to-collision at the time of TOR if the driver fails to take over, Fig. 3.3) and traffic densities, etc., 12

scenarios were designed with three levels of criticality (defined as the probability of involving in a collision if not attentive enough, decided mainly by LeadT and whether there are vehicles in the adjacent lane behind) (Table 3.2):

- Low-critical scenarios (TOR10: sudden startup of a truck ahead by the right side of the road, TOR11: merging of a vehicle from the acceleration lane at a relatively far distance, TOR12: obstacles by the left side of the lane);
- Medium-critical scenarios (TOR7: cut-in of a truck from the right side at a relatively far distance, TOR8: animals ahead, TOR9: obstacles in the middle of the lane ahead); and
- High-critical scenarios (TOR1: sudden braking of the vehicle ahead, TOR2: dropped cargo from a truck ahead, TOR3: merging of a vehicle from the acceleration lane at a near distance, TOR4: cut-in of a vehicle from the right side at a near distance, TOR5: approaching a slow moving truck gradually in a thick fog, TOR6: construction site ahead)

More specific definitions and schematic of the scenarios could be referenced to Appendix A.1.

DoM was chosen considering that eyes-on-road gazes rose steadily in the first 15 seconds and then stabilized while drivers divert their attention from NDRTs to roadway during automated driving [89, 103], hence, DoM<sub>S</sub>, DoM<sub>M</sub>, DoM<sub>L</sub> and DoM<sub>LL</sub> corresponded to the start-point, mid-point, end-point and steady-point of this process respectively. LeadT was chosen considering that the mean takeover lead time was  $6.37 \pm 5.36$  s with a mean reaction time of  $2.96 \pm 1.96$  s [16] and 3 s would be too short [65], hence, 5–6 s would be a bit critical whereas enough to avoid accidents for most drivers, and 7–8 s would be less critical. And results of pre-experiments also showed that if LeadT was less than 5 s, there would be a high probability of collision. Therefore, the two LeadTs were chosen as  $5.5 \pm 0.2$  s and  $7.5 \pm 0.2$  s, and denoted as LeadT<sub>S</sub> and LeadT<sub>L</sub> respectively. The  $\pm 0.2$  s is the error of manual calibration, which is less than 5 m if converted to meters given vehicle speed of 80 km/h.

**Table 3.2:** Summary of the takeover scenarios.

Criticality	Scenarios	DoM	LeadT
High	TOR1–TOR6	DoM <sub>S</sub> , DoM <sub>M</sub> , DoM <sub>L</sub>	LeadT <sub>S</sub>
Medium	TOR7–TOR9	DoM <sub>S</sub> , DoM <sub>M</sub> , DoM <sub>L</sub>	LeadT <sub>L</sub>
Low	TOR10–TOR12	DoM <sub>LL</sub>	LeadT <sub>L</sub>

\*DoM<sub>S</sub> = 0 s, DoM<sub>M</sub> = 5 s, DoM<sub>L</sub> = 10 s, DoM<sub>LL</sub> ≥ 15 s.

\*LeadT<sub>S</sub> =  $5.5 \pm 0.2$  s, LeadT<sub>L</sub> =  $7.5 \pm 0.2$  s.

The road was a three-lane highway, with the left lane as the emergency lane. In automated driving mode, the vehicle was set to be in the middle lane at about 80 km/h, except when the lane was blocked and a lane change or a takeover was necessary. Except for TOR5, all the other takeover scenarios were in daylight and sunny days. Overall, the traffic density was low except before and after some of the takeover events, where the participants needed to pay attention to the surroundings. Specifically, (1) in high-critical scenarios TOR1–TOR6, when the takeover request was issued, a vehicle in the right lane behind might be approaching. If the participant was not careful, rear-end collisions might happen. (2) In medium-critical scenarios TOR7–TOR9, when the takeover request was issued, a vehicle in the right lane behind may also be approaching. However, the risk of rear-end collision was relatively low. (3) In low-critical scenarios TOR10–TOR12, when the takeover request was issued, there was almost no rear-end collision concerns. In addition, the expected operations from the driver were also similar. Facing the emergent situation, the driver should slow down the vehicle to avoid collision with the obstacles or vehicles ahead. And when they felt safe enough, they should change lane to the right to pass the obstacles or vehicle to accomplish the whole process.

### 3.2.4 Experiment Design

Per results of several rounds of pre-experiments, when scenarios with the same criticality were taken repeatedly, the learning effect would be too strong, as the scenarios would be much more predictable. To avoid testing scenarios with the same criticality consecutively so as to reduce the predictability of the scenarios, the test scenarios were further divided into three blocks, with each block including 2 high-critical, 1 medium-critical and 1 low-critical scenarios in different orders (Table 3.3, where HC, MC and LC represent high-critical, medium-critical and low-critical respectively). Each scenario was assigned a different DoM. This results in an experimental design with within-subject factors DoM ( $DoM_S$ ,  $DoM_M$ ,  $DoM_L$  and  $DoM_{LL}$ ) as well as the within-in subject factors scenarios (12 scenarios). The specific arrangements when considering DoM at the same time can be referenced to Appendix A.2, according to which, 18 participants are needed to complete a round of experiments. As the low-critical scenarios had almost no safety concerns, they were also meant to be used as disturbance scenarios, so as to ease drivers' tension during the experiment. And the results analysis would also be mainly concentrated on high- and medium-critical scenarios after initial analysis.

**Table 3.3:** Test order of the takeover scenarios.

Block1				Block2				Block3			
HC	LC	HC	MC	HC	MC	HC	LC	MC	HC	LC	HC

\*HC: High-Critical, MC: Medium-Critical, LC: Low-Critical

### 3.2.5 Personality Tests

To test the big-five personality traits, the abbreviated version of the IPIP-NEO inventory—IPIP-NEO-120 was adopted, which included 120 items and was designed to measure exactly the same traits as the original version (IPIP-NEO-300, which included 300 items), but more efficiently with fewer items. Development of the IPIP-NEO-120 and its reliability and validity are described in [106]. And to obtain the scores of the big-five personality traits, the following two links can be referenced:

- <https://drj.virtualave.net/IPIP/ipipneo120.htm>
- <https://bigfive-test.com/>

The tests were taken online before the experiments, and it took 10-20 minutes to accomplish the whole test.

### 3.2.6 Experimental Procedures

Before the experiment, the participants were briefed on the purpose of the experiment and signed the consent form. After completing demographic questionnaires (Appendix A.3 and the big five personality tests, they were briefed on the driving simulator, the non-driving related task (SuRT) and the process of each scenario. Then, they were asked to sit on the driver's seat to adjust the seat and get the smart eye systems calibrated. Following which, they were asked to take 1–2 test scenarios to familiarized themselves with operations of the driving simulator, e.g, how to activate and deactivate the ADS, how to do the non-driving related task, etc. After the preparation, all participants were presented with 12 different takeover scenarios assigned per Table A.2 in the Appendix. Each scenario followed the procedures below (Fig. 3.3):

1. The participants started from manual driving mode;
2. After a few hundred meters, the participants were asked to press the button next to him/her to start the ADS, after which, the driving simulator was switched to automated driving mode;

3. After a period of time, the SuRT started with a 60 s' countdown, and the participants were asked to do the task with full attention;
4. TOR maybe issued during the task (DoM<sub>S</sub>, near the task boundary) or after the task has ended for 5 s (DoM<sub>M</sub>), 10 s (DoM<sub>L</sub>) or  $\geq 15$  s (DoM<sub>LL</sub>), during which the participants had some time to observe the surroundings (the length of the DoM was not known to the driver beforehand, however, they were instructed in advance to divert their attention back to the roadway after the NDRT has ended);
5. Then, TOR was issued, and the participants had about 5.5 s (LeadT<sub>S</sub>) or 7.5 s (LeadT<sub>L</sub>) to take over the vehicle;
6. After taking over, the participants were asked to stabilize the vehicle and further drive for a few hundred meters until they were asked to stop the vehicle by the side of the road;
7. Finally, they were asked to answer a simple questionnaire (Appendix A.4) to self-evaluate the takeover scenario and their takeover performance.

All the narrations were in Japanese. The whole experiment including the preparation took about 100 minutes. After the 12 scenarios, the participants were asked to fill out their bank accounts, so that they can get paid for 2100 yen for their participation.

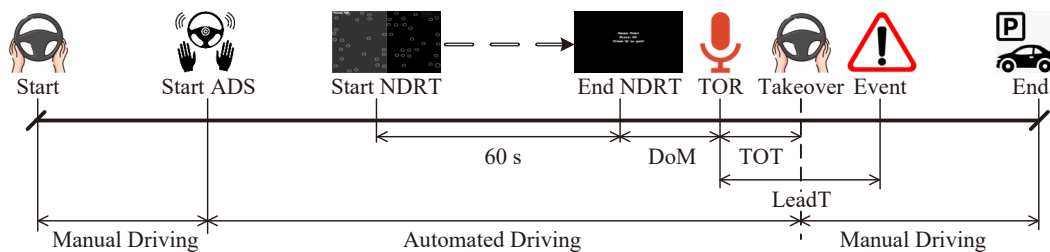


Figure 3.3: Illustration of procedures of each takeover scenario.

### 3.3 Analysis 1: Impact of Duration of Monitoring

#### 3.3.1 Research Questions of Analysis 1

Seeing that the driver's eye gazes are believed to have a great influence on takeover performance [36, 81], and together with the research results from the two-stage warning systems (Section 2.3.2), it is reasonable to hypothesize that performance of drivers would be improved if they were given a few seconds to divert their attention away from NDRT to roadway, even without an explicit



warning. Nonetheless, what remains questionable is whether the performance would be better as the interval before the TOR is further prolonged, i.e., whether there is a monotonously negative correlation between duration of monitoring (DoM, defined in Section 2.3.2 and Fig. 3.3) and takeover time. Considering that different reaction time was recorded due to different configurations of takeover scenarios [81], this relationship would also be examined under different scenarios with different criticality. Moreover, since eye movements and gaze behavior were found to be related with situation awareness of the driver [102], eye-tracking data would also be utilized to explore the different patterns of eye movements and gaze behavior given different DoM.

The general objective of this analysis, as a sum-up of the previous discussions, is to explore the impact of DoM before the TOR on takeover performance, with insights into gaze behavior and eye movements of drivers. Specifically, it tries to answer and discuss the following questions:

- Q1: What are the effects of DoM and takeover scenarios on takeover time? Are there any interacting effects between them?
- Q2: How are patterns of gaze behavior and eye movements (mainly pupil dilation and eyelid opening) affected by DoM?
- Q3: What is the relationship between takeover time and DoM, i.e., are they monotonously negatively correlated? If not, what might be the optimal time interval that might yield the optimal performance?
- Q4: What implications can we obtain regarding effects of eye movements and gaze behavior on takeover time?
- Q5: Is criticality alone enough to explain the effect of scenarios on takeover time?

Q1–Q2 are basic questions we intend to answer in Section 3.3.3 by simple statistical analysis of the results, and Q3–Q5 are more general questions that will be discussed in depth in Section 3.3.4 based on the results in Section 3.3.3. Regarding Q1, we suppose that effects of DoM and scenarios on takeover time would be statistically significant, however, interaction effects between DoM and scenarios are not expected. Regarding Q2, we hope to find statistically significant differences of eye gazes directed to the road center, right side of the road, and the cluster depending on the DoM, and also the overall gaze distribution of the three groups; at the same time, we also intend to discover statistically significant differences of eyelid opening and pupil dilation over time shortly after the TOR given different DoM. Regarding Q3, instead of a monotonously relationship

between DoM and takeover time, we aim to uncover a quasi-parabolic one, so that the optimal interval can be speculated that can be used to guide the HMI design. Regarding Q4, we desire to find evidence from the eye tracking data that support our conclusions made in Q3, so that we can utilize the eye tracking data to infer states of drivers in turn. And regarding Q5, we wish to perceive factors concerning scenarios other than criticality that might have essential influence on takeover performance, paving the way for subsequent studies.

### 3.3.2 Data Processing and Performance Metrics

Data were sampled in 60 Hz, and 432 takeovers were collected from the 36 participants in this analysis. As 8 takeovers were either earlier than the TOR or failed to be recorded, 424 recordings were valid for analyzing takeover time. For eye tracking data, excluding the failed takeovers and malfunction of the Smart Eye System (e.g., hardware failures), data from 381 TORs were successfully recorded, and segments of 15 s before and after the TOR were extracted for analysis.

#### Takeover Time

Takeover time refers to “time interval between onset of Rtl (request to intervene, the same as TOR) and user-initiated intervention or deactivation of an engaged automation function” [1]. In this experiment, it was defined as the time interval between the TOR and a steering torque  $\geq 0.8 \text{ N}\cdot\text{m}$  or a pedal travel  $\geq 20\%$  (including brake and accelerator pedals). The threshold for the pedals was relatively high compared with other researches, e.g., [101], since a fixed pedal travel was found to exist in the driving simulator. At the TOR, the speed of the vehicle was lowered a bit from 80 km/h to 75 km/h and then kept constant until any intervention was detected. Automated emergency braking was disabled.

#### Percentage of Area of Interest

For our purposes, each screen as shown in Fig. 3.1 was defined as an area of interest (AOI), namely the center screen, the right screen, the left screen, the right mirror, the left mirror, the distractor (infotainment), and the cluster (speedometer). Gazes that did not fall into the seven areas were counted as off-road glances. For convenience, the eight AOIs were denoted as CenterScreen, LeftScreen, RightScreen, LeftMirror, RightMirror, Dirstractor, Cluster and OffRoad respectively. For calculating the percentages of AOIs:



1. All takeovers with the same DoM were collected, resulting in three groups;
2. Each group were further sub-grouped per 1-second intervals, resulting in 30 sub-groups for each group;
3. Finally, the percentages of AOIs in each sub-groups were calculated, and finally were plotted as stacked histograms (Fig. 3.5).

### Heat Map and Gaze Entropy

To get an insight into the eye gaze distribution before and after the TOR, heat map would be a proper choice. To realize this, a feasible option would be the stacked scatter plots. To be specific:

1. The relative coordinates of the gaze intersection within each AOI were calculated respectively, with the lower left corner as the origin;
2. The scatter plots of eye gazes of different AOIs were plotted in the same area (Fig. 3.6, the size was determined by the size of the largest AOI, 2 m  $\times$  1.6 m in this research);
3. The area was segmented per size of 0.25 m  $\times$  0.2 m, resulting in a 8  $\times$  8 grid, i.e., 64 cells;
4. Gaze points in each cell were calculated, and finally were plotted into heat maps (Fig. 3.7).

Based on this, we could further calculate gaze entropy, which has been used to measure visual scanning randomness and uncertainty [102, 107]. Denoting each cell as  $(R_i, C_j)$ ,  $i, j = 1, \dots, 8$  and the probability of the gaze points falling into each cell as  $p(R_i, C_j)$ , the gaze entropy could be calculated per Eq. (3.1) [108],

$$H(G) = - \sum_{i=1}^8 \sum_{j=1}^8 p(R_i, C_j) \log_2 p(R_i, C_j). \quad (3.1)$$

### Pupil Dilation and Eyelid Opening

Since pupil dilation and eyelid opening were found by many to be related with situation awareness and alertness of the driver [102, 109, 110], the mean pupil dilation and eyelid opening were also calculated in 1-second intervals. To eliminate the effect of individual differences, the pupil dilation and eyelid opening were normalized according to Eq. (3.2),

$$X = \frac{X - \min(X)}{\max(X) - \min(X)}, \quad (3.2)$$

where  $X$  can be pupil dilation or eyelid opening. In this way, pupil dilation and eyelid opening were expressed as percentage of the range of pupil dilation and eyelid opening respectively, which were convenient for comparison between different participants.

### 3.3.3 Results and Analysis

#### Analyzing Methods

For takeover time, independent variables are DoM and scenarios, and dependent variable is takeover time. Since each participant experienced each DoM and scenario respectively, but not each combination of DoM and scenario, a **two-way ANOVA** was used for inspecting interaction effects between DoM and scenarios, with Tukey HSD tests for multiple comparisons if the results were statistically significant. And **one-way repeated measures ANOVAs** were utilized to examine effects of DoM and scenarios on takeover time respectively, followed by pairwise paired t-tests for multiple comparisons.

For eye tracking data, independent variable is DoM, and dependent variables are gaze behaviors (percentage of gazes and gaze counts directed to certain AOIs) and eye movements (eyelid opening and pupil dilation). Due to non-normality caused by zero gaze data in certain AOIs, non-parametric statistical tests were performed as suggested by [102]. Specifically, **Friedman tests** were conducted to 1) examine the differences of percentage of gazes directed to specific AOIs in certain time intervals according to the DoM, 2) compare the differences of gaze counts in specific AOIs sampled across different takeover scenarios given different DoM, and 3) explore the different patterns of eyelid opening and pupil dilation in certain time intervals depending on the DoM. Wilcoxon signed-rank tests were then applied for multiple comparisons. And **analysis of similarities (ANOSIM)** was performed for comparing overall gaze distribution of the three groups.

For all the tests, results were reported as statistically significant if the p-value was less than .05. All tests were performed in R-4.2.2. Specifically, repeated measures ANOVAs and Friedman tests were performed using *rstatix*-0.7.2, and ANOSIM was performed using *vegan*-2.6-4.

#### Effects of DoM and Scenarios (Q1)

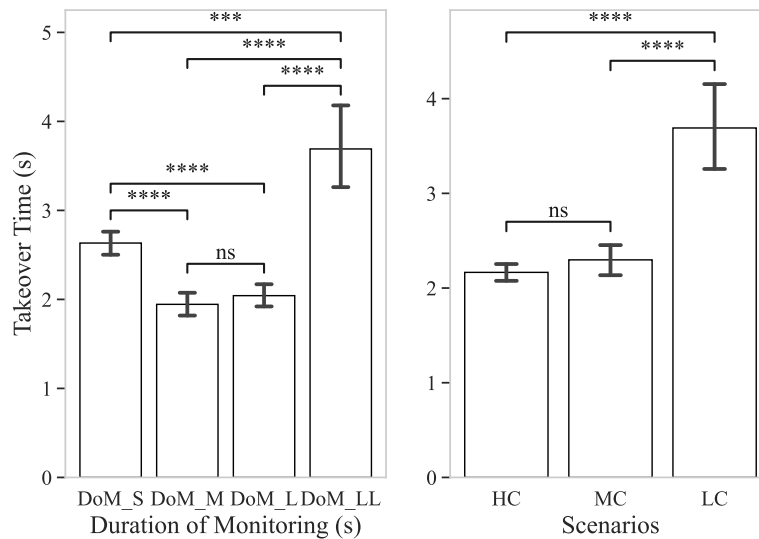
**Effects of DoM on Takeover Time** Since each participant experienced 3 scenarios for each DoM respectively, average values were taken before the analysis,

so that each participant underwent 4 different conditions. An one-way repeated measures ANOVA revealed that the average takeover time was statistically significantly different under different conditions ( $F(1.34, 45.58) = 39.408, p < .001, \eta^2 = .825$ ). Summary statistics grouped by DoM and criticality of the scenarios showed that there was a slight increase of the average takeover time in group DoM<sub>L</sub> compared with that in group DoM<sub>M</sub> both in high- and medium-critical scenarios (Table 3.4). Despite the tendency, post-hoc analysis with a Bonferroni adjustment revealed that the difference between DoM<sub>M</sub> and DoM<sub>L</sub> was not statistically significant ( $t(35) = -0.983, p.\text{adj} = 1.00$ , Fig. 3.4-Left), whereas all the other pairwise differences were statistically significant.

**Table 3.4:** Takeover time (mean and standard deviation) in scenarios with different criticality given different DoM.

Criticality	DoM <sub>S</sub>	DoM <sub>M</sub>	DoM <sub>L</sub>	DoM <sub>LL</sub>
HC	2.48(0.59)	1.94(0.65)	2.07(0.66)	
MC	2.93(0.83)	1.95(0.75)	2.00(0.71)	
LC				3.69(2.49)

\*HC: High-Critical, MC: Medium-Critical, LC: Low-Critical



**Figure 3.4:** Left: Takeover time given different DoM. Right: Takeover time in scenarios with different criticality. ns means non-significant.

**Effects of Scenarios on Takeover Time** Before going into details regarding each takeover scenario, it is better to check the effect of scenarios grouped by criticality. Since each participant experienced 6, 3 and 3 scenarios for each level of criticality respectively, average values were taken before the analysis, so that each participant underwent 3 different conditions. An one-way repeated measures

ANOVA revealed that the average takeover time was statistically significantly different under different conditions ( $F(1.07, 36.53) = 37.644$ ,  $p < .001$ ,  $\eta^2 = .352$ ). From Fig. 3.4-Right, we could observe that takeover time in medium-critical scenarios showed a slight increase than that in high-critical scenarios. However, the difference between them was not statistically significant ( $t(35) = -2.20$ ,  $p\text{-adj} = .104$ ), whereas all the other pairwise differences were statistically significant. As takeover time in low-critical scenarios was far different from that in the other two groups, and also because there was no interaction between low-critical scenarios and DoM<sub>S</sub>, DoM<sub>M</sub> and DoM<sub>L</sub>, we would concentrate on high- and medium-critical scenarios in the following analysis.

**Interaction Between DoM and Scenarios** A two-way ANOVA performed on DoM and scenarios for TOR1–TOR9 revealed that the interaction between DoM and scenarios was not statistically significant ( $F(16, 292) = 1.193$ ,  $p = .272$ ), whereas both DoM and scenarios had statistically significant effects on takeover time (DoM:  $F(2) = 37.67$ ,  $p < .001$ ; scenario:  $F(8) = 8.590$ ,  $p < .001$ ), which were in accordance with the results in Sections 3.3.3 and 3.3.3. The results were similar when performing two-way ANOVAs for high- and medium-critical scenarios respectively (Table 3.5).

Nonetheless, results of Tukey HSD post-hoc tests indicated that not all of the scenarios were statistically significantly different from each other. To be specific, differences were found between the following pairs: (TOR1, TOR4),  $p = .040$ ; (TOR1, TOR6),  $p < .001$ ; (TOR1, TOR7),  $p < .001$ ; (TOR2, TOR6),  $p = .003$ ; (TOR2, TOR7),  $p = .013$ ; (TOR3, TOR6),  $p = .002$ ; (TOR3, TOR7),  $p = .010$ ; (TOR4, TOR8),  $p = .004$ ; (TOR4, TOR9),  $p = .002$ ; (TOR6, TOR8),  $p < .001$ ; (TOR6, TOR9),  $p < .001$ ; (TOR7, TOR8),  $p < .001$ ; (TOR7, TOR9),  $p < .001$ . And the average and standard deviation of the takeover time in each scenario were summarized in Table 3.6. Results of one-way ANOVAs showed that the effects of DoM on takeover time were not statistically significant in TOR1, TOR3 and TOR4, whereas statistically significant in the rest of the scenarios.

**Table 3.5:** Two-way ANOVA results performed on DoM and scenarios for takeover time in high- and medium critical scenarios.

	sum_sq	df	F	PR(>F)	sum_sq	df	F	PR(>F)
	High-Critical				Medium-Critical			
DoM	11.633	2.0	16.329	<.001	21.696	2.0	23.05	<.001
Scenario	12.381	5.0	6.970	<.001	13.403	2.0	14.13	<.001
DoM*Scenario	3.239	10.0	0.912	.523	0.739	4.0	0.39	.816

**Table 3.6:** ANOVA results and takeover time (mean and standard deviation) in high- and medium-critical scenarios given different DoM.

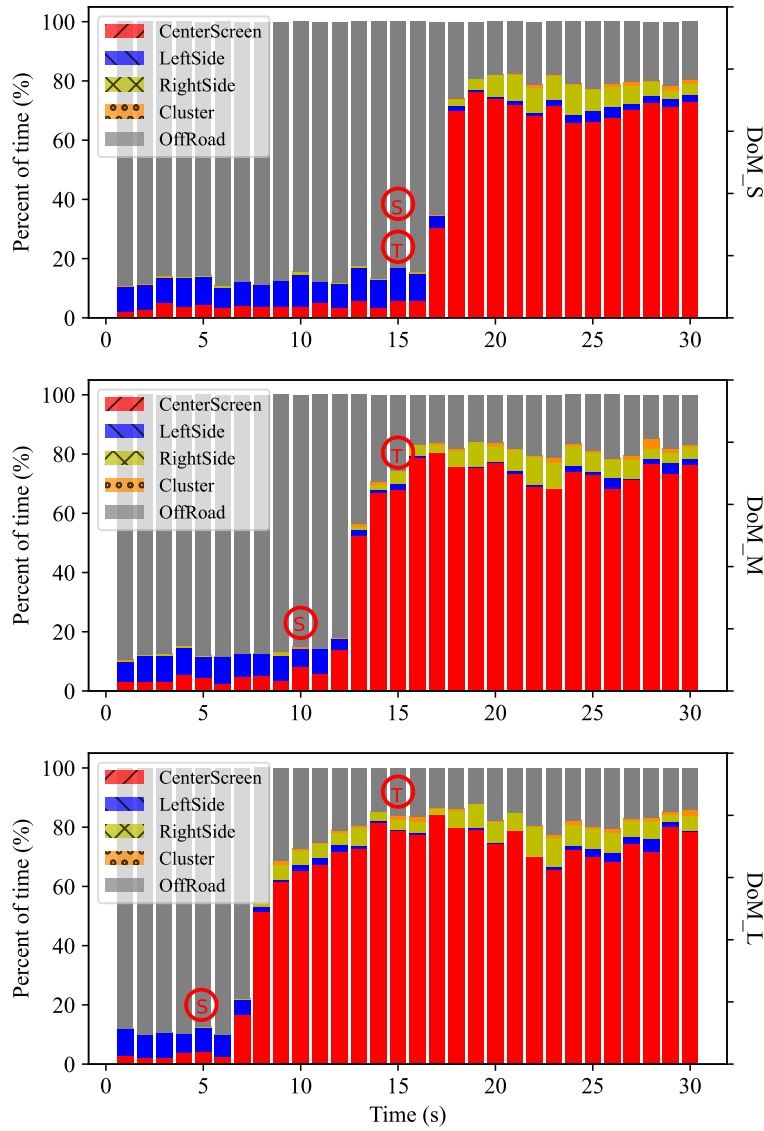
	DoM <sub>S</sub>	DoM <sub>M</sub>	DoM <sub>L</sub>	df	F	PR(>F)
TOR1	2.547(0.35)	2.411(0.61)	2.365(0.59)	2	0.386	.683
<b>TOR2</b>	2.757(0.41)	2.113(0.50)	2.122(0.47)	2	7.629	.002
TOR3	2.722(0.67)	2.117(0.56)	2.159(0.85)	2	2.862	.071
TOR4	2.288(0.67)	1.178(0.52)	1.868(0.64)	2	2.893	.070
<b>TOR5</b>	2.344(0.51)	1.869(0.57)	2.347(0.51)	2	3.336	.048
<b>TOR6</b>	2.244(0.74)	1.447(0.82)	1.495(0.57)	2	4.551	.018
<b>TOR7</b>	2.588(0.76)	1.413(0.43)	1.393(0.51)	2	16.570	<.001
<b>TOR8</b>	3.100(0.87)	2.180(0.49)	2.318(0.61)	2	6.226	.005
<b>TOR9</b>	3.108(0.80)	2.889(0.92)	2.281(0.61)	2	4.383	.021

### Gaze Behavior and Eye Movements (Q2)

**Percentage of Area of Interest** As expected, during the SuRT (Fig. 3.5, 0–15 s, 0–10 s and 0–5 s for DoM<sub>S</sub>, DoM<sub>M</sub> and DoM<sub>L</sub> respectively), the percentage of eye gazes of drivers directed to **CenterScreen** (red bars in Fig. 3.5) was below 5% (4.11%, 4.33% and 3.21% on average respectively). In the first 5 s after the **SuRT** has ended, eye gazes on CenterScreen for DoM<sub>S</sub> grew from 5.72% to 74.09% rapidly, whereas for DoM<sub>M</sub> and DoM<sub>L</sub>, the percentages rose only to 67.88% and 65.54% respectively. And Friedman test also did revealed statistically significant differences among the three groups ( $\chi^2(2) = 6.40$ ,  $p = .041$ ). However, in the first 10 s or 15 s after the SuRT has ended, the differences became gradually statistically insignificant ( $\chi^2(2) = 1.80$ ,  $p = .407$ ;  $\chi^2(2) = 0.533$ ,  $p = .766$ ). **This indicated the similarities of the gaze patterns on CenterScreen after the SuRT.**

On the other hand, in the first 5 s after the **TOR**, percentage of eye gazes directed to CenterScreen for DoM<sub>S</sub>, DoM<sub>M</sub> and DoM<sub>L</sub> all climbed to about 75% (74.09%, 77.05% and 74.51% respectively). And Friedman test also did not reveal statistically significant differences among the three groups ( $\chi^2(2) = 5.20$ ,  $p = .074$ ). However, the results became statistically significant when the time interval was extended to 10 s ( $\chi^2(2) = 7.40$ ,  $p = .025$ ) and 15 s ( $\chi^2(2) = 12.4$ ,  $p = .002$ ). **This manifested the differences of the gaze patterns on CenterScreen after the TOR.** Besides, it could be observed that after the percentage of gazes on CenterRoad reached a peak (76.58%, 80.40% and 84.34% respectively), it then stabilized after a slight decline, and the mean values in the last 10 s were 69.96%, 72.42% and 73.09% respectively.

For eye gazes directed to **RightMirror** and **RightScreen**, Friedman tests showed that the differences were basically not statistically significant in all three time intervals, i.e., 5 s, 10 s and 15 s after the TOR (RightMirror:  $\chi^2(2) = 8.40$ ,  $p =$



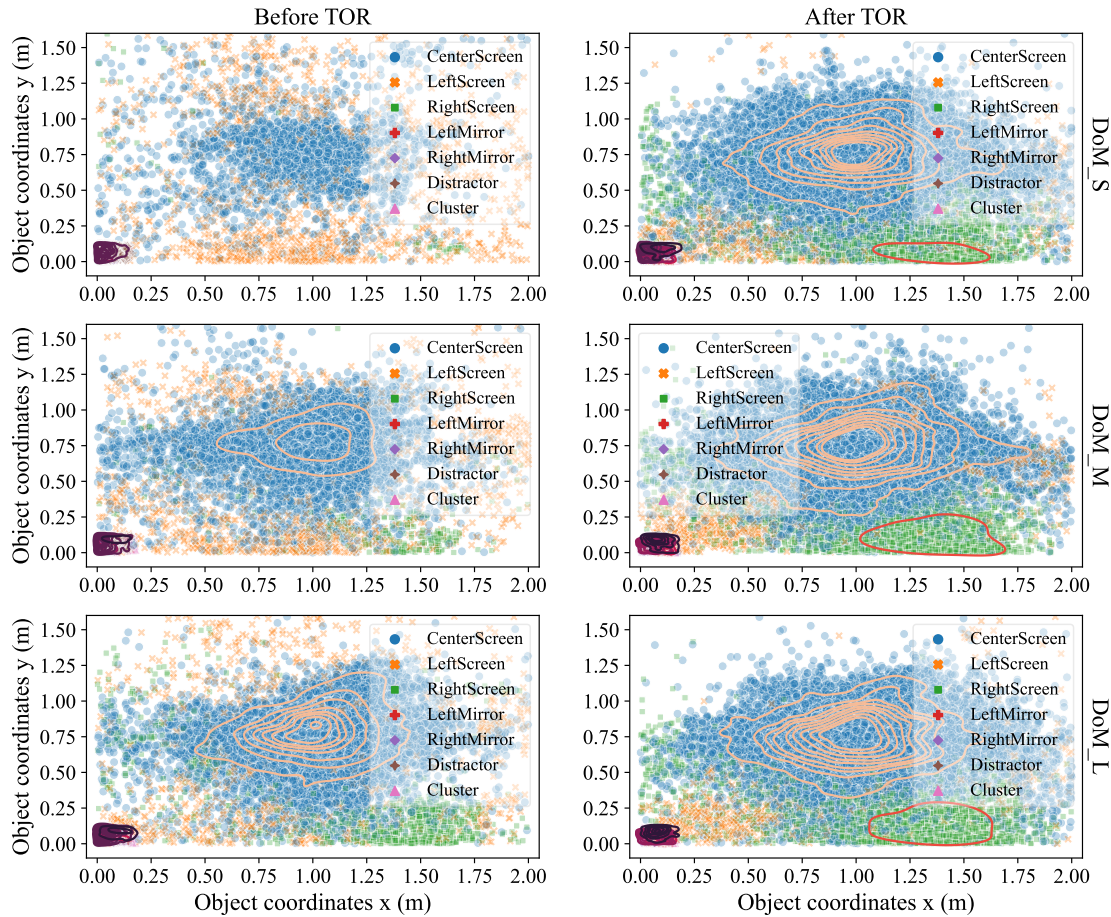
**Figure 3.5:** Percentage of eye gazes directed to the defined AOIs, where Right-Mirror and RightScreen are combined as RightSide, and LeftMirror, LeftScreen and Distractor are combined as LeftSide. (S) and (T) represent end of SuRT and TOR respectively. From top to bottom are short, medium, and long DoM respectively.

.015;  $\chi^2(2) = 0.600$ ,  $p = .741$ ;  $\chi^2(2) = 0.933$ ,  $p = .627$ ; RightScreen:  $\chi^2(2) = 2.80$ ,  $p = .247$ ;  $\chi^2(2) = 2.40$ ,  $p = .301$ ;  $\chi^2(2) = 8.13$ ,  $p = .017$ ). Moreover, since the two AOIs were partly overlapped (Fig. 3.1), making it kind of difficult to discriminate sometimes, and also for better visualization, it would be reasonable to combine the two areas for further analysis. For convenience, the two areas were denoted as **RightSide** (yellow bars in Fig. 3.5). Similarly, LeftMirror, LeftScreen and Distractor were denoted as LeftSide (blue bars in Fig. 3.5). And Friedman tests revealed no statistically significant differences in all three time intervals ( $\chi^2(2) = 5.20$ ,  $p = .074$ ;  $\chi^2(2) = 2.60$ ,  $p = .273$ ;  $\chi^2(2) = 1.73$ ,  $p = .420$ ). **This**



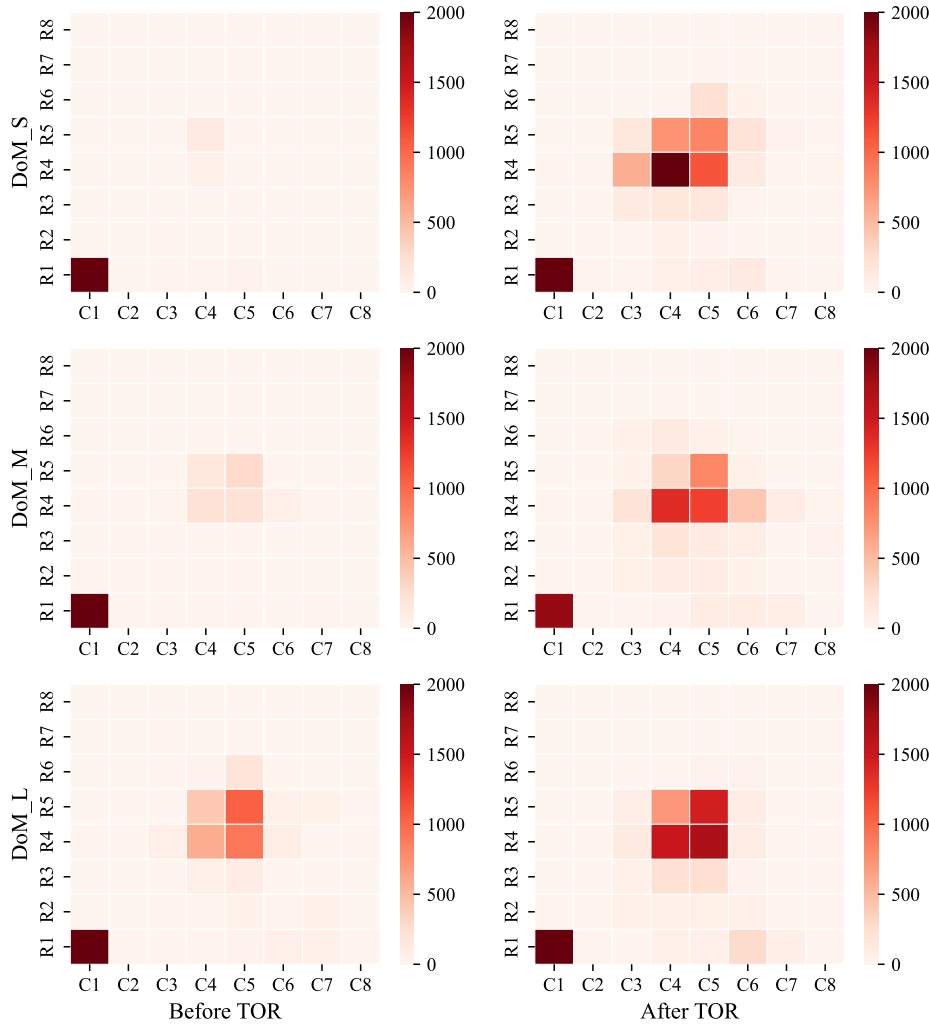
indicated the similarities of the gaze patterns on RightSide after the TOR. Finally, for eye gazes directed to **Cluster** (orange bars in Fig. 3.5), statistically significant results were found only in the 5 s and 10 s intervals ( $\chi^2(2) = 6.40$ ,  $p = .041$ ;  $\chi^2(2) = 6.20$ ,  $p = .045$ ). This implied different gaze patterns toward the cluster shortly after the TOR.

**Heat Map and Gaze Entropy** Obviously, before the TOR, eye gazes were more directed to CenterScreen with longer DoM (Fig. 3.6-Left and Fig. 3.7-Left). And after the TOR, eye gazes were mainly concentrated in CenterScreen and RightScreen (Fig. 3.6-Right and Fig. 3.7-Right), although the specific spatial property might be different. The left-bottom corner of each sub-figure in Fig. 3.6 and Fig. 3.7 were highly densified, since the coordinates were relative and LeftMirror, RightMirror, Distractor and Cluster shared the same area.



**Figure 3.6:** Stacked scatter plots of gaze points of drivers before and after the TOR given different DoM. From top to bottom are short, medium, and long DoM respectively.

Results of ANOSIM test revealed that the distributions of eye gazes after the TOR were indeed statistically significantly different depending on DoM



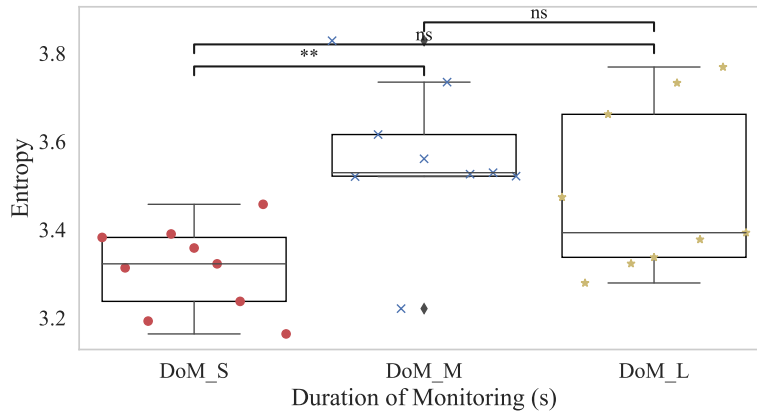
**Figure 3.7:** Heat maps of eye gazes of drivers before and after the TOR given different DoM. From top to bottom are short, medium, and long DoM respectively.

( $R = 0.183$ ,  $p = .004$ ). Furthermore, from the kernel density estimate plots in Fig. 3.6, and in combination with the heat maps in Fig. 3.7, most of the gaze points were concentrated in blocks (C3, R4), (C3, R5), (C4, R4), (C4, R5), (C5, R4), (C5, R5), (C6, R4), (C6, R5), (C6, R1) and (C1, R1). Therefore, Friedman tests were employed to compare differences of gaze counts after the TOR in the ten blocks across different test scenarios given different DoM. For convenience, the first eight blocks and the last two were denoted as Block<sub>Center</sub>, Block<sub>Right</sub> and Block<sub>LRDC</sub> respectively. Results revealed that **there were statistically significant differences in gaze counts depending on the length of DoM in all the three blocks defined** ( Block<sub>Center</sub>:  $\chi^2(2) = 6.222$ ,  $p = .045$ ; Block<sub>Right</sub>:  $\chi^2(2) = 6.222$ ,  $p = .045$ ; Block<sub>LRDC</sub>:  $\chi^2(2) = 13.556$ ,  $p = .001$ ). Nonetheless, Wilcoxon tests adjusted with bonferroni only reported statistically significant differences between DoM<sub>S</sub> and DoM<sub>L</sub> in Block<sub>Center</sub> ( $z = 1.00$ ,  $p_{\text{adj}} = .023$ ) and Block<sub>LRDC</sub>



( $z = 45.0$ ,  $p_{\text{adj}} = .012$ ), and between  $\text{DoM}_S$  and  $\text{DoM}_M$  in  $\text{Block}_{\text{LRDC}}$  ( $z = 45.0$ ,  $p_{\text{adj}} = .012$ ). Further tests performed on each individual block in  $\text{Block}_{\text{Center}}$  showed that, however, the differences in all the eight blocks were not statistically significant.

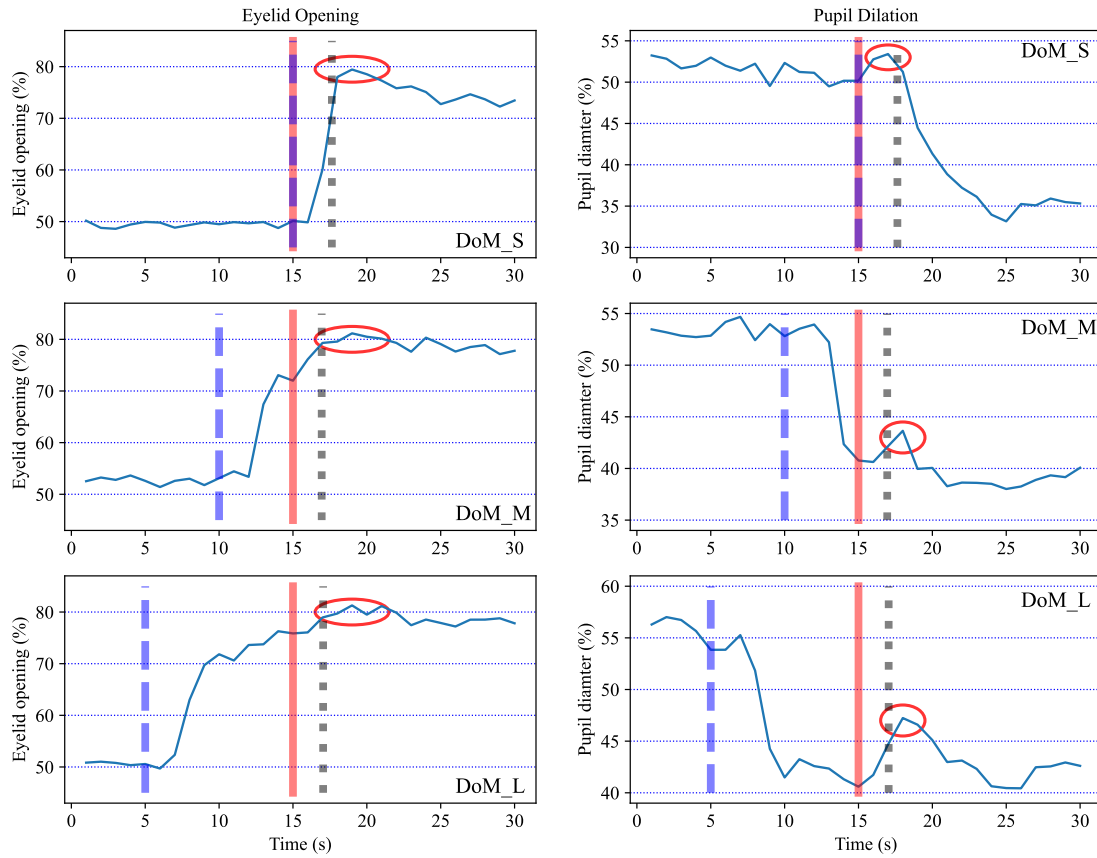
Finally, gaze entropy was calculated according to Eq. (3.1). Results of an one-way repeated measures ANOVA performed on DoM showed that **differences of gaze entropy depending on DoM were statistically significant** ( $F(2, 16) = 7.606$ ,  $p < .005$ ). However, the post-hoc tests adjusted with bonferroni only revealed a statistically significant difference between  $\text{DoM}_S$  and  $\text{DoM}_M$  ( $z = -3.74$ ,  $p_{\text{adj}} = .017$ , Fig. 3.8). Since gaze entropy reflected gaze dispersion, the gazes were more dispersed with longer DoM.



**Figure 3.8:** Box and scatter plots of gaze entropy given different DoM.

**Pupil Dilation and Eyelid Opening** Congruent with eye gazes on road center, eyelid opening of the driver grew steadily after the SuRT and before the TOR ( $\text{DoM}_M$ : 53.1%→72.0%;  $\text{DoM}_L$ : 50.6%→71.8%). Upon hearing the TOR, eyelid opening reached its peak within a short period of time ( $\text{DoM}_S$ : 79.4%;  $\text{DoM}_M$ : 81.2%;  $\text{DoM}_L$ : 81.3%, Fig. 3.9-Left), and was then kept at a high level compared with that during the SuRT. The average eyelid opening of the last 10 seconds was respectively 74.5%, 78.6% and 78.6%. As for pupil dilation, it decreased to a low level first before the TOR ( $\text{DoM}_M$ : 52.8%→40.8%;  $\text{DoM}_L$ : 53.8%→40.6%), and then reached a maximum after 2–3 seconds ( $\text{DoM}_S$ : 53.4%;  $\text{DoM}_M$ : 43.6%;  $\text{DoM}_L$ : 47.2%, Fig. 3.9-Right). Finally, pupil dilation continued to decrease to a relatively low level. And the average pupil dilation of the last 10 seconds was respectively 35.6%, 38.8% and 42.1%.

For 5 s, 10 s and 15 s after the TOR, results of Friedman tests were summarized in Table 3.7. The results revealed that, **in all the three time intervals, patterns of eyelid opening and pupil dilation were all statistically significantly different**,



**Figure 3.9:** Eyelid opening and pupil dilation before and after the TOR given different DoM. Dashed blue lines mark the end of the SuRT, solid red lines mark the TOR and dotted black lines mark the takeovers. Peaks after the TOR are marked in red circles.

although the results of pairwise comparisons were different according to the length of time intervals. Besides, we could also observe that patterns of eyelid opening and pupil dilation were most similar in the first 5 s after the TOR, and then parted with each other afterwards.

**Table 3.7:** Friedman test results performed on DoM for eyelid opening and pupil dilation 5 s, 10 s and 15 s after the TOR.

			5 s	10 s	15 s
Eyelid Opening	Friedman		$p = .022^*$	$p < .001^{***}$	$p < .001^{***}$
	Wilcoxon	S-M	$p.\text{adj} = .188$	$p.\text{adj} = .006^{**}$	$p.\text{adj} < .001^{***}$
		S-L	$p.\text{adj} = .188$	$p.\text{adj} = .006^*$	$p.\text{adj} < .001^{***}$
		L-M	$p.\text{adj} = 1.0$	$p.\text{adj} = 1.0$	$p.\text{adj} = 1.0$
Pupil Dilation	Friedman		$p = .022^*$	$p = .014^*$	$p < .001^{***}$
	Wilcoxon	S-M	$p.\text{adj} = .188$	$p.\text{adj} = 1.0$	$p.\text{adj} = 1.0$
		S-L	$p.\text{adj} = .936$	$p.\text{adj} = 1.0$	$p.\text{adj} = .363$
		L-M	$p.\text{adj} = .188$	$p.\text{adj} = .006^{**}$	$p.\text{adj} < .001^{***}$

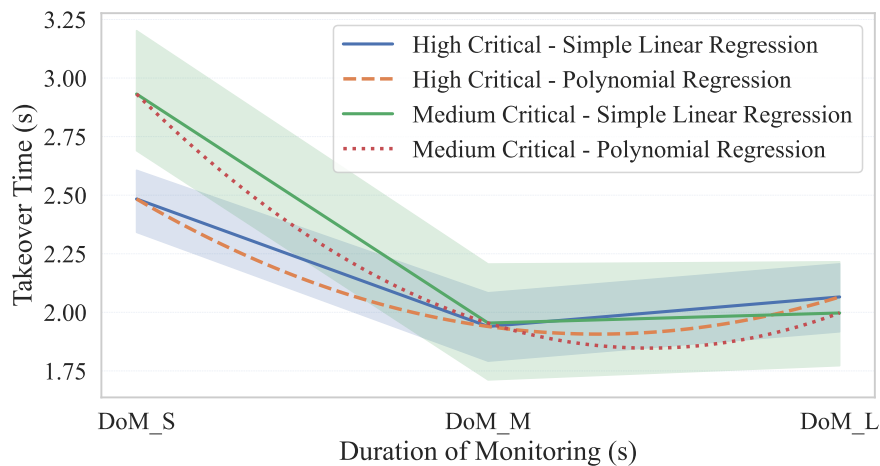
\*:  $< .05$ ; \*\*:  $< .01$ ; \*\*\*:  $< .001$ ; \*\*\*\*:  $< .0001$ .

### 3.3.4 General Discussions

#### Effects of Duration of Monitoring (Q3)

Results in Sections 3.3.3 and 3.3.3 suggested that impacts of DoM on takeover time were indeed statistically significant. However, as one of the main purposes of this analysis was to verify the relationship between the two, we would like to delve deeper into that and find a regression model to fit the data collected (Fig. 3.10). In both high- and medium-critical scenarios, we could observe that:

1. The relationship between DoM and takeover time was quasi-parabolic, and the minimum could be obtained somewhere between  $DoM_M$  and  $DoM_L$ , which was around 7 s.
2. The difference between  $DoM_M$  and  $DoM_L$  became less significant as the scenarios became less critical. This was straightforward in that, in medium-critical scenarios, drivers had more time to respond to the request, making the takeover not so urgent as that in high-critical scenarios.
3. Around  $DoM_M$ , we could also observe that the takeover time in medium- and high-critical scenarios almost overlapped. This suggested that there were some intervals where drivers' performance could be optimized regardless of the criticality of the scenarios.

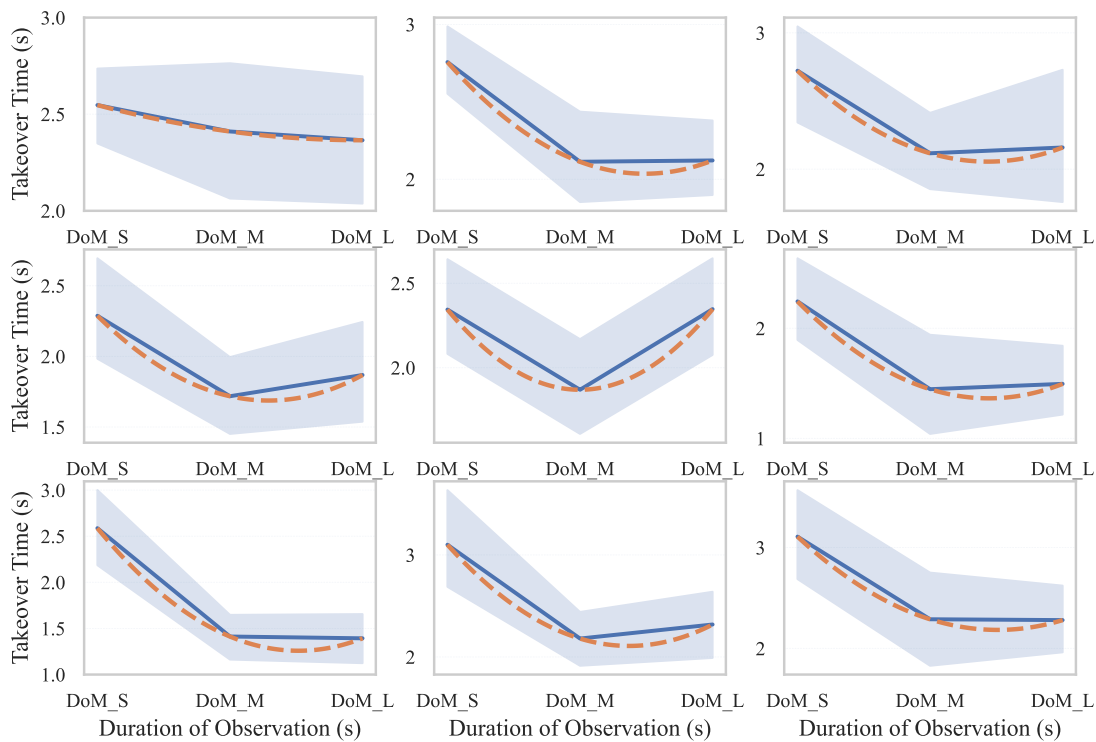


**Figure 3.10:** Regression model of takeover time over DoM in high- and medium-critical scenarios respectively.

Taking one step further, regression models for each individual scenario were summarized in Fig. 3.11. We could observe that, except for TOR1, all the others were in accordance with the conclusions above, although the minimum in TOR5 was a bit different. And curiously, takeover time in TOR1 almost remained

constant compared with others, which exemplified drivers' sensitivity to longitudinal behavior of other vehicles, especially in high speeds [111]. And the distinctiveness of TOR5 might be related with the weather conditions, which requires further verification through more specific experiments.

Furthermore, despite the quasi-parabolic relationship in TOR3 and TOR4 as shown in Fig. 3.11, results in Table 3.6 suggested that effects of DoM on takeover time in TOR1, TOR3 and TOR4 were actually not statistically significant. Although the difference between DoM<sub>M</sub> and DoM<sub>L</sub> became less significant as the scenarios became less critical, those results could possibly suggest that takeover time in medium-critical scenarios was more likely to be affected by DoM than that in high-critical scenarios, particularly between DoM<sub>S</sub> and DoM<sub>M</sub>. This could be attributed to the fact that drivers would have more time to gain situation awareness in medium-critical scenarios than that in high-critical ones, and there was no need to hurry in medium-critical scenarios even for DoM<sub>S</sub>.



**Figure 3.11:** Regression model of takeover time over DoM in medium- and high-critical scenarios, where the solid and the dashed lines are results of linear and polynomial regression, respectively.

In summary, longer DoM before the TOR did not necessarily lead to quicker response and better takeover performance, i.e., DoM and takeover time were neither monotonously positively nor negatively correlated. This was in coincidence with the results by [68]. A straightforward explanation for this relationship

could be found from the working principles of human beings' vision system. As introduced by [79], human beings actually had a two-stream vision system—vision-for-action and vision-for-identification, which were guided by two separate systems called dorsal and ventral visual streams respectively [112]. When the situation awareness was low, vision-for-action dominated, which operated in real time and converted visual information directly into action without involving much previous knowledge [113] or in the absence of consciousness [114]. As eye gazes of the driver were directed to road center and situation awareness was gradually built up, vision-for-identification began to take effects, which helped to improve performance based on prior experience and knowledge. However, as DoM continued to increase, there might appear two kind of problems. Firstly, the two vision systems had different visual fields, where cells in the vision-for-action stream had a large representation of the peripheral visual fields and cells in the vision-for-identification stream were centered around the fovea [115]. Hence, if eye gazes were too concentrated on road center, some information from the peripheral visual field might be lost, leading to degraded takeover performance in certain situations. Secondly, just as [116] mentioned in their study, “a warning must hold attention for the time necessary to encode and store the message contained in the warning” and “prevent attention from being distracted by and to other stimuli before the message is satisfactorily encoded”. When the DoM was too long, drivers were prone to be distracted by other stimuli [68, 116], slowing down drivers' response speed. Therefore, the length of the DoM should be appropriate.

#### Implications of Eye Tacking Data (Q4)

**Gaze Behavior** In a 60 s' time period after automation was disengaged, [89] found a significant increase of percentage road center (PRC) in the 5–10 s time interval compared to that in the 0–5 s. And following a sharp increase during the 15–20 s time interval, PRC was seen to drop again until around the 40–45 s. Similar patterns were also observed in our study, whereas in more compact time intervals. This could be attributed to the criticality of the test scenarios, and drivers were asked to direct their gazes away from the NDRT after the SuRT has ended. For both DoM<sub>M</sub> and DoM<sub>L</sub>, we could observe that eye gazes directed to CenterScreen grew steadily until the TOR was issued, at which point the percentages were 67.88% and 78.72% respectively. Upon hearing an audible message, the percentages reached peaks quickly within 2 s (80.40% and 84.34%), and then were seen to decrease until around 23 s. As for DoM<sub>S</sub>, although there

was a sharp increase of eye gazes directed to CenterScreen upon hearing the TOR, a few seconds were still necessary in order to direct attention to the road center and gain enough situation awareness for successful and safe takeovers. This explained partly the differences of the takeover time given different DoM, particularly between DoM<sub>S</sub> and DoM<sub>M</sub>, and also between DoM<sub>S</sub> and DoM<sub>L</sub>.

Since no statistically significant differences were found regarding percentage of eye gazes directed to CenterScreen in the 5 s, 10 s, and 15 s time intervals after the SuRT has ended depending on the three conditions (DoM<sub>S</sub>, DoM<sub>M</sub> and DoM<sub>L</sub>), gaze patterns since the first glance back to the roadway were presumably similar regardless of DoM. Similarly, gaze patterns directed to RightSide after the TOR were also found to be similar. What was different were the gaze patterns directed to CenterScreen after the TOR, where statistically significant differences were found.

Basically, a relatively high level of PRC helped the driver to make better and quicker decisions at the time of TOR. However, when PRC was too high, the effects might also be counterproductive, as was the case for DoM<sub>L</sub>. On one hand, too concentrated attention may slow down drivers' response speed to other tasks, e.g., TOR. As found out by [117], visual attention is "the selective control of information in the visual system, achieved in various ways by various processes". Since our attention is limited [118], when we choose to pay more attention to one task (e.g., the roadway), the attention left for the other tasks is destined to be more or less limited, as is the case observed in the relationship of takeover time with DoM. On the other hand, some peripheral information might be overlooked, e.g., rear mirrors, leading to hazardous situations in some scenarios. This is what is called inattention blindness, defined as "failure to report unexpected or unattended stimuli" [117]. This has also been observed during the experiment, where we found that compared with DoM<sub>S</sub> and DoM<sub>M</sub>, drivers were more prone to side or rear-end collisions with DoM<sub>L</sub>. The interview results by drivers involved in accidents also suggested that the main reason was neglecting information in the rear mirrors.

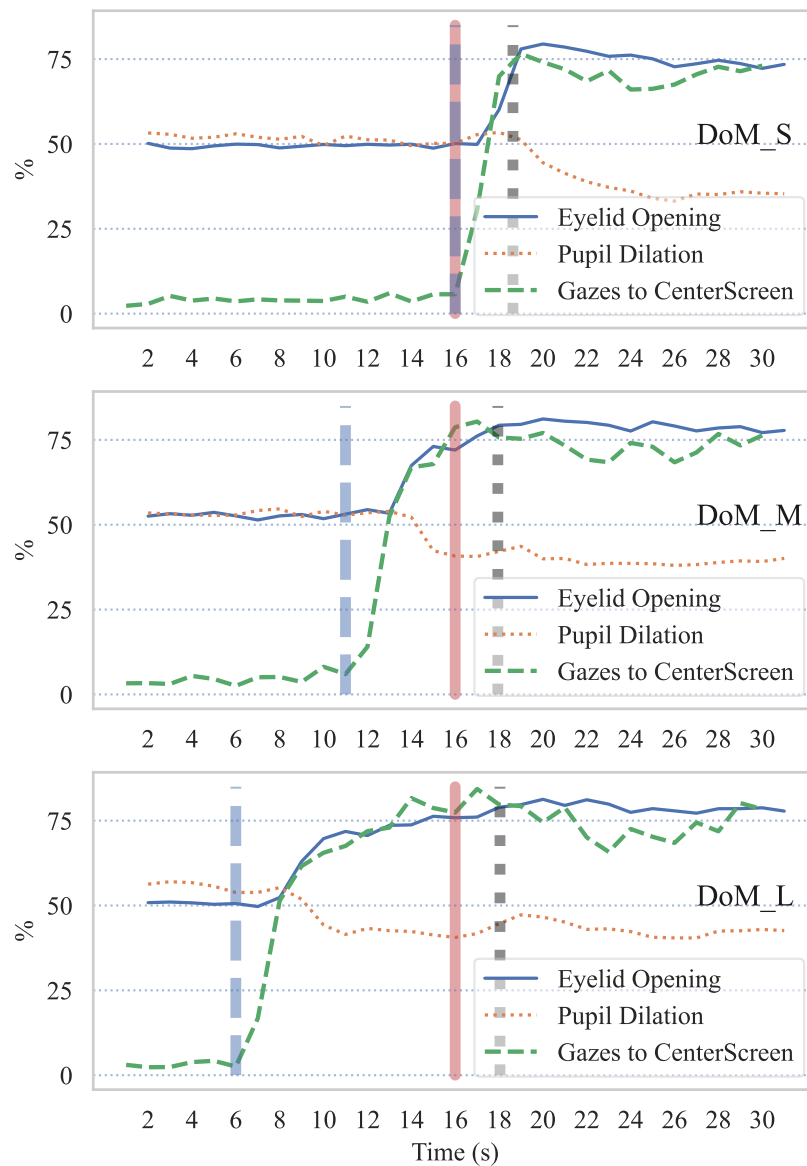
The stacked scatter plots (Fig. 3.6) and the heat maps (Fig. 3.7) revealed that statistically significant difference existed regarding the distribution of eye gazes depending on the DoM. More specifically, statistically significant differences could be found in Block<sub>Center</sub>, Block<sub>Right</sub> and Block<sub>LRDC</sub>, although pairwise difference was not found for Block<sub>Right</sub>. These results were consistent with the above analysis regarding percentage of eye gazes directed to CenterScreen. Intuitively from Fig. 3.7, eye gazes were more concentrated for DoM<sub>S</sub> than that for DoM<sub>M</sub> and DoM<sub>L</sub> (The deeper the color was, the more concentrated the eye gazes were).

This intuition became clearer from Fig. 3.8, where we could see that gaze entropy for DoM<sub>S</sub> was statistically significantly smaller than that for DoM<sub>M</sub>. Moreover, gaze entropy for DoM<sub>L</sub> seemed to vary a lot across different takeover scenarios, indicating greater effect of scenarios on distribution of eye gazes for DoM<sub>L</sub> than that for DoM<sub>S</sub> and DoM<sub>M</sub>. While gaze entropy for DoM<sub>S</sub> and DoM<sub>M</sub> fell approximately in [3.2, 3.4] and [3.5, 3.6] respectively, it varied between 3.2 and 3.8 for DoM<sub>L</sub>. As can be observed from the scatter plots in Fig. 3.8, in 6 out of the 9 scenarios, gaze entropy for DoM<sub>L</sub> was lower than that for DoM<sub>M</sub>, and the other 3 were just opposite. This was not desirable when proper concentration of attention was beneficial for safe takeovers. Moreover, it rendered takeover performance of the driver more difficult to predict under different takeover scenarios.

**Pupil Dilation and Eyelid Opening** Upon careful comparison between eyelid opening and eye gazes directed to CenterScreen (Fig. 3.12), we could observe that the trend of eyelid opening was very similar to eye gazes directed to CenterScreen, except with a delay about 2–3 s. Different from gazes to CenterScreen, eyelid opening patterns were also statistically significantly different in the first 5 s after the TOR, and then became more statistically significantly different afterwards. This was also obvious from Fig. 3.9-Left or Fig. 3.12. As eyelid opening reached a peak for DoM<sub>S</sub>, it then decreased to a relatively lower level, whereas almost no decline could be observed for both DoM<sub>M</sub> and DoM<sub>L</sub>. The trend of pupil dilation was just contrary to that of eyelid opening or gazes to CenterScreen (Fig. 3.12), and pupil dilation patterns were statistically significantly different during all the three time intervals (5 s, 10 s and 15 s after the TOR), except that the range of pupil dilation was narrower.

For DoM<sub>M</sub> and DoM<sub>L</sub>, the change of eyelid opening and pupil dilation took place in two stages (Fig. 3.9). The first stage started from the end of the SuRT, during which eyelid opening increased and pupil dilation decreased continuously until the TOR was issued. Then began the second stage, during which eyelid opening was maximized, whereas pupil dilation only reached a local maximum which decreased gradually afterwards. The reasons behind these changes were complicated [119–121], in our experiment, however, two main reasons might be attributable: The first was the variation of illumination, and the second was the arousal level of the driver. In the first stage, since eye gazes of the driver were directed from the panel inside the cabin to the forward screen, there was a sharp change of illumination, the first reason might be dominated. Upon entering the second stage, the driver has already become accustomed to the new





**Figure 3.12:** Eyelid opening, pupil dilation and eye gazes directed to Center-Screen before and after TOR given different DoM. Dashed blue lines mark the end of the SuRT, solid red lines mark the TOR and dotted black lines mark the takeovers.

illumination, the dominated reason would be the second one. An instant higher eyelid opening and pupil dilation level only indicated a higher arousal level [120] to the stimulus, so that the driver would be more responsive upon hearing the request to intervene. As for **DoM<sub>S</sub>**, the two reasons overlapped. Different from the pupil dilation under the other two conditions, the maximum of the pupil dilation was obtained before the driver has actually taken over the vehicle (Fig. 3.9). This was supported by the fact that time was needed in order for the eyes to get adapted to the sudden change of illumination [122], during which time, however, takeover was already necessary. This also explained partly why



the takeover performance of the driver was worst given DoM<sub>S</sub>, as the driver would need some time to get prepared before taking over the vehicle.

### Impacts of Takeover Scenarios (Q5)

In Section 3.3.4, it was mentioned that takeover time in medium-critical scenarios was more likely to be affected by DoM than that in high-critical scenarios. In this sense, criticality of scenarios could be considered as a moderator variable that acted upon the relationship between DoM and takeover time and modulate its strength accordingly. However, distinctiveness of scenarios TOR1 and TOR5 also suggested that the power of criticality for explaining effects of scenarios on takeover performance was kind of limited.

Apart from weather conditions related with TOR5, upon careful scrutinization, we could also find that, besides criticality, other factors such as urgency, predictability, and driver's response required [4] might also matter a lot for takeover performance of drivers. Take predictability of scenarios as an example, basically, it seemed that effects of DoM would be more significant when the scenarios became more predictable. Such was the case with TOR2 or TOR6 in high-critical scenarios, where the static obstacles ahead made the scenarios easier to predict, so that drivers could make easier decisions and tended to take quicker actions to avoid accidents given some time to observe the surroundings. In addition, curvature of the road might also make a difference, and it seemed that drivers would be more responsive in curved (TOR2) than that in straight roads (TOR9). As the database at hand is still relatively small, more data are necessary to verify these conclusions. What is certain is that criticality alone is not enough to explain effects of scenarios on takeover performance of drivers.

### 3.3.5 Summary of Analysis 1

Since eyes-on-road gazes have been found to increase steadily during the first few seconds [89, 103], it was hypothesized that the performance of drivers would be improved if they were given a few seconds to divert their attention away from NDRT to roadway. And the **main purpose** of this analysis was to investigate the impact of DoM before the TOR on takeover time, and further on, the relationship between them in different scenarios. Correspondingly, eye tacking data were also explored.

- Data collected from a set of takeover scenarios revealed that impacts of DoM on takeover time were indeed statistically significant, especially between DoM<sub>S</sub> and DoM<sub>M</sub>, and also between DoM<sub>S</sub> and DoM<sub>L</sub>.

- Although the difference between  $DoM_M$  and  $DoM_L$  was not statistically significant, a slight increase of takeover time for  $DoM_L$  compared with that for  $DoM_M$  could still be observed, resulting in a relationship close to a parabola. In view of this, the optimal DoM could be obtained somewhere between  $DoM_M$  and  $DoM_L$ , which was approximately 5–7 s.

This was further supported by the eye tracking data.

- Of the seven AOIs defined, eye gazes directed to CenterScreen were found to be statistically significantly different depending on DoM.
- Basically, higher percentage of eye gazes on CenterScreen at the time of TOR indicated better response of the driver. However, the benefit was not without limitation. As indicated by the results, too frequent gazes on road center would in turn be counterproductive.
- Calculation of gaze entropy showed that compared with  $DoM_S$ , gaze entropy for  $DoM_M$  was statistically significantly more dispersed. And gaze entropy for  $DoM_L$  was slightly greater than that for  $DoM_M$  and varied a lot across scenarios, which might not be favorable not only for safe takeovers but also for prediction of takeover performance.
- Patterns of eyelid opening and pupil dilation were found to be statistically significantly different, they could be good choices in predicting takeover time of drivers.

The final goal is to build a prediction model to predict drivers' takeover performance by taking various possible influencing factors into consideration. Results of this analysis suggest that DoM could be one of them, and gazes on CenterScreen, gaze entropy, eyelid opening and pupil dilation could also be good indicators to reflect drivers' state depending on DoM. Finally, for safety of the takeover, and also for good predictability of drivers, it is desirable that the length of DoM be appropriate, so that both the performance of the driver and the prediction model could be optimized.

## 3.4 Analysis 2: Impact of Personality

### 3.4.1 Hypotheses of Analysis 2

Although relationships of takeover behaviors with the big five personality traits are yet not clear, the relationships of manual driving behaviors with the big five personality traits have been researched a lot, which may shed light on studies

regarding automated driving. A meta-analysis by Akbari et al. [123], for example, revealed that risky driving behaviors were negatively and positively related to agreeableness and neuroticism, respectively. And conscientiousness was found to be marginally negatively associated with crashes and near-crashes by Ehsani et al. [124]. Besides, it was indicated that participants with higher openness traits tend to have less trust in ADS [125]. As trust in ADS was believed to significantly affect takeover performance [17], although no relationships between risky behaviors and openness was found during manual driving [123], it is highly possible that openness is also associated with takeover performance of drivers. As for extraversion, no effect of it seems to be found either on risky behaviors [123, 124] or reaction time [126]. Given these results, it is reasonable to hypothesize that:

- H1:** Statistically significant correlations exist between neuroticism (**H1-1**), agreeableness (**H1-2**), conscientiousness (**H1-3**) and takeover performance.
- H2:** There is the possibility that openness is statistically significantly (at least marginally significantly) associated with takeover performance.
- H3:** Obvious correlation between extraversion and takeover performance is not expected.

To verify these hypothesis, a driving simulator study involving 6 critical takeover scenarios were designed. Since effect of personality on takeover time was also found to be different for different levels of situation awareness of drivers [68], levels of situation awareness were also taken into consideration while designing the experiment. And it is reasonable to hypothesize that:

- H4:** Effects of personality on takeover performance under different levels of situation awareness would be different, i.e., the effects may be positive or negative, and may or may not be statistically significant depending on situation awareness of drivers.

Overall, there are two main purposes of this analysis: 1) Determine relationships between the big five personality and takeover performance in critical takeover situations; 2) Explore the correlations between the big five personality and takeover performance given different levels of situation awareness. This contributes to the existing literature in that: 1) It helps to understand effects of drivers' personality on takeover time and maneuvers shortly after the TOR during automated driving; 2) It helps to fulfill the ultimate goal to give some instructions on selecting influencing factors when building prediction models

for predicting takeover behaviors of drivers, so that the prediction model can be personalized, which is import to enhance safety and acceptability of the ADS.

### 3.4.2 Data Processing and Performance Metrics

Data from 288 takeovers in the high-critical takeover scenarios were utilized in this analysis. For takeover time, as 6 takeovers were either earlier than the TOR or failed to be recorded, 282 recordings were valid for analysis. For eye tracking data, excluding the failure cases, data from 258 TORs were successfully recorded, and segments of 15 s before and after the TOR were extracted for analysis. For lateral and longitudinal performance data, since one of the purposes was to analyze drivers' instinct takeover maneuvers shortly after the TOR, data segments of 8 s after the TOR were extracted for analysis. As a comparison, analysis results of data segments of 15 s after the TOR were also presented briefly. The performance metrics utilized were summarized in Table 3.8, where takeover time, maximum acceleration, maximum steering angle, and turn signal missing rates were safety-related measures, and the others were quality-related measures that may have potential mid- and long-term detrimental effects.

**Table 3.8:** Summary of the performance metrics (effects of personality).

Performance Metrics		Notation	Units
Takeover Time		TOT	s
Longitudinal	Max Acceleration	AccMax	m/s <sup>2</sup>
	Mean Speed	SpeedMean	km/h
	STD* of Speed	SpeedStd	
Lateral	Max Steering Angle	SteerMax	°
	Mean Steering Angle	SteerMean	
	STD of Steering Angle	SteerStd	
	Max Yaw Angle	YawMax	
	STD of Yaw Angle	YawStd	
Turn Signal		TurnSigRate*	%

\*STD stands for standard deviation.

\*TurnSigRate represents the missing rate of the turn signal.

### 3.4.3 Results and Analysis

As this analysis is one part of a series of studies and eye tracking data is not the emphasis, the detailed statistical analysis regarding eye movements and gaze behavior could be referenced to Fig. 3.5 in Section 3.3.3 and [127]. **Analysis of Variance (ANOVA)** tests were performed for inspecting effects of each

personality trait on each performance metric in this section, followed by **Tukey HSD post-hoc tests** for multiple comparisons if the results were significant. Results were reported as statistically significant if the p-value was less than .05. If the p-value lies between .05 and .10, it was reported as marginally significant [128, 129] as a reference (possibly due to the size of the dataset [18]). All tests were performed in R-4.3.1. As a reference, turn signal missing rates against each personality trait were summarized.

### ANOVA Results

For ease of analysis, the participants were grouped according to the value of personality traits per criterion in Table 3.9 (left). Since the highest score was 100 and no scores were less than 40, the big-five personality traits were firstly grouped in 10-scores intervals, resulting in six groups. And the number of participants in each group were summarized in the first column on the left. For relative balance of the number of participants in each group, if there were too few participants in certain groups, the adjacent groups were further combined. And the results were summarized in the second column on the right, where the circled numbers were the cells that were combination of the corresponding rows in the previous column.

**Table 3.9:** Grouping criterion and grouping results of the participants per values of the personality traits (each participated in 6 scenarios).

Value	Notation	N		E		O		A		C	
$\leq 50$	Low 1 (L1)	7	7	2	⑦	2		0	-	1	⑦
(50,60]	Low 2 (L2)	8	8	5		1	⑥	0	-	6	
(60,70]	Medium 1 (M1)	14	14	16	16	3		6	6	9	9
(70,80]	Medium 1 (M2)	12	12	11	11	14	14	12	12	14	14
(80,90]	High 1 (H1)	4		8	8	18	18	16	16	10	10
$> 90$	High 2 (H2)	3	⑦	6	6	10	10	14	14	8	8

N, E, Q, A, and C represent Neuroticism, Extraversion, Openness, Agreeableness and Conscientiousness, respectively.

On basis of the groups, ANOVA results for inspecting effects of each personality trait on each metric were summarized in Table 3.10 (data segments of 8 s after the TOR). For each metric in the table, the first line is the test result overall, and the second to the fourth lines are test results depending on DoM. The results revealed that:

- For **TOT**, effects of extraversion and openness were found to be statistically significant ( $F(4, 277) = 2.748, p = .029, \eta^2 = 0.038$ ;  $F(3, 278) = 6.145, p <$

.001,  $\eta^2 = 0.062$ ), especially for DoM<sub>L</sub> ( $F(4, 85) = 3.518$ ,  $p = .011$ ,  $\eta^2 = 0.142$ ;  $F(3, 86) = 2.982$ ,  $p = .036$ ,  $\eta^2 = 0.094$ ).

- For **SpdMean** and **SpdStd**, only effects of neuroticism were found to be statistically significant ( $F(4, 283) = 4.666$ ,  $p = .001$ ,  $\eta^2 = 0.062$ ;  $F(4, 283) = 3.951$ ,  $p = .004$ ,  $\eta^2 = 0.053$ ). Similar to TOT, the effects were also found to be statistically significant for DoM<sub>L</sub> ( $F(4, 91) = 2.884$ ,  $p = .027$ ,  $\eta^2 = 0.112$ ;  $F(4, 91) = 2.846$ ,  $p = .028$ ,  $\eta^2 = 0.111$ ).
- For **SteerStd**, effects of neuroticism and agreeableness were found to be statistically significant ( $F(4, 283) = 3.294$ ,  $p = .012$ ,  $\eta^2 = 0.044$ ;  $F(3, 284) = 4.209$ ,  $p = .006$ ,  $\eta^2 = 0.043$ ). Although effect of extraversion was only found to be marginally significant overall, effects of both extraversion and agreeableness on SteerStd were statistically significant for DoM<sub>M</sub> ( $F(4, 91) = 3.049$ ,  $p = .021$ ,  $\eta^2 = 0.118$ ;  $F(3, 92) = 3.043$ ,  $p = .033$ ,  $\eta^2 = 0.090$ ).
- For **YawStd**, effects of neuroticism, agreeableness and conscientiousness were found to be statistically significant ( $F(4, 283) = 3.348$ ,  $p = .011$ ,  $\eta^2 = 0.045$ ;  $F(3, 284) = 3.715$ ,  $p = .012$ ,  $\eta^2 = 0.038$ ;  $F(4, 283) = 2.652$ ,  $p = .034$ ,  $\eta^2 = 0.036$ ) overall. However, while inspecting the effects for each DoM respectively, the results were only marginally significant.
- Finally, for **AccMax**, **SteerMax**, **SteerMean** and **YawMax**, effects of personality were almost not statistically significant, except for a few cases when DoM was long (AccMax-Extraversion:  $F(4, 91) = 2.635$ ,  $p = .039$ ,  $\eta^2 = 0.104$ ; SteerMean-Extraversion:  $F(4, 91) = 2.553$ ,  $p = .044$ ,  $\eta^2 = 0.101$ ; YawMax-Agreeableness:  $F(3, 92) = 3.411$ ,  $p = .021$ ,  $\eta^2 = 0.100$ ).

We could observe that different performance metrics seemed to be affected by different personality traits overall. In summary,

- Takeover time was mainly affected by extraversion and openness.
- Longitudinal performance (SpdMean and SpdStd) was mainly affected by neuroticism.
- Lateral performance (SteerStd and YawStd) was mainly affected by neuroticism and agreeableness, with effect of agreeableness more significant than neuroticism.

Besides, we could also observe that effects of personality were more significant for DoM<sub>M</sub> and DoM<sub>L</sub> than that for DoM<sub>S</sub>, indicating that the performance of drivers were similar in emergent situations, where the drivers' situation awareness was comparatively low.

**Table 3.10:** Summary of ANOVA results (effects of personality).

	N		E		O		A		C	
	F	p	F	p	F	p	F	p	F	p
<b>TOT</b>	1.41	.232	2.75	.029*	6.15	.001***	1.76	.155	1.31	.265
DoM <sub>S</sub>	1.55	.194	0.05	.995	3.43	.020*	0.63	.596	0.07	.992
DoM <sub>M</sub>	0.32	.863	1.15	.338	1.87	.140	1.46	.230	1.28	.286
DoM <sub>L</sub>	0.46	.766	3.52	.011*	2.98	.036*	1.22	.307	0.97	.431
<b>AccMax</b>	1.26	.285	1.90	.110	0.50	.681	0.33	.812	0.48	.754
DoM <sub>S</sub>	0.67	.61	0.43	.790	0.33	.802	0.04	.989	0.17	.952
DoM <sub>M</sub>	1.51	.21	0.04	.997	1.00	.397	0.34	.795	0.42	.797
DoM <sub>L</sub>	1.24	.30	2.64	.039*	1.17	.326	1.85	.143	0.53	.714
<b>SpdMean</b>	4.67	.001**	1.57	.182	0.76	.517	1.02	.383	1.17	.325
DoM <sub>S</sub>	1.44	.226	0.92	.455	0.65	.582	0.06	.979	0.59	.668
DoM <sub>M</sub>	1.25	.295	2.43	.054◇	2.01	.118	1.89	.137	1.05	.387
DoM <sub>L</sub>	2.88	.027*	0.14	.968	0.63	.598	0.94	.424	1.01	.407
<b>SpdStd</b>	3.95	.004**	1.48	.208	0.45	.717	0.54	.658	1.24	.293
DoM <sub>S</sub>	1.19	.319	0.93	.452	0.59	.623	0.14	.938	0.58	.680
DoM <sub>M</sub>	1.15	.336	3.14	.018*	1.76	.160	1.19	.319	1.18	.323
DoM <sub>L</sub>	2.85	.028*	0.24	.915	0.86	.465	0.73	.535	0.43	.785
<b>SteerMax</b>	1.72	.145	1.62	.170	1.43	.235	1.26	.288	1.72	.145
DoM <sub>S</sub>	1.66	.166	0.76	.553	0.22	.884	0.71	.551	0.66	.621
DoM <sub>M</sub>	1.25	.296	1.62	.175	0.98	.406	1.03	.384	0.60	.666
DoM <sub>L</sub>	1.06	.382	1.59	.184	1.43	.240	2.62	.055◇	0.88	.478
<b>SteerMean</b>	0.69	.598	0.66	.624	0.41	.750	0.06	.979	0.42	.798
DoM <sub>S</sub>	1.40	.242	0.47	.754	0.75	.527	0.66	.582	1.39	.243
DoM <sub>M</sub>	0.61	.660	1.31	.274	0.73	.539	0.15	.930	1.44	.229
DoM <sub>L</sub>	1.20	.317	2.55	.044*	2.19	.095◇	1.55	.206	0.85	.495
<b>SteerStd</b>	3.29	.012*	2.06	.087◇	1.30	.275	4.21	.006**	2.38	.052◇
DoM <sub>S</sub>	2.09	.088◇	0.73	.574	0.88	.456	1.76	.160	0.38	.820
DoM <sub>M</sub>	2.02	.099◇	3.05	.021*	1.24	.300	3.04	.033*	0.97	.430
DoM <sub>L</sub>	0.75	.562	0.53	.712	0.39	.758	1.96	.125	1.24	.300
<b>YawMax</b>	2.13	.077◇	1.38	.240	1.39	.247	1.61	.182	1.87	.117
DoM <sub>S</sub>	1.67	.164	1.11	.359	0.21	.887	0.58	.630	1.05	.384
DoM <sub>M</sub>	1.45	.225	1.66	.165	1.25	.295	0.71	.551	0.53	.715
DoM <sub>L</sub>	1.09	.366	1.65	.170	1.29	.282	3.41	.021*	1.04	.389
<b>YawStd</b>	3.35	.011*	1.82	.125	1.29	.277	3.72	.012*	2.65	.034*
DoM <sub>S</sub>	1.54	.197	1.21	.313	0.70	.555	1.29	.282	0.82	.517
DoM <sub>M</sub>	2.17	.079◇	2.91	.026*	1.94	.128	2.50	.064◇	0.77	.546
DoM <sub>L</sub>	0.87	.487	0.92	.458	0.20	.894	2.61	.056◇	1.50	.210

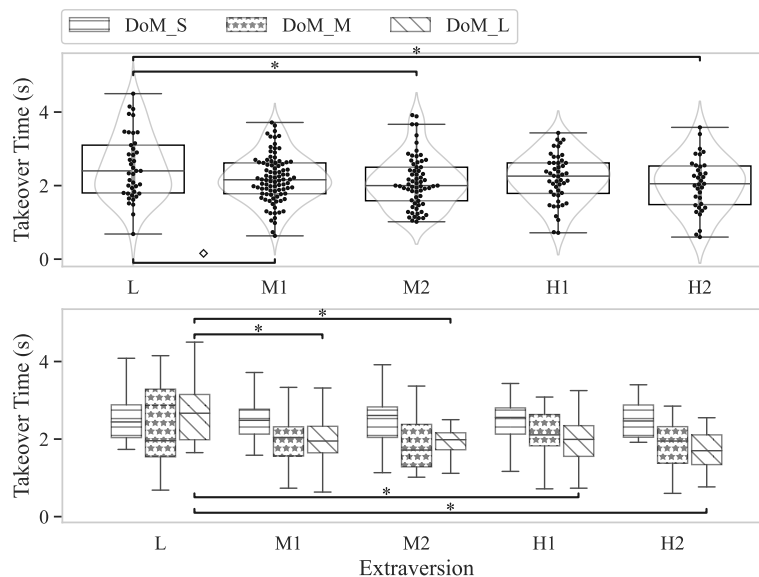
◇: &lt; 0.10; \*: &lt; 0.05; \*\*: &lt; 0.01; \*\*\*: &lt; 0.001.

Furthermore, the same analysis was performed for data segments of 15 s after the TOR. Results revealed that similar effects of neuroticism and agreeableness on SpdMean and YawStd could be found (SpdMean-Neuroticism (Overall): F(4,

283) = 3.210,  $p = .013$ ,  $\eta^2 = 0.043$ ; SpdMean-Neuroticism (DoM<sub>L</sub>):  $F(4, 91) = 3.411$ ,  $p = .012$ ,  $\eta^2 = 0.130$ ; YawStd-Neuroticism (Overall):  $F(4, 283) = 3.918$ ,  $p = .004$ ,  $\eta^2 = 0.052$ ; YawStd-Neuroticism (DoM<sub>M</sub>):  $F(4, 91) = 3.490$ ,  $p = .011$ ,  $\eta^2 = 0.133$ ; YawStd-Agreeableness:  $F(3, 284) = 3.962$ ,  $p = .009$ ,  $\eta^2 = 0.040$ ). However, the effects on SpdStd and SteerStd were not longer so significant. This suggested that although long-term effects still existed, personality traits seemed to impact instant more than stable maneuvers during takeovers. This could possibly be attributed to scenario constraints on drivers' maneuvers. Hence, the remaining analysis would focus on data segments of 8 s after the TOR.

### Takeover Time

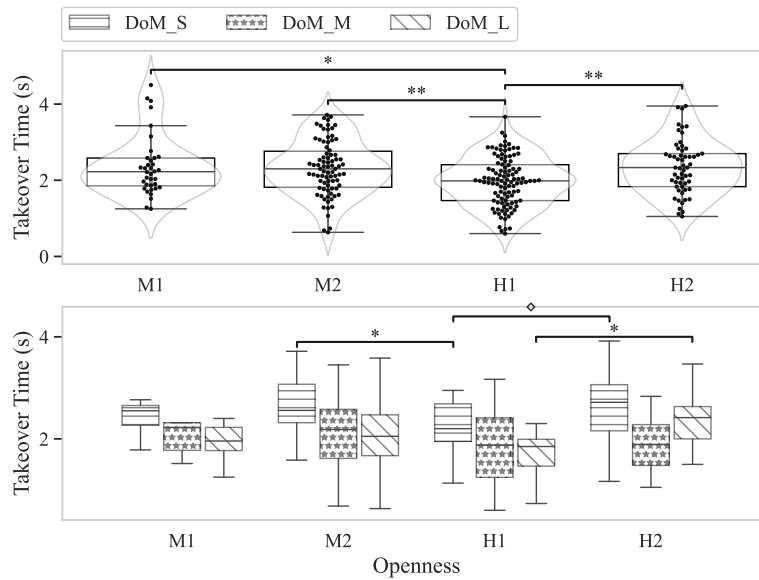
For **extraversion**, Tukey HSD tests post-hoc revealed that (Fig. 3.13), overall, takeover time in group L was statistically significantly longer than that in group H2 and M2 (L-H2: diff = 0.472,  $p_{\text{adj}} = .033$ ; L-M2: diff = 0.406,  $p_{\text{adj}} = .034$ ), and marginally significantly longer than that in group M1 (L-M1: diff = 0.333,  $p_{\text{adj}} = .093$ ). Considering only DoM<sub>L</sub>, it could also be found that takeover time in group L was statistically significantly longer than all the other groups (L-H1: diff = 0.686,  $p_{\text{adj}} = .047$ ; L-H2: diff = 0.861,  $p_{\text{adj}} = .015$ ; L-M1: diff = 0.656,  $p_{\text{adj}} = .029$ ; L-M2: diff = 0.730,  $p_{\text{adj}} = .016$ ). Thus, it could be concluded that **lower extraversion would lead to longer takeover time**.



**Figure 3.13:** Above: Takeover time of different levels of extraversion. Below: Takeover time of different levels of extraversion given different DoM.

For **openness**, Tukey HSD post-hoc tests revealed that (Fig. 3.14), overall, takeover time in group H1 was statistically significantly shorter than that in





**Figure 3.14:** Above: Takeover time of different levels of openness. Below: Takeover time of different levels of openness given different DoM.

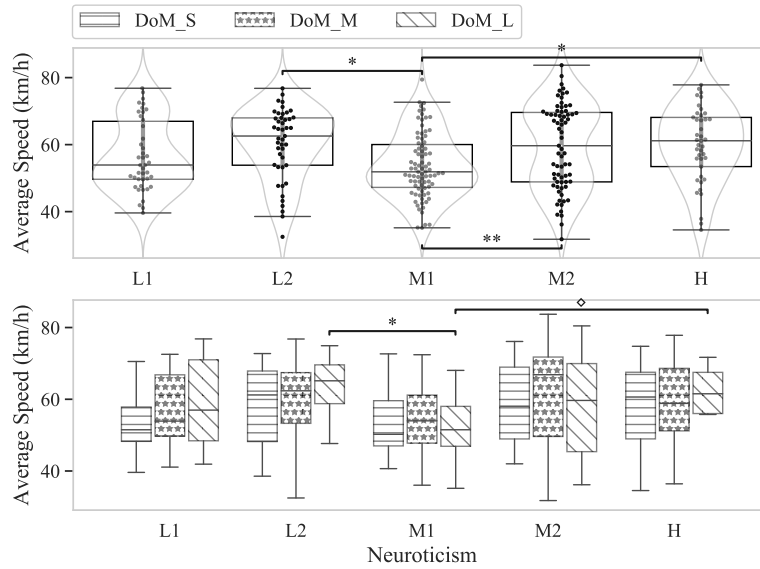
other groups (H1-M1:  $\text{diff} = -0.396$ ,  $p.\text{adj} = .018$ ; H1-M2:  $\text{diff} = -0.346$ ,  $p.\text{adj} = .005$ ; H1-H2:  $\text{diff} = -0.376$ ,  $p.\text{adj} = .006$ ). With DoM<sub>L</sub>, statistically significant difference was only found between groups H2 and H1 (H1-H2:  $\text{diff} = -0.560$ ,  $p.\text{adj} = .028$ ); and with DoM<sub>S</sub>, it was found that takeover time in group H1 was statistically significantly shorter than that in group M2 (H1-M2:  $\text{diff} = -0.391$ ,  $p.\text{adj} = .045$ ), and marginally significantly shorter than that in group H2 (H1-H2:  $\text{diff} = -0.417$ ,  $p.\text{adj} = .058$ ). Thus, it could be concluded that **appropriate openness would help to decrease the takeover time**.

### Longitudinal Performance

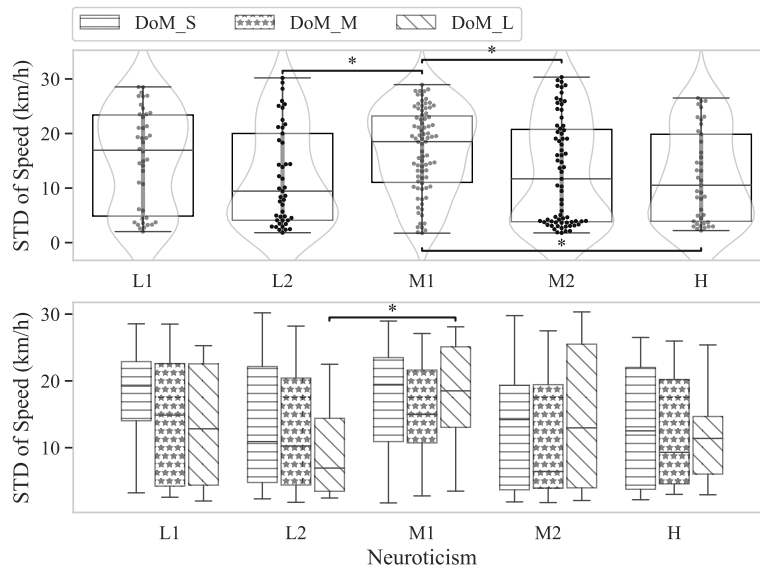
For longitudinal performance, mean speed (SpdMean) and standard deviation of speed (SpdStd) were considered, which were found to be mainly affected by neuroticism. Tukey HSD post-hoc tests revealed that (Figs. 3.15 and 3.16):

- Mean speed in group M1 was statistically significantly lower than that in groups L2, M2, and H (M1-L2:  $\text{diff} = -6.570$ ,  $p.\text{adj} = .011$ ; M1-M2:  $\text{diff} = -5.989$ ,  $p.\text{adj} = .008$ ; M1-H:  $\text{diff} = -6.937$ ,  $p.\text{adj} = .010$ );
- STD of speed in group M1 was statistically significantly higher than that in groups L2, M2, and H (M1-L2:  $\text{diff} = 4.800$ ,  $p.\text{adj} = .020$ ; M1-M2:  $\text{diff} = 3.844$ ,  $p.\text{adj} = .048$ ; M1-H:  $\text{diff} = 4.930$ ,  $p.\text{adj} = .024$ ).

Therefore, it could be concluded that **medium neuroticism led to relatively lower speeds shortly after the TOR, whereas with larger standard deviations**.



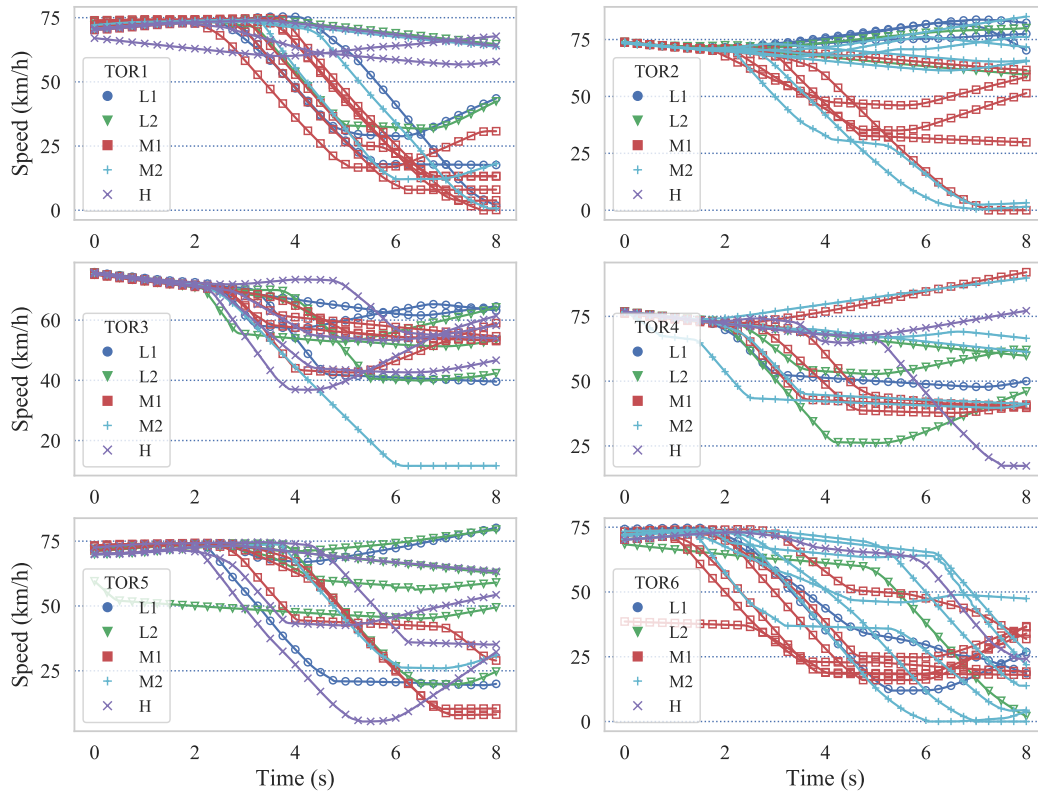
**Figure 3.15:** Above: Mean speed of different levels of neuroticism. Below: Mean speed of different levels of neuroticism given different DoM.



**Figure 3.16:** Above: STD of speed of different levels of neuroticism. Below: STD of speed of different levels of neuroticism given different DoM.

It is more straightforward to plot the speed profiles of different levels of neuroticism in each takeover scenario (Fig. 3.17). Besides the above conclusions, it could be observed that, except for TOR3, drivers with medium neuroticism overall tended to decelerate quicker and resulted in a lower speed shortly after the TOR compared with others, especially with drivers in groups L1 and H. Actually, as proposed in [130] (introduced in Section 4.4), using dynamic-time-warping-based (DTW-based) k-means clustering, drivers' takeover maneuvers can be classified into three categories in consideration of speed, acceleration,

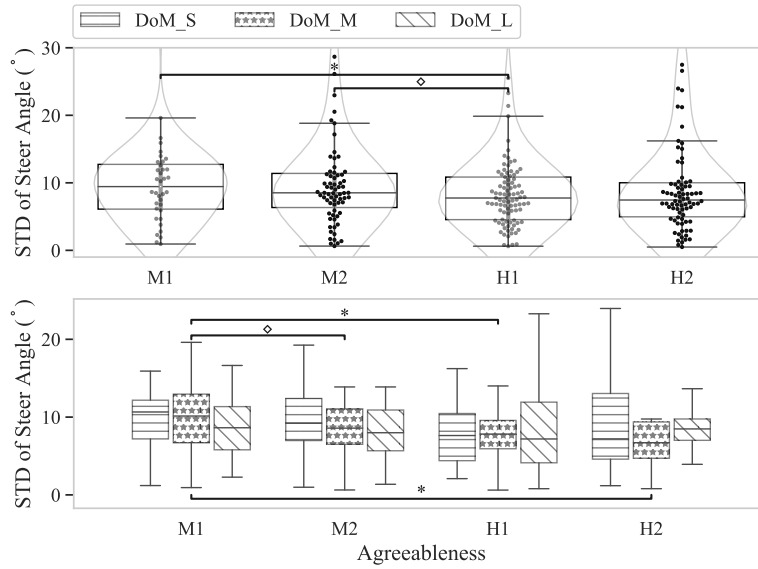
steering wheel angle and yaw rates, which may be attributed to effects of drivers' personality traits.



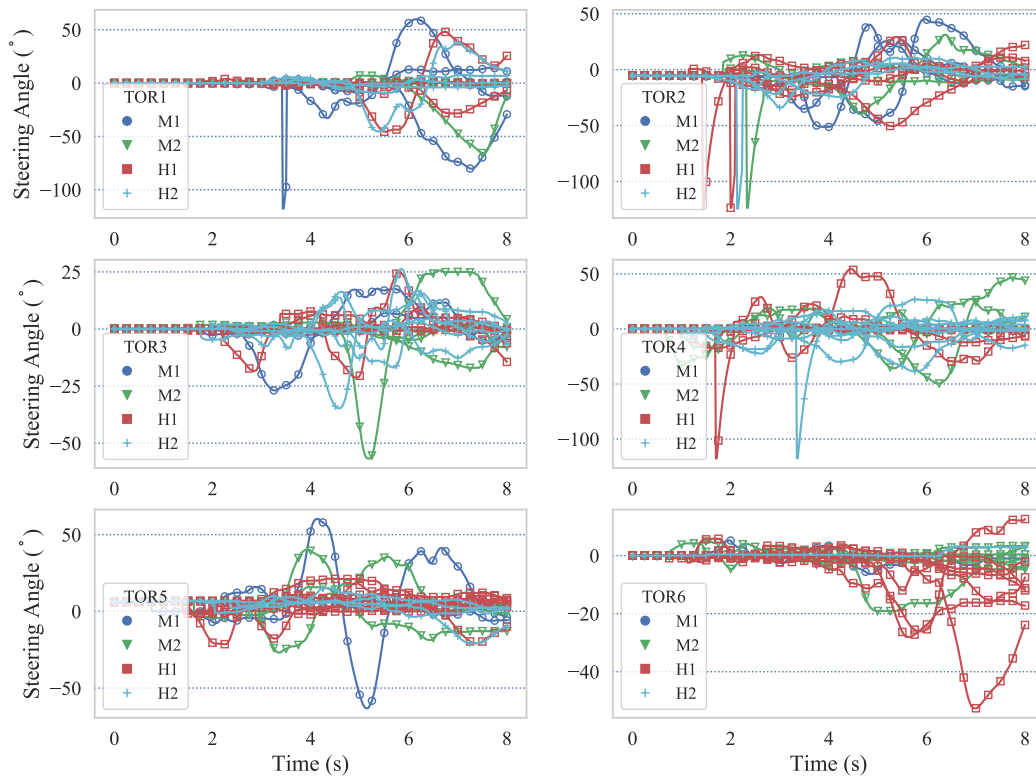
**Figure 3.17:** Speed profile of different levels of neuroticism given DoM<sub>L</sub> in each takeover scenario.

### Lateral Performance

For lateral performance, only standard deviation of steering wheel angle (Steer-Std) was considered, which was found to be mainly affected by agreeableness (since effects of neuroticism were not statistically significant when considered in groups). Tukey HSD post-hoc tests revealed that (Fig. 3.18), STD of steering angle in group H1 was statistically significantly smaller than that in group M1 (H1-M1:  $\text{diff} = -4.269$ ,  $p.\text{adj} = .006$ ), and marginally significantly smaller than that in group M2 (H1-M2:  $\text{diff} = -2.362$ ,  $p.\text{adj} = .100$ ). Further looking into the group of DoM<sub>M</sub>, it could be found that STD of steering angle in group M1 was statistically significantly larger than that in groups H1 and H2 (M1-H1:  $\text{diff} = 5.757$ ,  $p.\text{adj} = .047$ ; M1-H2:  $\text{diff} = 6.407$ ,  $p.\text{adj} = .025$ ), and marginally significantly larger than that in group M2 (M1-M2:  $\text{diff} = 5.542$ ,  $p.\text{adj} = .077$ ). Therefore, it could be concluded that **higher agreeableness would lead to relatively smaller standard deviation of steering angles shortly after the TOR**. However, the effect of agreeableness lower than 60 was still not clear from the data collected.



**Figure 3.18:** Above: STD of steering angle of different levels of agreeableness. Below: STD of steering angle of different levels of agreeableness given different DoM.



**Figure 3.19:** Steering profile of different levels of agreeableness given DoM<sub>M</sub> in each takeover scenario.

The steering profiles were plotted in Fig. 3.19. The relationships were not as clear as those of speed profiles in Fig. 3.17. Smaller standard deviation indicated relatively more stable operations. Besides, we could observe that drivers with

higher agreeableness (groups H1 and H2) tended to make the first steering operation earlier than the others (groups M1 and M2). However, it took a longer time to reach the maximum. That was in corresponding to smaller standard deviation of the steering wheel angles.

### Turn Signal Missing Rates

Turn signal missing rates in the first 8 seconds after TOR were plotted in Fig. 3.20. It can be observed that (from the tendency of the line plots): 1) higher neuroticism, openness and conscientiousness led to higher turn signal missing rates; 2) and medium agreeableness seemed to also result in higher turn signal missing rates; 3) whereas, effects of extraversion was not quite obvious.

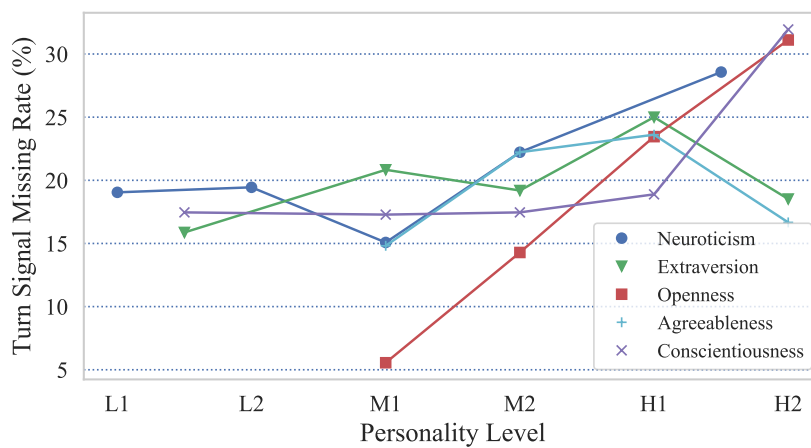


Figure 3.20: Turn signal missing rates for each personality trait.

### Summary

The above results were summarized in Table 3.11, where +, −, U and ∩ represent positive, negative, u-shape (similar to parabolic) and n-shape (similar to negative-parabolic) relationships, respectively.

Table 3.11: Summary of the analysis results (effects of personality).

	TOT	SpdMean	SpdStd	SteerStd	TurnSigRate
N		U	∩		+
E	−				
O	U				+
A				−	∩
C					+

N, E, Q, A, and C represent Neuroticism, Extraversion, Openness, Agreeableness and Conscientiousness, respectively.

### 3.4.4 General Discussions

#### Role of DoM (H4)

From Table 3.10, we could observe that, except for openness, effects of personality traits on takeover performance were basically not statistically significant given DoM<sub>S</sub>, where drivers were left no time to gain enough situation awareness. In contrast, effects of personality on takeover performance were most significant given DoM<sub>L</sub> (either statistically or marginally). This suggested that personal preference was only significant when drivers has gained certain levels of situation awareness. Therefore, hypothesis **H4** was **supported**. In a work regarding effect of stress and personality on driving behavior, Ge et al. [131] analyzed the mediating effects of personality traits, and found that anger (a subitem of the personality trait neuroticism, referring to an individual's perception of the situation as annoying or frustrating) mediated the relationship between stress and dangerous driving behavior (e.g., aggressive driving). Our work contributed to the literature in that it has found DoM as a moderator variable between drivers' personality traits and takeover performance, where there existed a stronger relationship for longer DoM.

#### Effect of Neuroticism (H1-1)

Results of Section 3.4.3 suggested that neuroticism mainly affects longitudinal performance, especially speed behaviors, and the effects were most obvious when DoM was long. As suggested by Matthews [132], highly neurotic person usually have high awareness of threat, it made sense that people with higher neuroticism tended to have a greater impetus to step on the brake and slow down the vehicle facing stimulus. Curious was that, the relationship seemed to be nonlinear. Our results showed that drivers with medium level of neuroticism was found to decelerate earlier than the other groups, resulting in lower average speeds and higher standard deviation shortly after TOR. Since lower speeds in emergent situations were good guarantees for safety, although risky driving behaviors were found to be positively related to neuroticism during manual driving [123], drivers with medium neuroticism seemed to exemplify safer takeover behaviors in the longitudinal direction. Therefore, hypothesis **H1-1** was **supported**. However, this was not the case when it came to using of turn signals, where drivers' turn signal missing rate was positively related with neuroticism shortly after TOR. This could possibly pose potential threats to the driver when a vehicle is approaching from behind while the driver is changing lanes to avoid the obstacles.

### Effect of Agreeableness (H1-2)

Results of Section 3.4.3 suggested that agreeableness mainly affects lateral performance, especially steering behaviors, and the effects were most obvious given DoM<sub>M</sub>. Since individuals high in agreeableness were believed to be less aggressive than those low in agreeableness after being exposed to hostile prime words [133], it made sense that drivers with higher agreeableness (groups H2 and H1) tended to make the first steering operation earlier (to avoid being too close to the obstacles) and had more stable operations (smaller standard deviation). This was in alignment with the study by Akbari et al. [123], which showed that risky driving behaviors were negatively related to agreeableness. In this sense, drivers with higher agreeableness seemed to exemplify safer takeover behaviors in the lateral direction. Therefore, hypothesis **H1-2** was **supported**. As for turn signal missing rates, medium agreeableness seemed to result in higher turn signal missing rates than the other groups shortly after TOR. Higher agreeableness appears to be able to improve such inadvisable maneuvers.

### Effect of Conscientiousness (H1-3)

For conscientiousness, no significant results can be observed from Table 3.10. Therefore, hypothesis **H1-3** was **negated**. As mentioned by DeYoung et al. in [134], “conscientiousness appears to reflect the tendency to maintain motivational stability within the individual, to make plans and carry them out in an organized and industrious manner”, indicating that conscientiousness would mainly affect planned rather than spontaneous behaviors. Since takeover was more of immediate urges than long-term goals, it made sense that effects of conscientiousness was not significant in emergent situations. This was actually in contradiction with the results given by Koposov et al. [135], where a positive correlation was found between reaction times in a shooting game and conscientiousness. Considering that the same tasks were repeated in their study and all the scenarios in our study were repeated only once, conscientious might make a difference in the takeover performance after the same scenarios were repeated. However, to verify this, more experiments are necessary.

### Effect of Extraversion and Openness (H2 & H3)

Results of Section 3.4.3 implied that openness and extraversion mainly affect takeover time, and the effects were most obvious when DoM was long. For extraversion, our results revealed a negative relationship between extraversion and takeover time. Hence, hypothesis **H3** was **negated**. This was supported by



Brebner and Cooper's model of extraversion [136], where it was indicated that extraverts have shorter reaction times than introverts in simple discrimination and response tasks. Nevertheless, contradictory results were also reported by Hummel et al. [126] and Casal et al. [137], where no statistically significant relations between extraversion and simple reaction time were found. Again, the differences could find its explanation in the theory of S-analysis and R-organization, a unified model of extraversion and introversion [136]. In this theory, since extroverts generate excitation from R-organization but inhibition from S-analysis, they are believed to be able to make faster and more frequent responses of a motor nature than introverts do. In simple tasks requiring simple response as in our case, it was anticipated that extraverts should make faster responses than introverts do, as they "generate excitation from R-organization but inhibition from S-analysis" [136].

For openness, our results revealed that a proper level of openness would help to decrease the takeover time. Therefore, hypothesis **H2** was **supported**. Generally, people with a high level of openness has been believed to be intellectually curious, willing to try new things, and more likely to engage in risky behaviors [138]. Hence, it was counterintuitive that participants with higher openness were found to have less trust in ADS [125]. However, this may also partly explain why drivers with medium level of openness tended to have quicker takeovers. Curious was that, the relationship seemed to be nonlinear. Therefore, higher openness does not necessarily lead to quicker and safer takeovers. Since the relationship between openness and movement has not been emphasized enough yet, more research are necessary to verify this conclusion. As for turn signal missing rates, similar to neuroticism and conscientiousness, drivers' turn signal missing rate was positively related with openness in the first 8 s after TOR, which may bring about safety concerns in complicated situations.

### 3.4.5 Summary of Analysis 2

In this analysis, we attempted to explore the impact of big five personality traits on takeover performance of drivers given different duration of monitoring before takeover requests. Main results were summarized in Table 3.11.

Overall, the results revealed that different personality traits affected takeover performance in different aspects. Specifically,

- extraversion and openness mainly affects takeover time;
- neuroticism mainly affects longitudinal performance, especially mean and standard deviation of speeds shortly after takeover requests;



- agreeableness mainly affects lateral performance, especially standard deviation of steering angles shortly after takeover requests;
- effects of personality are most significant when drivers have gained certain levels of situation awareness;
- regarding using of turn signals shortly after takeover requests, it was found that turn signal missing rates were positively related with neuroticism, openness and conscientiousness, respectively, and negatively quadratically related with agreeableness.

In manual driving, results given by Akbari et al. [123], Corr et al. [139] and Settles et al. [140] indicated that high neuroticism, low agreeableness and low conscientiousness were linked to riskier behaviors and higher accident involvement rates. Corresponding to our study, shortly after takeover requests, both lower and higher neuroticism led to higher average speeds, and lower agreeableness led to higher standard deviation of steering angles, whereas effects of conscientiousness were not significant. It seems that manual driving behaviors and takeover behaviors during automated driving are not strictly corresponded. Therefore, conclusions obtained from manual driving require to be reverified before adapted to automated driving.

### 3.5 Summary of Chapter 3

Chapter 3 aims to fulfill Objective 2, and the main results and conclusions are:

- Experiments were designed and conducted to collect data regarding drivers' responses and takeover behaviors during different takeover scenarios.
- Impacts of two new factors—duration of monitoring before the TOR and drivers' personality traits, were explored, with deep insights of eye-tracking data.
- Detailed conclusions of the main results can be referenced to Sections 3.3.5 and 3.4.5, respectively.

Based on the collected data and analyzing results, in Chapter 4, some models will be built to predict takeover drivers' takeover behaviors before the TOR, so that optimal measures can be taken by the ADS based on the predicted results.

## Chapter 4

# Modeling Takeover Behaviors

### 4.1 Dataset Preparation

#### 4.1.1 Raw Data

576 takeovers were collected from 48 participants, where gaze attributes were calculated  $x$  seconds ( $x$  was up to 15 s) before the TOR in 3 seconds time interval (3 s, 6 s, 9 s, 12 s, 15 s), and metrics of takeover quality were calculated 15 s after the TOR.

#### 4.1.2 Independent Variables

Based on the review results in Chapter 2 and the research results in Chapter 3, a list of the independent variables that will be considered in the modeling process are summarized in Table 4.1, along with a short description. The variables are categorized into driver attributes, system attributes and scenario attributes, with each emphasizing on different aspects.

##### Driver Attributes

**Driver Attributes 1** Drivers attributes 1 are driver-related factors that were obtained through questionnaires before the experiment. Drivers' age, gender, years of driving, and experience with ADS were obtained through a Google form online, where 10 simple questions were involved (Appendix A.3). And drivers' personality traits were obtained through the website <https://bigfive-test.com/>, where 120 items were involved. A sample of the results of the personality test can be accessed through <https://bigfive-test.com/result/62590e5c72715a0009a7963f>. Results per coding criterion in Table 4.1 were summarized in Table 4.2.

**Driver Attributes 2** Driver attributes 2 are driver-related factors that were obtained through the SuRT during the experiment. Hand-in-use refers to drivers'

**Table 4.1:** Independent variables for modeling takeover behaviors.

Independent Variables		Description and Coding
Driver Attributes 1	Age	1: 18–23; 2: 24–30; 3: > 30
	Gender	1: male; 2: female
	Driving Experience	1: < 1; 2: 1–3; 3: 3–5; 4: 5–10; 5: > 10
	Experience With ADS	0: no experience; 1: experience with ADAS; 2: experience with ADS
	Personality	1: ≤ 50; 2: ≤ 60; 3: ≤ 70; 4: ≤ 80; 5: ≤ 90; 6: > 90; per big-five personality tests
Driver Attributes 2	Reaction Speed	<b>float</b> ; average NDRT scores
	Hand in Use	1: left; 2: right or both; while performing the NDRT
Driver Attributes 3	Eyes-on-Road (EoR)	0: eyes-off-road; 1: eyes-on-road; at the moment of TOR
	Proportion of AOIs	<b>float</b> ; proportion of gazes on road/ left side/right side.
	Fixation	<b>int</b> ; number of fixations
	Blink	<b>int</b> ; number of blinks
	Pupil Diameter	<b>float</b> ; mean, std, and amplitude
	Eyelid Opening	<b>float</b> ; mean, std, and amplitude
System Attributes	Gaze Dispersion	<b>float</b> ; measured by gaze entropy
	DoM	1: short (0 s); 2: medium (5 s); 3: long (10 s); 4: long+ (15+ s)
	TB/LeadT	1: short (5 s); 2: long (7 s)
Scenario Attributes	Weather	1: sunny; 2: foggy
	Road	1: straight; 2: curve
	Takeover Scenario	<b>int</b> ; 1–12
	Mental Demand	<b>float</b> ; NASA-TLX
	Physical Demand	<b>float</b> ; NASA-TLX
	Temporal Demand	<b>float</b> ; NASA-TLX
	Scenario Predictability	<b>float</b> ; NASA-TLX
	Scenario Criticality	<b>float</b> ; NASA-TLX

hand in use while performing the SuRT, which was obtained by playing back the videos recorded. Although it was anticipated that the same driver would use the same hands to do the NDRT even in different scenarios out of habits, we have also observed that some drivers changed their hands for several times across different scenarios. Therefore, we reviewed all the 576 video recordings one by one. This resulted in 308 takeovers where drivers used their left hand, and 268 takeovers where drivers used their right or both hands, respectively.

Although there are standard tests to test drivers' reaction speed. In this research, drivers' reaction speed was obtained by measuring the average time

**Table 4.2:** Coding results of driver attributes 1.

Independent Variables		Coding Results
Age		1: 26; 2: 16; 3: 6
Gender		1: 38; 2: 10
Driving Experience		1: 12; 2: 14; 3: 11; 4: 6; 5: 5
Experience with ADS		0: 8; 1: 26; 2: 14
Personality	Neuroticism	1: 7; 2: 8; 3: 14; 4: 12; 5: 7
	Extraversion	1: 7; 2: 16; 3: 11; 4: 8; 5: 6
	Openness	1: 6; 2: 14; 3: 18; 4: 10
	Agreeableness	1: 6; 2: 12; 3: 16; 4: 14
	Conscientiousness	1: 7; 2: 9; 3: 14; 4: 10; 5: 8

required by the driver to complete the SuRT for a single time. An excerpt of the raw data obtained can be referenced to Appendix B.1, where the first column is the timestamp recording drivers' operations, and the final two columns are number of correct times and wrong times of drivers doing the task. The average reaction time is calculated as the total duration of the task (difference of the last row and first row in the timestamp column) divided by the number of times drivers doing the task (sum of the last row in the last two columns). Besides, drivers' correction rates on performing the NDRT were also considered, which are calculated as the number of correct times divided by the total number of times drivers doing the task. Since each driver participated in 12 takeover scenarios, the final average reaction speed and correct rates are averaged through the 12 files. And the Python script can also be referenced to Appendix B.1.

**Driver Attributes 3** Driver attributes 3 are drivers gaze behaviors during the experiments. Eyes-on-road at the moment of TOR, fixation and blink can be obtained from the raw data directly. Proportion of AOIs, pupil diameter, and eyelid opening were calculated  $x$  seconds before the TOR in three seconds time interval, where  $x$  can be 3 s, 6 s, 9 s, 12 s, and 15 s. Before the calculation, pupil diameter and eyelid opening were first processed via Eq. (3.2). Gaze dispersion was calculated via Eq. (3.1).

### System Attributes

As discussed in Section 2.3.1, system-related factors involve level of automation, NDRTs, and HMI, etc. Although they all have been found to affect drivers' takeover behaviors, for simplicity of the experiments, NDRT, modality of TOR, level of automation, and fidelity of the driving simulator were kept constant in this research. Only duration of monitoring before the TOR and time budget were

considered as variables to adjust urgency of the scenarios and states of drivers before the TOR.

### Scenario Attributes

Except traffic density and vehicle speeds, the other variables regarding takeover scenarios have received relatively fewer attention. Regarding weather conditions, except TOR5, all the other scenarios were in sunny days. Regarding road structures, 6 were straight roads (TOR1, TOR4, TOR6, TOR8, TOR9, TOR10), 4 were curve roads (TOR2, TOR5, TOR7, TOR12), and 2 were ramp roads with vehicle merging from the left side of the road (TOR3, TOR11).

Since scenario attributes (urgency, predictability, criticality and complexity) are not easy to define beforehand, we asked the participant to answer a short questionnaire (Appendix A.4) after each takeover in the latter half of the experiments, where the first five questions were drivers rating on different aspects of the scenarios, and the last three questions were drivers' self-ratings on their performance in completing the tasks. Finally, the average scores were utilized in evaluating the scenarios. An excerpt of the NASA-TLX data and the Python script to process the data can be referenced to Appendix B.2.

### 4.1.3 Dependent Variables

Takeover time [17, 37, 63, 96], takeover quality [95, 97, 99], and takeover readiness [94, 97, 141] have been three of the most frequently researched dependent variables in the literature. Takeover time and takeover readiness are relatively easy to define, hence, they are also considered in this research. Takeover quality, however, can usually be defined differently depending on the scenarios. Therefore, instead of takeover quality, a new term is defined in this research—takeover style, which refers to drivers takeover maneuvers shortly after the TOR (longitudinal and lateral maneuvers). More detailed discussions can be referenced to Sections 4.2, 4.3 and 4.4, respectively.

### 4.1.4 Data Cleaning

System attributes can easily be determined, and do not need special preprocessing. Driver attributes 1 and 2 are also relatively easy to handle, which have been summarized in Table 4.2 and 4.4 respectively. Driver attributes 3, however, is not that easy to handle, since missing recordings and low quality recordings need to be dealt with following the steps below.

**Table 4.3:** Dependent variables for modeling takeover behaviors.

Dependent Variables		Comments
Takeover Time		
takeover readiness		Whether there is a crash.
	Speed	m/s
	Acceleration	m/s <sup>2</sup>
Takeover Style	Steering Wheel Angle	°
	Yaw Rate	°/s

1. Sometimes, it can be difficult to discriminate between left screen, left mirror and distractor, they are denoted as LeftSide for convenience; similarly, right screen and right mirror are denoted as RightSide. This results in 4 AOIs—CenterScreen, LeftSide, RightSide, and Others.
2. Since eye tracking data are missing from 61 takeovers (G2N02: TOR2 & TOR7; G2N03: TOR8; G2N04: TOR2; G2N05: TOR4 & TOR10–12; G2N07: TOR1; G4N04: all; G4N06: TOR12; G4N07: TOR10; G4N11: TOR5; G4N12: all; G4N14: all; G5N06: TOR6; G5N07: all), they were deleted from the dataset.
3. Considering that frequency of blinks is relatively low and number of fixations is not very informative, these two features are also removed from the dataset.

In deicing scenario attributes, except performance, all the other seven questions of the questionnaire as shown in Appendix A.4 are utilized, and the results are summarized in Table 4.6. Besides, a predefined scenario criticality is also considered as discussed in Section 3.3, which is different from drivers' ratings on scenario criticality.

Finally, we obtain 515 samples and 38 features for modeling takeover behaviors, including 25 driver-related features, 2 system-related features, and 11 scenario-related features, respectively.

Below is a list of the column names:

```

1 'Age', 'Gender', 'YearsDriving', 'ExperienceADS',
2 'Neuroticism', 'Extraversion', 'Openness', 'Agreeableness', 'Conscientiousness',
3 'HandInUse', 'avgSpd', 'avgCorrectRate', 'DoM', 'TB',
4
5 'EoR', 'CenterScreen_xs', 'Others_xs', 'LeftSide_xs', 'RightSide_xs',
6 'EyelidOpeningMean_xs', 'EyelidOpeningStd_xs', 'EyelidOpeningAmp_xs',
7 'PupilDiameterMean_xs', 'PupilDiameterStd_xs', 'PupilDiameterAmp_xs',
8 'NumFixations_xs', 'GazeEntropy_xs',
9
10 'Scenario', 'Scenario_C', 'Weather', 'Road', 'Mental_Demand', 'Physical_Demand',
11 'Temproal_Demand', 'Predictability', 'Criticality', 'Effort', 'Frustration'

```

**Table 4.4:** Summary of drivers' reaction speed and correction rates.

Participant	avgSpd	avgCorrectRate	Participant	avgSpd	avgCorrectRate
G2N01	0.991725	0.995516	G4N04	0.439187	0.482439
G2N02	1.017722	0.989080	G4N05	0.920621	0.987500
G2N03	0.946391	0.974286	G4N06	0.879372	0.986631
G2N04	0.741174	0.985344	G4N07	1.104026	0.996581
G2N05	0.751500	0.985092	G4N08	0.762827	0.944828
G2N06	0.844444	0.998691	G4N09	0.690114	0.991675
G2N07	0.923378	0.997222	G4N10	1.564531	1.000000
G2N08	0.848687	0.982051	G4N11	0.890120	0.987755
G2N09	0.831099	0.966165	G4N12	1.147169	0.996528
G2N10	1.429546	0.978070	G4N13	0.917598	0.961057
G2N11	1.023785	0.990551	G4N14	0.825383	0.971357
G2N12	1.024183	0.992188	G4N15	0.831956	0.978535
G2N13	0.833207	0.991162	G5N01	0.863545	0.992095
G2N14	1.429964	0.995565	G5N02	1.082905	0.991763
G2N15	1.088770	1.000000	G5N03	0.853464	0.976654
G2N16	0.727301	0.957001	G5N04	0.725760	0.964758
G2N17	1.029301	0.987441	G5N05	0.750423	0.981962
G2N18	0.713754	0.979212	G5N06	1.407347	0.991398
G3N09	0.862733	0.997382	G5N07	1.024451	0.995231
G3N10	0.822102	0.997519	G5N08	0.977498	0.994074
G3N11	1.010681	0.993893	G5N09	0.950444	0.985401
G4N01	0.831801	0.970037	G5N10	0.927186	0.988780
G4N02	0.861871	0.988387	G5N11	1.606755	0.963145
G4N03	0.839045	0.987469	G5N12	0.759982	0.972125

GxNxx is number of the participant, where Gx represents Group x, and Nxx represent number of the participant in that group.

avgSpd and avgCorrectRate represent average speed and average correction rate while doing the SuRT respectively.

## 4.2 Modeling Takeover Time

### 4.2.1 Problem Statement and Objectives

The definition of takeover time can be referenced to Section 3.3.2. Since takeover time from 4 and 7 takeovers were missing and earlier than the TOR respectively, they were deleted from the dataset. The rest of the data are plotted in the histogram of Fig. 4.1, where the red line is the kernel density estimation. We can see that except a few of the samples (23 out of 565), the data are approximately normally distributed. And a descriptive statistics of the remaining samples is summarized in Table 4.7.

Since takeover time is a continuous variable, the objective of this section is to model takeover time as a regression problem based on the independent variables summarized in Table 4.1.2. In this research, the model is developed based on

**Table 4.5:** Summary of driver-related features.

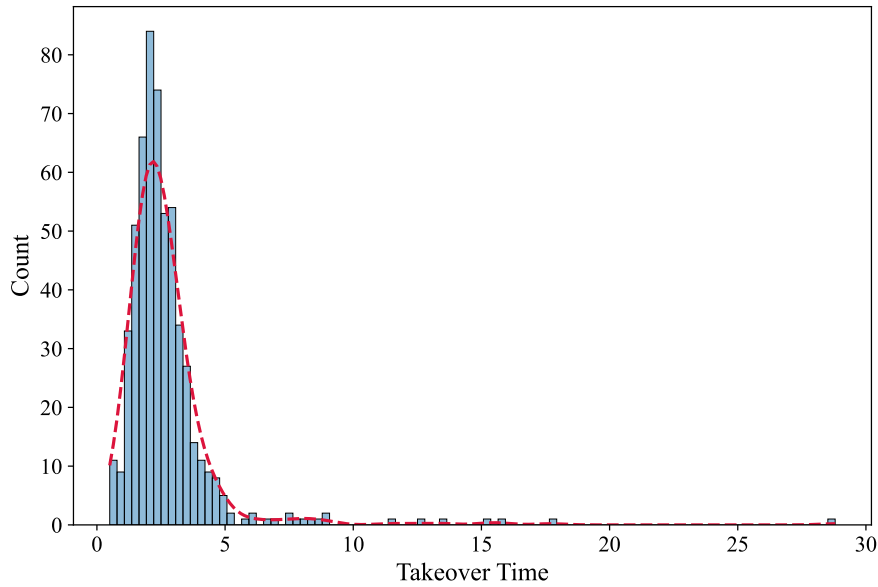
Driver Attributes 1	Age Gender YearsDriving ExperienceADS	Neuroticism Extraversion Openness Agreeableness Conscientiousness
	Personality	
Driver Attributes 2	SuRT performance HandInUse	avgSpd avgCorrecRate
	EoR	
Driver Attributes 3	Proportion of AOIs	CenterScreen_xs LeftSide_xs RightSide_xs Others_xs
		EyelidOpeningMean_xs EyelidOpeningStd_xs EyelidOpeningAmp_xs
	Eyelid Opening	PupilDiameterMean_xs PupilDiameterStd_xs PupilDiameterAmp_xs
	Pupil Diameter	
	GazeEntropy_xs NumFixations_xs	

**Table 4.6:** Drivers' ratings on scenario attributes.

	Mental	Physical	Temporal	Predictability	Criticality	Effort	Frustration
1	10.852	8.741	10.926	10.185	12.333	10.741	9.667
2	11.370	8.222	11.037	10.704	11.037	11.370	7.481
3	11.852	9.815	12.222	9.963	13.704	11.444	8.815
4	12.037	8.889	12.296	11.630	13.259	11.556	8.370
5	11.926	9.333	11.556	11.667	13.444	12.296	9.222
6	11.333	9.963	12.370	9.815	13.704	11.889	9.000
7	9.259	7.296	8.519	8.185	9.333	9.037	6.407
8	11.889	9.630	9.963	12.741	12.593	11.296	7.519
9	9.222	7.815	9.185	8.259	10.148	10.593	6.963
10	6.296	4.481	4.704	6.185	6.333	7.519	6.074
11	7.074	6.259	5.519	7.148	7.333	9.407	6.185
12	8.074	6.593	8.444	8.519	9.000	9.111	6.111

XGBoost (eXtreme Gradient Boosting), which is an algorithm that is recommended for classification and regression problems with tabular data. The results are further explained using SHAP (SHapley Additive exPlanations). Combined with XGBoost, they can provide both good performance and explainability of





**Figure 4.1:** Histogram of the takeover time.

**Table 4.7:** Descriptive statistics of takeover time.

count	565.0	25% percentile	1.817
mean	2.690	50% percentile	2.333
std	2.022	75% percentile	3.000
min	0.500	max	28.80

the prediction. As comparisons, results of linear regression, k-nearest neighbors, support vector machine (with different kernels), decision tree, and random forest will also be provided.

### 4.2.2 Basics of XGBoost and SHAP

The XGBoost model is a learning algorithm composed of an ensemble of decision trees. It is especially useful for dealing with tabular data. In some circumstances, it is even better than deep-learning-based methods [142]. Besides, this method is able to handle missing values inherently and build trees using parallel computing, making it very efficient in dealing with large datasets.

#### Objective Function

For a given dataset with  $n$  examples and  $m$  features  $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}$  ( $|\mathcal{D}| = n, \mathbf{x}_i \in \mathbb{R}^m, y_i \in \mathbb{R}$ ), it uses  $K$  additive functions to predict the output.

$$\hat{y}_i = \phi(\mathbf{x}_i) = \sum_{k=1}^K f_k(\mathbf{x}_i), \quad f_k \in \mathcal{F} \quad (4.1)$$

where  $\mathcal{F} = \{f(\mathbf{x}) = w_{q(\mathbf{x})}\} (q : \mathbb{R}^m \rightarrow T, w \in \mathbb{R}^T)$  is the space of regression trees (CART).  $q$  represents the structure of each tree that maps an example to the corresponding leaf index.  $T$  is the number of leaves in the tree. Each  $f_k$  corresponds to an independent tree structure  $q$  and leaf weights  $w$ .

For a given example, we will use the decision rules in the trees (given by  $q$ ) to classify it into the leaves and calculate the final prediction by summing up the score in the corresponding leaves (given by  $w$ ). To learn the set of functions used in the model, we minimize the following regularized objective.

$$L(\phi) = \sum_i l(y_i, \hat{y}_i) + \sum_k \Omega(f_k), \quad \Omega(k) = \gamma T + \frac{1}{2} \lambda \|w\|^2 \quad (4.2)$$

### Gradient Tree Boosting

The model is trained in an additive manner. Formally, let  $\hat{y}_i^{(t)}$  be the prediction of the  $i$ th instance at the  $t$ th iteration, we will need to add  $f_t$  to minimize the objective:

$$L^{(t)} = \sum_{i=1}^n l(y_i, \hat{y}_i^{(t-1)} + f_t(\mathbf{x}_i)) + \Omega(f_t) \quad (4.3)$$

This means that we greedily add the  $f_t$  that most improves the model. Second order approximation can be used to quickly optimize the objective:

$$L^{(t)} \approx \sum_{i=1}^n \left[ l(y_i, \hat{y}_i^{(t-1)}) + g_i f_t(\mathbf{x}_i) + \frac{1}{2} h_i f_t^2(\mathbf{x}_i) \right] + \Omega(f_t) \quad (4.4)$$

where  $g_i = \partial_{\hat{y}_i^{(t-1)}} l(y_i, \hat{y}_i^{(t-1)})$  and  $h_i = \partial_{\hat{y}_i^{(t-1)}}^2 l(y_i, \hat{y}_i^{(t-1)})$  are first and second order gradient statistics on the loss function. Remove the constant terms yields the simplified objective:

$$\tilde{L}^{(t)} = \sum_{i=1}^n \left[ g_i f_t(\mathbf{x}_i) + \frac{1}{2} h_i f_t^2(\mathbf{x}_i) \right] + \Omega(f_t) \quad (4.5)$$

Define  $I_j = \{i | q(\mathbf{x}_i) = j\}$  as the instance set of leaf  $j$  (For all  $\{\mathbf{x}_i\}_{i=1}^n$  belongs to leaf  $j$ , the subscripts of  $\mathbf{x}_i$  are gathered into  $I_j$ ). We can further rewrite the objective as

$$\tilde{L}^{(t)} = \sum_{i=1}^n \left[ g_i f_t(\mathbf{x}_i) + \frac{1}{2} h_i f_t^2(\mathbf{x}_i) \right] + \gamma T + \frac{1}{2} \lambda \|w\|^2 \quad (4.6)$$

$$= \sum_{j=1}^T \left[ \left( \sum_{i \in I_j} g_i \right) w_j + \frac{1}{2} \left( \sum_{i \in I_j} h_i + \lambda \right) w_j^2 \right] + \gamma T \quad (4.7)$$

For a fixed structure  $q(\mathbf{x})$ , we can compute the optimal weight  $w_j^*$  of leaf  $j$  by

$$w_j^* = \frac{\sum_{i \in I_j} g_i}{\sum_{i \in I_j} h_i + \lambda} \quad (4.8)$$

and calculate the corresponding optimal value by

$$\tilde{L}^{(t)}(q) = -\frac{1}{2} \sum_{j=1}^T \frac{\left(\sum_{i \in I_j} g_i\right)^2}{\sum_{i \in I_j} h_i + \lambda} + \gamma T \quad (4.9)$$

This equation can be used as a scoring function to measure the quality of a tree structure  $q$ . This score is like the impurity score for evaluating decision trees, except that it is derived for a wider range of objective functions.

Usually it is impossible to enumerate all the possible tree structures  $q$ . A greedy algorithm that starts from a single leaf and iteratively adds branches to the tree is used instead. Assume that  $I_L$  and  $I_R$  are the instance sets of left and right nodes after the split. Let  $I = I_L \cup I_R$ , then the loss reduction after the split is given by

$$L_{split} = \frac{1}{2} \left[ \frac{(\sum_{i \in I_L} g_i)^2}{\sum_{i \in I_L} h_i + \lambda} + \frac{(\sum_{i \in I_R} g_i)^2}{\sum_{i \in I_R} h_i + \lambda} - \frac{(\sum_{i \in I} g_i)^2}{\sum_{i \in I} h_i + \lambda} \right] - \gamma \quad (4.10)$$

This formula is used to evaluate the split candidates.

### Shapley Value

SHAP values are a way to explain how each feature impacts the model's prediction. This method is based on game theory and assigns an importance value to each feature in a model. The sign of the SHAP values indicates whether there is a positive or negative impact, and the magnitude indicates how strong the effect is. It is defined using Eq. (4.11),

$$\phi_i(val) = \sum_{S \subseteq \{1, \dots, p\} \setminus \{j\}} \frac{|S|!(p - |S| - 1)!}{p!} (val(S \cup \{j\}) - val(S)) \quad (4.11)$$

where  $S$  is the subset of the features in the model and  $p$  is the number of features.  $val(S)$  is calculated by marginalizing over features that are not included in  $S$  using Eq. (4.12),

$$val(S) = \int \hat{f}(x_1, \dots, x_p) dP_{x \notin S} - E_X(\hat{f}(X)) \quad (4.12)$$

Similarly, the interaction effect is defined using Eq. (4.13),

$$\begin{aligned} \phi_{i,j}(val) = & \sum_{S \subseteq \{1, \dots, p\} \setminus \{i,j\}} \frac{|S|!(p - |S| - 2)!}{p!} \\ & \times (val(S \cup \{i, j\}) - val(S \cup \{i\}) - val(S \cup \{j\}) + val(S)) \end{aligned} \quad (4.13)$$

### 4.2.3 Results and Discussion

#### Training Parameters and Metrics

In this research, 80% and 20% of the data were used for training and testing. The XGBoost regressor was trained with 10-fold cross validation strategy and repeated for 5 times. The parameters were optimized using the grid search strategy, and the optimized parameters included `learning_rate`, `max_depth`, `subsample`, and `n_estimators`.

To evaluate performance of the models, three metrics were utilized, including mean absolute error (MAE, Eq. (4.14)), root mean squared error (RMSE, Eq. (4.15)), and  $R^2$  score ( $R^2$ , Eq. (4.16)).

1. **Mean Absolute Error:** MAE corresponds to the expected value of the absolute error loss or  $l_1$ -norm loss. Suppose  $\hat{y}_i$  and  $y_i$  are the predicted value and true value of the  $i$ th sample, then MAE estimated over  $n$  samples is

$$MAE(y, \hat{y}) = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (4.14)$$

2. **Root Mean Squared Error:** RMSE corresponds to the expected value of the squared quadratic error or  $l_2$ -norm loss. Suppose  $\hat{y}_i$  and  $y_i$  are the predicted value and true value of the  $i$ th sample, then RMSE estimated over  $n$  samples is

$$RMSE(y, \hat{y}) = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (4.15)$$

3.  **$R^2$  Score:**  $R^2$  score represents the proportion of variance that can be explained by the independent variables in the model. Suppose  $\hat{y}_i$  and  $y_i$  are the predicted value and true value of the  $i$ th sample, then  $R^2$  estimated over  $n$  samples is

$$R^2(y, \hat{y}) = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (4.16)$$

### XGBoost Regressor Performance

The model was trained using Python 3.7 and xgboost 1.5.3. A 10-fold cross-validation was run 5 times to yield the final results, which was summarized in Table 4.8. In the table, the subscript *\_R* represents “regressor”, the subscripts *\_xx* represent *xx* s before the TOR, the subscript *\_MC* represents datasets with only medium critical and high critical scenarios (TOR1–TOR9), and the subscript *\_HC* represents datasets with only high critical scenarios (TOR1–TOR6).

**Table 4.8:** XGBoost regressor performance given different scenarios and window size.

	MAE	RMSE	$R^2$
xgb_R_3	0.648	0.964	0.473
xgb_R_6	0.702	1.121	0.288
xgb_R_9	0.696	1.063	0.360
xgb_R_12	0.683	0.997	0.437
xgb_R_15	0.666	1.010	0.422
xgb_R_3_MC	0.489	0.612	0.502
xgb_R_6_MC	0.481	0.622	0.486
xgb_R_9_MC	0.485	0.628	0.475
xgb_R_12_MC	0.493	0.636	0.461
xgb_R_15_MC	0.486	0.625	0.480
xgb_R_3_HC	0.398	0.490	0.395
xgb_R_6_HC	0.403	0.490	0.394
xgb_R_9_HC	0.418	0.504	0.359
xgb_R_12_HC	0.415	0.491	0.392
xgb_R_15_HC	0.407	0.489	0.397

From the results, we can observe that the regressor performs better when we concentrate only on medium-critical and high-critical scenarios, and the MAE decreases from 0.648 s to 0.481 s when we exclude low-critical scenarios from the dataset, with RMSE decreases from 0.964 s to 0.612 s. That probably because most of the long takeover times were recorded in low-critical scenarios, where there is no risk of crash even when drivers did not takeover at all. Furthermore, when we consider only the high-critical scenarios, the MAE further decreases to 0.398 s, with RMSE decreases to 0.490 s. However, compared with dataset with both medium- and high-critical scenarios, dataset with only high-critical scenarios results in lower  $R^2$  score, meaning that less variance could be explained by the model.

Moreover, we can also observe that as the time window get longer, the performance of the regressor is also compromises, with the best performance recorded when the length of the time window is 3 s or 6 s. Specifically, for the

full dataset, the best performance is recorded when the time window is 3 s, with the MAE, RMSE and  $R^2$  being 0.648 s, 0.964 s, and 0.473 respectively; for the dataset excluding low-critical scenarios, the best MAE is recorded when the time window is 6 s, whereas the best RMSE and  $R^2$  are recorded when the time window is 3 s, with the MAE, RMSE and  $R^2$  being 0.481 s, 0.612 s and 0.502 respectively; and for dataset with only high-critical scenarios, the best MAE is recorded when the time window is 3 s, whereas the best RMSE and  $R^2$  are recorded when the time window is 15 s, although they are not much different from that when the time window is 3 s.

Considering the three metrics, MAE and RMSE show us how precise is the regressor, and  $R^2$  shows us how predictive is it. Hence, in considering performance of the model, we would like a model with smaller MAE and RMSE whereas larger  $R^2$ . Hence, although dataset with only high-critical scenario yields the highest precision, its  $R^2$  is not good enough. Therefore, in the latter comparison with other models and analysis, we will consider window size of 3 s and dataset with both medium- and high-critical scenarios, and the results are summarized in Table 4.9.

Before training, categorical variables were turned into one-hot vectors, for example, for categorical variables with values 1, 2, and 3, they were turned into [1, 0, 0], [0, 1, 0] and [0, 0, 1], respectively. And for numerical variables were scaled using min-max scaler to avoid bias toward large values in the variable. For training k-nearest neighbors, support vector machine, decision tree and random forest, again, we adopted the grid search method, and the optimized parameters were respectively:

- k-Nearest Neighbors: `n_neighbors` (3, 4, 5, 6, 7, 8, 9), `weights` (uniform, distance)
- Support Vector Machine: `kernel` (linear, polynomial, rbf), `degree` (2, 3, 4, 5), `C` (1, 10, 100)
- Decision Tree: `criterion` (absolute\_error, squared\_error), `splitter` (random, best), `max_depth` (3, 4, 5, 6)
- Random Forest: `criterion` (absolute\_error, squared\_error), `n_estimators` (100, 200, 300, 400), `max_depth` (3, 4, 5, 6)

And the best parameters were summarized in Table 4.10.

We can see that linear regression models perform the worst in any cases, with linear SVM a little better than linear regression model, suggesting the non-linearity of the problem. Polynomial SVM performs even better than XGBoost

when it comes to MAE, however, this is at the cost of larger RMSE and smaller  $R^2$ , although the differences are not so significant compared with other models. In total, we can say that polynomial SVM performs almost as good as XGBoost, with XGBoost a bit better than polynomial SVM. Besides, KNN performs as good as RBF SVM when it comes to MAE, whereas RBF SVM performs better than with respect to the other two metrics. Finally, tree-based method like decision tree and random forest are not so good as XGBoost in all aspects.

**Table 4.9:** XGBoost regressor vs. baseline regression models.

	MAE	RMSE	$R^2$
XGBoost	0.489	<b>0.612</b>	<b>0.502</b>
Linear Regression	0.532	0.659	0.422
K-Nearest Neighbor	0.524	0.682	0.382
Support Vector Machine (Linear)	0.533	0.651	0.436
Support Vector Machine (Polynomial)	<b>0.487</b>	0.615	0.497
Support Vector Machine (RBF)	0.509	0.638	0.458
Decision Tree	0.561	0.765	0.222
Random Forest	0.525	0.680	0.384

**Table 4.10:** Best parameters for regression models.

Model	Best Parameters
XGBoost	learning_rate: 0.01, max_depth: 3, n_estimators: 440, subsample: 0.7
K-Nearest Neighbor	n_neighbors: 9 weights: distance
Support Vector Machine (Linear)	C: 1
Support Vector Machine (Polynomial)	degree: 2 C: 1
Support Vector Machine (RBF)	C: 1
Decision Tree	criterion: squared_error, splitter: best, max_depth: 3
Random Forest	criterion: absolute_error, n_estimators: 400, max_depth: 6

## SHAP Explanation–Global

**Feature Importance** SHAP summary bar plot (Fig. 4.2) shows the global importance of each feature, which is taken to be the mean absolute value for that feature over all the given samples. In comparison, SHAP summary bee-swarm

plot (Fig. 4.3) shows the global importance of each feature and the distribution of effect sizes. Each dot represents a SHAP value for its instance of one feature. The  $x$  position represents the SHAP value of that feature, indicating small or large effect of the feature on the prediction from left to right; color is used to display the original value of one feature, with red representing high value of that feature; dots pile up to show the density (distribution) of that feature in the dataset; and the  $y$  position represents the importance of the feature.

We can see that among all the 38 features, CenterScreen is the most important feature in predicting takeover time, followed by avgCorrectionRate, Physical\_Demand, Temporal\_Demand, Neuroticism, Scenario, GazeEntropy, Effort, DoM, Conscientiousness, Extraversion, etc. Specifically, a high level of percentages of EoR was found to reduce the takeover time by about 0.3 s, and a low level of EoR increased the takeover time by 0.4–0.5 s. A lower average correction rate of drivers performing the SuRT reduced the takeover time by about 0.2 s, which indicated a lower level of concentration on the NDRT compared with the driving task. Lower physical demand and higher temporal demand were also found to reduce takeover time by about 0.4 s and 0.2–0.3 s respectively, suggesting that drivers were quicker at taking over the vehicle when the scenarios were relatively simple and time-pressured. Higher neuroticism was also found to increase the takeover time by almost 1.2 s, however, the this effects may not be statistically significant considering that the effects were found in only a few samples. Overall, it seems that drivers' gaze behaviors, personality traits, and scenario characteristics were more important in deciding takeover over time compared with other factors. Curiously, effect of time budget was not found to be very significant, perhaps because of the small difference between the two time budgets in this experiment. Moreover, we can also observe that more variations exist in the more important features (longer tails) than the less important ones, indicating the interaction effects with other features.

**Main Effects of a Certain Feature** To further understand the effects of the features on takeover time, we can explore the main and interaction effects of the most important features using dependence scatter plots. In the main effects plots (Fig. 4.4), each dot is a single prediction from the dataset; the  $x$  axis is the value of the feature; the  $y$  axis is the SHAP value for the feature, representing how much knowing that feature's value changes the output of the model for that sample's prediction; and the light gray area at the bottom of the plot is a histogram showing the distribution of data values. In the main effect plot for CenterScreen, we can see that when the percentage of EoR is lower than 20%, takeover time



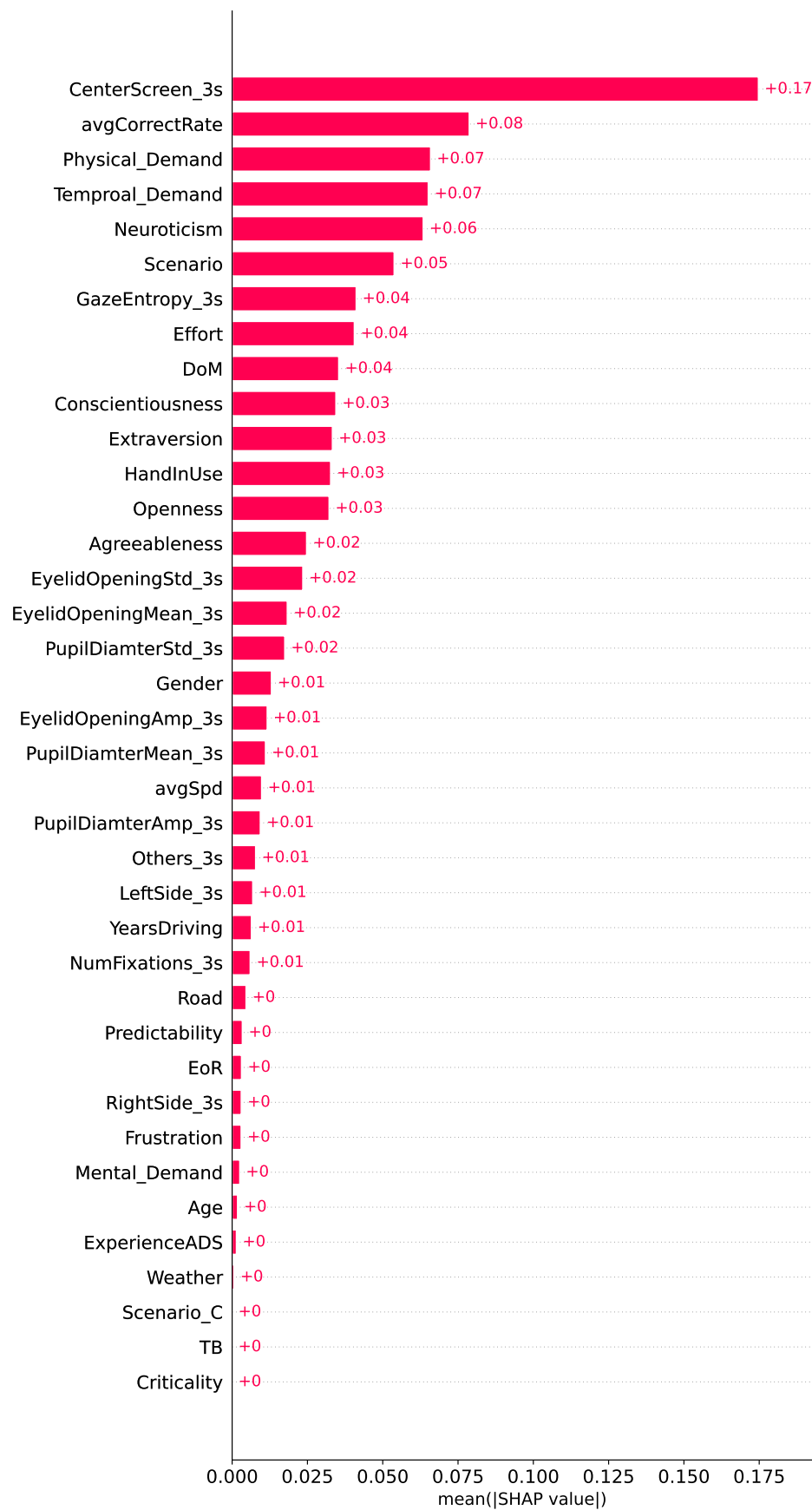


Figure 4.2: SHAP summary bar plot for takeover time.

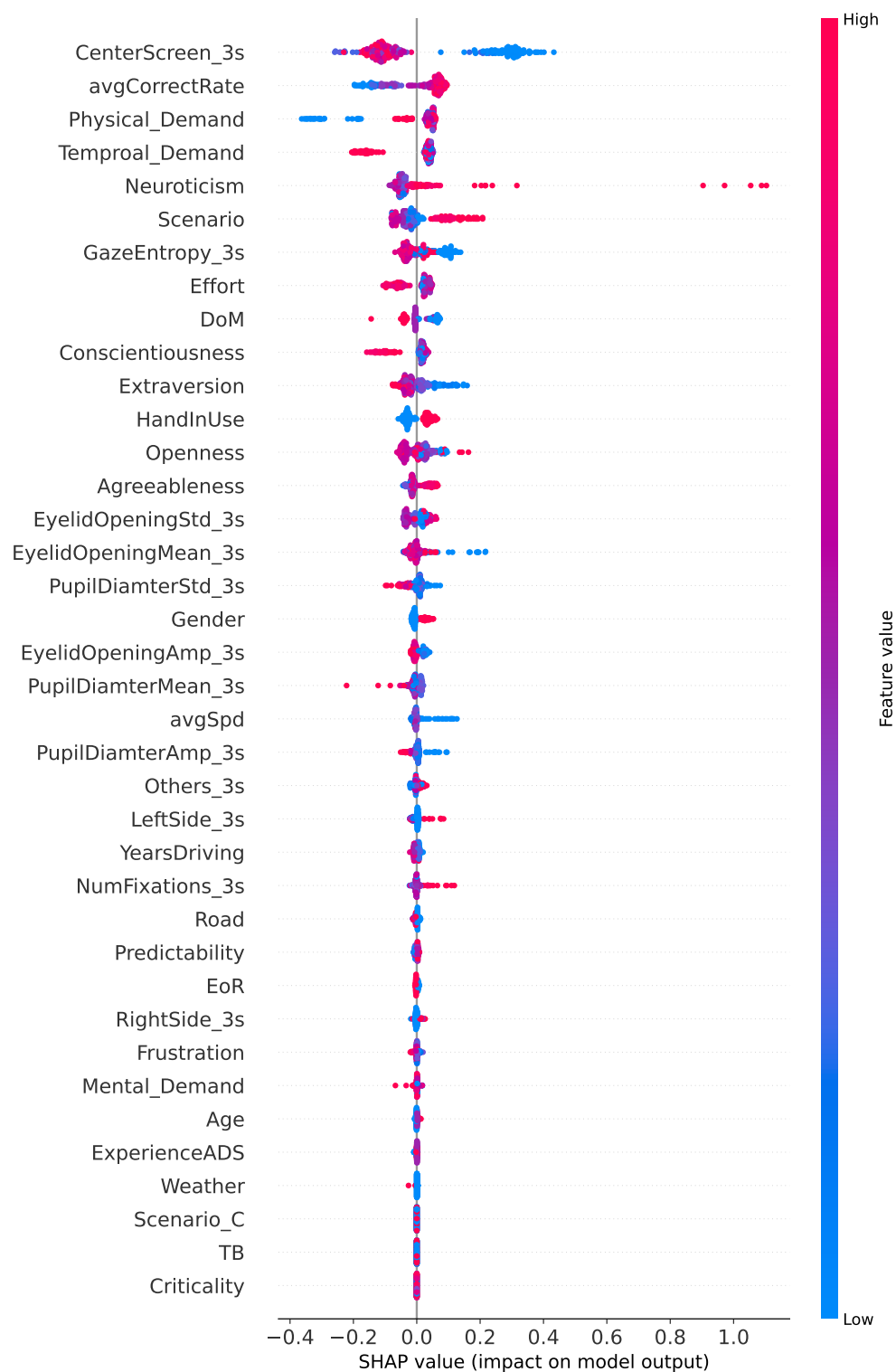
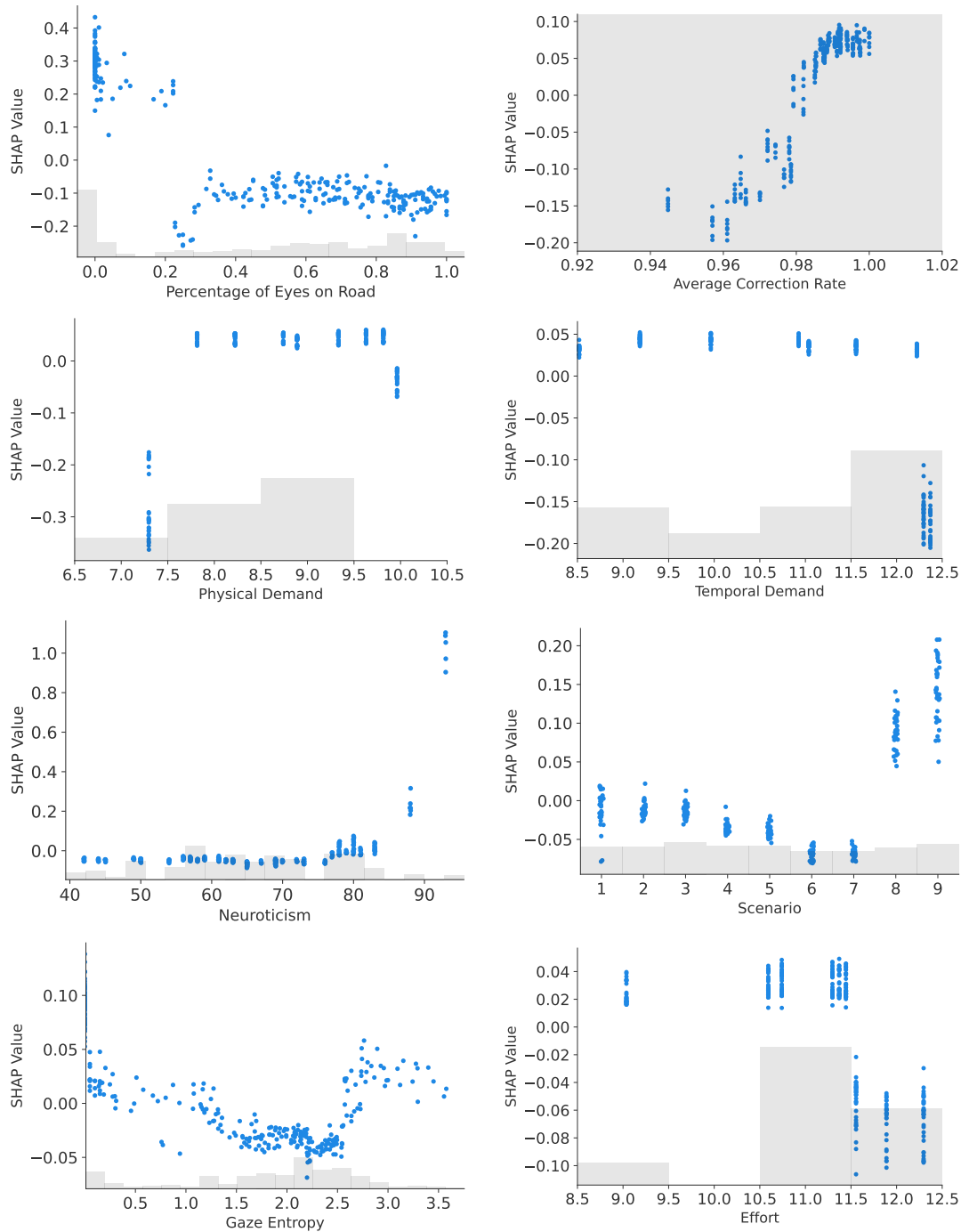


Figure 4.3: SHAP summary beeswarm plot for takeover time.



**Figure 4.4:** SHAP main effects scatter plots for takeover time.

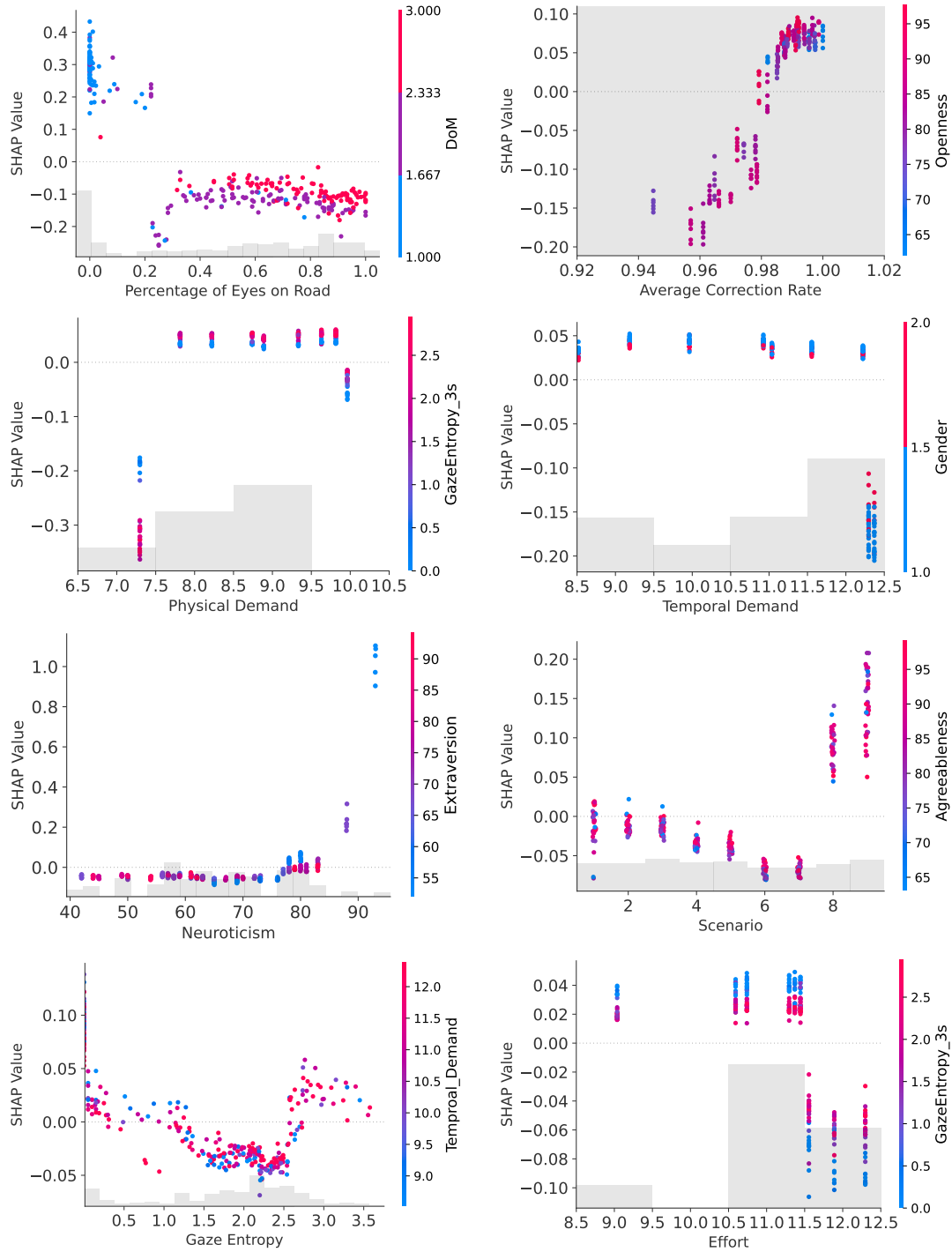
is increased by over 0.2 s. And above this threshold, takeover time started to decrease, which varies between 0.1 s and 0.2 s. For average correction rates, a linear relationship could be observed, and an increase of 1% of the average correction rate could approximately increase the takeover time by 0.06 s. For gaze entropy, a decrease of takeover time could be observed until the gaze entropy is over about 2.25, then it starts to increase until saturated after about 2.65. This suggests that to improve drivers' performance on takeover time, eye gazes of

drivers should neither be too concentrated (small gaze entropy) or too spread out (large gaze entropy), which will be further discussed in Section 5.2. As for physical demand, mental demand, neuroticism, and effort, similar to percentage of EoR, a certain threshold could be observed that decides whether the takeover time is increased or decreased. Finally, for takeover scenarios, takeover time in scenarios TOR8 and TOR9 are shown to be longer than the other scenarios, and the quickest takeover could be observed in TOR6 and TOR7.

**Interaction Effects Between Two Features** The vertical dispersion in Fig. 4.4 shows that the same value of a certain feature can have a rather different impact on the model's output, suggesting that there exists non-linear interaction effects between different features. To show which feature may be driving these interaction effects, we can color the main effects plot by another feature. If an interaction effect is present between two features, it will show up as a distinct vertical pattern between the two features. The results are plotted in Fig. 4.5, the left  $y$ -axis and the  $x$ -axis are the same as that in Fig. 4.4, and the right  $y$ -axis represents the second feature that has interaction effects with the first feature on the  $x$ -axis. Take the percentage of EoR as an example, we can see that short duration of monitoring is highly related with lower percentage of EoR. For those with percentage of EoR greater than 20%, medium duration is more likely to lead to quicker takeovers compared with long duration of monitoring, which coincides with the conclusions in Section 3.3. For physical demand, effect of gaze entropy seems to be different according to physical demand of the scenarios. Specifically, when physical demand is low, smaller gaze entropy leads to longer takeover time, whereas when physical demand is high, smaller gaze entropy leads to shorter takeover time. Similar phenomenon can be observed in the plots of temporal demand, gaze entropy, effort, etc. This suggests that requirements on drivers' behaviors vary according to urgency, complexity and criticality of scenarios.

### SHAP Explanation–Local

In Figs. 4.2 and 4.3, we can see the importance of all the features in predicting takeover time, however, this is not necessarily true for individual samples. To display explanations for individual predictions, we can utilize so-called waterfall plots. The bottom of a waterfall plot starts as the expected value of the model output, and then each row shows how each feature contributes to the prediction. The red and blue represent the positive and negative contribution of each feature to the prediction. Figs. 4.6, 4.7, 4.8 are three examples, showing how each feature



**Figure 4.5:** SHAP interaction effects scatter plots for takeover time.

affects drivers' takeover time differently for short, medium and long predicted takeover time, respectively.

In the first example (Fig.4.6), we can see that almost all the features have a negative impact on takeover time. Specifically, having a 23.9% of percentage of EoR reduces the takeover time by 0.23 s, a lower average correction rate reduces the takeover time by 0.17 s, and a high conscientiousness score also

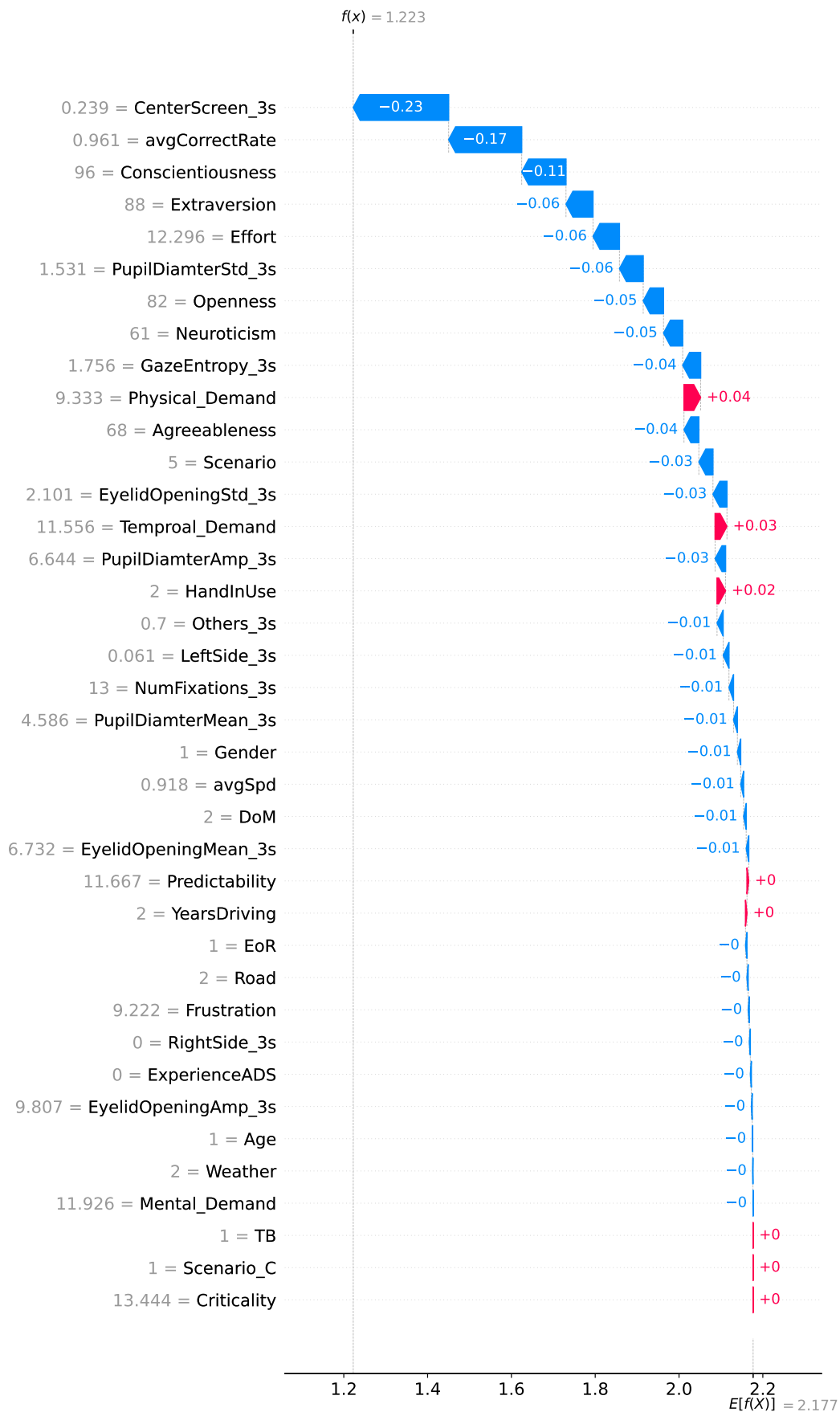


Figure 4.6: SHAP individual explanation for takeover time prediction 1.223 s.

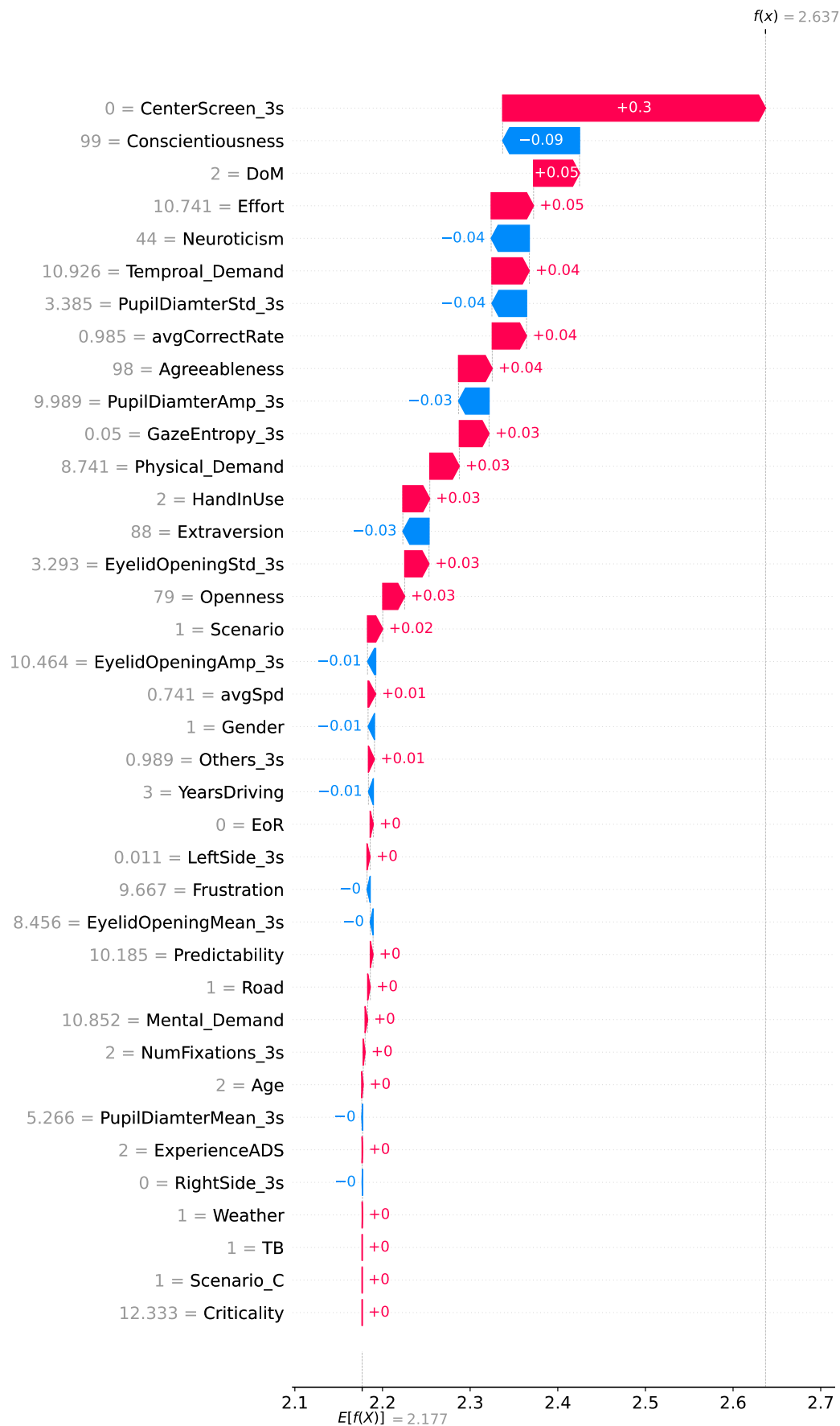
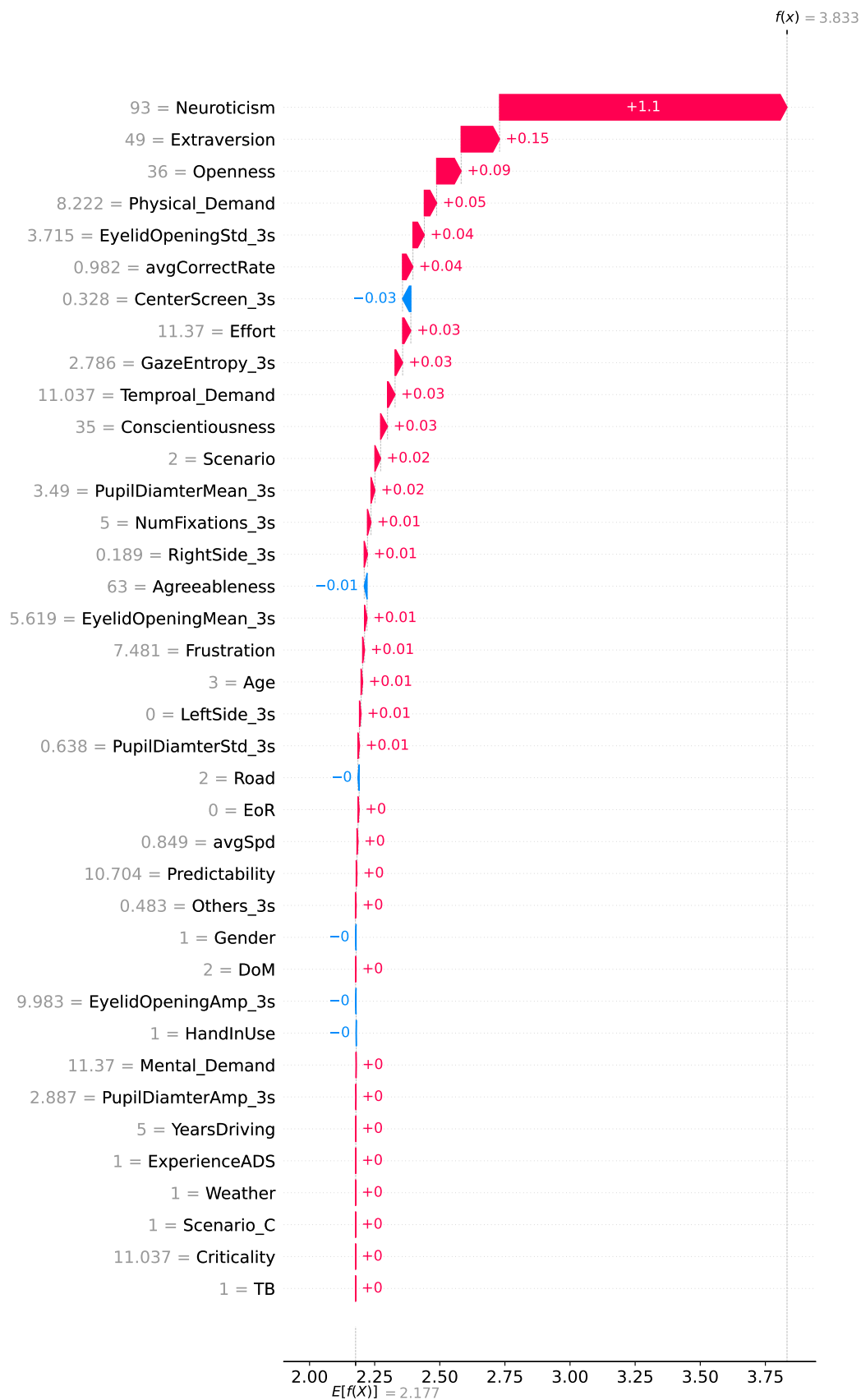


Figure 4.7: SHAP individual explanation for takeover time prediction 2.637 s.



**Figure 4.8:** SHAP individual explanation for takeover time prediction 3.833 s.



reduces the takeover time by 0.11 s. In the second example (Fig. 4.7), although a high conscientiousness score reduces the takeover time by 0.09 s, a very low percentage of EoR leads to an increase of takeover time by 0.3 s. And in the third example (Fig. 4.8), a high neuroticism seems to dominates in predicting the driver's takeover time, leading to an increase of 1.1 s. Hence, although CenterScreen is found to be most important in predicting drivers' takeover time overall. For individual cases, the dominant features may be quite different, depending largely on drivers' preferences and scenario characteristics.

#### 4.2.4 Summary

Considering that takeover time is a continuous variable, the objective of this section is to model takeover time as a regression problem based on 38 independent variables as summarized in Table 4.1.2. Results show that:

- Among the list of models, XGBoost regressor is best at predicting takeover time, with the minimal MAE, minimal RMSE and the maximal  $R^2$ .
- SHAP explainer is utilized to understand the effects of each feature on takeover time by providing global and local explanations. Global explanation provides an overview of the importance ranking of the variables. Overall, it seems that drivers' gaze behaviors, personality traits, and scenario characteristics were more important in deciding takeover over time.
- Main effects and interaction effects plots (Figs. 4.4 and 4.5) gave us a more depth insight on how each of the features affects takeover time. It seems that most of the effects are highly nonlinear, which explains why linear regression models are not well-performed compared with non-linear regression models, especially tree-based models.
- Local explanation is also explored for individual explanations. It works by comparing the predicted value to an expected value as a reference, showing how each feature contributes to the prediction by pushing the results close to or further from the expected value.

### 4.3 Modeling Takeover Readiness

#### 4.3.1 Problem Statement and Objectives

Compared with takeover time, a uniform definition of takeover readiness seems to be lacking in the literature. In [97], driver readiness was labeled according to

whether a participant initiated the takeover action by pressing the two buttons mounted on the steering wheel upon receiving a TOR. In [143], driver readiness was labeled according to whether the vehicle was in automated or manual mode during the experiments. And in [141], driver readiness was classified into five different takeover maneuvers. However, even if drivers have initiated the takeover actions, it does not necessarily mean that drivers are with high takeover readiness.

Therefore, in this research, we have adopted a different definition regarding driver readiness, and it was labeled according to whether there was a crash or near-crash after the TOR. In this way, drivers did not take over or take over too late are all classified into low takeover readiness. This is realized by playing back all the 576 videos recorded during the experiment (the process can be referenced to Section 5.2). Finally, 56 and 520 samples are labeled as low and high takeover readiness respectively.

Since takeover readiness is a categorical variable, the objective of this section is to model takeover readiness as a classification problem based on the independent variables summarized in Table 4.1.2. Same as that in Section 4.2, the model is developed based on XGBoost. Similarly, the results are further explained using SHAP (SHapley Additive exPlanations). As comparisons, results of logistic regression, k-nearest neighbors, support vector machine (with different kernels), decision tree, and random forest will also be provided.

### 4.3.2 Oversampling and Undersampling Technique

#### Purpose of Oversampling and Undersampling

Since the dataset obtained is very unbalanced, with 520 and 56 samples in the positive and negative classes respectively, training the classifier would be very challenging, especially when we are interested in the classifier's performance on the minority class (crashes/near crashes). One approach to address such kind of problem is to oversample the minority class. The simplest approach is to duplicate existing samples in the minority class, however, this examples will not add new information to the model. Instead, we can synthesize new samples from the existing examples in the dataset. This technique is called synthetic minority oversampling technique (SMOTE) [144]. Besides oversampling, it is suggested in [144] that it is better to combine oversampling the minority with undersampling the majority. Therefore, in this section, we will try different ratios of oversampling and undersampling, and compare their performance on the classification task.

### Synthetic Minority Oversampling Technique

SMOTE works by selecting samples that are close in the feature space, drawing a line between the examples in the feature space and drawing a new sample at a point along the line. At a high level, the algorithm can be described as:

1. A random sample from the minority class is first chosen.
2.  $k$  of the nearest neighbors of the sample are found (default 5 in [144]).
3. Take difference between the sample and one of its neighbor.
4. Multiply the difference by a random number between 0 and 1.
5. Add this difference to the sample to generate a new synthetic sample in feature space.
6. Continue with the next neighbor until the number is enough.

More detailed discussions can be referenced to [144], as a reference, the pseudo code is summarized in Algorithm 1 (Modified from [144]). In implementing the algorithm, we have used the library `imbalance-learn`, where SMOTE can be realized in a few lines of codes.

### Random Undersampling Technique

Random undersampling involves randomly selecting examples from the majority class to delete from the training dataset. Obviously, a modest amount of oversampling can be applied to the minority class to improve the bias towards the minority samples, whilst a modest amount of undersampling can also be applied to the majority class to improve the bias toward the majority samples. Theoretically, This can results in improved performance compared to performing only one of the techniques. However, whether this will improve the performance of the model or not depends on the property of the dataset.

## 4.3.3 Results and Discussion

### Training Parameters and Metrics

Similarly, 80% and 20% of the data were used for training and testing respectively. And the training parameters of XGBoost classifier are the same as that in XGBoost regressor, with 10-fold cross validation strategy adopted in the training process. However, the evaluation metrics are totally different. In evaluating performance of a classifier, 4 metrics were utilized, including accuracy (Eq. (4.17) or (4.21)),

**Algorithm 1:** Synthetic Minority Oversampling Technique.

---

**Input:** number of minority class samples  $T$ ; Amount of SMOTE  $N\%$ ;  
 Number of nearest neighbors  $k$   
**Output:**  $N/100 * T$  synthetic samples

```

1 if  $N < 100$  then
2   Randomize the  $T$  samples
3    $T = (N/100) * T$ 
4 end
5  $N = (\text{int})(N/100)$ 
6  $\text{numAttrs} = \text{Number of attributes}$ 
7  $\text{Sample}[][]$ : array for original minority class
8  $\text{newIndex}$ : keep count of the number of synthetic samples
9  $\text{Synthetic}[][]$ : array for synthetic samples
10 for  $i \leftarrow 1, \dots, T$  do
11   Compute  $k$  nearest neighbors for  $i$ , and save the indices in  $\text{nnArray}$ 
12   while  $N \neq 0$  do
13     Choose a random number between 1 and  $k$ , call it  $\text{nn}$ .
14     for  $\text{attr} \leftarrow 1, \dots, \text{numAttrs}$  do
15        $\text{dif} = \text{Sample}[\text{nnArray}[\text{nn}]][\text{attr}] - \text{Sample}[i][\text{attr}]$ 
16        $\text{gap} = \text{random number between 0 and 1}$ 
17        $\text{Synthetic}[\text{newIndex}][\text{attr}] = \text{Sample}[i][\text{attr}] + \text{gap} * \text{dif}$ 
18     end
19      $\text{newIndex}++$ 
20      $N = N - 1$ 
21   end
22 end

```

---

precision (Eq. (4.18)), recall (Eq. (4.19)) and F1 score (Eq. (4.20)). Besides, confusion matrix will also be reported for the best cases. For binary classifications, confusion matrix is equivalent to truth table.

1. **Accuracy:** Accuracy computes the fraction or the count of correct predictions. Suppose  $\hat{y}_i$  and  $y_i$  are the predicted value and true value of the  $i$ th sample respectively, then accuracy calculated over  $n$  samples is defined as:

$$\text{accuracy}(y, \hat{y}) = \frac{1}{n} \sum_{i=1}^n \mathbf{1}(y_i = \hat{y}_i) \quad (4.17)$$

2. **Precision, Recall and F1 Score:** Precision is defined as the ability of the classifier not to label a sample as positive that is negative, and recall is defined as the ability of the classifier to find all the positive examples. F1 score is defined as a weighted harmonic mean of the precision and recall. Considering the truth table in Table 4.11 (where “positive” and

**Table 4.11:** Truth table for binary classification.

Predicted Class	Actual Class	
	TP (True Positive) FN (False Negative)	FP (False Positive) TN (True Negative)

“negative” refer to the classifier’s prediction), Precision, recall and F1 score can be calculated using Eqs. (4.18), (4.19) and (4.20) respectively. Besides, accuracy can also be calculated using Eq. (4.21).

$$\text{precision} = \frac{TP}{TP + FP} \quad (4.18)$$

$$\text{recall} = \frac{TP}{TP + FN} \quad (4.19)$$

$$\text{F1 score} = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \quad (4.20)$$

$$\text{accuracy} = \frac{TP + TN}{TP + FP + FN + TN} \quad (4.21)$$

### XGBoost Classifier Performance

The model was trained using Python 3.7 and xgboost 1.5.3. Similar to XGBoost regressor, a 10-fold cross-validation was run 5 times to yield the final results, which was summarized in Table 4.12. In the table, the subscript <sub>C</sub> represent “classifier”, the subscripts <sub>xx</sub> represents xx s before the TOR, the subscript <sub>Oxx</sub> represents the ratio of the minority class and the majority class, and the subscript <sub>Uxx</sub> represents the ratio of the majority class to be undersampled.

From the results, we can observe that with the increasing of the ratio of oversampling, the performance of the classifier also increases a lot. As the ratio increases from 50% to 80%, the best accuracy increases from 0.948 to 0.963, whereas the precision decreases from 0.980 to 0.939. The most significant improvement could be observed from recall, which is increased from 0.889 to 0.969. However, as the ratio is further increased to 100%, the improvement seems to not be that significant. The accuracy only increases from 0.963 to 0.967, with no improvement of recall, although precision increased from 0.939 to 0.979. As we further add undersampling to the dataset, contrary to the anticipation, the accuracy and recall decrease from 0.964 and 0.969 to 0.944 and 0.929 respectively, although precision increases a bit from 0.939 to 0.956.

Moreover, we can also observed that effect of window size in the classification problem is not so obvious as that in the regression problem. Still, the best

**Table 4.12:** XGBoost classifier performance given ratios of oversampling and window size.

	Accuracy	Precision	Recall	F1 Score
xgb_C_3_O50	0.926	0.958	0.852	0.902
xgb_C_6_O50	0.948	0.980	0.889	0.932
xgb_C_9_O50	0.933	0.979	0.852	0.911
xgb_C_12_O50	0.941	0.960	0.889	0.923
xgb_C_15_O50	0.933	0.979	0.852	0.911
xgb_C_3_O80	0.963	0.939	0.969	0.954
xgb_C_6_O80	0.951	0.924	0.953	0.938
xgb_C_9_O80	0.944	0.910	0.953	0.931
xgb_C_12_O80	0.944	0.899	0.969	0.932
xgb_C_15_O80	0.951	0.912	0.969	0.939
xgb_C_3_O100	0.967	0.969	0.969	0.969
xgb_C_6_O100	0.967	0.969	0.969	0.969
xgb_C_9_O100	0.967	0.969	0.969	0.969
xgb_C_12_O100	0.967	0.979	0.959	0.969
xgb_C_15_O100	0.961	0.959	0.969	0.964
xgb_C_3_O80_U20	0.938	0.955	0.914	0.934
xgb_C_6_O80_U20	0.938	0.942	0.928	0.935
xgb_C_9_O80_U20	0.917	0.939	0.886	0.912
xgb_C_12_O80_U20	0.944	0.956	0.929	0.942
xgb_C_15_O80_U20	0.944	0.956	0.929	0.942

performance is still obtained when the window size is 3 s or Specifically, for oversampling ratio 50%, the best performance is recorded when the time window is 6 s, with accuracy, precision, recall and F1 score being 0.948, 0.980, 0.889 and 0.932 respectively; for oversampling ratio 80%, the best performance is recorded when the time window is 3 s, with accuracy, precision, recall and F1 score being 0.963, 0.939, 0.969 and 0.954 respectively; for oversampling ratio 100%, the best accuracy and recall are obtained when the time window is 3 s, whereas the best precision is obtained when the time window is 12 s, also the difference is not very significant; and for oversampling followed by undersampling, the best performance is obtained when the time window is 12 s, with accuracy, precision, recall and F1 score being 0.944, 0.956, 0.929 and 0.942 respectively.

In our research, crashes/near-crashes are defined as positive class. Hence, precision measures the classifier's ability to not label a sample as crash that is non-crash, and recall measures the classifier's ability to find all the crashes. Considering this, compared with precision, we are more concerned with recall, since classifying a crash as a non-crash is more dangerous than classifying a non-crash as a crash. In considering performance of the model, we would like a model to have higher accuracy and recall, whereas with modest precision.

Therefore, in the latter comparison with other models and analysis, we will consider window size of 3 s and oversampling ratio of 80%, and the results are summarized in Table 4.13.

Before training, the data were preprocessed similarly as that in regression model, with categorical variables turned into one-hot vectors and numerical variables scaled with min-max scaler. For training k-nearest neighbors, support vector machine, decision tree and random forest, again, we adopted the grid search method. The optimized parameters of KNN and SVM were the same as that in regression models, and the optimized parameters for decision tree and random forest were respectively:

- Decision Tree: criterion (gini, entropy), splitter (random, best), max\_depth (3, 4, 5, 6)
- Random Forest: criterion (gini, entropy), n\_estimators (100, 200, 300, 400), max\_depth (3, 4, 5, 6)

And the best parameters were summarized in Table 4.14.

Similar to regression models, we can see that logistic regression model and linear SVM perform the worst in any cases, with linear SVM better than logistic regression model in any aspects, suggesting the non-linearity of the problem. Overall, tree-based methods (decision tree, random forest, and XGBoost) perform better than the other models, with XGBoost the best in all the aspects, and its confusion matrix is

		Predicted Class	
		0	1
Actual Class	0	94	4
	1	2	62

Curiously, when it comes to recall, KNN, SVM and random forest perform as well as XGBoost, however, with respect to other metrics, their performances are much worse, especially in terms of precision.

### SHAP Explanation–Global

**Feature Importance** We can see that different from the regression model, Predictability is the most important feature in predicting takeover readiness, followed by Frustration, Scenario\_C, Neuroticism, Road, Mental\_Demand, Gender, Criticality, Effort, avgSpd, etc. Note that in Fig. 4.10, the units on the  $x$ -axis are log-odds units, defined by Eq. (4.22), where  $p$  is the probability of crash. Hence,

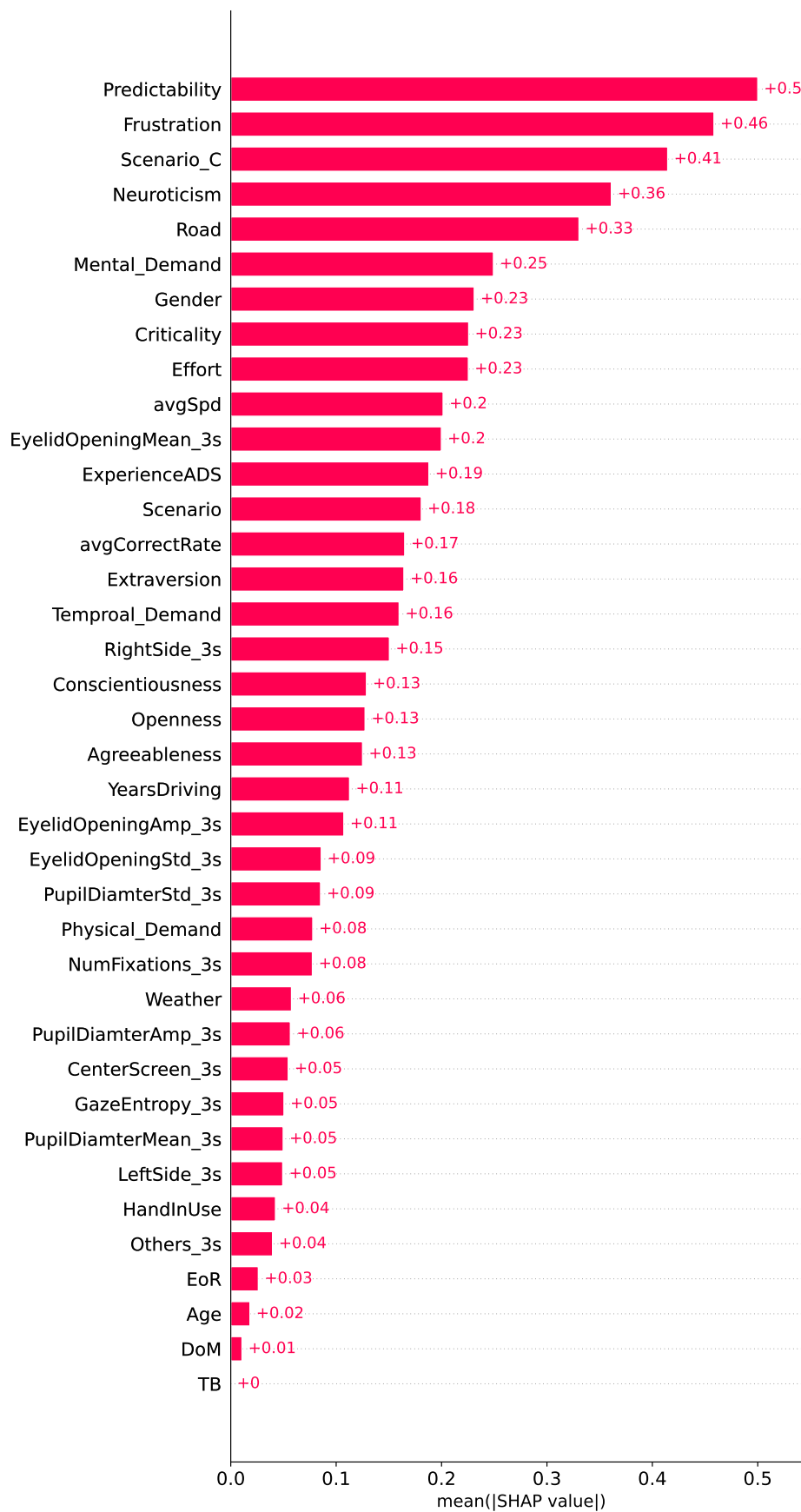
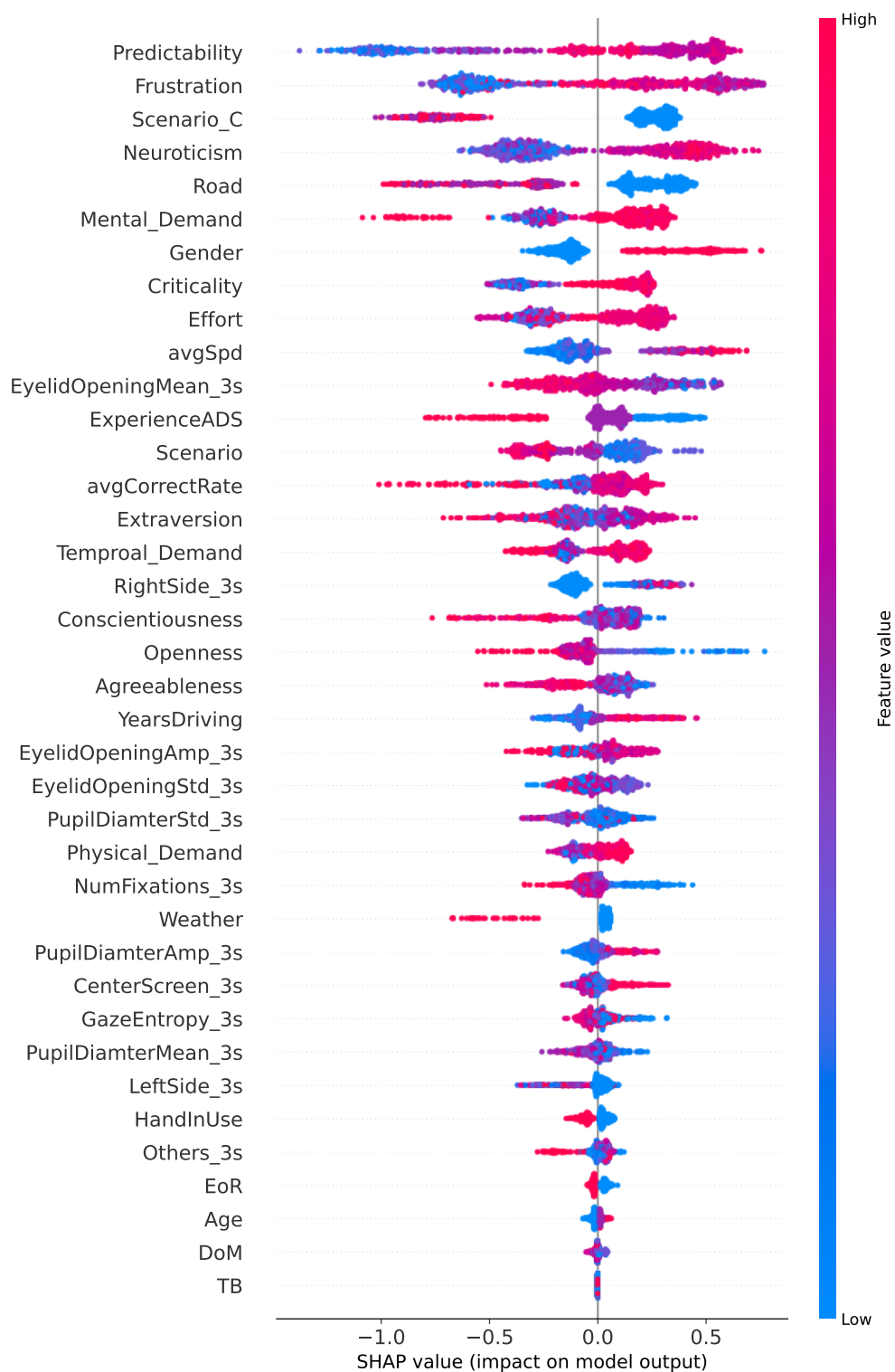


Figure 4.9: SHAP summary bar plot for takeover readiness.





**Figure 4.10:** SHAP summary beeswarm plot for takeover readiness.

**Table 4.13:** XGBoost classifier vs. baseline classification models.

	Accuracy	Precision	Recall	F1 Score
XGBoost	<b>0.963</b>	<b>0.939</b>	<b>0.969</b>	<b>0.954</b>
Logistic Regression	0.833	0.728	0.922	0.814
K-Nearest Neighbor	0.895	0.805	<b>0.969</b>	0.879
Support Vector Machine (Linear)	0.852	0.738	<b>0.969</b>	0.838
Support Vector Machine (Polynomial)	0.889	0.795	<b>0.969</b>	0.873
Support Vector Machine (RBF)	0.889	0.795	<b>0.969</b>	0.873
Decision Tree	0.870	0.795	0.906	0.847
Random Forest	0.901	0.824	0.953	0.884

**Table 4.14:** Best parameters for regression models.

Model	Best Parameters
XGBoost	learning_rate: 0.02, max_depth: 5, n_estimators: 350, subsample: 0.7
K-Nearest Neighbor	n_neighbors: 6 weights: uniform
Support Vector Machine (Linear)	C: 10
Support Vector Machine (Polynomial)	degree: 4 C: 1
Support Vector Machine (RBF)	C: 10
Decision Tree	criterion: gini, splitter: best, max_depth: 6
Random Forest	criterion: gini, n_estimators: 200, max_depth: 6

negative values imply probabilities of less than 0.5 that the driver will cause a crash, and the more negative, the less likely a crash will happen.

$$\text{logit}(p) = \log \left( \frac{p}{1-p} \right) \quad (4.22)$$

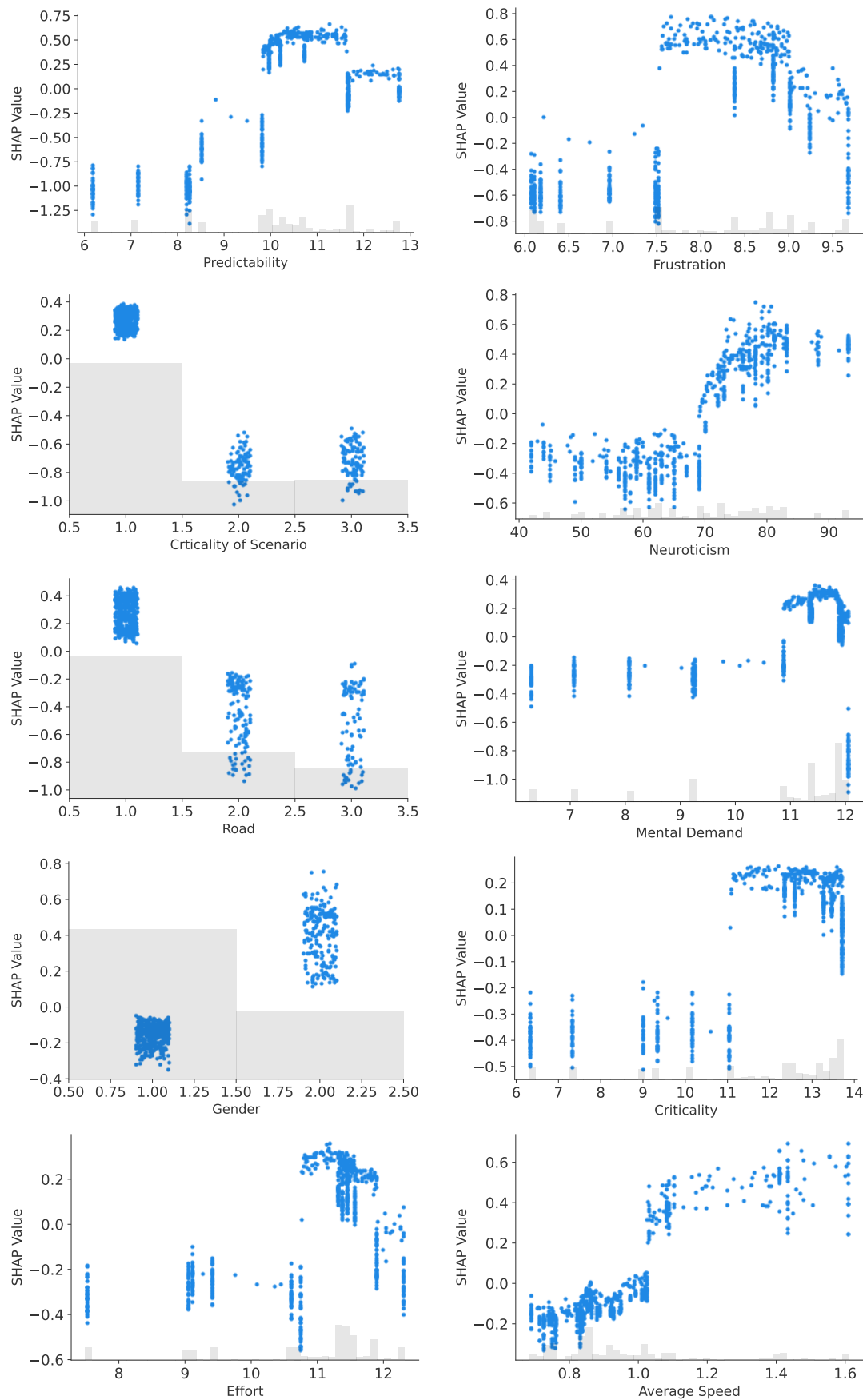
Contrary to the common belief, a higher level of predictability of the scenarios seems to contribute to a higher crash rate. Similarly, a higher level of frustration felt by the driver, a higher level of effort required from the driver, and a higher level of criticality and mental demand of the scenario are also shown to increase the crash rate. Besides, drivers with higher level of neuroticism in their personality traits tend to cause more crash than those with lower levels of neuroticism. And the crash rates of female drivers seem to be higher than male drivers. Finally, slower reaction speed (larger acgSpd) also contributes to higher

crash rates.

Overall, it seems that scenario characteristics (Predictability, Frustration, Scenario\_C, Mental\_Demand, Criticality, Effort), drivers' personality (Neuroticism) and reaction speed (avgSpd), and type of the road (Road) were more important in deciding takeover over time compared with other factors. Curiously, crashes are more often in straight roads compared with curve roads and ramp roads, perhaps because of the poor vision due to blocking of the vehicles ahead.

**Main Effects of a Certain Feature** Similar to the regression model, to further understand the effects of each feature on takeover readiness, we can explore the main and interaction effects of the most important features using dependence scatter plots. For each of the features, some thresholds could be observed from Fig. 4.11. For example, in the main effect plot for Predictability, we can see that when the level of predictability is lower than 8, the log-odds could be reduced by about 0.75–1.25; when the level of predictability is higher than 8 but lower than 10, the log-odds could be reduced by about 0.25–0.75; and when the level of predictability is above 10, the log-odds would be increased by about 0.25–0.50. This suggests that when the scenarios are highly predictable, drivers may be more careless, leading to higher crash rates. In the main effect plot for Neuroticism, we can also see that when drivers' neuroticism is lower than 70, the log-odds could be reduced by about 0.2–0.6, and above this threshold, the log-odds would be increased up to 0.6, suggesting that drivers with higher neuroticism are prone to crashes during emergent situations. Similar conclusions could be obtained from other main effect plots.

**Interaction Effects Between Two Features** The vertical dispersion in Fig. 4.11 shows that the same value of a certain feature can have a rather different impact on the model's output, suggesting that there exists non-linear interaction effects between different features. To show which feature may be driving these interaction effects, we can color the main effects plot by another feature like that in the regression model, and the results are plotted in Fig. 4.12. Take the neuroticism of drivers as an example, we can see that for drivers with lower levels of neuroticism, if they have less years of driving experience, they are more likely to reduce the probability of crashes, whereas for drivers with higher levels of neuroticism, they are more prone to crashes if they have more years of driving experience. For criticality of scenarios, if the criticality is high, crashes are more likely to happen in straight roads, whereas when the criticality is low, crashes are more likely to happen in curve roads. Similar phenomenon can be observed in



**Figure 4.11:** SHAP main effects scatter plots for takeover readiness.

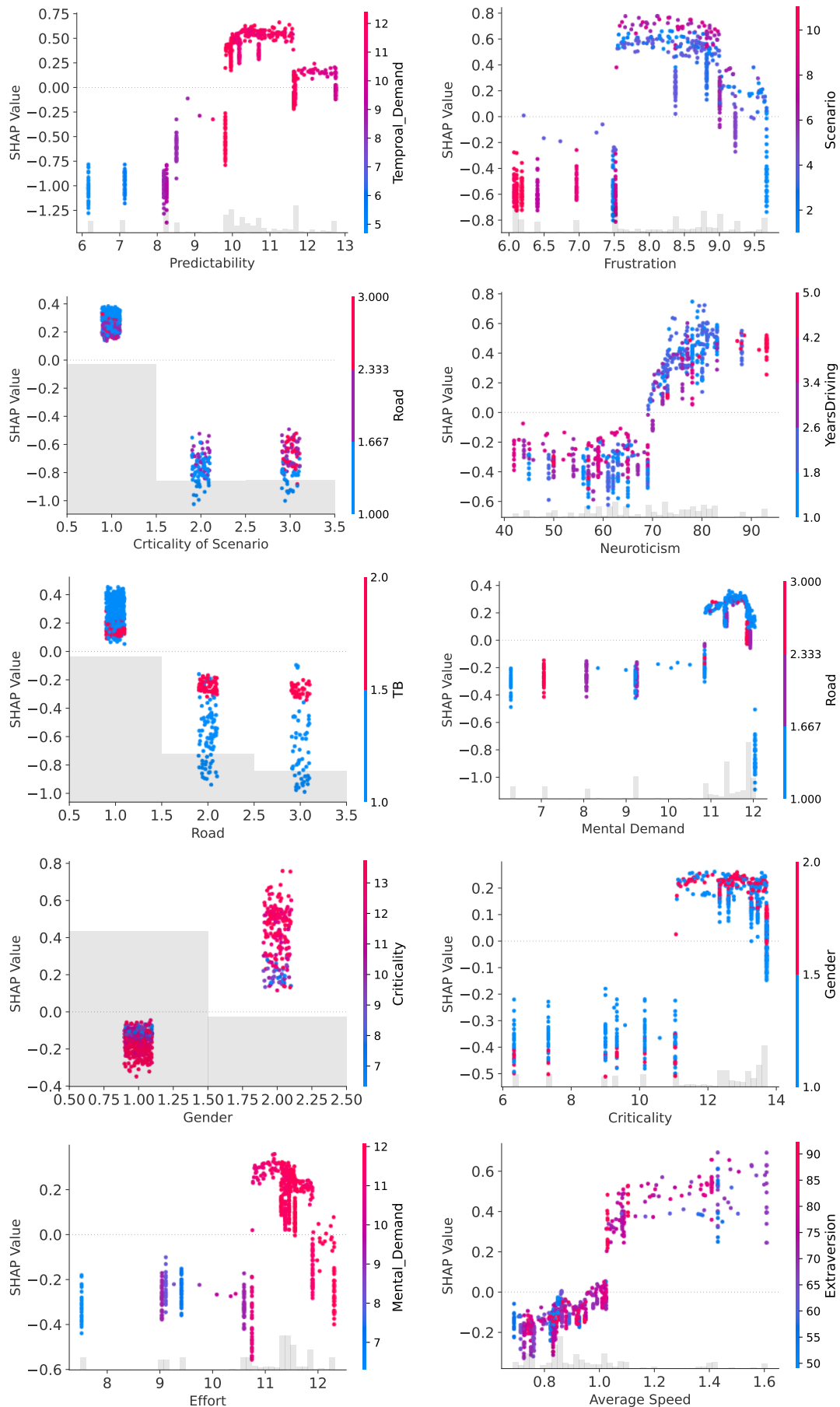


Figure 4.12: SHAP interaction effects scatter plots for takeover readiness.

other plots, such that effect of one feature varies according to the value of another feature. This complicates the problem, making a non-linear model indispensable.

### SHAP Explanation–Local

Similar to the regression model, we can utilize waterfall plots to display explanations for individual predictions. Figs. 4.13 and 4.14 are three examples, showing how each feature affects drivers' takeover readiness differently for high and low takeover readiness, respectively.

In the first example (Fig. 4.13), we can see that almost all the import features have a positive impact on takeover readiness (negative impact on crash probability). Specifically, although a medium level of neuroticism increases the log-odds by 0.42, having a low criticality of scenario reduces the log-odds by 0.75, an average correction rate reduces the log-ratio by 0.68, and a lower level of frustration (7.519) reduces the log-odds by 0.62.

On the contrary, in the second example (Fig. 4.14), we can see that almost all the important features have a negative impact on takeover readiness (positive impact on crash probability). Specifically, gender increased the log-odds by 0.62, a relatively high level of neuroticism (80.412) increased the log-odds by 0.56, a high level of frustration (8.555) and predictability (11.097) increased the log-odds by 0.51 and 0.47 respectively, and the type of the road also increased the log-odds by 0.36.

Therefore, although Predictability is found to be most important in predicting drivers' takeover time overall. For individual cases, the dominant features may be quite different, depending largely on drivers' personality and scenario characteristics.

### 4.3.4 Summary

Considering that takeover readiness is a categorical variable, the objective of this section is to model takeover readiness as a classification problem based on 38 independent variables as summarized in Table 4.1.2 (the same as that in the regression model). Results show that:

- Among the list of models, XGBoost classifier is best at predicting takeover readiness, with the highest accuracy, precision, recall and F1 score.
- Similarly, SHAP explainer is utilized to understand the effects of each feature on takeover readiness by providing global and local explanations. Global explanation provides an overview of the importance ranking of the

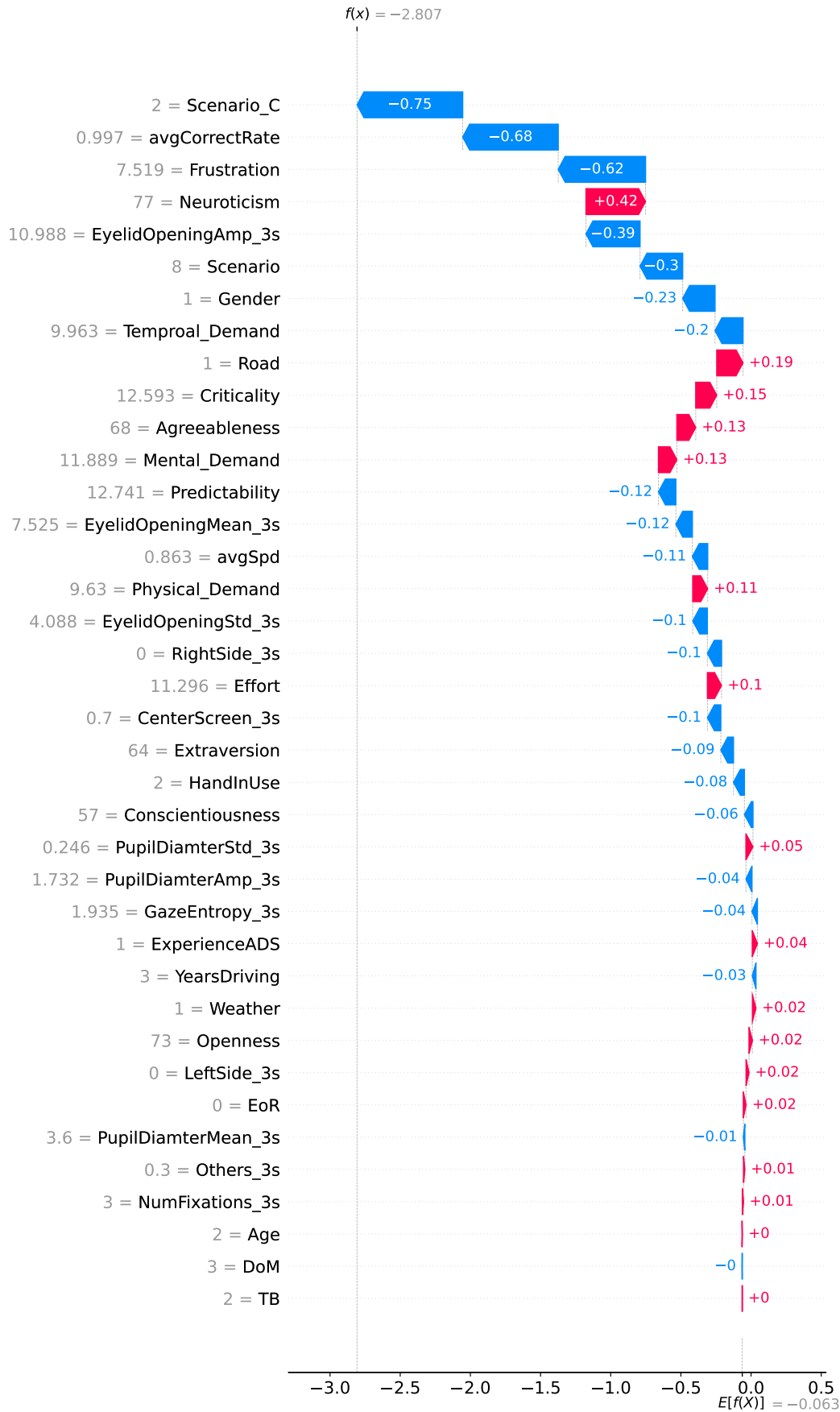
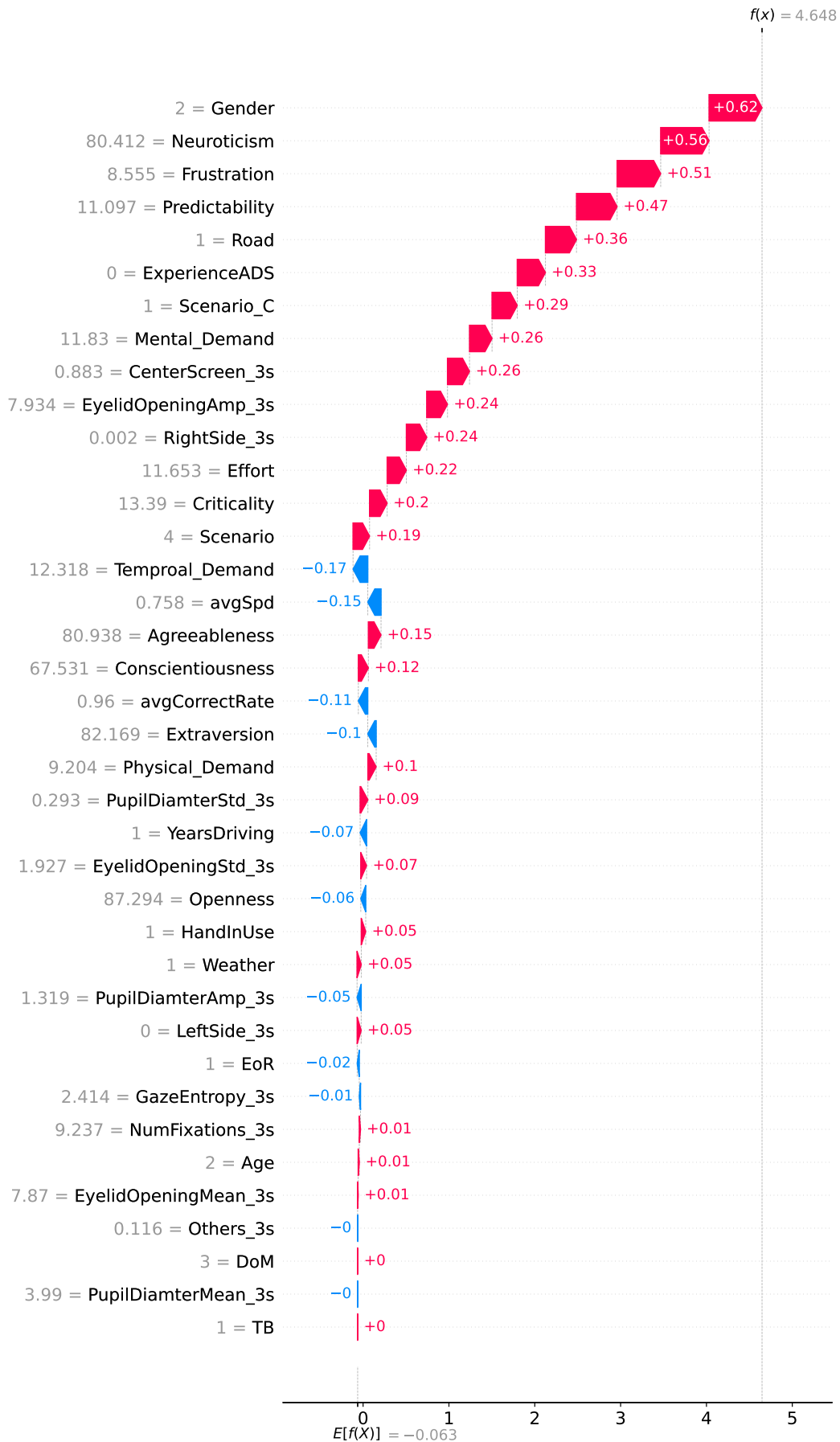


Figure 4.13: SHAP individual explanation for takeover readiness prediction (high).



**Figure 4.14:** SHAP individual explanation for takeover readiness prediction (low).



variables. Overall, it seems that scenario characteristics, drivers' personality and reaction speed (avgSpd), and type of the road are more important in deciding takeover readiness compared with other factors.

- Main effects and interaction effects plots gave us a more depth insight on how each of the features affects takeover readiness. Results show that most of the effects are highly nonlinear, with certain threshold that decides how takeover readiness is affected by that feature.
- Local explanation was also explored for individual explanations. Two examples which represent high and low takeover readiness are analyzed. It can be observed that individual explanations can be very different from global explanations, depending on drivers' personality and scenario characteristics.

## 4.4 Modeling Takeover Style

### 4.4.1 Problem Statement and Objectives

Takeover style refers to characteristic of drivers' takeover maneuvers shortly after the takeover request. Instead of using certain parameters like mean speed, maximum acceleration, maximum steering wheel angle, maximum yaw rate, etc., to evaluate the quality of drivers' takeovers like that in [95, 99], we decided to classify drivers' takeover maneuvers shortly after the TOR into different driving styles. Many methods are possible for this task, like k-mean clustering was followed by a kNN classifier [145], SVM clustering was followed by decision trees [146] and semisupervised SVM [147].

Nevertheless, to apply these methods, a group of features need to be extracted first. Since drivers' speed and steering profiles are essentially time series data, that would result in loss of useful information. Hence, time series clustering methods would be desirable [148]. A good method for automotive application would be multivariate dynamic time warping (MDTW) [149], which is suitable and effective in extracting structural information from data based on the clustering of system behavior time series. Since DTW is basically a distance metric, we can integrate it into k-means clustering and apply the DTW-based k-means clustering to the data collected to extract patterns takeover maneuvers of drivers. Besides, considering DoM (discussed in Section 3.3) before the TOR has been found to affect takeover time [127], whether it will affect patterns of evasive maneuvers will also be investigated.

Here are the objectives of this section: (1) Based on the data collected, DTW-based k-means clustering will be designed and applied to learn from drivers patterns of takeover maneuvers in case of emergency during automated driving; (2) Based on the clustering results, effects of DoM and scenarios on patterns of takeover maneuvers will be analyzed.

## 4.4.2 DTW-Based Clustering

### Data Processing and Feature Selection

All 12 takeover scenarios were adopted in this analysis. Data were recorded in 60 Hz using the driving simulator, where the vehicle's longitudinal and lateral maneuvers were all logged into a .csv file. Takeovers that resulted in crashes were not included. For ease of analysis, data of 15 seconds after the TOR has been issued were excerpted and saved for analysis.

As we are interested in drivers' evasive maneuvers during takeover, what matter most should be the vehicle's speed and steering profiles, therefore, longitudinal speeds and steering wheel angle should be included in the features. Moreover, since risk levels of drivers' driving styles have been found to typically be related to vehicle's speed and acceleration [150], and also to capture drivers' accelerating and braking behaviors, longitudinal acceleration and yaw rate should also be considered. Hence, a four-dimensional time series data could be formulated, and the sequence could be mathematically represented as

$$s = [(v_1, a_1, \delta_1, \dot{\psi}_1), (v_2, a_2, \delta_2, \dot{\psi}_2), \dots], \quad (4.23)$$

where  $v, a, \delta$  and  $\dot{\psi}$  represent speed, acceleration, steering wheel angle, and yaw rate respectively, and the subscripts denote the time step.

Finally, since ranges of different features differ a lot, each feature is further normalized to  $[0, 1]$  according to Eq. (4.24), so that different features are comparable.

$$f = \frac{f - f_{\min}}{f_{\max} - f_{\min}}, f = v, a, \delta, \dot{\psi} \quad (4.24)$$

### DTW-Based K-Means Clustering

Given a set of observations  $(s_1, s_2, \dots, s_n)$  and  $k$  clusters, the objective of the k-means algorithm is to find a partition of the data so as to minimize the within-cluster variances [151]. The algorithm starts by selecting  $k$  cluster centers (centroids) randomly, and then repeat the following two steps iteratively until the

algorithm has converged:

- Assignment step: each observation is assigned to its nearest centroid  $c_i$  based on a distance function  $\text{dist}(c_i, s_j)$ ,  $i = 1, \dots, k$ ,  $j = 1, \dots, n$ ;
- Update step: update the centroids based on the observations in the cluster, typically,  $c_i = \sum_{j=1}^{N_i} s_j / N_i$ , where  $N_i$  is the size of the cluster  $i$  and  $\sum_{i=1}^k N_i = n$ .

The basic form of k-means algorithm uses a Euclidean distance function, if samples to be clustered are time series sequences, this measure attempts to find similarity between two observations based on the correspondence between their time indices. This is problematic in that it will not be able to compare sequences with different lengths. Besides, it will fail to identify two sequences with similar patterns whereas with a small time shift in the time domain. DTW is a method that can hopefully resolve these issues. Instead of one-to-one correspondence, it developed a one-to-many match so that the troughs and peaks with the same pattern can be perfectly matched, making it suitable for measuring similarity between two time series sequences.

Let us denote two time series  $\mathbf{x}$ ,  $\mathbf{y}$  with length  $t$  and  $m$  as  $\mathbf{x} = [x_1, \dots, x_t]^T$ ,  $\mathbf{y} = [y_1, \dots, y_m]^T$ . To align  $\mathbf{x}$  and  $\mathbf{y}$ , a  $t$ -by- $m$  matrix is constructed, where the  $(i, j)$  element contains the distance  $\text{dist}(x_i, y_j)$  between two points  $x_i$  and  $y_j$ , and every element  $(i, j)$  corresponds to the alignment between the points  $x_i$  and  $y_j$ . To obtain the optimal alignment of  $\mathbf{x}$  and  $\mathbf{y}$ , a warping path  $W$  as a set of  $K$  matrix elements  $w_k = (i, j)_k$  is created as:

$$W = w_1, w_2, \dots, w_K, \quad \max(t, m) \leq K < t + m - 1 \quad (4.25)$$

And it must satisfy three constraints: (1) Boundary conditions:  $w_1 = (1, 1)$  and  $w_K = (t, m)$ ; (2) Continuity: If  $w_k = (a, b)$  is given, then  $w_{k-1} = (a', b')$  must satisfy  $a - a' \leq 1$  and  $b - b' \leq 1$ ; (3) Monotonicity: If  $w_k = (a, b)$  is given, then  $w_{k-1} = (a', b')$  must satisfy with  $a - a' \geq 0$  and  $b - b' \geq 0$ . To achieve the optimal alignment, the problem comes down to minimizing the warping cost (Eq. (4.26)) in terms of the number and magnitude of elements  $w_k$  satisfying the three constraints:

$$\text{DTW}(\mathbf{x}, \mathbf{y}) = \min \left\{ \sum_{k=1}^K w_k \right\} \quad (4.26)$$

This warping path can be found using dynamic programming

$$\gamma(i, j) = d(x_i, y_j) + \min\{\gamma(i-1, j-1), \gamma(i-1, j), \gamma(i, j-1)\} \quad (4.27)$$

where  $\gamma(i, j)$  is the cumulative distance of the cell  $(i, j)$ .

For multidimensional time series, different methods have been introduced in the literature. Basically, they can be divided into two categories: independent DTW and dependent DTW, denoted as  $\text{DTW}_I$  and  $\text{DTW}_D$  respectively. In the first category, each dimension of the sequences is warped independently, and then, the warping costs on all dimensions are summed using Eq. (4.28),

$$\text{DTW}_I(\mathbf{X}, \mathbf{Y}) = \sum_{l=1}^v c_l \cdot \text{DTW}(\mathbf{x}_l, \mathbf{y}_l) \quad (4.28)$$

where  $v$  is the dimension of the time series, and  $c_j$  is a factor used to weight each dimension.

In the second category, multivariate time series are treated as a single series with  $v$ -dimensional vectors. In this case, only a single warping is conducted using Eq. (4.29) with a cost function  $\zeta(\mathbf{x}_i, \mathbf{y}_j)$ .

$$\text{DTW}_D(\mathbf{X}, \mathbf{Y}) = \text{DTW}(\mathbf{X}, \mathbf{Y}), \text{ with } \zeta(\mathbf{x}_i, \mathbf{y}_j) \quad (4.29)$$

Most frequently, cost function appears to be in the form of Eq. (4.30),

$$\zeta(\mathbf{x}_i, \mathbf{y}_j) = \left( \sum_{l=1}^v c_l |x_{il} - y_{jl}|^p \right)^{1/p} \quad (4.30)$$

where  $v$  is the dimension of the time series, in our case,  $v = 4$ . And for  $c_j = 1$  and  $p = 2$ , Eq. (4.30) becomes the Euclidean distance, which is adopted in this research.

Besides, instead of initializing the  $k$  cluster centers randomly, after initializing the first centroid as a random selection of one of the observations, the following two steps are repeated until  $k$  cluster centers has been selected:

1. Calculate the sum of the distances between each observation and all the centroids;
2. Select the next centroid randomly, with a probability proportional to the total distance to the centroids.

Combined, the whole algorithm is summarized in Algorithm 2, where the calculation of  $\text{DTW}_D(\mathbf{x}, \mathbf{y})$  can be referenced to Algorithm 3. Algorithm 2 is no different from the normal k-means++ clustering, except that the distance between two samples is calculated using Algorithm 3. It is to be noted that it can be time consuming to run the program when the time series are long. However, some techniques can be applied to make the program more efficient.

**Algorithm 2:** DTW-based k-means++ clustering.

---

**Input:** data  $\mathbf{X} = \{\mathbf{x}_n\}_{n=1}^N$ ; number of clusters  $K$ ; max iteration  $I_{\max}$   
**Output:** cluster means  $\{\mu_k\}_{k=1}^K$

```

1 # Initialization
2  $\mu_1 \leftarrow \text{RandomChoice}(\mathbf{X})$ 
3 for  $k \leftarrow 2, \dots, K$  do
4   for  $n \leftarrow 1, \dots, N$  do
5      $d_n \leftarrow \min_{k' < k} \text{DTW}_D(\mathbf{x}_n, \mu_{k'})$ 
6   end
7   for  $n \leftarrow 1, \dots, N$  do
8      $p_n \leftarrow d_n^2 / \sum_{n'} d_{n'}^2$ 
9   end
10   $\mu_k \leftarrow \text{RandomChoice}(\mathbf{X}, p = (p_1, \dots, p_N))$ 
11 end
12 # Main loop
13 repeat
14   # Assignment step
15   for  $n \leftarrow 1, \dots, N$  do
16      $c_n \leftarrow \arg \min_j \text{DTW}_D(\mathbf{x}_n, \mu_j)$ 
17   end
18   # Update step
19   for  $k \leftarrow 1, \dots, K$  do
20      $\mu_k \leftarrow \frac{\sum_{n=1}^N 1\{c_n = k\} \mathbf{x}_n}{\sum_{n=1}^N 1\{c_n = k\}}$ 
21   end
22 until none of the  $\mu_k$  changes or iteration  $> I_{\max}$ ;

```

---

**Algorithm 3:** Dynamic Time Warping (DTW).

---

**Input:** sequences  $\mathbf{x}$  and  $\mathbf{y}$  of length  $m$  and  $n$   
**Output:** minimum distance  $\text{DTW}[m, n]$

```

1  $\text{DTW} \leftarrow \text{Matrix}(m, n)$ 
2 for  $i \leftarrow 1, \dots, m$  do
3   for  $j \leftarrow 1, \dots, n$  do
4      $\text{DTW}[i, j] \leftarrow \infty$ 
5   end
6    $\text{DTW}[0, 0] \leftarrow 0$  for  $i \leftarrow 1, \dots, m$  do
7     for  $j \leftarrow 1, \dots, n$  do
8        $\text{cost} \leftarrow \zeta(\mathbf{x}_i, \mathbf{y}_j)$ 
9        $\text{DTW}[i, j] \leftarrow \text{cost} + \min(\text{DTW}[i-1, j-1],$ 
10                                      $\text{DTW}[i-1, j],$ 
11                                      $\text{DTW}[i, j-1])$ 
12     end
13   end
14 end

```

---

### 4.4.3 Results and Discussions

Since data were recorded in 60 Hz, 15 s would result in a time series of dimension 4 and length 900. And in consideration of both feature representations and interpretability of the resulted clusters,  $K$  was chosen as 3 using the elbow curve method, which was in line with the results in [152], where driving behaviors were classified as “conservative”, “normal” and “risky” ones. However, in this analysis, we simply denoted them as Type I, Type II and Type III respectively. Finally, the max iteration was set as 10, as the loops usually converges within 10 iterations.

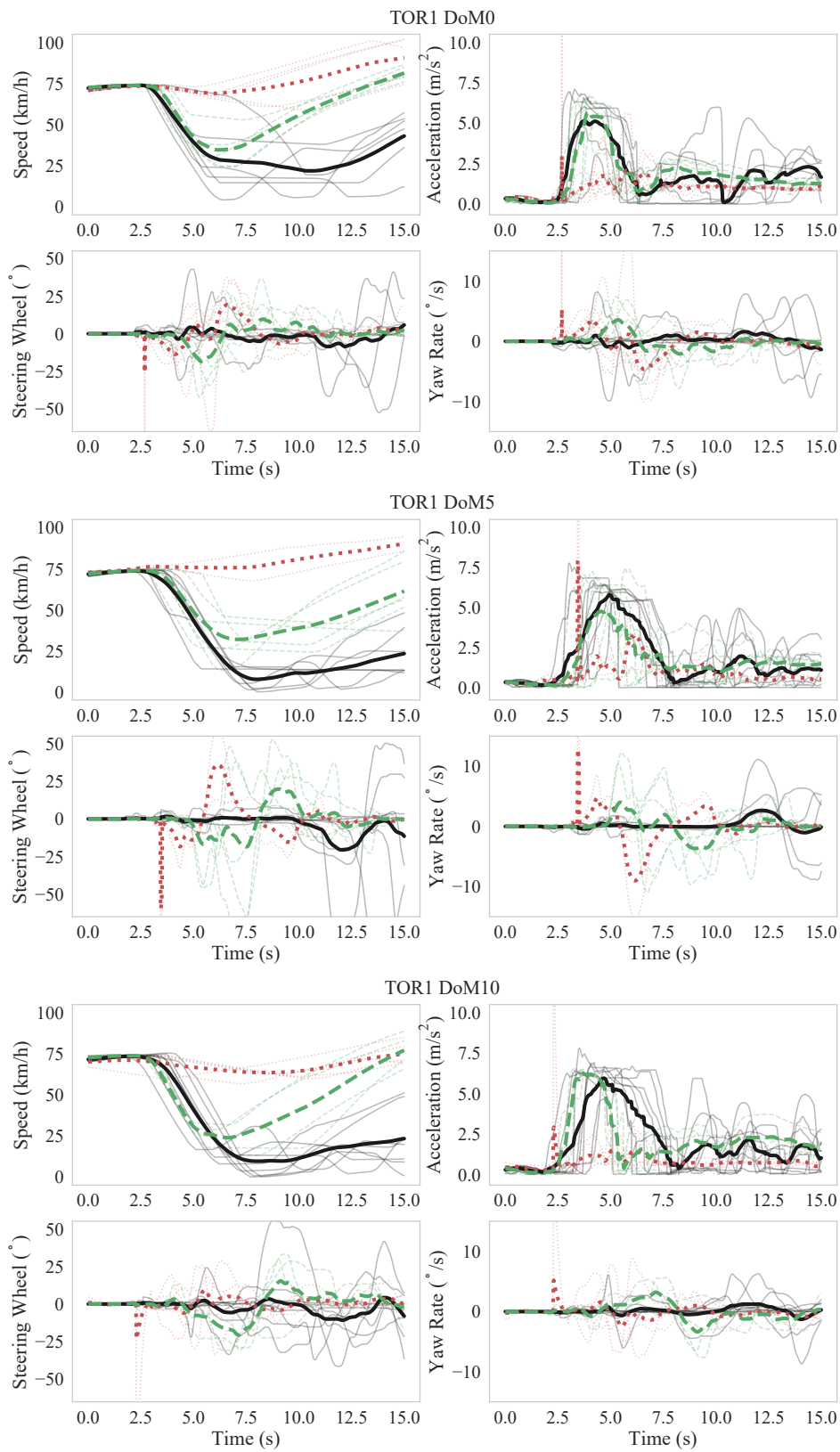
To investigate impacts of DoM on drivers’ patterns of evasive maneuvers, Algorithm 2 was firstly conducted for each scenario with different DoMs. As an example, the results of TOR1 were summarized below, and results of the rest can be downloaded from <https://github.com/c-huang-tty/MRM/tree/main/Results>. Then, the algorithm was conducted for each scenario regardless of DoMs.

#### Effects of DoM on Patterns of Evasive Maneuvers

Speed, acceleration, steering, and yaw rate profiles of the first 15 s after the TOR has been issued in TOR1 were plotted in Fig. 4.15, where the three patterns (Type I, Type II, and Type III) of evasive maneuvers were represented as thin dotted (red), dashed (green) and solid (black) lines respectively, with the corresponding thick lines indicating the averages of the profiles. Besides, descriptive statistics (mean, min, and max) of each feature were summarized in Table 4.15, where Spd, Acc, Steer, and Yaw represent speed, acceleration, steering wheel angle, and yaw rate respectively.

Overall, DoM appeared to affect driver’s takeover time more than patterns of evasive maneuvers. And irrespective of DoMs, three patterns of maneuvers could be similarly distinguished. Specifically,

- drivers with Type I evasive maneuvers tended to pass the vehicle ahead at higher speeds and lower accelerations, correspondingly, they also preferred to make quicker lane change maneuvers;
- drivers with Type II maneuvers seemed to be more cautious, before changing lanes, they would slow down the vehicle a bit first, and then started to accelerate and perform the lane change maneuvers at the same time;
- and the most cautious drivers were those with Type III maneuvers, they would decelerate to a low speed, and kept at that speed for a period before they started to accelerate and finish the lane change maneuvers.



**Figure 4.15:** Patterns of evasive maneuvers in TOR1 with different DoMs. Dotted, dashed, and solid lines represent Type I, Type II and Type III patterns of evasive maneuvers respectively.

**Table 4.15:** Descriptive statistics of each feature for each driving style in TOR1 given different DoMs.

Indices	DoM <sub>S</sub>			DoM <sub>M</sub>			DoM <sub>L</sub>		
	I	II	III	I	II	III	I	II	III
Spd <sub>mean</sub>	75.90	59.70	40.96	79.03	51.55	33.43	68.08	51.07	33.87
Spd <sub>min</sub>	68.99	34.54	21.78	72.55	31.93	7.55	63.60	23.86	9.55
Spd <sub>max</sub>	90.65	81.38	73.92	90.14	73.98	73.63	74.95	77.20	73.53
Acc <sub>mean</sub>	1.01	1.75	1.78	0.95	1.66	1.80	0.78	2.05	1.98
Acc <sub>max</sub>	2.98	5.69	5.15	7.94	4.83	5.81	2.92	6.26	5.95
Steer <sub>mean</sub>	0.33	-0.14	-1.50	-0.10	-0.65	-3.20	0.31	-0.47	-2.07
Steer <sub>min</sub>	-25.75	-18.44	-8.53	-59.65	-18.55	-20.60	-23.17	-20.63	-10.76
Steer <sub>max</sub>	19.28	9.72	5.82	36.15	19.83	1.54	9.30	15.30	4.33
Yaw <sub>mean</sub>	-0.05	-0.04	0.16	0.00	-0.01	0.23	-0.07	-0.06	0.10
Yaw <sub>min</sub>	-4.68	-2.11	-1.38	-9.01	-3.68	-1.05	-2.48	-3.38	-1.29
Yaw <sub>max</sub>	5.35	3.57	1.59	12.84	4.01	2.65	5.19	3.22	1.24

Spd: m/s, Acc: m/s<sup>2</sup>, Steer: °, Yaw: °/s.

Moreover, it can be read from Table 4.15 and Fig. 4.15 that drivers with Type II and Type III maneuvers exhibited similar maximum accelerations, whereas with different durations, where the latter seemed to be longer than the former, resulting in lower speeds. As for the steering behavior, it can be observed from Fig. 4.15 that drivers with Type III maneuvers started to make obvious steering only about 10 s after the TOR has been issued. However, it was 2–7 s for drivers with Type I and Type II maneuvers, with the former (2–5 s) quicker than the latter (5–7 s). And the trend of yaw rate was in opposite to the steering angles, with Type III smoother than the others.

### Effects of Scenarios on Patterns of Evasive Maneuvers

Since DoM appeared to impact takeover time more than patterns of maneuvers, we could further apply Algorithm 2 to each scenario regardless of DoMs, and the number of drivers' maneuvers belonging to each type of takeover maneuvers were summarized in Table 4.16. Besides, the profiles were plotted in Figs. 4.16, 4.17, 4.18 and 4.19. Similarly, the descriptive statistics were summarized in Table 4.17 and 4.18.

With respect to speeds, it could be observed that in the first 15 s after the TOR, except for TOR3 and TOR4, speed of the three types of maneuvers in the rest of the scenarios all exhibited similar patterns, with the overall speed of Type II smaller and greater than that of Type I and Type III respectively. When it comes to 8 s after the TOR, all the scenarios exhibit similar patterns. However, amplitudes of speeds and accelerations seemed to vary across different



**Table 4.16:** Clustering results of the takeover style.

Scenario	Type I	Type II	Type III
TOR1	13	21	14
TOR2	15	3	30
TOR3	25	15	8
TOR4	26	13	9
TOR5	22	15	11
TOR6	10	11	27
TOR7	19	13	16
TOR8	10	21	17
TOR9	5	32	11
TOR10	16	3	28
TOR11	19	9	19
TOR12	11	22	14

Numbers in the table are numbers of samples in each group.

scenarios. Typically, the overall speeds in low-critical scenarios TOR10–TOR12, medium-critical scenario TOR7 and high-critical scenario are higher than that in other scenarios. Pattern of acceleration in TOR5 was similar to that in TOR1, whereas with slightly smaller amplitudes. In TOR2 and TOR4, accelerations of Type III were much larger than that of Type I and Type II, with the latter two had similar amplitudes whereas opposite directions. Besides, accelerations of Type III maneuvers in TOR 2 were much larger than that in TOR4. In TOR6, acceleration of Type I was at first smaller, and then greater than that of Type II and Type III. Therefore, scenarios appeared to affect amplitudes of speeds and patterns of accelerations greatly, although the resulting patterns of speeds were kind of similar to each other.

With respect to steering behaviors, similar patterns can still be observed in most of the scenarios, where steering operations were firstly observed in Type I, then in Type II, and lastly in Type III maneuvers. However, the amplitudes of steering angles as well as yaw rates were different across different scenarios, with the greatest steering wheel angle and yaw rate observed in TOR6. The others were kind of similar except for steering angles of Type III maneuvers in TOR2, where large steering angles were recorded. Hence, scenarios appeared to affect amplitudes of steering wheel angle and patterns of yaw rate greatly, whereas the resulting patterns of steering behaviors were kind of similar, especially in regard to the timing of the lane change maneuvers.

Overall, regardless of scenarios, application of Algorithm 2 would result in three distinctive patterns of evasive maneuvers. It appeared that drivers' acceleration and deceleration behaviors (both longitudinal and lateral) were greatly

affected by scenarios, however, the resulting patterns of speed and steering behaviors seemed to be similar irrespective of scenarios. This was most obvious for speed profiles, where drivers with Type I maneuvers would pass the obstacles in high speeds, drivers with Type II maneuvers tended to decelerated and then accelerate quickly before passing the obstacles, and drivers with Type III maneuvers would pass the obstacles in low speeds. In TOR3, it seemed that most drivers has chosen not to change lanes, however, their speed profiles also followed similar patterns.

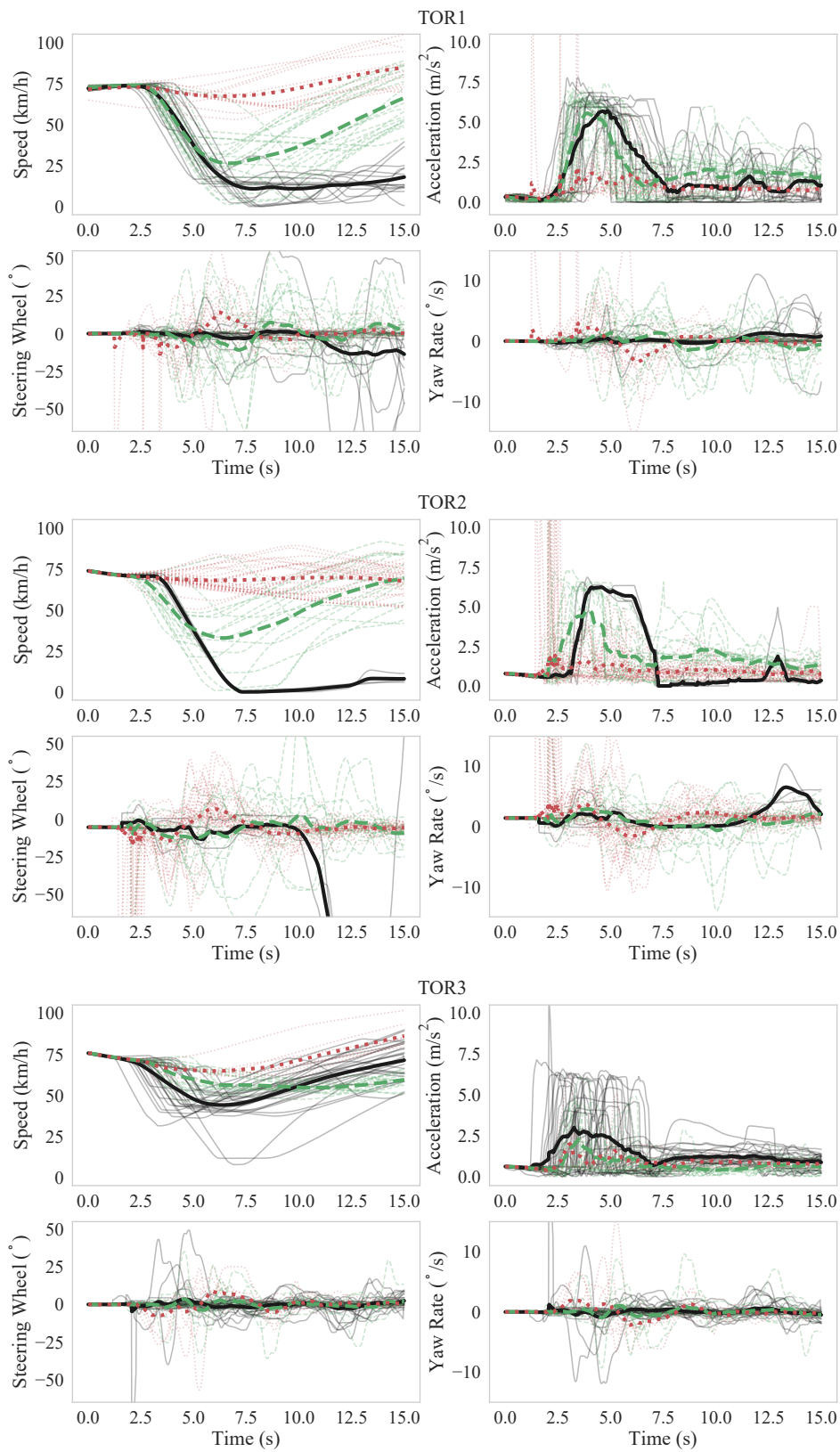
#### 4.4.4 Summary

Considering that drivers' maneuvers were actually temporal sequences, time series clustering methods were considered in this section, specifically, DTW-based k-means clustering method. Application of the algorithm resulted in three patterns of evasive maneuvers irrespective of DoMs and scenarios, and the basic conclusions were:

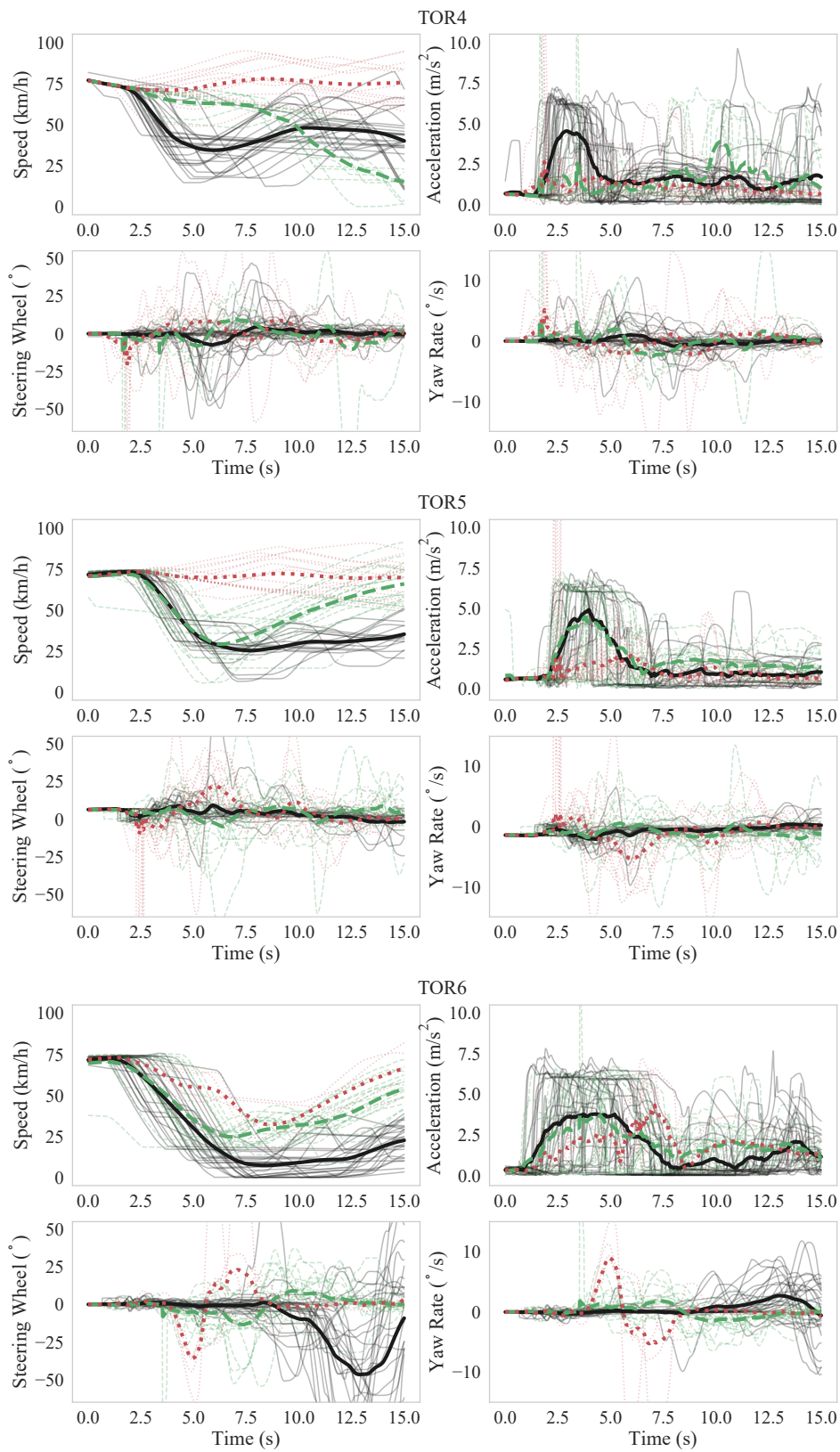
- DoM appeared to impact takeover time more than patterns of evasive maneuvers;
- With regard to longitudinal maneuvers, scenarios appeared to affect amplitudes of speeds and patterns of accelerations greatly, although the resulting patterns of speeds were kind of similar to each other, with the overall speed of Type II smaller and greater than that of Type I and Type III respectively;
- With regard to lateral maneuvers, scenarios appeared to affect amplitudes of steering wheel angle and patterns of yaw rate greatly, whereas the resulting patterns of steering behaviors were kind of similar to each other, with steering operations firstly observed in Type I, then in Type II, and lastly in Type III maneuvers.

Specifically,

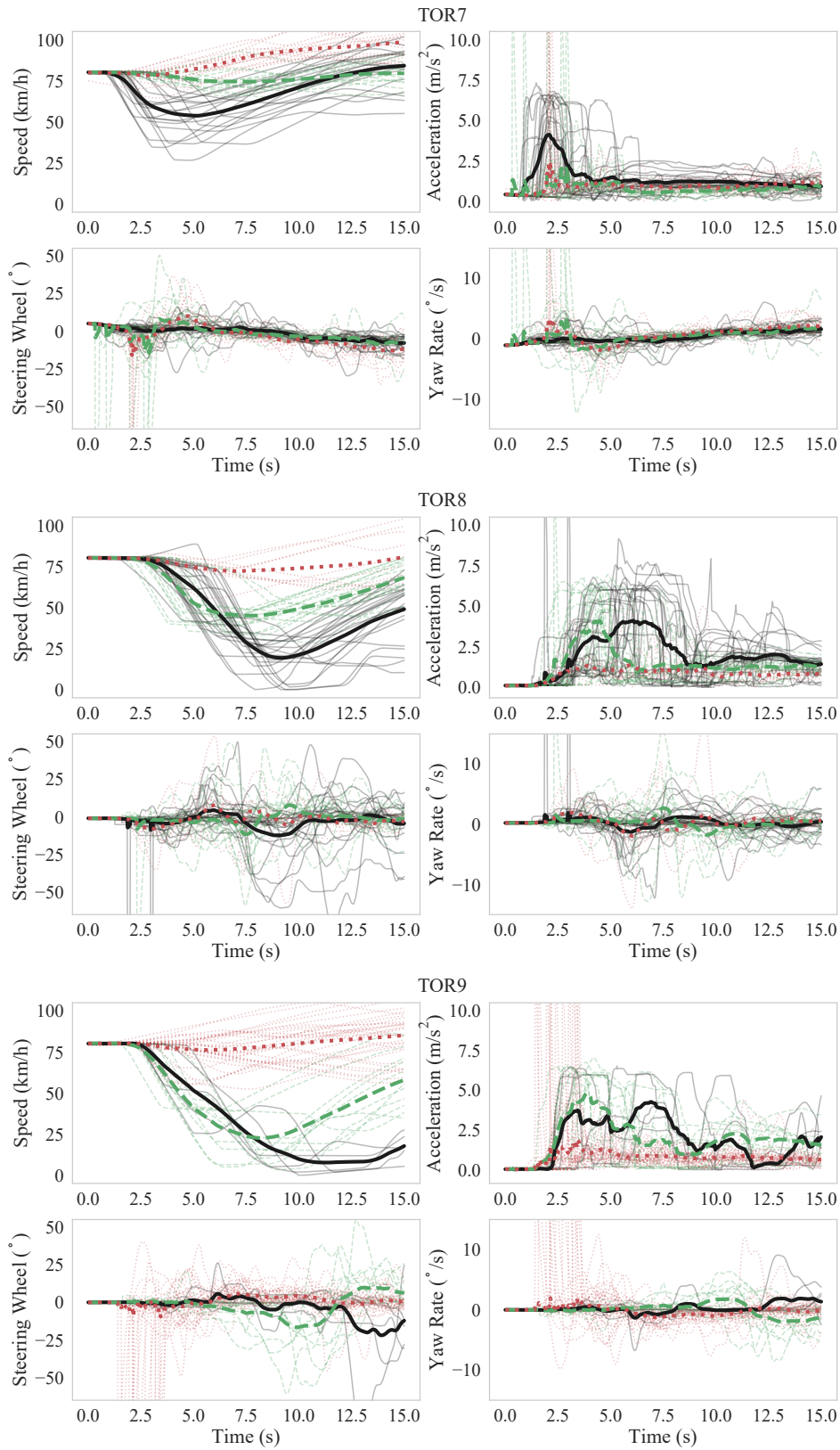
- drivers with Type I maneuvers preferred to pass the vehicle ahead at higher speeds and lower accelerations, correspondingly, they also preferred to make quicker lane change maneuvers;
- drivers with Type II maneuvers were more cautious, before changing lanes, they would slow down the vehicle a bit first, and then started to accelerate and perform the lane change maneuvers at the same time;
- drivers with Type III maneuvers would decelerate the vehicle to a low speed, and then kept at that speed for a period of time before they started to accelerate and finish the lane change maneuvers.



**Figure 4.16:** Patterns of evasive maneuvers in scenarios TOR1, TOR2 and TOR3. Dotted, dashed, and solid lines represent Type I, Type II and Type III patterns of evasive maneuvers respectively.

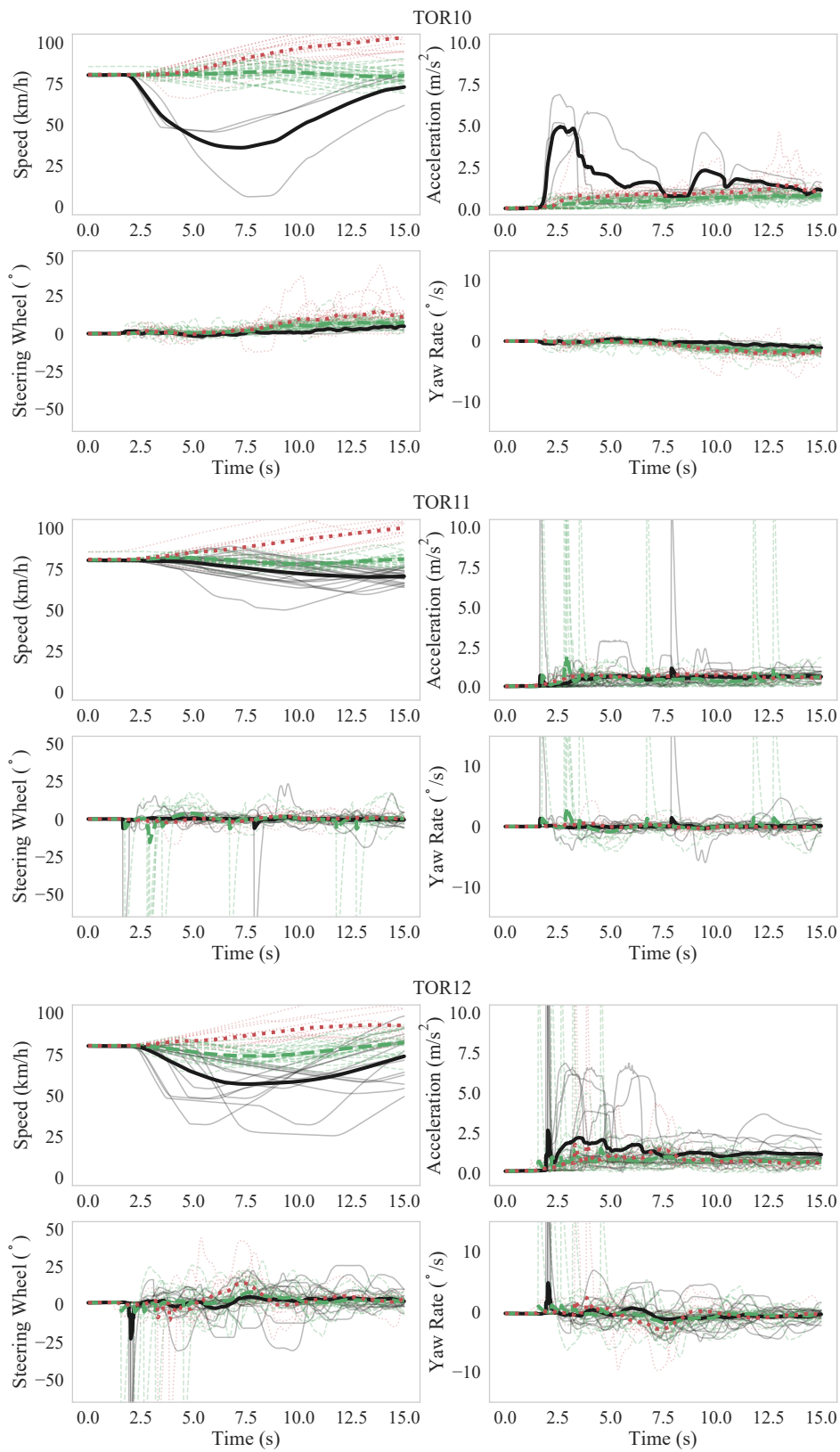


**Figure 4.17:** Patterns of evasive maneuvers in scenarios TOR4, TOR5 and TOR6. Dotted, dashed, and solid lines represent Type I, Type II and Type III patterns of evasive maneuvers respectively.



**Figure 4.18:** Patterns of evasive maneuvers in scenarios TOR7, TOR8 and TOR9. Dotted, dashed, and solid lines represent Type I, Type II and Type III patterns of evasive maneuvers respectively.





**Figure 4.19:** Patterns of evasive maneuvers in scenarios TOR10, TOR11 and TOR12. Dotted, dashed, and solid lines represent Type I, Type II and Type III patterns of evasive maneuvers respectively.

**Table 4.17:** Descriptive statistics of each feature for each driving style in scenarios TOR1–TOR6.

Indices		Spd <sub>mean</sub> (m/s)	Spd <sub>min</sub> (m/s)	Spd <sub>max</sub> (m/s)	Acc <sub>mean</sub> (m/s <sup>2</sup> )	Acc <sub>max</sub> (m/s <sup>2</sup> )	Steer <sub>mean</sub> (°)	Steer <sub>min</sub> (°)	Steer <sub>max</sub> (°)	Yaw <sub>mean</sub> (°/s)	Yaw <sub>min</sub> (°/s)	Yaw <sub>max</sub> (°/s)
TOR1	Type I	72.95	67.26	84.72	0.93	2.23	0.23	-13.60	14.18	-0.04	-3.52	3.11
	Type II	49.13	26.27	73.93	1.93	5.58	-0.54	-10.88	7.44	-0.02	-1.48	1.38
	Type III	31.91	10.78	73.61	1.77	5.71	-3.40	-14.11	1.55	0.25	-0.33	1.37
TOR2	Type I	69.33	67.58	73.69	0.99	1.92	-5.75	-16.72	6.63	1.40	-1.57	3.95
	Type II	53.72	32.83	73.80	1.87	4.77	-5.98	-12.49	2.50	1.05	-0.65	2.94
	Type III	26.05	0.00	73.76	1.65	6.24	-34.59	-138.30	-0.91	1.62	-0.00	6.53
TOR3	Type I	71.88	64.74	86.11	0.89	2.01	0.07	-7.31	8.71	0.01	-2.19	1.88
	Type II	60.53	54.60	75.74	0.75	2.27	-0.16	-3.82	3.52	0.04	-0.94	1.02
	Type III	59.08	44.03	75.64	1.36	3.02	-0.21	-4.67	3.43	0.04	-0.87	1.22
TOR4	Type I	74.57	70.86	77.89	1.11	2.63	0.09	-21.11	8.85	0.01	-2.16	5.21
	Type II	52.02	15.24	76.74	1.48	3.92	-0.68	-12.99	9.97	0.03	-2.54	3.39
	Type III	48.49	34.22	76.87	1.72	4.54	-0.25	-7.51	4.57	0.00	-0.84	1.02
TOR5	Type I	70.84	69.28	73.29	1.04	2.06	4.68	-10.23	22.05	-1.10	-5.38	2.23
	Type II	51.80	29.14	71.54	1.81	4.44	4.09	-5.05	8.57	-0.98	-2.04	0.56
	Type III	40.90	25.12	73.12	1.50	4.89	3.63	-2.28	9.12	-0.68	-2.10	0.38
TOR6	Type I	53.81	32.28	72.97	1.76	4.32	0.29	-35.94	23.06	-0.02	-5.24	8.98
	Type II	42.87	24.64	70.42	1.93	3.75	-0.65	-13.39	9.07	0.02	-1.51	1.86
	Type III	28.26	7.25	72.50	1.76	3.78	-10.45	-46.59	1.18	0.52	-0.58	2.71

**Table 4.18:** Descriptive statistics of each feature for each driving style in scenarios TOR7–TOR12.

Indices		Spd <sub>mean</sub> (m/s)	Spd <sub>min</sub> (m/s)	Spd <sub>max</sub> (m/s)	Acc <sub>mean</sub> (m/s <sup>2</sup> )	Acc <sub>max</sub> (m/s <sup>2</sup> )	Steer <sub>mean</sub> (°)	Steer <sub>min</sub> (°)	Steer <sub>max</sub> (°)	Yaw <sub>mean</sub> (°/s)	Yaw <sub>min</sub> (°/s)	Yaw <sub>max</sub> (°/s)
TOR7	Type I	88.07	78.26	98.51	0.95	2.40	-3.17	-15.96	10.00	0.54	-2.05	3.42
	Type II	77.37	74.49	80.23	0.80	2.04	-1.75	-14.51	7.99	0.37	-1.95	3.05
	Type III	68.90	53.72	84.14	1.39	4.13	-1.30	-7.97	5.03	0.25	-1.10	1.59
TOR8	Type I	76.08	72.16	80.70	0.83	1.36	-1.68	-8.04	7.64	0.37	-1.96	1.98
	Type II	59.56	45.03	80.24	1.43	4.09	-1.29	-11.71	7.76	0.28	-1.77	2.67
	Type III	48.35	19.27	80.22	1.92	4.10	-2.95	-12.39	4.43	0.25	-1.31	1.91
TOR9	Type I	79.83	76.23	85.25	0.79	1.88	0.71	-9.53	6.00	-0.14	-1.29	2.05
	Type II	46.91	22.63	80.27	1.94	4.81	-2.18	-16.65	9.85	0.02	-1.96	1.79
	Type III	36.88	7.69	80.22	1.85	4.27	-3.50	-21.94	5.66	0.30	-1.33	1.90
TOR10	Type I	90.18	80.19	102.64	0.82	1.46	5.07	-0.76	15.07	-0.84	-2.42	0.15
	Type II	80.64	79.23	82.21	0.51	0.84	3.19	-0.02	7.54	-0.71	-1.69	0.00
	Type III	56.19	35.79	80.20	1.68	4.95	1.11	-1.87	5.10	-0.26	-1.17	0.39
TOR11	Type I	88.89	80.63	99.83	0.59	0.81	-0.03	-2.36	2.01	0.02	-0.32	0.56
	Type II	79.96	77.81	81.96	0.47	1.96	-0.23	-15.67	3.83	0.04	-0.88	3.14
	Type III	75.28	70.06	80.20	0.50	1.17	-0.13	-6.20	1.45	0.03	-0.37	1.47
TOR12	Type I	86.53	80.11	92.81	0.74	1.77	2.89	-10.46	14.11	-0.56	-2.85	2.35
	Type II	77.52	73.91	82.18	0.72	1.52	2.18	-5.13	8.24	-0.50	-1.99	1.15
	Type III	66.40	56.79	80.04	1.20	2.66	1.86	-22.83	5.03	-0.45	-1.24	4.84



## 4.5 Summary of Chapter 4

Chapter 4 aims to fulfill Objective 3, and the main results and conclusions are:

- Based on the data collected in Chapter 3, 38 features were extracted to model drivers' takeover behaviors in three respects, including takeover time, takeover readiness, and takeover style. And they were modeled as regression, classification, and clustering problems, respectively.
- Considering that XGBoost is the best for modeling tabular data, takeover time and takeover readiness were modeled using XGBoost regressor and XGBoost classifier, respectively. Comparison with other baseline methods also showed that they yielded the best performance among a list of models.
- Considering that drivers' maneuvers were temporal sequences, time series clustering methods were considered, specifically, dynamic-time-warping (DTW)-based k-means clustering, which resulted in three distinctive types of maneuvers irrespective of DoM and scenarios.
- The detailed conclusions of the main results can be referenced to Sections 4.2.4, 4.3.4, and 4.4.4, respectively.

## Chapter 5

# Main Results and Unsafe Takeover Behaviors

### 5.1 Main Results Discussions

Takeover is a challenging task in that drivers may need a longer time to shift their attention back to the driving tasks in some situations, especially when they are involved in NDRTs. In [13], it was stipulated that “sufficient time” should be provided to the driver, so that the driver would have enough time to take over the vehicle. However, a sufficient time for a vigilant driver may not be sufficient for a sleepy driver anymore. Besides, there may also be the situation where the system has issued a takeover request, yet the driver fails to take over control of the vehicle. Therefore, sufficient time would be quite different in accordance with states and preferences of the driver. This denotes the necessity for ADS to constantly monitor the driver and predict driver takeover behaviors and adapt the system to ensure safe takeovers.

To build such prediction models, a wide variety of factors would be necessary. For a systematic overview, a thorough discussion was provided in Chapter 2, where the factors were classified into system-, scenario-, and human-related factors, and the main conclusions were summarized in Table 2.5. Regarding system-related factors, a vast majority of the studies were centered around HMI design, which has been found to be very helpful if the HMI is well designed [18–22]. By analyzing failure cases during the experiment, some of the suggestions would also be provided in Section 5.2.3. Regarding scenario-related factors, it seems that except for time budget, traffic density and vehicle speeds, the other factors/elements concerning scenarios were not well researched in the previous studies. Hence, a series of scenarios and a questionnaire were designed in this study to have a glimpse of the impacts of various aspects of scenarios on takeover behaviors. These results were analyzed briefly in Section 3.3 and also applied

during modeling takeover behaviors in Chapter 4. Finally, regarding human-related factors, consistent conclusions seem to be missing in many of the factors, such as age, driving experience, gender, etc. This could be attributed to the different setting of experiments in different studies, suggesting the complexity of such kind of problems.

In all the previous studies, effects of the time interval before the TOR and drivers' personality seem to be less emphasized. To fill these gaps, a series of experiments were designed and implemented, which involved 48 participants in total. Regarding impacts of the time interval before the TOR, denoted as DoM, it was found out that the effects of it are indeed statistically significant. However, instead of a linear relationship, regression models revealed that the relationship between DoM and takeover time was quasi-parabolic, and the minimum could be obtained somewhere between  $DoM_M$  and  $DoM_L$ , which was around 7 s. Hence, longer DoM before the TOR did not necessarily lead to quicker response and better takeover performance, i.e., DoM and takeover time were neither monotonously positively nor negatively correlated. This was in coincidence with the results by [68] and also agreed with the working principles of human beings' vision system. Regarding impacts of drivers' personality, it was found out that effects of personality traits on takeover performance were basically not statistically significant given  $DoM_S$ . This suggested that drivers would only exhibit preferences for certain takeover maneuvers only when they have gained enough situation awareness. Otherwise, their instant behaviors would be similar to each other. Besides, different personality traits seem to affect drivers' takeover behaviors in different aspects. This is reasonable since different aspects of personality are generally considered to be related with certain personal qualities, which has been studied broadly in the field of psychology.

Based on the previous research, 38 independent variables were identified from the dataset, including 12 categorical variables and 26 numerical variables. Different from the previous studies like [2, 17, 37, 63, 95–97], more scenario-related factors and drivers' personality were also involved in the modeling process, so that the prediction can be more personalized and more responsive to variations of takeover scenarios. Drivers' takeover behaviors were modeled in three aspects, including takeover time, takeover readiness, and takeover style. Regarding takeover time, since takeover time is a continuous variable, it was modeled as a regression problem, achieving a mean absolute error of less than 0.5 s. Regarding takeover readiness, since takeover readiness is a categorical variable, it was modeled as a classification problem, achieving an accuracy of over 95%, along with a recall of over 95%. And regarding takeover style, since

drivers' maneuver is time series data, it was modeled as a clustering problem, achieving three distinctive types of driving styles. However, it is to be noted that the XGBoost model utilized in modeling takeover time and readiness reveals only statistical relationships between the independent and dependent variables. To obtain causal relationships between them, other more advanced methods would be necessary. Besides, in this research, drivers' takeover styles were only analyzed after the data have been collected, however, for real time applications, online identification of drivers' takeover style would be necessary, which would be perfected in the future studies.

Finally, during the experiments, some crashes and near-crashes have been observed. In the literature, most of the studies has focused on how to deliver information more effectively, whereas how HMIs are related with crashes and how HMIs can be designed to reduce these crashes during takeovers seems to be not that informative. Therefore, in the rest part of this chapter, the primary purpose is to analyze the crash data collected from several takeover scenarios and summarize the typical unsafe behaviors during takeovers, which will then be utilized to obtain some guidance on HMI design to reduce potential crashes as far as possible. Of special concerns are drivers' gaze behaviors shortly after the TOR. And this aims to fulfill Objective 4.

## 5.2 Failure Cases Discussions

### 5.2.1 Data Processing and Annotation

Eye tacking data from 162 takeovers collected from 27 participants were utilized in this analysis. Since data from 4 participants and 6 takeovers of the left 23 participants failed to be recorded, 132 recordings were valid for analysis. Firstly, we performed a manual annotation to determine whether the recordings were related to crashes or near-crashes (Phase 1). Then, using the crash-related videos, we developed a codebook that guided the thematic analysis of the possible unsafe behaviors that might lead to the crashes or near-crashes (Phase 2). Below are elaborations of the two phases.

#### Phase 1: Identifying Crashes and Near-Crashes

Phase I of the annotation process dealt with identifying crashes and near-crashes during the experiment. Two researchers reviewed the data and discussed together to build a shared understanding of what a crash or a near-crash should be.

Based on the initial annotation and in reference to [153], the following definition for crashes and near-crashes were adopted.

**Definition 1 (Crash)** Any contact with an object, either moving or fixed, at any speed, including other vehicles, obstacles on or off the roadway, roadside barriers, animals, etc.

**Definition 2 (Near-Crash)** Any circumstance that requires a rapid, evasive maneuver by the driver to avoid a crash that almost causes a crash. A rapid, evasive maneuver is defined as steering, braking, accelerating, or any combination of control that approaches the limits of the vehicle capabilities. By “almost”, we mean the distance between the ego vehicle and the object is less than, e.g., 0.3 m.

Two researchers (two authors of [154]) reviewed the 132 recordings independently and annotated them according to Definition 1 and Definition 2. By calculating Cohen’s Kappa score [155, 156], we found that there was an almost perfect agreement between the two annotators ( $\kappa = 0.952$ ). For the few samples with disagreement, the two annotators solved the disagreements by discussing the recordings and reached conclusions on whether the recording belonged to crashes or near-crashes.

## Phase 2: Categorizing Crashes and Near-Crashes

Since one of the purposes of this research was to categorize drivers’ unsafe behaviors into several categories, in this phase, we designed a codebook that guided us towards understanding the possible reasons behind the crashes and near-crashes. To build the codebook and perform annotations, two researchers independently checked the recordings of crashes and near-crashes as determined by Phase I and produced initial codes using thematic coding [157], which is a “method for identifying, analyzing and reporting patterns within data”. Then, the two annotators discussed these initial codes and went through several iterations until the codebook has reached stability.

After several rounds of discussion, we agreed to classify the unsafe behaviors into five categories, denoted as C1 to C5, respectively. Specifically, C1 refers to neglect of mirrors while overtaking the vehicles or obstacles ahead; C2 refers to use of mirrors at improper time, i.e., looking into the mirrors for too late, too short, or too long; C3 refers to improper judgement of the situations even though mirrors were glanced at, e.g., change lanes when the driver should not; C4 refers to unreasonable allocation of attention, gazing at certain areas for too long so as to neglecting important information in other areas; and C5 refers to

improper vehicle speeds relative to vehicles in front or behind. While annotating the recordings, the following rules were referenced:

1. Check if the vehicle's speed was proper enough (rate subjectively by the annotators) to keep safe distance to other vehicles, i.e., neither too fast or too slow relative to the vehicles in front or behind. If not, the recording should be classified as C5.
2. Check if the driver looked into mirrors (by checking dynamic eye gazes of the driver captured by the high-speed camera) while overtaking the vehicles or obstacles ahead. If not, the recording should be classified as C1; otherwise,
  - (a) Check if drivers have used mirrors properly, i.e., mirrors were glanced at before the lane changes instead of during or after the lane change, and at proper frequencies. If not, the recording should be classified as C2.
  - (b) Check if drivers have concentrated on certain areas of interests for too short or too long so as they neglected importance information in other areas of interests, e.g., drivers were too focused on the right side of the road so as to neglect the obstacles ahead, or drivers were too absorbed in the roadway ahead so as to neglect the traffic situations on both sides of the road. If yes, the recording should be classified as C4.
  - (c) Check if drivers made improper judgements based on information they observed from mirrors, i.e., whether they changed lanes when they actually should not, e.g., the vehicle behind was approaching at high speeds and the distance between that vehicle and the ego vehicle was not enough for safe lane changes. Since this can hardly be observed just based on the recordings, we have also referenced the interview results after the collisions have happened, where drivers were asked if they have observed the vehicles behind. If yes, the recording should be classified as C3.

Based on the rules above, each annotator coded the crashes and near-crashes independently. Then, the interrater reliability was measured using Cohen's Kappa score. Overall, we find substantial agreement between the two annotators ( $\kappa = 0.911$ ). For the one sample with disagreement (annotated as C1 and C3 by the two annotators respectively), the two annotators solved the disagreements by discussing the recordings and finally decided to assign it to C1.

## 5.2.2 Analysis of Safe and Unsafe Takeover Behaviors

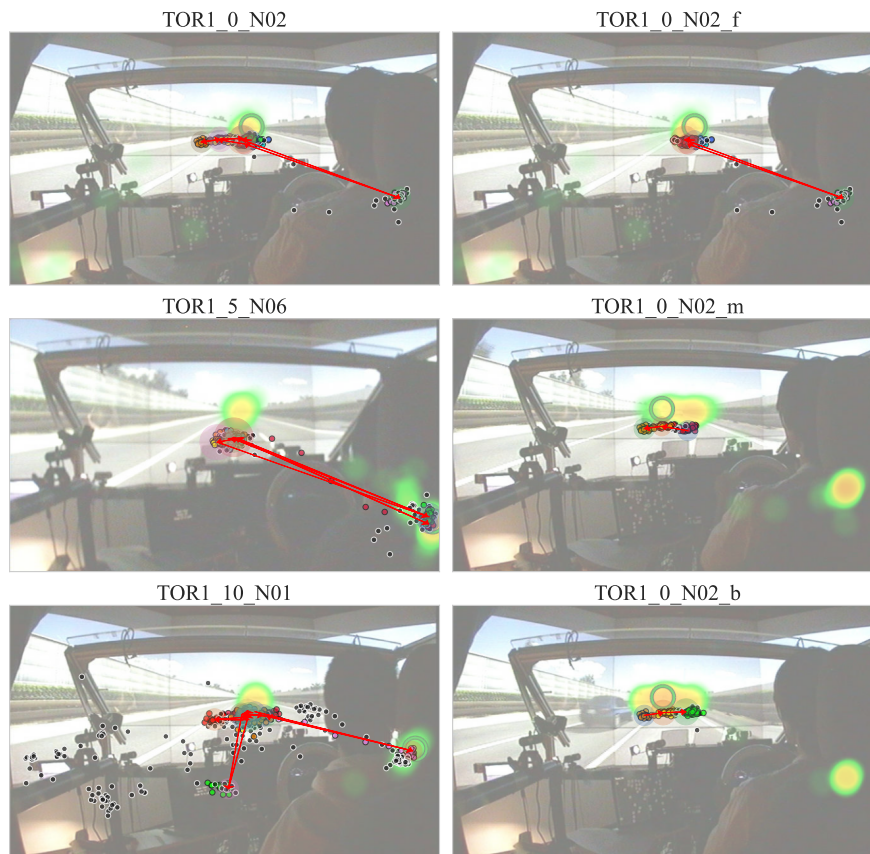
### Safe Takeover Behaviors

Even though the specific maneuvers and gaze behaviors of drivers differ from scenario to scenario, one thing they have in common is that if drivers would like to pass the slow-moving vehicles or the obstacles ahead, they have to change lanes to the right and then change back again at suitable time after passing the vehicles or obstacles. During this process, discreet drivers would slow down and look at the mirrors before they can safely change lanes. As an example, typical eye gazes of drivers in scenario TOR1 under each DoM were plotted in Fig. 5.1-Left, and subprocesses of the first case were plotted in Fig. 5.1-Right. It can be observed that, before changing lanes to the right, drivers would look at the right mirror first to figure out the traffic situations behind (1st subfigure in Fig. 5.1-Right), and then divert their attention toward the roadway to carry on the maneuvers (2nd and 3rd subfigures in Fig. 5.1-Right). And with longer DoM, eye gazes were more allocated to right mirror or other areas of interests. At the same time, eye gazes were also more dispersed, especially for DoM<sub>L</sub>, which were in alignment with the results in [127].

As defined in ISO 15007:2020 [156], fixation refers to “short temporal holds of movements that keep alignment of the eyes to a particular point within an AOI which falls on the fovea for a given time period”. Points of different colors in Fig. 5.1 represent different fixations, and the bigger circles represent the center of the fixations, with size of the circles representing the length of duration and arrows pointing the moving directions. Typically, individual fixations last from 100 ms to 2000 ms [156]. Herein, we set the threshold to be 300 ms (default value of MAPPS), i.e., if a gaze point directed toward a certain area is less than 300 ms, it is regarded as an isolated point, otherwise, it is considered as belonging to a fixation. In the example in Fig. 5.1, for eye gazes toward right mirror, a fixation of 600 ms was recorded for DoM<sub>S</sub>, two fixations of 433 ms and 499 ms were recorded for DoM<sub>M</sub>, and a fixation of 433 ms was recorded for DoM<sub>L</sub>. For DoM<sub>L</sub>, an additional fixation of 484 ms toward the infotainment system was recorded, demonstrating the inattentiveness of drivers when DoM became too long.

From the data collected, we could find that, no matter how long the DoM is, for the takeovers to be safe, it is important that drivers slow down and take a look at the mirrors in proper ways (at proper time and for proper durations of time) before taking further actions. In a study in [158], it was found that most of the mirror-checking actions had durations between 0.5 and 1 s during normal driving. And lane change maneuvers induced significantly higher number of





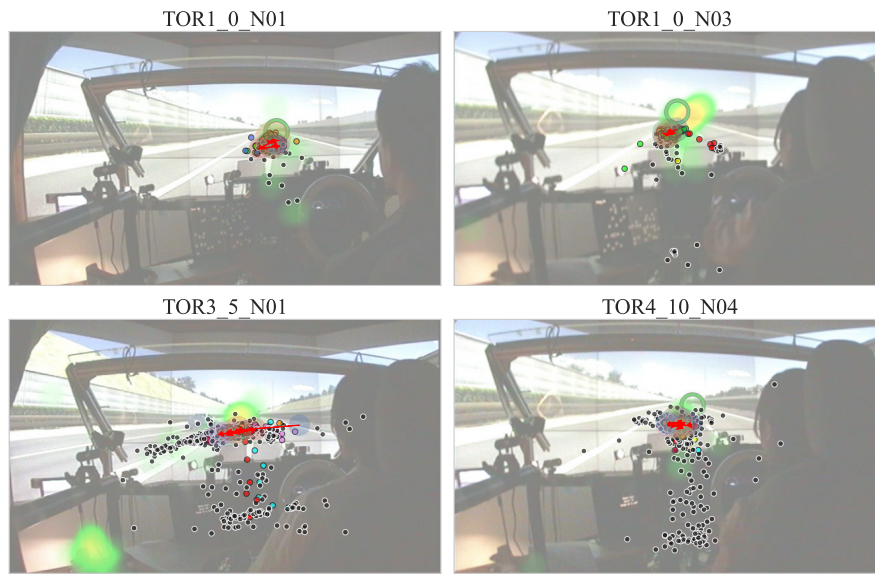
**Figure 5.1:** Left: Eye gazes of drivers in scenario TOR1 under each DoM. From above to below to right were DoM<sub>S</sub>, DoM<sub>M</sub> and DoM<sub>L</sub> respectively. Different colors represent different fixations, and the bigger circles represent the center of the fixations. Right: Subprocesses of eye gazes of drivers with safe takeover behaviors. From above to below are eye gazes before, during and after lane change.

mirror checking actions and longer mirror-checking durations than turn and straight maneuvers. Besides, it was indicated in [44] that total eyes-off-road durations of greater than 2 seconds significantly increased individual crash risks. Failing to cater to these conditions may lead to high probability of crashes or near-crashes. By analysing the crashes and near-crashes, five typical unsafe takeover behaviors were recognized. And below were elaborations of these unsafe behaviors and how they could be helpful in HMI design to enhance comfort and safety of automated vehicles.

### C1—Neglect of Mirror

Of the 15 crashes recorded, 4 crashes (2 for TOR1, 1 for TOR3 and 1 for TOR4, respectively) happened because drivers forgot to glance at the mirrors before changing lanes, leading to rear-end collisions with the vehicle in the right lane



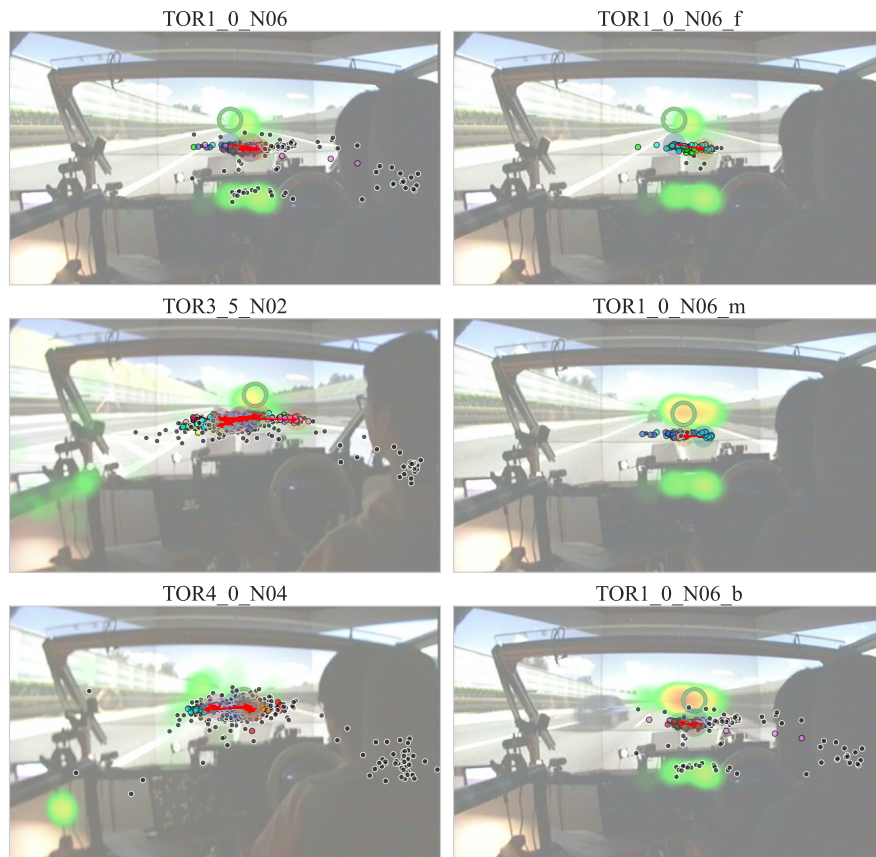


**Figure 5.2:** Typical eye gazes of drivers with C1 unsafe behaviors.

behind. And eye gazes of the 4 cases before and after the lane change were plotted in Fig. 5.2. We can see that, irrespective of the length of the time left for drivers to observe the surroundings ( $DoM_S$ ,  $DoM_M$  and  $DoM_L$  for the first and last two subfigures, respectively), drivers in the four scenarios still forgot to check the traffic situations in the right lane behind before taking any actions (e.g., change lanes), making the crashes unavoidable. Moreover, we could also notice that with longer  $DoM$ , drivers' eye gazes became more dispersed. This is beneficial for mastering more information surrounding the ego vehicle. However, if the eye gazes are not allocated in a reasonable way, it may also incur safety concerns, especially for non-experienced drivers. Neglect of information in the mirrors belongs to this category. In other situations, drivers become so concentrated around certain areas that they neglect some of the most vital information ahead, even when they have no intention to change lanes. Such kind of safety concerns were classified as C4 unsafe behaviors, which would be discussed below.

### C2—Use of Mirror at Improper Time

In some cases, drivers indeed glanced at the mirrors while changing lanes, however, since either the duration or timing was not quite appropriate, crashes sometimes happened under such circumstances. Of the 15 crashes, 3 such cases (1 for TOR1, 1 for TOR3 and 1 for TOR4, respectively) were recorded, and eye gazes before and after the lane change maneuvers were plotted in Fig. 5.3-Left. To take a closer look, the process was broken down into 3 subprocesses, i.e., before, during and after lane change, with eye gazes of the first case plotted in

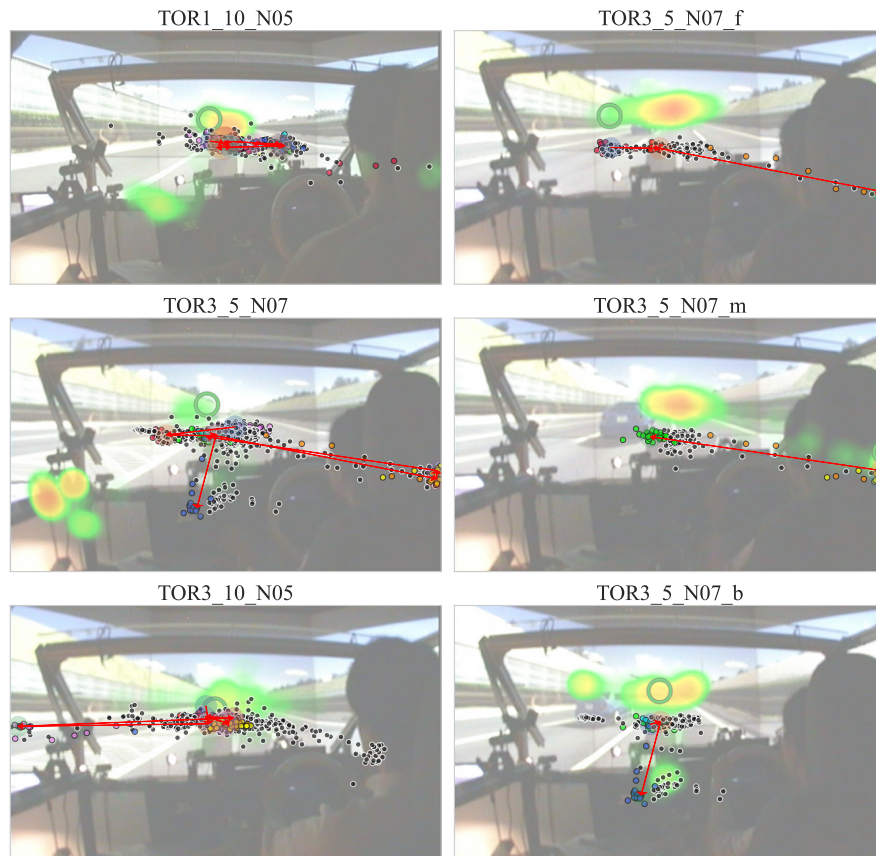


**Figure 5.3:** Left: Typical eye gazes of drivers with C2 unsafe behaviors. Right: Subprocesses of eye gazes of drivers with C2 unsafe behaviors. From above to below are eye gazes before, during and after lane change.

Fig. 5.3-Right. We can see that before and during the lane change process, the driver did not take a look at the mirrors at all. Only when the lane change was about to be accomplished did the driver take several quick glances at the right mirror. At this point, it is already too late to avoid any potential crashes. This is in direct contrast with that in Fig. 5.1-Right.

### C3—Improper Judgment

Sometimes, even when drivers glanced at the mirrors before taking any actions, crashes may also happen because of their improper judgements of the traffic situations at the time. For example, drivers may choose to change lanes when they actually should not, leading to rear-end collisions. Such is the case with five of the crashes (2 for TOR1 and 3 for TOR3, respectively). Correspondingly, typical eye gazes before and after the lane change maneuvers were plotted in Fig. 5.4-Left. Similarly, subprocesses of eye gazes of the second case were plotted in Fig. 5.4-Right. We could observe that even though the driver has noticed the vehicle behind, s/he failed to discern the speed of the vehicle correctly and chose



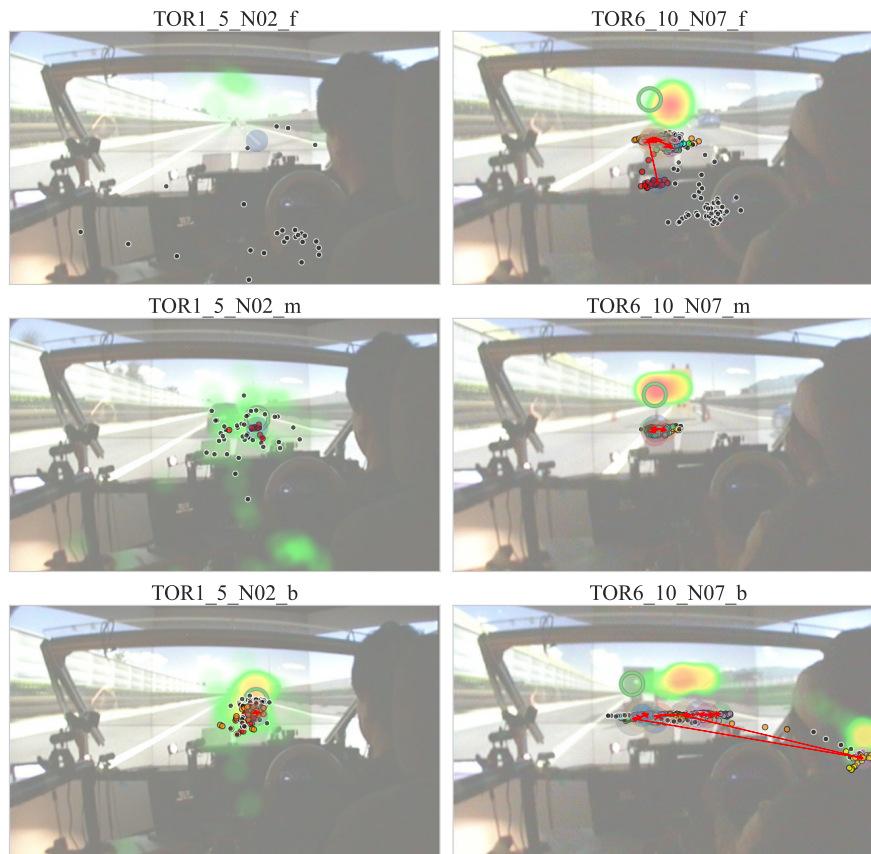
**Figure 5.4:** Left: Typical eye gazes of drivers with C3 unsafe behaviors. Right: Subprocesses of eye gazes of drivers with C3 unsafe behaviors. From above to below are eye gazes before, during and after lane change.

to change lanes immediately, resulting in rear-end collisions. Hence, it would be helpful if the driver could be reminded of the current situations, so that s/he could undertake safer maneuvers to avoid potential crashes.

#### C4—Improper Attention Allocation

As briefly mentioned in the above section, if drivers' eye gazes are not allocated reasonably, there might be safety concerns. It may be either because the eye gazes are too concentrated or too dispersed, or due to the improper timing of the eye gazes directed to certain areas. And out of the 15 crashes, 2 such crashes (1 for TOR1 and 1 for TOR6) were recorded. In the first case (Fig. 5.5-Left), it seemed that the driver's eyes-off-road glances were too long before colliding with the vehicle ahead. Three seconds before the crash, driver's eye gazes were still around speedometer and right side of the road (1st subfigure). It was not until about one second before the collision did the driver divert his attention back to the roadway (2nd subfigure), when it was already too late to avoid the crash. And after the collision, driver's eye gazes became more concentrated on





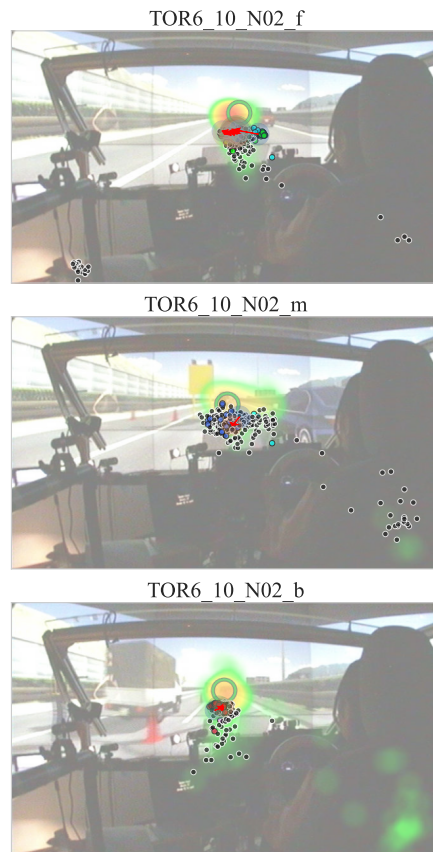
**Figure 5.5:** Left (Case 1): Subprocesses of eye gazes of drivers with C4 unsafe behaviors before and after colliding with the vehicle. Right (Case 2): Subprocesses of eye gazes of drivers with C4 unsafe behaviors before and after colliding with the obstacles.

the center of the road (3rd subfigure), exhibiting the driver's nonalertness from the TOR to the collision.

In the second case, different from the first one, it seemed that the driver has been too concentrated on the center of the road, so as to neglect the surrounding traffic situations. As can be seen from Fig. 5.5-Right, no matter 10 seconds (1st subfigure) or 3 seconds (2nd subfigure) before the collision, the driver has always focused on the roadway and had no intention of changing lanes. When the driver figured out the situation, the safety cones have already been too near to be avoided. Since the right lane has already been occupied, the driver has no choice but to change to the emergency lane. From that point on, the driver started to check the right mirror in order to safely get back to the right lane (3rd subfigure).

### C5—Improper Vehicle Speed

The last crash (TOR6) and most of the near-crashes could be attributed to too fast vehicle speeds. In this case, the driver has noticed the vehicles in the right



**Figure 5.6:** Subprocesses of eye gazes of drivers with C5 unsafe behaviors before and after colliding with the vehicle ahead.

lane and also the obstacles ahead. However, since the vehicle speed was too fast, the driver was not able to gain enough time to safely finish the lane change maneuvers and resulted a rear-end collision with the vehicle in the right lane ahead. As can be seen from Fig. 5.6, before changing lanes, the driver indeed glanced at the right mirror (1st subfigure) briefly and confirmed that the distance with the vehicle behind was enough to make a safe lane change. What she forgot to do was to slow down the vehicle so that the distance with the vehicle ahead would also be safe enough. During the lane change, the driver checked a few more times in the right mirror (2nd subfigure), however, collision with the vehicle ahead was already unavoidable. Therefore, it is important to remind the driver to slow down if s/he is too nervous to remember to step on the brake pedal or even step on the wrong pedal to avoid such kind of crashes and near-crashes.

### 5.2.3 Implication for Safe Takeovers

For one thing, it appears that even if we have given drivers more time to observe the surroundings ( $DoM_M$  and  $DoM_L$ ), the number of crashes has not decreased

as expected (4, 6 and 5 crashes for DoM<sub>S</sub>, DoM<sub>M</sub> and DoM<sub>L</sub>, and 4, 2 and 3 near-crashes for DoM<sub>S</sub>, DoM<sub>M</sub> and DoM<sub>L</sub>, respectively). What is more important seems to be drivers' maneuvers and gaze behaviors shortly after the TOR, where drivers' attentions should be allocated reasonably so that enough information surrounding the ego vehicle could be obtained and speed of the vehicle could be well controlled.

For another, although we have classified the unsafe behaviors into five categories and allocated the crashes correspondingly, it is not to say that the allocated crashes were solely due to the corresponding reasons. For example, for crashes allocated to C4 unsafe behaviors, if the vehicle speeds were not so fast, drivers would still have enough time to brake the vehicle and avoid the frontal collisions. And for crashes allocated to C2, we can also say that drivers' eye gazes were not allocated reasonably so as to forgot to allocate attention to the right mirror before changing lanes. We can simply understand them as the main reasons that led to the crashes and utilize them to guide the HMI design to increase safety of the automated vehicles. Moreover, there may exist other reasons that have possibly caused the crashes, such as the looked-but-failed-to-see events discussed by [159]. Since these events were not observed in the experiment, they were not discussed in this analysis. However, considering the possibilities, we may also utilize them to guide the HMI designs.

Overall, in order for the takeovers to be safe enough, good cooperation between drivers' gaze behavior and maneuvers (braking, acceleration, and steering) would be essential. Otherwise, the risk of colliding with other vehicles or obstacles would be unavoidable. It seemed that in emergent situations that required takeovers, some drivers tended to have difficulty in allocating attentions reasonably, which appeared to have less to do with the time left for drivers to observe the surroundings. Some had difficulty in controlling vehicle speeds while concentrating on surrounding traffic situations (C5); some forgot to take a look at the mirrors before changing lanes (C1); some failed to allocate eye gazes reasonably so as to glance at certain areas too late (C2) or stare at certain areas for too short or too long (C4); and some cannot make proper judgements either because the drivers were too nervous or the time left was too short (C4).

To address the above safety concerns, we may as well consider the following while designing the HMI of conditionally automated vehicles to improve comfort of drivers as well as safety of takeovers in emergent situations:

- Remind drivers to slow down if they forget to step on the brake pedal when they should, and to make matters worse, when they step on the gas pedal by mistake (C5). In more emergent situations, activate the autonomous

emergency braking (AEB) system and notify the drivers when the system chooses to do so in order that drivers understand what the system is doing at the time.

- Remind drivers to take a look at the mirrors before changing lanes if they forget to do so and have the intention to change lanes directly (using light, sound, and vibration, etc.). This is especially useful when there are obstacles in front and vehicles approaching from behind at the same time. What is more important is that drivers should be reminded to use mirrors reasonably, so that mirrors can be used at proper time and for proper duration of time (C1 & C2).
- Remind drivers to take a glance at other areas of interests when they have stared at certain areas for too long when they should not, so that they can better keep track of the overall traffic situations and avoid any unnecessary crashes. A good approach would be to use head-up displays to guide eye gazes of drivers, so that drivers could understand when and where to look at when they are overly nervous and stressful (C4).
- Remind drivers the traffic situations that are critical to safety of the automated vehicles when they seem to overlook certain information or misunderstand the situations at the time, such as when drivers choose to change lanes when the vehicles in the right lane behind are approaching in fast speeds. Some drivers are prone to make such mistakes under stressful situations (C4).

## Chapter 6

# Conclusions

### 6.1 Key Findings

The main goal of this research was to model drivers' takeover behaviors by integrating a variety of factors, so that drivers' takeover behaviors could be predicted in advance and systems could adapt their strategies accordingly to ensure safe takeovers. To achieve this goal, Objective 1–Objective 4 were proposed in Section 1.2, which corresponded to the contents in Chapter 2–Chapter 5, respectively. And below are the methods and key findings in each chapter.

1. In Chapter 2, a thorough literature review was conducted by classifying the factors that influence drivers' takeover performance into system-, scenario-, and human-related factors, along with a complete framework for evaluation and prediction of takeover behaviors. Although a lot of factors have already been researched, it was found out that impacts of the time interval before the TOR (denoted as DoM) and drivers' personality seemed to receive little attention. Besides, effects of scenarios have also not been analyzed good enough.
2. In Chapter 3, an experiment was designed to collect drivers' takeover behaviors under different conditions, along with analysis of the impacts of DoM and personality on takeover performance. Results showed that:
  - Impacts of DoM was indeed statistically significant, however, the relationship was more approximate to parabolic than linear, with the optimal interval obtained between 5–7 s;
  - Effects of different aspects of personality traits as obtained by the IPIP-NEO-120 inventory seemed to affect drivers' takeover performance in different ways. Specifically, extraversion and openness mainly affects takeover time; neuroticism mainly affects longitudinal performance; and agreeableness mainly affects lateral performance.



3. In Chapter 4, based on the results of Chapters 2 and 3, 38 features were identified in the dataset, and drivers' takeover behaviors were modeled in three different aspects, including takeover time, takeover readiness, and takeover style.
  - Takeover time and takeover readiness were modeled as regression and classification problems, respectively. Among all the models trained and tested, XGBoost model yielded the best performance in both of the problems. Specifically, XGBoost regressor obtained a mean absolute error of 0.489 s and a  $R^2$  score of 0.502; and XGBoost classifier obtained an accuracy of 96.3% and a recall of 96.9%.
  - The most important features in predicting takeover time and readiness were also analyzed using SHAP explainer. Results showed that, for XGBoost regressor, drivers' gaze behaviors, personality traits, and scenario characteristics seemed to be more important compared with other features; and for XGBoost classifier, scenario characteristics, drivers' personality and reaction speed, and type of the road seemed to be more important compared with other factors.
  - Takeover styles was modeled based on drivers' actual maneuvers, since the maneuvers were time series data, this problem was constructed as a clustering problem. Results showed that irrespective of DoM and scenarios, three distinctive takeover styles could be similarly identified, which could be utilized in personalizing system behaviors when takeovers during automated driving is necessary.
4. In Chapter 5, some failure cases during the experiments were analyzed, and five patterns of unsafe behaviors were recognized that may have caused the crashes. Correspondingly, several suggestions were put forward, which we believe can be utilized for more advanced HMI design and to mitigate these patterns of crashes.

## 6.2 Main Contributions

Overall, the main contributions of this research are as follows:

1. A systematic literature review was conducted regarding factors that affect drivers' takeover behaviors, along with a complete architecture for evaluation and prediction of drivers' takeover behaviors, indicating the challenges and gaps exist in the previous studies.

2. Two new factors (including duration of monitoring and drivers' personality) impacting drivers' takeover performance that have not been studied in the previous studies were identified, with an in-depth insight into eye tracking data before and after takeovers, and the results could be referenced to [127] and [160], respectively.
3. Drivers' takeover behaviors were modeled in three aspects, including takeover time, takeover readiness and takeover style. Although models regarding takeover time were quite common in the literature, takeover readiness and takeover style have not been well modeled in the previous studies, which could be used as a benchmark in the latter studies, and results regarding takeover style could be referenced to [130].
4. Failure cases (crashes and near-crashes) during the experiments were extracted and analyzed, with five patterns of unsafe behaviors identified, which we believe could be utilized to obtain some guidance on HMI design to reduce potential crashes as far as possible, and the results could be referenced to [154].

The application of this research has broad implications for the development and enhancement of automated driving systems. Here are some potential applications:

1. Improved Design of Automated Driving Systems

- This research provides a systematic understanding of the factors influencing drivers' takeover behaviors, offering valuable insights for designing more effective automated driving systems.
- The identified features and modeling techniques can be directly applied to enhance the prediction of takeover time, readiness, and maneuver styles, contributing to the development of more reliable and adaptive automated driving systems.
- **For example:** If the predicted takeover quality is too low, or the predicted takeover time exceeds the safety limit, ADS can warn the driver to engage less in NDRT.

2. Enhancement of Human-Machine Interface (HMI)

- The suggestions derived from the analysis of crashes and near-crashes data in chapter 5 offer practical guidance for designing advanced HMIs.

- By incorporating the findings into HMI design, manufacturers and developers can work towards mitigating patterns of unsafe behaviors and improving overall system safety.
- **For example:** When the driver has stared at certain areas for too long, head-up displays can be utilized to guide drivers' gazes.

### 3. Real-World Implementation and Testing

- The predictive models and clustering techniques developed in this research can be integrated into real-world automated driving systems for practical testing and validation.
- By implementing the models in field trials, researchers and industry professionals can assess their effectiveness in real-world scenarios and further refine the technology.
- **For example:** Instead of specific scenarios, considering the characteristics of the scenarios would help to assess effectiveness of the models in real-world scenarios more effectively.

### 4. Regulatory and Policy Considerations

- The research results, especially those related to the factors affecting takeover performance, could influence the development of regulations and policies for automated vehicles.
- Policymakers may use the findings to establish guidelines for system manufacturers, encouraging the implementation of features that enhance takeover predictability and safety.
- **For example:** The results of this research can be utilized to establish guidelines for features to be included while designing the automated driving systems to enhance takeover predictability and safety.

## 6.3 Future Works

From academic and industrial point of view, this research can still be improved in a few ways.

1. The participants recruited in the experiments have been concentrated mainly on young drivers less than 30 years old, limiting the generality of the conclusions to other age groups. This could be complemented in the following stages of the research by recruiting a comparable number

of participants from other age groups, especially the middle-age (40–50 years old) and old-age groups (60–70 years old), which requires long-term studies.

2. Another limitation is related with the methods for measuring personality. This research has utilized the 120-item IPIP-NEO inventory, which is easy and convenient to yield the five factor model. However, some other tests may also be available, such as the California Psychological Inventory and the Sixteen Personality Factor Questionnaires. Despite their differences in the number of factors, these tests are actually related. To further investigate the influence of certain aspects of personality, it might be preferable to assess it by several tests together.
3. The driving simulator is a moving base simulator, however, the scenarios may still be different from real vehicles and may cause some deviations. For real applications, and also to verify the conclusions, experiments on real vehicles are essential. Besides, the scenarios designed in the experiments have been mostly high-way scenarios. To generalize the conclusions to city road, experiments involving urban scenarios would also be necessary.
4. Compared with other studies, drivers' personality and more scenario-related factors have been involved in modeling takeover behaviors, whereas heart-rate indices and galvanic skin response indices have not been taken into consideration. For industrial applications, incorporating these signals may not be quite realistic at present. However, for academic research, it would be a good way to enhance predictability of the models.
5. It is to be noted that the XGBoost model utilized in modeling takeover time and readiness reveals only statistical relationships between the independent and dependent variables. To obtain causal relationships between them, other more advanced methods would be necessary, e.g., probabilistic graphical model. Besides, in this research, drivers' takeover styles were only analyzed after the data have been collected, however, for real time applications, online identification of drivers' takeover style would be necessary, which would be perfected in the future studies.

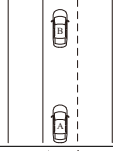

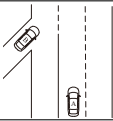
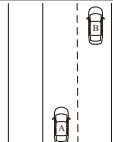
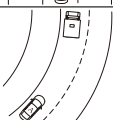
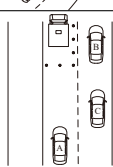
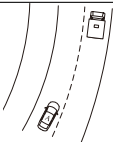
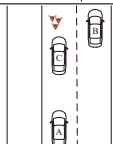
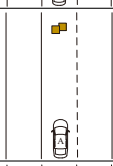
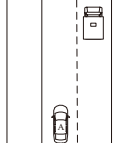
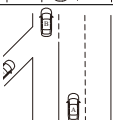
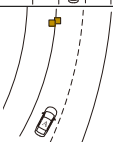
## Appendix A

# Experiment Design

### A.1 Description of Takeover Scenarios

1. The vehicle B ahead brakes suddenly from 80 km/h to 30 km/h.
2. Cargoes drop down from the truck ahead, and the truck slows down to pick them up.
3. The vehicle B in the acceleration lane merges in from the left side suddenly while the ego vehicle is approaching.
4. The Vehicle B ahead cuts in from the right side suddenly while the ego vehicle is approaching.
5. Thick fog arises and a slow-moving truck gradually becomes clear while the ego vehicle is approaching.
6. The ego vehicle approaches the construction site while the other vehicles are passing by slowly.
7. The truck ahead cuts in from the right side at a relatively far distance while the ego vehicle is approaching (compared with TOR4).
8. Several deer walk slowly in the middle of the road while the ego vehicle is approaching, and the other vehicles slow down while passing by.
9. Obstacles are detected in the middle of the lane while the ego vehicle is approaching.
10. A truck in the right lane ahead starts up suddenly while the ego vehicle is approaching.
11. The vehicle B in the acceleration lane merges in from the left side at a relatively far distance while the ego vehicle is approaching (compared with TOR3), and then moves at a relatively low speed.
12. Obstacles are detected by the side of the lane while the ego vehicle is approaching.

**Table A.1:** Description of the takeover scenarios.

Scenarios	Number	DoM	LeadT	Schematic	Description
Sudden braking	TOR1	S/M/L	S		1
Dropped cargo	TOR2	S/M/L	S		2
Near merging (left)	TOR3	S/M/L	S		3
Near cut-in (right)	TOR4	S/M/L	S		4
Thick fog	TOR5	S/M/L	S		5
Construction site	TOR6	S/M/L	S		6
Far cut-in (right)	TOR7	S/M/L	L		7
Animals	TOR8	S/M/L	L		8
Obstacles (lane middle)	TOR9	S/M/L	L		9
Sudden startup	TOR10	LL	L		10
Far merging (left)	TOR11	LL	L		11
Obstacles (lane side)	TOR12	LL	L		12

## A.2 Experimental Arrangements

Table A.2: Summary of the experimental arrangements.

		Block1				Block2				Block3			
N01	TOR	1	10	2	7	3	8	4	11	9	5	12	6
	DoM	S	LL	M	L	L	S	M	LL	M	L	LL	S
N02	TOR	2	11	3	8	4	9	5	12	7	6	10	1
	DoM	S	LL	M	L	L	S	M	LL	M	L	LL	S
N03	TOR	3	12	4	9	5	7	6	10	8	1	11	2
	DoM	S	LL	M	L	L	S'	M	LL	M	L	LL	S
N04	TOR	4	10	5	7	6	8	1	11	9	2	12	3
	DoM	S	LL	M	L	L	S	M	LL	M	L	LL	S
N05	TOR	5	11	6	8	1	9	2	12	7	3	10	4
	DoM	S	LL	M	L	L	S	M	LL	M	L	LL	S
N06	TOR	6	12	1	9	2	7	3	10	8	4	11	5
	DoM	S	LL	M	L	L	S	M	LL	M	L	LL	S
N07	TOR	1	10	2	7	3	8	4	11	9	5	12	6
	DoM	M	LL	L	S	S	M	L	LL	L	S	LL	M
N08	TOR	2	11	3	8	4	9	5	12	7	6	10	1
	DoM	M	LL	L	S	S	M	L	LL	L	S	LL	M
N09	TOR	3	12	4	9	5	7	6	10	8	1	11	2
	DoM	M	LL	L	S	S	M	L	LL	L	S	LL	M
N10	TOR	4	10	5	7	6	8	1	11	9	2	12	3
	DoM	M	LL	L	S	S	M	L	LL	L	S	LL	M
N11	TOR	5	11	6	8	1	9	2	12	7	3	10	4
	DoM	M	LL	L	S	S	M	L	LL	L	S	LL	M
N12	TOR	6	12	1	9	2	7	3	10	8	4	11	5
	DoM	M	LL	L	S	S	M	L	LL	L	S	LL	M
N13	TOR	1	10	2	7	3	8	4	11	9	5	12	6
	DoM	L	LL	S	M	M	L	S	LL	S	M	LL	L
N14	TOR	2	11	3	8	4	9	5	12	7	6	10	1
	DoM	L	LL	S	M	M	L	S	LL	S	M	LL	L
N15	TOR	3	12	4	9	5	7	6	10	8	1	11	2
	DoM	L	LL	S	M	M	L	S	LL	S	M	LL	L
N16	TOR	4	10	5	7	6	8	1	11	9	2	12	3
	DoM	L	LL	S	M	M	L	S	LL	S	M	LL	L
N17	TOR	5	11	6	8	1	9	2	12	7	3	10	4
	DoM	L	LL	S	M	M	L	S	LL	S	M	LL	L
N18	TOR	6	12	1	9	2	7	3	10	8	4	11	5
	DoM	L	LL	S	M	M	L	S	LL	S	M	LL	L

<sup>1</sup> N + number represent the number of the participants.

<sup>2</sup> TOR + number i represent the ith takeover scenario.

<sup>3</sup> DoM represents duration of monitoring, with S, M, L, and LL represent short, medium, long, and very long, respectively.

### A.3 Pre-Experiment Questionnaire

1. Number of Participants: Pre-assigned, e.g. G2N01
2. Age:
3. Gender: Male or female
4. Years of Driving:
5. Frequency of Driving (Days of driving every month on every average):
6. Type of Driving
  - ☐ Driving on the city road
  - ☐ Driving on the highway
  - ☐ Short distance driving ( $< 50$  km)
  - ☐ Medium distance driving ( $50 \sim 200$  km)
  - ☐ Long distance driving ( $> 200$  km)
7. Experience with driver-assistance systems. If Yes:
  - ☐ Adaptive Cruise Control (ACC)
  - ☐ Anti-lock Braking System (ABS)
  - ☐ Autonomous Emergency Braking (AEB)
  - ☐ Automated Parking System (APS)
  - ☐ Head Up Display (HUD)
  - ☐ Driver Monitoring System (DMS)
  - ☐ Blind Spot Detection (BSD)
  - ☐ Collision Avoidance System (CAS)
  - ☐ Forward Collision Warning (FCW)
  - ☐ Crosswind Stabilization
  - ☐ Electronic Stability Control (ESC)
  - ☐ Track Control System (TCS)
  - ☐ Hill Descent Control
  - ☐ Hill Start Assistance
  - ☐ Lane Departure Warning (LDW)
  - ☐ Lane Change Assistance (LCA)
  - ☐ Lane Keeping Assistance (LKA)
  - ☐ Others



8. Experience with automated driving systems. If Yes:

- ☐ Driving Simulator
- ☐ Experimental Automated Vehicles
- ☐ Private Vehicles

9. Whether having a regular sleep or not:

10. Hours of sleeping every day on average:

## A.4 Modified NASA-TLX After Each Scenario

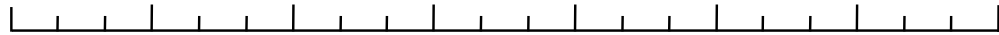
Mental Demand 精神的な要求



Low 低い

High 高い

Physical Demand 身体的な要求



Low 低い

High 高い

Temporal Demand 時間的な要求



Low 低い

High 高い

Predictability 予測性



Easy 簡単

Difficult 難しい

Criticality クリティカル・リスク



Low 低い

High 高い

Performance パフォーマンス



Good 良い

Poor 悪い

Effort 努力



Low 低い

High 高い

Frustration 不満



Low 低い

High 高い

**1. Mental Demand**

How much mental/perceptual activity was required? (Thinking, deciding, calculating, remembering, viewing, searching, etc.)

Was there a lot of demands or was it easy? Was it Simple or complex? Was it coercive or lenient?

**2. Physical Demand**

How much physical activity did you require? (push, pull, turn, control, move, etc.)

Was it Simple or complex? Was it tough or easy? Was it slow or active? Is it loose or intense? Was it easy or difficult?

**3. Temporal Demand**

How much time pressure did you feel regarding the progress, timing of tasks, and speed of progress?

Did you do it slowly and relaxingly? Or did you do it in a hurry and frantically?

**4. Predictability**

How predictable is it? Is it easy or difficult to predict? (The occurrence of an event, the direction and speed of movement of people and things around you, etc.)

**5. Criticality**

How safety critical is it? Is there a high risk or danger if we do not take over in a timely manner?

**6. Performance**

How well do you think you have performed in accomplishing the required task and achieving the goals?

**7. Effort**

How hard did you work to accomplish the required task at your level of ability? (mentally and physically)

**8. Frustration**

How did you feel while performing the task, such as anxiety, disappointment, irritability, stress, or satisfaction and relaxation?



## Appendix B

# Data Processing

## B.1 SuRT Data Processing

**Table B.1:** An excerpt of SuRT raw data.

UTC_datetime	is_left	select_left	key	correct	wrong
20220331_105019.108351	TRUE	TRUE	enter	1	0
20220331_105020.575714	FALSE	TRUE	right	1	0
20220331_105021.009493	FALSE	FALSE	enter	2	0
20220331_105021.859022	FALSE	FALSE	enter	3	0
20220331_105022.636263	FALSE	FALSE	enter	4	0
20220331_105023.487726	FALSE	FALSE	enter	5	0
20220331_105024.761830	TRUE	FALSE	left	5	0
20220331_105025.074507	TRUE	TRUE	enter	6	0
20220331_105026.042353	TRUE	TRUE	enter	7	0
20220331_105026.951141	FALSE	TRUE	right	7	0
20220331_105027.212473	FALSE	FALSE	enter	8	0
20220331_105028.705191	TRUE	FALSE	left	8	0
20220331_105028.920590	TRUE	TRUE	enter	9	0
20220331_105029.674851	FALSE	TRUE	right	9	0
20220331_105029.954518	FALSE	FALSE	enter	10	0
20220331_105031.965873	TRUE	FALSE	left	10	0
20220331_105032.277178	TRUE	TRUE	enter	11	0
20220331_105033.717110	TRUE	TRUE	enter	12	0
20220331_105034.528439	TRUE	TRUE	enter	13	0
20220331_105036.904758	TRUE	TRUE	enter	14	0
20220331_105037.680551	TRUE	TRUE	enter	15	0
20220331_105039.049063	FALSE	TRUE	right	15	0
20220331_105039.358535	FALSE	FALSE	enter	16	0
20220331_105040.510241	FALSE	FALSE	enter	17	0
20220331_105041.253634	FALSE	FALSE	enter	18	0
20220331_105042.233354	TRUE	FALSE	left	18	0
20220331_105042.502367	TRUE	TRUE	enter	19	0
20220331_105043.214919	FALSE	TRUE	right	19	0
20220331_105043.478853	FALSE	FALSE	enter	20	0
20220331_105044.179210	TRUE	FALSE	left	20	0

---

```

1 participants = ['G2N01', 'G2N02', 'G2N03', 'G2N04', 'G2N05', 'G2N06',
2               'G2N07', 'G2N08', 'G2N09', 'G2N10', 'G2N11', 'G2N12',
3               'G2N13', 'G2N14', 'G2N15', 'G2N16', 'G2N17', 'G2N18',
4               'G3N09', 'G3N10', 'G3N11', 'G4N01', 'G4N02', 'G4N03',
5               'G4N04', 'G4N05', 'G4N06', 'G4N07', 'G4N08', 'G4N09',
6               'G4N10', 'G4N11', 'G4N12', 'G4N13', 'G4N14', 'G4N15',
7               'G5N01', 'G5N02', 'G5N03', 'G5N04', 'G5N05', 'G5N06',
8               'G5N07', 'G5N08', 'G5N09', 'G5N10', 'G5N11', 'G5N12']
9
10 names = ['UNIX_sec', 'UTC_datetime', 'is_left', 'select_left', 'key', 'correct', 'wrong']
11 columns1 = ['num_correct', 'num_wrong', 'time', 'Spd', 'correct_rate']
12 columns2 = ['avgSpd', 'avgCorrectRate']
13
14 avgSpd_avgCorrectRate = []
15
16 for p in range(len(participants)):
17     df = []
18     for dirname, _, filenames in os.walk('SuRT-Data/' + participants[p]):
19         for filename in filenames:
20             df_temp = pd.read_csv(os.path.join(dirname, filename), sep=',',
21                                   names = names, skiprows = [0])
22             df.append(df_temp)
23
24     temp = []
25     for i in range(len(df)):
26         num_correct = df[i].tail(1)['correct'].values[0]
27         num_wrong = df[i].tail(1)['wrong'].values[0]
28
29         # calculate duration of the SuRT
30         time_start_h = float(df[i].head(1)['UTC_datetime'].values[0][9:11])
31         time_end_h = float(df[i].tail(1)['UTC_datetime'].values[0][9:11])
32         time_start_m = float(df[i].head(1)['UTC_datetime'].values[0][11:13])
33         time_end_m = float(df[i].tail(1)['UTC_datetime'].values[0][11:13])
34         time_start_s = float(df[i].head(1)['UTC_datetime'].values[0][13:])
35         time_end_s = float(df[i].tail(1)['UTC_datetime'].values[0][13:])
36
37         hours = time_end_h - time_start_h
38         minutes = time_end_m - time_start_m
39         seconds = time_end_s - time_start_s
40         time = hours * 3600 + minutes * 60 + seconds
41         spd = time / (num_correct + num_wrong)
42         correct_rate = num_correct / (num_correct + num_wrong)
43
44         temp.append([num_correct, num_wrong, time, spd, correct_rate])
45
46     df_temp = pd.DataFrame(temp, columns = columns1)
47
48     total_time = df_temp['time'].sum()
49     total_number = df_temp['num_correct'].sum() + df_temp['num_wrong'].sum()
50     avgSpd = total_time / total_number
51     avgCorrectRate = df_temp['num_correct'].sum() / total_number
52
53     avgSpd_avgCorrectRate.append([avgSpd, avgCorrectRate])
54
55 avgSpd_avgCorrectRate = pd.DataFrame(avgSpd_avgCorrectRate, columns = columns2)

```

---

## B.2 NASA-TLX Data Processing

**Table B.2:** An excerpt of NASA-TLX raw data.

Scenario	Mental	Physical	Temporal	Predictability	Criticality
TOR1	14	3	15	18	18
TOR2	6	5	8	6	9
TOR3	12	3	12	6	12
TOR4	12	6	12	15	15
TOR5	12	9	12	15	15
TOR6	9	8	13	6	11
TOR7	9	6	9	9	12
TOR8	15	6	15	18	15
TOR9	6	7	4	4	14
TOR10	3	3	3	3	6
TOR11	3	3	3	3	3
TOR12	7	6	4	3	4

```

1 participants = ['G4N01', 'G4N02', 'G4N03', 'G4N04', 'G4N05', 'G4N06', 'G4N07',
2               'G4N08', 'G4N09', 'G4N10', 'G4N11', 'G4N12', 'G4N13', 'G4N14',
3               'G4N15', 'G5N01', 'G5N02', 'G5N03', 'G5N04', 'G5N05', 'G5N06',
4               'G5N07', 'G5N08', 'G5N09', 'G5N10', 'G5N11', 'G5N12']
5
6 names = ['Scenario', 'Mental Demand', 'Physical Demand', 'Temporal Demand',
7          'Predictability', 'Criticality', 'Performance', 'Effort', 'Frustration']
8
9 df = []
10 for dirname, _, filenames in os.walk('TLX-Data/'):
11     for filename in filenames:
12         df_temp = pd.read_excel(os.path.join(dirname, filename), names = names)
13         df.append(df_temp)
14
15 TOR_scores = []
16 for TOR in range(12):                                # 12 takeover scenarios
17     temp = pd.DataFrame()
18     for participant in range(len(df)):                  # 48 participants
19         data_temp = df[participant].iloc[TOR]
20         temp = pd.concat([temp, data_temp], axis = 1)
21     TOR_scores.append(temp.transpose())
22
23 avgScores = pd.DataFrame()
24 for i in range(len(TOR_scores)):
25     avgScores = pd.concat([avgScores, TOR_scores[i].mean(axis=0)], axis = 1)
26
27 avgScores.columns = ['TOR1', 'TOR2', 'TOR3', 'TOR4', 'TOR5', 'TOR6',
28                     'TOR7', 'TOR8', 'TOR9', 'TOR10', 'TOR11', 'TOR12']

```

## B.3 Eye Gazes Data Processing

The complete code can be downloaded from <https://github.com/c-huang-tty/PhDThesis/tree/main/Code>. Below is a reference of the APIs in `dataProcessing_EyeGazes.py`.

---

```

1  def EyesOnRoad():
2      """
3      Whether drivers' gazes are on road at the moment of takeover request.
4
5      Returns
6      -----
7      df : DataFrame
8
9      """
10
11  def excel2pkl():
12      """
13      Convert excel format to pickle format.
14
15      Returns
16      -----
17      None.
18
19      """
20
21  def ReadFile(participant, tor):
22      """
23      Parameters
24      -----
25      participant : str
26      tor : str
27
28      Returns
29      -----
30      data : DataFrame
31
32      """
33
34  def GazeCountAOIs(participant, tor, window_size):
35      """
36      Count gaze points in different AOIs given certain participant,
37      tor and window size.
38
39      Parameters
40      -----
41      participant : str
42      tor : str
43      window_size : int (15, 12, 9, 6, 3)
44
45      Returns
46      -----
47      AOI_gaze_count : DataFrame
48
49      """

```



```

50
51 def BlinkCount(window_size):
52     '''
53     Count the number of blinks given certain window size.
54
55     Parameters
56     -----
57     window_size : int
58
59     Returns
60     -----
61     df : DataFrame
62
63     '''
64
65 def FixationCount(window_size):
66     '''
67     Count the number of fixatios given certain window size.
68
69     Parameters
70     -----
71     window_size : int
72
73     Returns
74     -----
75     df : DataFrame
76
77     '''
78
79 def PupilDiameterEyelidOpening_Mean_Std_Amplitude(participant, tor, window_size):
80     '''
81     Calculate mean, std, and amplitude of pupil diameter and eyelid opening
82     given certain participant, tor, and window size.
83
84     Parameters
85     -----
86     participant : str
87     tor : str
88     window_size : int (15, 12, 9, 6, 3)
89
90     Returns
91     -----
92     pupil_diameter_eyelid_opening : numpy array
93
94     '''
95
96 def PercentageOfEyesOnRoad(window_size):
97     '''
98     Calculate percentage of eye gazes in different AOIs
99     given certain window size.
100
101     Parameters
102     -----
103     window_size : int (15, 12, 9, 6, 3)
104
105     Returns
106     -----

```

```
107         df : DataFrame
108
109     '''
110
111 def PupilDiameterEyelidOpening(window_size):
112     '''
113     Process pupil diameter and eyelid opening given certain window size.
114
115     Parameters
116     -----
117     window_size : int
118
119     Returns
120     -----
121     df : DataFrame
122
123     '''
124
125 def dataProcessGrid(window_size):
126     '''
127     Count number of eye gazes in grids of 8x8.
128
129     Parameters
130     -----
131     window_size : int
132
133     Returns
134     -----
135     df : array of arrays
136         - [Participant: str, TOR: str, GazeCount: Dataframe]
137
138     '''
139
140 def GazeEntropy(window_size):
141     '''
142     Calculate gaze entropy given certain window size.
143
144     Parameters
145     -----
146     window_size : int
147
148     Returns
149     -----
150     df : DataFrame
151
152     '''
```

---

# Bibliography

- [1] Road vehicles–human performance and state in the context of automated driving–part 1: Common underlying concepts. Standard ISO/TR 21959-1:2020, ISO, Geneva, CH, 2020.
- [2] C. Braunagel, D. Geisler, W. Rosenstiel, and E. Kasneci. Online recognition of driver-activity based on visual scanpath classification. *IEEE Intell. Transp. Syst. Mag.*, 9(4):23–36, 2017. doi: 10.1109/MITS.2017.2743171.
- [3] Road vehicles–human performance and state in the context of automated driving–part 2: Considerations in designing experiemnts to investigate transition process. Standard ISO/TR 21959-1:2020, ISO, Geneva, CH, 2020.
- [4] Christian Gold, Frederik Naujoks, Jonas Radlmayr, Hanna Bellem, and Oliver Jarosch. Testing scenarios for human factors research in level 3 automated vehicles. In Neville A Stanton, editor, *Advances in Human Aspects of Transportation*, pages 551–559, Cham, 2018. Springer International Publishing. doi: 10.1007/978-3-319-60441-1\_54.
- [5] S. Singh. Critical reasons for crashes investigated in the national motor vehicle crash causation survey. Technical report, NHTSA, Washington, DC, USA, February 2015.
- [6] Wikipedia. Death of elaine herzberg, June 2022. URL [https://en.wikipedia.org/wiki/Death\\_of\\_Elaine\\_Herzberg](https://en.wikipedia.org/wiki/Death_of_Elaine_Herzberg).
- [7] Road vehicles–functional safety–part 1: Vocabulary. Standard ISO 26262-1:2018, ISO, Geneva, CH, 2018.
- [8] Road vehicles–safety of the intended functionality. Standard ISO/PAS 21448:2019, ISO, Geneva, CH, 2019.
- [9] Road vehicles–cybersecurity engineering. Standard ISO/SAE 21434:2021, ISO, Geneva, CH, 2021.
- [10] IEEE standard for assumptions in safety-related models for automated driving systems. Standard IEEE 2846:2022, IEEE, 2022.

- [11] National Center for Statistics and Analysis. Traffic safety facts research note: Distracted driving 2019. Technical report, NHTSA, Washington, DC, April 2021.
- [12] M. Maurer, J. C. Gerdes, B. Lenz, and H. Winner. *Autonomous Driving: Technical, Legal and Social Aspects*. Springer, Berlin, Germany, 2016.
- [13] Taxonomy and definitions for terms related to driving automation systems for on-road motor vehicles. Standard ISO/SAE PAS 22736:2021, ISO, Geneva, CH, 2021.
- [14] Zhenji Lu, Riender Happee, Christopher D.D. Cabrall, Miltos Kyriakidis, and Joost C.F. de Winter. Human factors of transitions in automated driving: A general framework and literature survey. *Transp. Res. F: Traffic Psychol. Behav.*, 43:183–198, 2016. doi: 10.1016/j.trf.2016.10.007.
- [15] C. Huang, Y. Liu, L. Li, and Z. Chen. Safety oriented state transitions in level 3 automated driving systems: A general framework. In *2019 IEEE 4th Int. Conf. Adv. Robot. Mechatro. (ICARM)*, pages 918–923, 2019. doi: 10.1109/ICARM.2019.8833761.
- [16] Alexander Eriksson and Neville A. Stanton. Takeover time in highly automated vehicles: Noncritical transitions to and from manual control. *Hum. Factors*, 59(4):689–705, 2017. doi: 10.1177/001872081668583.
- [17] Mengxia Jin, Guangquan Lu, Facheng Chen, Xi Shi, Haitian Tan, and Junda Zhai. Modeling takeover behavior in level 3 automated driving via a structural equation model: Considering the mediating role of trust. *Accid. Anal. Prev.*, 157:106156, 2021. ISSN 0001-4575. doi: 10.1016/j.aap.2021.106156.
- [18] Soyeon Kim, René van Egmond, and Riender Happee. Effects of user interfaces on take-over performance: A review of the empirical evidence. *Information*, 12(4), 2021. doi: 10.3390/info12040162.
- [19] Frederik Naujoks, C. Mai, and A. Neukum. The effect of urgency of takeover requests during highly automated driving under distraction. In *Proceedings of the 5th International Conference on Applied Human Factors and Ergonomics AHFE 2014*, pages 2099–2106, 01 2014.
- [20] Ioannis Politis, Stephen Brewster, and Frank Pollick. 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*, AutomotiveUI '15, pages 3–10, New York, NY, USA, 2015. Association for Computing Machinery. ISBN 9781450337366. doi: 10.1145/2799250.2799262.
- [21] Sebastiaan Petermeijer, Fabian Doubek, and Joost de Winter. 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, 2017. doi: 10.1109/SMC.2017.8122827.
- [22] Alexander Eriksson, Sebastiaan M. Petermeijer, Markus Zimmermann, Joost C. F. de Winter, Klaus J. Bengler, and Neville A. Stanton. Rolling out the red (and green) carpet: Supporting driver decision making in automation-to-manual transitions. *IEEE Trans. Hum.-Mach. Syst.*, 49(1): 20–31, 2019. doi: 10.1109/THMS.2018.2883862.
- [23] S. Langlois and B. Soualmi. Augmented reality versus classical HUD to take over from automated driving: An aid to smooth reactions and to anticipate maneuvers. In *2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC)*, pages 1571–1578, 2016. doi: 10.1109/ITSC.2016.7795767.
- [24] Simon Ulbrich, Till Menzel, Andreas Reschka, Fabian Schuldt, and Markus Maurer. Defining and substantiating the terms scene, situation, and scenario for automated driving. In *2015 IEEE 18th International Conference on Intelligent Transportation Systems*, pages 982–988, 2015. doi: 10.1109/ITSC.2015.164.
- [25] Christian Gold, Moritz Körber, David Lechner, and Klaus Bengler. Taking over control from highly automated vehicles in complex traffic situations: The role of traffic density. *Hum. Factors*, 58(4):642–652, 2016. doi: 10.1177/0018720816634226.
- [26] J. Wan and C. Wu. The effects of lead time of take-over request and nondriving tasks on taking-over control of automated vehicles. *IEEE Trans. Human-Mach. Syst.*, 48(6):582–591, 2018. doi: 10.1109/THMS.2018.2844251.
- [27] Frederik Naujoks, Dennis Befelein, Katharina Wiedemann, and Alexandra Neukum. A review of non-driving-related tasks used in studies on automated driving. In *Adv. Hum. Aspects Transp.*, pages 525–537, Berlin, Heidelberg, 2018. Springer. doi: 10.1007/978-3-319-60441-1\_52.

- [28] Yrvann Emzivat, Javier Ibanez-Guzman, Philippe Martinet, and Olivier H. Roux. Dynamic driving task fallback for an automated driving system whose ability to monitor the driving environment has been compromised. In *2017 IEEE Intelligent Vehicles Symposium (IV)*, pages 1841–1847, 2017. doi: 10.1109/IVS.2017.7995973.
- [29] Jing Yu and Feng Luo. Fallback strategy for level 4+ automated driving system. In *2019 IEEE Intelligent Transportation Systems Conference (ITSC)*, pages 156–162, 2019. doi: 10.1109/ITSC.2019.8917404.
- [30] Chao Huang, Liang Li, Yahui Liu, and Lingyun Xiao. Robust observer based intermittent forces estimation for driver intervention identification. *IEEE Trans. Veh. Technol.*, 69(4):3628–3640, 2020. doi: 10.1109/TVT.2020.2975366.
- [31] Xin Wang, Longxiang Guo, and Yunyi Jia. Online sensing of human steering intervention torque for autonomous driving actuation systems. *IEEE Sens. J.*, 18(8):3444–3453, 2018. doi: 10.1109/JSEN.2018.2805381.
- [32] Anthony D. McDonald, Hananeh Alambeigi, Johan Engström, Gustav Markkula, Tobias Vogelpohl, Jarrett Dunne, and Norbert Yuma. Toward computational simulations of behavior during automated driving takeovers: A review of the empirical and modeling literatures. *Hum. Factors*, 61(4):642–688, 2019. doi: 10.1177/0018720819829572.
- [33] Bo Zhang, Joost de Winter, Silvia Varotto, Riender Happee, and Marieke Martens. Determinants of take-over time from automated driving: A meta-analysis of 129 studies. *Transp. Res. Part F Psychol. Behav.*, 64:285–307, 2019. doi: 10.1016/j.trf.2019.04.020.
- [34] Sónia Soares, António Lobo, Sara Ferreira, Liliana Cunha, and António Couto. Takeover performance evaluation using driving simulation: a systematic review and meta-analysis. *Eur. Transp. Res. Rev.*, 13(47), 2021. doi: 10.1186/s12544-021-00505-2.
- [35] Bradley W. Weaver and Patricia R. DeLucia. A systematic review and meta-analysis of takeover performance during conditionally automated driving. *Hum. Factors*, 64(7):1227–1260, 2022. doi: 10.1177/0018720820976476.
- [36] Kathrin Zeeb, Axel Buchner, and Michael Schrauf. What determines the take-over time? an integrated model approach of driver take-over after

- automated driving. *Accid. Anal. Prev.*, 78:212–221, 2015. doi: 10.1016/j.aap.2015.02.023.
- [37] Sol Hee Yoon, Seul Chan Lee, and Yong Gu Ji. Modeling takeover time based on non-driving-related task attributes in highly automated driving. *Appl. Ergon.*, 92:103343, 2021. ISSN 0003-6870. doi: 10.1016/j.apergo.2020.103343.
- [38] Niklas Strand, Josef Nilsson, I.C. MariAnne Karlsson, and Lena Nilsson. Semi-automated versus highly automated driving in critical situations caused by automation failures. *Transp. Res. F: Traffic Psychol. Behav.*, 27: 218–228, 2014. doi: 10.1016/j.trf.2014.04.005.
- [39] A. Hamish Jamson, Natasha Merat, Oliver M.J. Carsten, and Frank C.H. Lai. Behavioural changes in drivers experiencing highly-automated vehicle control in varying traffic conditions. *Transp. Res. Part C: Emerging Technol.*, 30:116–125, 2013. doi: 10.1016/j.trc.2013.02.008.
- [40] Mark Vollrath, Susanne Schleicher, and Christhard Gelau. The influence of cruise control and adaptive cruise control on driving behaviour—a driving simulator study. *Accid. Anal. Prev.*, 43(3):1134–1139, 2011. doi: 10.1016/j.aap.2010.12.023.
- [41] Claudia Wege, Sebastian Will, and Trent Victor. Eye movement and brake reactions to real world brake-capacity forward collision warnings—a naturalistic driving study. *Accid. Anal. Prev.*, 58:259–270, 2013. doi: 10.1016/j.aap.2012.09.013.
- [42] Joost C.F. De Winter, Riender Happee, Marieke H. Martens, and Neville A. Stanton. Effects of adaptive cruise control and highly automated driving on workload and situation awareness: A review of the empirical evidence. *Transp. Res. F: Traffic Psychol. Behav.*, 27:196–217, 2014. doi: 10.1016/j.trf.2014.06.016. Vehicle Automation and Driver Behaviour.
- [43] Frederik Naujoks, Christian Purucker, and Alexandra Neukum. Secondary task engagement and vehicle automation – comparing the effects of different automation levels in an on-road experiment. *Transp. Res. F: Traffic Psychol. Behav.*, 38:67–82, 2016. doi: 10.1016/j.trf.2016.01.011.
- [44] S.G. Klauer, T. A. Dingus, V. L. Neale, J.D. Sudweeks, and D.J. Ramsey. The impact of driver inattention on near-crash/crash risk: An analysis using

- the 100-car naturalistic driving study data. Technical Report DOT HS 810 594, NHTSA, Washington, DC, USA, April 2006.
- [45] V.L. Neale, S.G. Klauer, R.R. Knipling, T.A. Dingus, G.T. Holbrook, and A. Petersen. The 100 car naturalistic driving study, phase I–experimental design. Technical Report DTNH22-00-C-07007, NHTSA, Washington, D.C., U.S.A, December 2002.
- [46] T. A. Dingus, S.G. Klauer, and V. L. Neale. The 100 car naturalistic driving study, phase II–results of the 100-car field experiment. Technical Report DTNH22-00-C-07007, NHTSA, Washington, D.C., U.S.A, April 2006.
- [47] Natasha Merat, A. Hamish Jamson, Frank C. H. Lai, and Oliver Carsten. Highly automated driving, secondary task performance, and driver state. *Hum. Factors*, 54(5):762–771, 2012. doi: 10.1177/0018720812442087.
- [48] Johannes Beller, Matthias Heesen, and Mark Vollrath. Improving the driver–automation interaction: An approach using automation uncertainty. *Hum. Factors*, 55(6):1130–1141, 2013. doi: 10.1177/0018720813482327.
- [49] Philipp Kerschbaum, Lutz Lorenz, and Klaus Bengler. Highly automated driving with a decoupled steering wheel. *Proc. Hum. Factors Ergon. Soc. Annu. Meet.*, 58(1):1686–1690, 2014. doi: 10.1177/1541931214581352.
- [50] Jonas Radlmayr, Christian Gold, Lutz Lorenz, Mehdi Farid, and Klaus Bengler. How traffic situations and non-driving related tasks affect the take-over quality in highly automated driving. *Proc. Hum. Factors Ergon. Soc. Annu. Meet.*, 58(1):2063–2067, 2014. doi: 10.1177/15419312145814.
- [51] Tsang-Wei Lin, Sheue-Ling Hwang, and Paul A. Green. Effects of time-gap settings of adaptive cruise control (acc) on driving performance and subjective acceptance in a bus driving simulator. *Saf. Sci.*, 47(5):620–625, 2009. doi: 10.1016/j.ssci.2008.08.004.
- [52] Sebastiaan M. Petermeijer, David A. Abbink, and Joost C. F. de Winter. Should drivers be operating within an automation-free bandwidth? Evaluating haptic steering support systems with different levels of authority. *Hum. Factors*, 57(1):5–20, 2015. doi: 10.1177/0018720814563602.
- [53] Moritz Körber, Andrea Cingel, Markus Zimmermann, and Klaus Bengler. Vigilance decrement and passive fatigue caused by monotony in automated driving. *Procedia Manuf.*, 3:2403–2409, 2015. doi: 10.1016/j.promfg.



- 2015.07.499. 6th International Conference on Applied Human Factors and Ergonomics (AHFE 2015) and the Affiliated Conferences, AHFE 2015.
- [54] Kathrin Zeeb, Manuela Härtel, Axel Buchner, and Michael Schrauf. 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. *Transp. Res. F: Traffic Psychol. Behav.*, 50:65–79, 2017.
- [55] Oliver Carsten, Frank C. H. Lai, Yvonne Barnard, A. Hamish Jamson, and Natasha Merat. Control task substitution in semiautomated driving: Does it matter what aspects are automated? *Hum. Factors*, 54(5):747–761, 2012. doi: 10.1177/0018720812460246.
- [56] Matti Schwalk, Niko Kalogerakis, and Thomas Maier. Driver support by a vibrotactile seat matrix – recognition, adequacy and workload of tactile patterns in take-over scenarios during automated driving. *Procedia Manuf.*, 3:2466–2473, 2015. ISSN 2351-9789. doi: 10.1016/j.promfg.2015.07.507. 6th International Conference on Applied Human Factors and Ergonomics (AHFE 2015) and the Affiliated Conferences, AHFE 2015.
- [57] Moritz Körber, Christian Gold, David Lechner, and Klaus Bengler. The influence of age on the take-over of vehicle control in highly automated driving. *Transp. Res. F: Traffic Psychol. Behav.*, 39:19–32, 2016.
- [58] A. Bourrelly, C. Jacobé de Naurois, A. Zran, F. Rampillon, J. Vercher, and C. Bourdin. Long automated driving phase affects take-over performance. *IET Intell. Transp. Syst.*, 13(8):1249–1255, 2019.
- [59] Tanja Fuest, Lenja Sorokin, Hanna Bellem, and Klaus Bengler. Taxonomy of traffic situations for the interaction between automated vehicles and human road users. In Neville A Stanton, editor, *Advances in Human Aspects of Transportation*, pages 708–719, Cham, 2018. Springer International Publishing.
- [60] Na Du, Jinyong Kim, Feng Zhou, Elizabeth Pulver, Dawn M. Tilbury, Lionel P. Robert, Anuj K. Pradhan, and X. Jessie Yang. Evaluating effects of cognitive load, takeover request lead time, and traffic density on drivers' takeover performance in conditionally automated driving. In *12th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, AutomotiveUI '20, pages 66—73, New York, NY, USA, 2020. Association for Computing Machinery. doi: 10.1145/3409120.3410666.

- [61] Barbara Metz, Andreas Landau, and Mariana Just. Frequency of secondary tasks in driving – results from naturalistic driving data. *Saf. Sci.*, 68:195–203, 2014. doi: 10.1016/j.ssci.2014.04.002.
- [62] Changxu Wu, Dekuang Yu, Amy Doherty, Tianyi Zhang, Leo Kust, and Gang Luo. An investigation of perceived vehicle speed from a driver’s perspective. *PLoS ONE*, 12(10):1–11, 10 2017. doi: 10.1371/journal.pone.0185347.
- [63] Christian Gold, Riender Happee, and Klaus Bengler. Modeling take-over performance in level 3 conditionally automated vehicles. *Accid. Anal. Prev.*, 116:3–13, 2018. doi: 10.1016/j.aap.2017.11.009. Simulation of Traffic Safety in the Era of Advances in Technologies.
- [64] Xiaomei Tan and Yiqi Zhang. The effects of takeover request lead time on drivers’ situation awareness for manually exiting from freeways: A web-based study on level 3 automated vehicles. *Accid. Anal. Prev.*, 168: 106593, 2022. doi: 10.1016/j.aap.2022.106593.
- [65] Sandra Epple, Fabienne Roche, and Stefan Brandenburg. The sooner the better: Drivers’ reactions to two-step take-over requests in highly automated driving. *Proc. Hum. Factors Ergon. Soc. Annu. Meet.*, 62(1):1883–1887, 2018. doi: 10.1177/1541931218621428.
- [66] Z. Lu, B. Zhang, A. Feldhütter, R. Happee, M. Martens, and J.C.F. De Winter. Beyond mere take-over requests: The effects of monitoring requests on driver attention, take-over performance, and acceptance. *Transp. Res. Part F Psychol. Behav.*, 63:22–37, 2019. doi: 10.1016/j.trf.2019.03.018.
- [67] Shu Ma, Wei Zhang, and Zhen Yang. Take over gradually in conditional automated driving: The effect of two-stage warning systems on situation awareness, driving stress, takeover performance, and acceptance. *Int. Hum.-Comput. Interact.*, 37(4):352–362, 2021. doi: 10.1080/10447318.2020.1860514.
- [68] Wei Zhang, Yilin Zeng, and Zhen Yang. Optimal time intervals in two-stage takeover warning systems with insight into the drivers’ neuroticism personality. *Front. Psychol.*, 12, 2021. doi: 10.3389/fpsyg.2021.601536.
- [69] David F. Hultsch, Stuart W. S. MacDonald, and Roger A. Dixon. Variability in reaction time performance of younger and older adults. *J. Gerontol.: Psychol. Sci.*, 57B(2):101–115, 2002.

- [70] Anstey KJ, Wood J, Lord S, and Walker JG. Cognitive, sensory and physical factors enabling driving safety in older adults. *Clin Psychol Rev.*, 25(1): 45–65, 2005.
- [71] Shuo Li, Phil Blythe, Weihong Guo, and Anil Namdeo. Investigating the effects of age and disengagement in driving on driver’s takeover control performance in highly automated vehicles. *Transp. Plan. Technol.*, 42(5): 470–497, 2019. doi: 10.1080/03081060.2019.1609221.
- [72] Melanie Karthaus, Edmund Wascher, Michael Falkenstein, and Stephan Getzmann. The ability of young, middle-aged and older drivers to inhibit visual and auditory distraction in a driving simulator task. *Transp. Res. F: Traffic Psychol. Behav.*, 68:272–284, 2020. doi: 10.1016/j.trf.2019.11.007.
- [73] Monika Ucinska, Ewa Odachowska, Kamila Gasiorek, and Mikolaj Kruszewski. Age and experience in driving a vehicle and psychomotor skills in the context of automation. *Open Eng.*, 11(1):453–462, 2021. doi: 10.1515/eng-2021-0045.
- [74] Lisa J. Molnar, Anuj K. Pradhan, David W. Eby, Lindsay H. Ryan, Renée M. St. Louis, Jennifer Zakrajsek, Brittany Ross, Brian T. Lin, Chen Liang, Bethany Zalewski, and Liang Zhang. Age-related differences in driver behavior associated with automated vehicles and the transfer of control between automated and manual control: A simulator evaluation. Technical Report 2015-08, Mobility Transformation Center, Ann Arbor, U.S.A, May 2017.
- [75] Brandon J. Pitts and Nadine Sarter. What you don’t notice can harm you: Age-related differences in detecting concurrent visual, auditory, and tactile cues. *Hum. Factors*, 60(4):445–464, 2018. doi: 10.1177/0018720818759102.
- [76] Hallie Clark and Jing Feng. Age differences in the takeover of vehicle control and engagement in non-driving-related activities in simulated driving with conditional automation. *Accid. Anal. Prev.*, 106:468–479, 2017. doi: 10.1016/j.aap.2016.08.027.
- [77] Yanbin Wu, Ken Kihara, Kunihiro Hasegawa, Yuji Takeda, Toshihisa Sato, Motoyuki Akamatsu, and Satoshi Kitazaki. Age-related differences in effects of non-driving related tasks on takeover performance in automated driving. *J. Saf. Res.*, 72:231–238, 2020. ISSN 0022-4375. doi: 10.1016/j.jsr.2019.12.019.

- [78] William J. Horrey and Christopher D. Wickens. In-vehicle glance duration: Distributions, tails, and model of crash risk. *Transp. Res. Rec.*, 2018(1):22–28, 2007. doi: 10.3141/2018-04.
- [79] Trent Victor. *Keeping Eye and Mind on the Road*. PhD thesis, Uppsala Univ., Uppsala, Sweden, 2005.
- [80] Trent W. Victor, Joanne L. Harbluk, and Johan A. Engström. Sensitivity of eye-movement measures to in-vehicle task difficulty. *Transp. Res. F: Traffic Psychol. Behav.*, 8(2):167–190, 2005. doi: 10.1016/j.trf.2005.04.014.
- [81] Alexander Lotz, Nele Russwinkel, and Enrico Wohlfarth. Response times and gaze behavior of truck drivers in time critical conditional automated driving take-overs. *Transp. Res. Part F Psychol. Behav.*, 64:532–551, 2019. ISSN 1369-8478. doi: 10.1016/j.trf.2019.06.008.
- [82] Moritz Körber, Eva Baseler, and Klaus Bengler. Introduction matters: Manipulating trust in automation and reliance in automated driving. *Appl. Ergon.*, 66:18–31, 2018. doi: 10.1016/j.apergo.2017.07.006.
- [83] Sebastian Hergeth, Lutz Lorenz, Roman Vilimek, and Josef F. Krems. Keep your scanners peeled: Gaze behavior as a measure of automation trust during highly automated driving. *Hum. Factors*, 58(3):509–519, 2016. doi: 10.1177/0018720815625744. PMID: 26843570.
- [84] F. Walker, J. Wang, M.H. Martens, and W.B. Verwey. Gaze behaviour and electrodermal activity: Objective measures of drivers’ trust in automated vehicles. *Transp. Res. Part F Psychol. Behav.*, 64:401–412, 2019. doi: 10.1016/j.trf.2019.05.021.
- [85] Oliver Jarosch and Klaus Bengler. Is it the duration of the ride or the non-driving related task? what affects take-over performance in conditional automated driving? In Sebastiano Bagnara, Riccardo Tartaglia, Sara Albolino, Thomas Alexander, and Yushi Fujita, editors, *Proceedings of the 20th Congress of the International Ergonomics Association (IEA 2018)*, pages 512–523, Cham, 2019. Springer International Publishing. doi: 10.1007/978-3-319-96074-6\_54.
- [86] J. Gonçalves, R. Happee, and K. Bengler. Drowsiness in conditional automation: Proneness, diagnosis and driving performance effects. In *2016 IEEE 19th Int. Conf. Intell. Transp. Syst. (ITSC)*, pages 873–878, 2016. doi: 10.1109/ITSC.2016.7795658.

- [87] Anna Feldhütter, Christian Gold, Sonja Schneider, and Klaus Bengler. How the duration of automated driving influences take-over performance and gaze behavior. In *Adv. Ergon. Des. Syst., Prod. Processes*, pages 309–318, Berlin, Heidelberg, 2017. Springer. doi: 10.1007/978-3-662-53305-5\_22.
- [88] Francesca Favaro, Sky Eurich, Syeda Rizvi, Sumaid Mahmood, and Nazanin Nader. Analysis of disengagements in semi-autonomous vehicles: Drivers’ takeover performance and operational implications. Technical report, NHTSA, Washington, DC, USA, June 2019.
- [89] Natasha Merat, A. Hamish Jamson, Frank C.H. Lai, Michael Daly, and Oliver M.J. Carsten. Transition to manual: Driver behaviour when resuming control from a highly automated vehicle. *Transp. Res. F: Traffic Psychol. Behav.*, 27:274–282, 2014. doi: 10.1016/j.trf.2014.09.005. Vehicle Automation and Driver Behaviour.
- [90] Moritz Körber and Klaus Bengler. Potential individual differences regarding automation effects in automated driving. In *Proceedings of the XV International Conference on Human Computer Interaction, Interacción ’14*, New York, NY, USA, 2014. Association for Computing Machinery. ISBN 9781450328807. doi: 10.1145/2662253.2662275.
- [91] Moritz Körber, Thomas Weißgerber, Christoph Blaschke, Mehdi Farid, and Luis Kalb. Prediction of take-over time in highly automated driving by two psychometric tests. *Dyna*, 82:195–201, 10 2015.
- [92] J. Nilsson, P. Falcone, and J. Vinter. Safe transitions from automated to manual driving using driver controllability estimation. *IEEE Trans. Intell. Transp. Syst.*, 16(4):1806–1816, 2015. doi: 10.1109/TITS.2014.2376877.
- [93] T. Yamada, H. Irie, M. Kunitake, E. Nagano, and S. Sakai. Estimating driver’s readiness by understanding driving posture. In *2018 IEEE Int. Conf. Consum. Electron. (ICCE)*, pages 1–4, 2018. doi: 10.1109/ICCE.2018.8326115.
- [94] N. Deo and M. M. Trivedi. Looking at the driver/rider in autonomous vehicles to predict take-over readiness. *IEEE Trans. Intell. Veh.*, 5(1):41–52, 2020. doi: 10.1109/TIV.2019.2955364.
- [95] Na Du, Feng Zhou, Elizabeth M. Pulver, Dawn M. Tilbury, Lionel P. Robert, Anuj K. Pradhan, and X. Jessie Yang. Predicting driver takeover performance in conditionally automated driving. *Accid. Anal. Prev.*, 148:105748, 2020. doi: 10.1016/j.aap.2020.105748.

- [96] Jackie Ayoub, Na Du, X. Jessie Yang, and Feng Zhou. Predicting driver takeover time in conditionally automated driving. *IEEE Trans. Intell. Transp. Syst.*, 23(7):9580–9589, 2022. doi: 10.1109/TITS.2022.3154329.
- [97] Erfan Pakdamanian, Shili Sheng, Sonia Bae, Seongkook Heo, Sarit Kraus, and Lu Feng. Deeptake: Prediction of driver takeover behavior using multimodal data. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, New York, NY, USA, 2021. Association for Computing Machinery. doi: 10.1145/3411764.3445563.
- [98] Kathrin Zeeb, Axel Buchner, and Michael Schrauf. Is take-over time all that matters? the impact of visual-cognitive load on driver take-over quality after conditionally automated driving. *Accid. Anal. Prev.*, 92:230–239, 2016.
- [99] C. Braunagel, W. Rosenstiel, and E. Kasneci. Ready for take-over? A new driver assistance system for an automated classification of driver take-over readiness. *IEEE Intell. Transp. Syst. Mag.*, 9(4):10–22, 2017. doi: 10.1109/ITS.2017.2743165.
- [100] Brian P. Bailey and Joseph A. Konstan. On the need for attention-aware systems: Measuring effects of interruption on task performance, error rate, and affective state. *Comput. Hum. Behav.*, 22(4):685–708, 2006. doi: 10.1016/j.chb.2005.12.009.
- [101] Philipp Wintersberger, Andreas Riener, Clemens Schartmüller, Anna-Katharina Frison, and Klemens Weigl. 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*, AutomotiveUI '18, pages 53–65, New York, NY, USA, 2018. ACM. doi: <https://dl.acm.org/doi/10.1145/3239060.3239085>.
- [102] Nade Liang, Jing Yang, Denny Yu, Kwaku O. Prakah-Asante, Reates Curry, Mike Blommer, Radhakrishnan Swaminathan, and Brandon J. Pitts. Using eye-tracking to investigate the effects of pre-takeover visual engagement on situation awareness during automated driving. *Accid. Anal. Prev.*, 157: 106143, 2021. doi: 10.1016/j.aap.2021.106143.
- [103] Hidde Van der Meulen, Andrew L. Kun, and Christian P. Janssen. Switching back to manual driving: How does it compare to simply driving away after parking? In *Proceedings of the 8th International Conference*

- on Automotive User Interfaces and Interactive Vehicular Applications, Automotive'UI 16, pages 229–236, New York, NY, USA, 2016. ACM. doi: <https://dl.acm.org/doi/10.1145/3003715.3005452>.
- [104] Pål Ulleberg and Torbjørn Rundmo. Personality, attitudes and risk perception as predictors of risky driving behaviour among young drivers. *Saf. Sci.*, 41(5):427–443, 2003.
- [105] ISO. Road vehicles–Ergonomic aspects of transport information and control systems–Calibration tasks for methods which assess driver demand due to the use of in-vehicle systems. Standard ISO/TS 14198:2019, 2019. URL <https://www.iso.org/standard/71509.html>.
- [106] John A. Johnson. Measuring thirty facets of the five factor model with a 120-item public domain inventory: Development of the IPIP-NEO-120. *J. Res. Pers.*, 51:78–89, 2014. doi: 10.1016/j.jrp.2014.05.003.
- [107] Leandro L. Di Stasi, Carolina Diaz-Piedra, Héctor Rieiro, José M. Sánchez Carrión, Mercedes Martin Berrido, Gonzalo Olivares, and Andrés Catena. Gaze entropy reflects surgical task load. *Surg. Endosc.*, 30:5034–5043, 2016. doi: 10.1007/s00464-016-4851-8.
- [108] C. E. Shannon. A mathematical theory of communication. *Bell Syst. Tech. J.*, 27(3):379–423, 1948. doi: 10.1002/j.1538-7305.1948.tb01338.x.
- [109] J. C. F. de Winter, Y. B. Eisma, C. D. D. Cabrall, P. A. Hancock, and N. A. Stanton. Situation awareness based on eye movements in relation to the task environment. *Cogn. Tech. Work*, 21:99–111, 2019. doi: 10.1007/s10111-018-0527-6.
- [110] Punitkumar Bhavsar, Babji Srinivasan, and Rajagopalan Srinivasan. Quantifying situation awareness of control room operators using eye-gaze behavior. *Comput. Chem. Eng.*, 106:191–201, 2017. ISSN 0098-1354. doi: 10.1016/j.compchemeng.2017.06.004.
- [111] P. Anusree Anand, Priyanka Atmakuri, and Viswa Sri Rupa Anne. Calibration of vehicle-following model parameters using mixed traffic trajectory data. *Transp. in Dev. Econ.*, 5, 09 2019. doi: 10.1007/s40890-019-0086-4.
- [112] Leslie G. Ungerleider and Mortimer Mishkin. Two cortical visual systems. In David J. Ingle, Melvyn A. Goodale, and Richard J. W. Mansfield, editors, *Analysis of Visual Behavior*, pages 549–586. MIT Press, Cambridge, MA, 1982. ISBN 0-262-09022-8.

- [113] Melvyn A Goodale and David A Westwood. An evolving view of duplex vision: separate but interacting cortical pathways for perception and action. *Curr. Opin. Neurobiol.*, 14(2):203–211, 2004. doi: 10.1016/j.conb.2004.03.002.
- [114] C. Koch and F. Crick. The zombie within. *Nature*, 411:893, 2001. doi: 10.1038/35082161.
- [115] Melvyn A Goodale and G Keith Humphrey. The objects of action and perception. *Cognition*, 67(1):181–207, 1998. doi: 10.1016/S0010-0277(98)00017-1.
- [116] Michael S. Wogalter and S. David Leonard. Attention capture and maintenance. In Michael S. Wogalter, David M. Dejoy, and Kenneth R. Laughery, editors, *Warnings and Risk Communication*, chapter 7, pages 113–138. Taylor & Francis, London, 2005.
- [117] Ronald A. Rensink. Visual attention. In L. Nagel, editor, *Encyclopedia of Cognitive Science*, pages 509–515. Macmillan, 2002.
- [118] Reuven Dukas. Causes and consequences of limited attention. *Brain Behav. Evol.*, 63:197–210, 2004. doi: 10.1159/000076781.
- [119] Eckhard H. Hess and James M. Polt. Pupil size as related to interest value of visual stimuli. *Science*, 132(3423):349–350, 1960. doi: 10.1126/science.132.3423.349.
- [120] Margaret M. Bradley, Laura Miccoli, Miguel A. Escrig, and Peter J. Lang. The pupil as a measure of emotional arousal and autonomic activation. *Psychophysiology*, 45(4):602–607, 2008. doi: 10.1111/j.1469-8986.2008.00654.x.
- [121] Michel Guillon, Kathryn Dumbleton, Panagiotis Theodoratos, Marine Gobbe, C. Benjamin Wooley, and Kurt Moody. The effects of age, refractive status, and luminance on pupil size. *Optom. Vis. Sci.*, 93(9):1093–1100, 2016. doi: 10.1097/OPX.0000000000000893.
- [122] D. G. Stavenga, J. A. J. Numan, J. Tinbergen, and J. W. Kuiper. Insect pupil mechanisms. *J. Comp. Physiol.*, 113:73–93, 1977. doi: 10.1007/BF00610454.
- [123] Maryam Akbari, B. Lankarani Kamran, Seyed Taghi Heydari, Seyed Abbas Motevalian, Reza Tabrizi, Zohreh Asadi-Shekari, and Mark J. M. Sullman. Meta-analysis of the correlation between personality characteristics and



- risky driving behaviors. *J. Inj. Violence Res.*, 11(2):107–122, 2019. doi: 10.5249/jivr.v11i2.1172.
- [124] Johnathon P. Ehsani, Kaigang Li, Bruce G. Simons-Morton, Cheyenne Fox Tree-McGrath, Jessamyn G. Perlus, Fearghal O’Brien, and Sheila G. Klauer. Conscientious personality and young drivers’ crash risk. *J. Saf. Res.*, 54: 83.e29–87, 2015.
- [125] Wenmin Li, Nailang Yao, Yanwei Shi, Weiran Nie, Yuhai Zhang, Xiangrong Li, Jiawen Liang, Fang Chen, and Zaifeng Gao. Personality openness predicts driver trust in automated driving. *Automot. Innovation*, 3:3–13, 2020.
- [126] H. Hummel and D. Lester. Extraversion and simple reaction time. *Percept. Mot. Skills*, 45:1236, 1977.
- [127] Chao Huang, Bo Yang, and Kimihiko Nakano. Impact of duration of monitoring before takeover request on takeover time with insights into eye tracking data. *Accid. Anal. Prev.*, 185:107018, 2023. doi: 10.1016/j.aap.2023.107018.
- [128] Laura Pritschet, Derek Powell, and Zachary Horne. Marginally significant effects as evidence for hypotheses: Changing attitudes over four decades. *Psychol. Sci.*, 27(7):1036–1042, 2016. doi: 10.1177/0956797616645672.
- [129] Anton Olsson-Collentine, Marcel A. L. M. van Assen, and Chris H. J. Hartgerink. The prevalence of marginally significant results in psychology over time. *Psychol. Sci.*, 30(4):576–586, 2019. doi: 10.1177/0956797619830326.
- [130] Chao Huang, Yang Bo, and Kimihiko Nakano. Learning from drivers patterns of evasive maneuvers in case of emergency using dynamic-time-warping-based clustering. In *2023 IEEE 26th International Conference on Intelligent Transportation Systems (ITSC)*, 2023.
- [131] Yan Ge, Weina Qu, Caihong Jiang, Feng Du, Xianghong Sun, and Kan Zhang. The effect of stress and personality on dangerous driving behavior among chinese drivers. *Accid. Anal. Prev.*, 73:34–40, 2014. doi: 10.1016/j.aap.2014.07.024.
- [132] Gerald Matthews. Personality and information processing: a cognitive-adaptive theory. In Gregory J. Boyle, Gerald Matthews, and Donald H. Saklofske, editors, *The SAGE Handbook of Personality Theory and Assessment*:

- Volume 1–Personality Theories and Models*, chapter 2, pages 56–79. SAGE Publications, Thousand Oaks, CA: Sage, 2008.
- [133] Brian P. Meier, Benjamin M. Wilkowski, and Michael D. Robinson. Bringing out the agreeableness in everyone: Using a cognitive self-regulation model to reduce aggression. *J. Exp. Soc. Psychol.*, 44(5):1383–1387, 2008. doi: 10.1016/j.jesp.2008.05.005.
- [134] Colin G. DeYoung and Jeremy R. Gray. Personality neuroscience: explaining individual differences in affect, behaviour and cognition. In Philip J. Corr and Gerald Matthews, editors, *The Cambridge Handbook of Personality Psychology*, chapter 20, pages 323–346. Cambridge University Press, New York, 2009.
- [135] Denis Kuposov, Maria Semenova, Andrey Somov, Andrey Lange, Anton Stepanov, and Evgeny Burnaev. Analysis of the reaction time of esports players through the gaze tracking and personality trait. In *IEEE Int. Symp. Ind. Electron. (ISIE)*, pages 1560–1565, 2020. doi: 10.1109/ISIE45063.2020.9152422.
- [136] John Brebner. Personality theory and movement. In Bruce D. Kirkcaldy, editor, *Individual Differences in Movement*, chapter 2, pages 27–41. MTP Press, Hingham, USA, 1985.
- [137] G. Buena Casal, V.E. Caballo, E. García Cueto, and P. Flores Cubos. Attention and reaction time differences in introversion-extraversion. *Pers. Individ. Differ.*, 11(2):195–197, 1990. doi: 10.1016/0191-8869(90)90015-J.
- [138] R. R. McCrae and P. T. Costa. Validation of the five-factor model of personality across instruments and observers. *J. Pers. Social Psychol.*, 52(1):81–90, 1987.
- [139] Philip J. Corr and Gerald Matthews. Editors’ introduction to parts I to VIII. In Philip J. Corr and Gerald Matthews, editors, *The Cambridge Handbook of Personality Psychology*, pages xliii–liv. Cambridge University Press, New York, 2009.
- [140] Regan E Settles, Sarah Fischer, Melissa A Cyders, Jessica L Combs, Rachel L Gunn, and Gregory T Smith. Negative urgency: a personality predictor of externalizing behavior characterized by neuroticism, low conscientiousness, and disagreeableness. *J. Abnorm. Psychol.*, 121(1):160–72, 2012. doi: 10.1037/a0024948.

- [141] Mahdi Bonyani, Mina Rahmanian, Simindokht Jahangard, and Mahdi Rezaei. Dipnet: Driver intention prediction for a safe takeover transition in autonomous vehicles. *IET Intell. Transp. Syst.*, n/a(n/a):1–15, 2023. doi: 10.1049/itr2.12370.
- [142] Ravid Shwartz-Ziv and Amitai Armon. Tabular data: Deep learning is not all you need, 2021.
- [143] Shantanu Gupta, Rohit Mishra, Yu-Hao Chang, Zheng Ma, Fenglong Ma, and Yiqi Zhang. Modeling driver takeover intention in automated vehicles with attention-based cnn algorithm. *Proc. Hum. Factors Ergon. Soc. Annu. Meet.*, 66(1):1607–1611, 2022. doi: 10.1177/1071181322661303.
- [144] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. SMOTE: Synthetic minority over-sampling technique. *J. Artif. Intell. Res.*, 16:321–357, jun 2002. doi: 10.1613/jair.953.
- [145] Liu Yang, Rui Ma, H. Michael Zhang, Wei Guan, and Shixiong Jiang. Driving behavior recognition using eeg data from a simulated car-following experiment. *Accid. Anal. Prev.*, 116:30–40, 2018. doi: 10.1016/j.aap.2017.11.010.
- [146] Yulin Ma, Zhixiong Li, and Yicheng Li. Driving style estimation by fusing multiple driving behaviors: a case study of freeway in china. *Cluster Comput.*, 22:8259—8269, 2019. doi: 10.1007/s10586-018-1739-5.
- [147] Wenshuo Wang, Junqiang Xi, Alexandre Chong, and Lin Li. Driving style classification using a semisupervised support vector machine. *IEEE Trans. Human-Mach. Syst.*, 47(5):650–660, 2017. doi: 10.1109/THMS.2017.2736948.
- [148] Ali Javed, Byung Suk Lee, and Donna M. Rizzo. A benchmark study on time series clustering. *Mach. Learn. Appl.*, 1:100001, 2020. doi: 10.1016/j.mlwa.2020.100001.
- [149] Uwe Moser and Dieter Schramm. Multivariate dynamic time warping in automotive applications: A review. *Intell. Data Anal.*, 23(3):535–553, 2019. doi: 10.3233/IDA-184130.
- [150] Mazen Danaf, Maya Abou-Zeid, and Isam Kaysi. Modeling anger and aggressive driving behavior in a dynamic choice-latent variable model. *Accid. Anal. Prev.*, 75:105–118, 2015. doi: 10.1016/j.aap.2014.11.012.
- [151] Wikipedia. k-means clustering, 2006. URL [https://en.wikipedia.org/wiki/K-means\\_clustering](https://en.wikipedia.org/wiki/K-means_clustering). [Accessed April 28, 2023].

- [152] Ying Yao, Xiaohua Zhao, Yiping Wu, Yunlong Zhang, and Jian Rong. Clustering driver behavior using dynamic time warping and hidden markov model. *J. Intell. Transp. Syst.*, 25(3):249–262, 2021. doi: 10.1080/15472450.2019.1646132.
- [153] Feng Guo, Sheila G. Klauer, Michael T. McGill, and Thomas A. Dingus. Task 3-evaluating the relationship between near-crashes and crashes: Can near-crashes serve as a surrogate safety metric for crashes? Technical Report DOT HS 811 382, NHTSA, Washington, DC, USA, October 2010.
- [154] Chao Huang, Bo Yang, and Kimihiko Nakano. Where drivers are looking at during takeover: Implications for safe takeovers during conditionally automated driving. *Traffic Inj. Prev.*, 24(7):599–608, 2023. doi: 10.1080/15389588.2023.2224910.
- [155] Marry L. McHugh. Interrater reliability: the kappa statistic. *Biochem Med*, 22(3):276–282, 2012. doi: 10.11613/BM.2012.031.
- [156] ISO. Road vehicles—measurement and analysis of driver visual behavior with respect to transport information and control systems. Standard ISO 15007:2020, 2020. URL <https://www.iso.org/standard/63220.html>.
- [157] Virginia Braun and Victoria Clarke. Using thematic analysis in psychology. *Qual. Res. Psychol.*, 3(2):77–101, 2006. doi: 10.1191/1478088706qp063oa.
- [158] Nanxiang Li and Carlos Busso. Detecting drivers’ mirror-checking actions and its application to maneuver and secondary task recognition. *IEEE Trans. Intell. Transp. Syst.*, 17(4):980–992, 2016. doi: 10.1109/TITS.2015.2493451.
- [159] Mai-Britt Herslund and Niels O Jørgensen. Looked-but-failed-to-see-errors in traffic. *Accid. Anal. Prev.*, 35(6):885–891, 2003. doi: 10.1016/S0001-4575(02)00095-7.
- [160] Chao Huang, Bo Yang, and Kimihiko Nakano. Impact of personality on takeover time and maneuvers shortly after takeover request. unpublished, 2023.