Modeling human driver in takeover scenarios

A neuropsychological model informed by physiological dynamics

DISSERTATION

submitted in partial fulfillment
of the requirements for the degree

Doktor-Ingenieur (Dr.-Ing.)
Doctor of Engineering

at the

Faculty of electrical engineering and information technology,
TU Dortmund University

by

Khazar Dargahi Nobari, M.Sc.

Dortmund, Germany

First examiner: Univ.-Prof. Dr.-Ing. Prof. h.c. Dr. h.c. Torsten Bertram
Second examiner: Prof. Dr. Meike Jipp

Date of submission: January 13, 2025
Date of examination: June 24, 2025

Abstract

The present dissertation introduces a holistic driver model aimed at predicting driver state and reaction during the transition from automated driving to manual driving. Common driver models are typically confined to specific scenarios with limited situational criteria and simplified driver states, often reducing the complexity of the driving situations to a single variable and the human decision-making process to a few measuring factors. These limitations hinder the generalizability of such models to real-world, complex driving environments, where a wide array of variables interact dynamically. Addressing these shortcomings, this research proposes a novel driver model that accounts for a broad spectrum of driver states, integrating emotional and cognitive structures to reflect the nature of real-world driving better.

The driver model is grounded in two core theories: the adaptive control of thought-rational cognitive architecture and the theory of constructed emotion. The cognitive architecture provides a robust computational framework for predicting driver performance and reactions. However, its application has been limited to scenarios involving drivers in an emotionally neutral state, overlooking the impact of human affect. On the other hand, the theory of constructed emotion posits that emotions are not fixed responses but are dynamic constructs shaped by the interplay of cognitive, social, and physiological factors.

By leveraging these theories, the proposed model offers a more holistic approach, capable of handling a wide range of driver states, considering affect dynamics and emotional variations. The model integrates the driving context with the driver state to predict subsequent states and reactions more accurately, using a generative machine learning algorithm inspired by neuroscientific insights into the human brain. Consequently, the model represents an advancement over the adaptive control of thought-rational cognitive architecture by incorporating the complexity of affective process into its predictive capabilities. While this dissertation focuses on takeover situations, the model's design is inherently generalizable, allowing for future applications across various driving scenarios.

To validate the proposed model, a comprehensive dataset, manD 1.0, is collected through a precisely designed subject study using a driving simulator. This dataset captures a range of driving situations, variations in driver state, and detailed information about vehicle state, all synchronized to provide a rich source of training data. From this dataset, segments involving takeover situations are extracted for model training and assessment. Given that individual differences, such as past experiences and physiological characteristics, can significantly influence driver behavior, the model is individualized for each person by training it on individual-specific data. A comparative analysis with a baseline model demonstrates that incorporating a psychological architecture enhances the accuracy of predicting driver states and reactions, improves the model's robustness against variability, and increases its generalizability across diverse driving situations. This model holds potential for use in driver assistance systems, particularly in critical situations where predicting driver state and reaction could prevent severe safety consequences.

Contents

Nomenclature

General symbols

i	Counting index
k	Upper limit of counting index
t	Time

Specific symbols

$\mathbf{A}_i$	Coefficient matrix of VAR model for the lag i
α	Decay parameter in ACT-R model
α_i	i-th attention weight in attention mechanism
a_{max}	Maximum differencing order
$a_{m,t}$	Activation of memory m at time t in ACT-R model
β	A paremeter configuring sharpness of softmax function
$\mathbf{b}_\cdot$	Bias vector of an LSTM cell with subscripts f for forget gate, x for input gate, c for cell state, and o for output gate
$b_{m,t}$	Base-level activation of memory m at time t in ACT-R model
C	Constant term in decay rate of ACT-R model
C_j	Cost to achieve a goal by selecting the production rule j in ACT-R model
$\mathbf{c}_t$	LSTM cell state vector at time t
$\tilde{\mathbf{c}}_t$	Candidate cell state vector of LSTM at time $t-1$
$\mathbf{c}_v$	Verctor of VAR intercepts
d	Single decay rate parameter
D	Number of variables
d_{hid}	Hidden size of the LSTM
d_{trace_i}	Decay rate parameter of the i-th trace in ACT-R model
d_{in}	Size of BLSTM input vector
d_{lat}	Dimension of latent space in VAE and AE
$\mathbf{d}_o$	Distance to object
$\mathcal{D}_{\text{test}}$	Test set
$\mathcal{D}_{\text{train}}$	Training set
$\mathbf{f_m}$	Memory feature
$\mathbf{f_r}$	Required features for retrieval in ACT-R model
$\mathbb{F}_r$	Set of retrieval features
$\mathbf{f}_t$	Forget gate vector of LSTM at time step t
G	Value of the current goal in ACT-R model
g_0	Gravitational force
$\mathbf{h}^{\text{BLSTM}}$	Sequence of hidden states of BLSTM network
$\mathbf{h}_{\text{final}}$	Encoder's final hidden state
$\mathbf{H}$	Set of all hidden states of BLSTM network

$\mathbf{h}^{\text{Hop}}$	OUtput of Hopfield layer
$\mathbf{h}_t$	Hidden state vector of LSTM at time t
$\mathbf{h}'_t$	Hidden state vector of backward LSTM at time t in BLSTM layer
$\mathbf{h}_t^{\text{BLSTM}}$	Hidden state vector of BLSTM at time t
h_v	Prediction horizon of VAR model
j	Index of the production rule in ACT-R model
$\mathbf{k}_i$	i-th key vector in attention mechanism
$\mathbf{K}$	Set of key vectors for attention mechanism
L	Number of trials in Bayesian optimization
m	Memory chunk in ACT-R model
med_{stat}	Median
M	Maximum norm of the patterns in modern Hopfield network
MP	Mismatch penalty scaling factor in ACT-R model
m_{stat}	Mean value
μ	Index of stored pattern in Hopfield network
μ_i	Mean value of the i-th element in latent space of VAE
$\boldsymbol{\mu}_z$	Vector of mean values for latent space $\mathbf{z}$ of VAE
n	Number of times the memory has been accessed in ACT-R model
N	Number of contextual features in ACT-R model
N_n	Number of neurons in Hepfield network
N_p	Number of patterns in Hepfield network
$\mathbb{O}$	Number of available matching production rules in ACT-R model
$\mathbf{o}_t$	Output vector of LSTM at time step t
P	Likelihood of obtaining the observed data under the null hypothesis
p	vector autoregressive order
P_j	Probability of achieving goal by selecting the production rule j in ACT-R model
p_m	Partial matching penalty for memory m in ACT-R model
p_{pred}	Binary variable reflecting predicted driver reaction by engaging pedals
p_{true}	Binary variable reflecting driver reaction by engaging pedals
q	Contextual cue in ACT-R model
$\mathbf{q}_{att}$	Query vector of attention mechanism
$\mathbb{Q}$	Set of contextual cue in ACT-R model
r	Pearson correlation coefficient
σ_{stat}	Standard deviation
σ_i^2	Variance f the i-th element
s_m	Spreading activation of memory m from contextual cues in ACT-R model
s_{pred}	Binary variable reflecting predicted driver reaction with steering
$s_{q\to m}$	Spreading activation of memory m from contextual cue q in ACT-R model
s_{true}	Binary variable reflecting driver reaction with steering
τ	Close estimation of the time-to-collision in the concept of Lee [Lee76]

θ	Angle subtended by an obstacle on the driver's retina in the concept of Lee [Lee76]
θ_i	Threshold of neuron i in Hopfield network
trace_i	Memory trace of index i in ACT-R model
T	Sequence length
$Theta$	Hyperparameter distributions for Bayesian optimization
t_{trace_i}	Encoding time of the i-th memory trace in ACT-R model
U_j	Utility of the production rule j
$\mathbf{u}_t$	Vector of error terms in VAR model
v	Noise of production rule selection in ACT-R model
$\boldsymbol{\sigma^2}_z$	Vector of variances for latent space $\mathbf{z}$ of VAE
v_{ego}	Speed of the ego-vehicle
v_{front}	Speed of the front vehicle
$\mathbf{v}_i$	i-th value vector in attention mechanism
$\mathbf{V}$	Set of value vectors for attention mechanism
w_{ij}	Weight of connection between neuron i and neuron j in Hopfield network
$\mathbf{W}_\cdot$	Weight matrix of an LSTM cell with subscripts f for forget gate, x for input gate, c for cell state, and o for output gate
$\mathbf{w}_q$	Attentional weight of the cue q in ACT-R model
x_{ego}	Longitudinal position of the ego-vehicle
x_{front}	Longitudinal position of the front vehicle
ξ_i	State of neuron i in Hopfield network
ξ_i^μ	State of neuron i in μ-th pattern in Hopfield network
$\mathbf{x}_t$	Input vector of LSTM at time t
$\mathbf{X}$	Sequence of inputs
$\mathbf{X}^{\text{rev}}$	Sequence of inputs to backward LSTM in BLSTM
y_{ego}	Lateral position of the ego-vehicle
y_{front}	Lateral position of the front vehicle
$\mathbf{y}_t$	Vector of endogenous variables at time t
$\mathbf{Y}$	Stored patterns in Hopfield network
$\mathbf{z}$	Latent-space representation of autoencoder

Functions

$\delta(\mathbf{f_r}, \mathbf{f_m})$	Dissimilarity between the required features and the corresponding memory features
E	Energy function of Hopfield network
$\odot(\cdot)$	Element-wise multiplication of the vectors
$\text{LSE}(\cdot)$	Log sum exponential function
$\phi(\cdot)$	Softmax function
$\text{score}(\mathbf{q}_{att}, \mathbf{k}_i)$	Compatibility function between q and k_i in attention mechanism
$\psi(\cdot)$	Sigmoid activation function
$\tanh(\cdot)$	Hyperbolic tangent activation function

Abbreviations and acronyms

R^2	r-squared

ACT-R	adaptive control of thought-rational
ADF	augmented Dickey-Fuller
AE	autoencoder
AE-BLSTM-HA	autoencoder with BLSTM and Hopfield layers in encoder and decoder structure and attention mechanism
ANN	artificial neural network
ANOVA	analysis of variance
API	application programming interface
BLSTM	bidirectional long short-term memory
BLSTM-HA	BLSTM and Hopfield layers with attention mechanism
BMI	body mass index
BVP	blood volume pulse
DES	Differential Emotions Scale
DLPFC	dorsolateral prefrontal cortex
DVI	driver-vehicle interaction
ECG	electrocardiogram
EDA	electrodermal activity
EEG	electroencephalogram
EPIC	executive-process interactive control
fMRI	functional magnetic resonance imaging
HAM	Human Associative Memory
HMI	human-machine interface
HMM	hidden Markov model
HR	heart rate
HRV	heart rate variability
IBI	interbeat interval
KL	Kullback-Leibler
LOGO-CV	leave-one-group-out cross-validation
LSTM	long short-term memory
MAE	mean absolute error
manD 1.0	human dimension in automated driving - version 1.0
MAPE	mean absolute percentage error
NDRT	non-driving related task
OLS	ordinary least squares
PPG	photoplethysmogram
PPI	peak-to-peak interval
RMSE	root mean square error
RMSSD	root mean square of successive differences
RNN	recurrent neural network
SAE	society of automotive engineers
SCL	skin conductance level
SCR	skin conductance response
SDNN	standard deviation of normal-to-normal intervals
TOR	takeover request
TTC	time-to-collision

VAE	variational autoencoder
VAE-BLSTM-A	variational autoencoder with BLSTM in encoder and decoder structure and attention mechanism
VAE-BLSTM-H	variational autoencoder with BLSTM and Hopfield layers in encoder and decoder structure
VAE-BLSTM-HA	variational autoencoder with BLSTM and Hopfield layers in encoder and decoder structure and attention mechanism
VAE-LSTM-HA	variational autoencoder with LSTM and Hopfield layers in encoder and decoder structure and attention mechanism
VAR	vector autoregressive
VLPFC	ventrolateral prefrontal cortex
WMSE	weighted mean squared error

Glossary

affect	The experience of feeling or emotion encompassing both emotions and moods
automated system	A system responsible for dynamic control of the vehicle
driving behavior	The behavior of the driver that influences or indicates the driving trajectory, such as steering
degeneracy	The ability for many independent mechanisms to produce the same functional outcome
driver model	A functional model of driver that focuses on the actual behavior of driver
driver state	Driver's condition at a given moment, characterized in this research by the integration of interoceptive, exteroceptive, and proprioceptive signals
driver state factor	Signal or influence that contribute to and describe the driver state in detail
driver state feature	Quantitative measure derived from a driver state factor, used as input for a driver model
emotion	A complex psychological state that involves three distinct components: a subjective experience, a physiological response, and a behavioral or expressive response
exteroception	Brain's ability to sense, interpret, and integrate signals originating from exteroceptors
interoception	Brain's ability to sense, interpret, and integrate signals originating from within the body
proprioception	Sense of the relative positioning of body parts
situation awareness	Perception and understanding of the environment and situation
stationarity	A property of a time series where its statistical characteristics, like mean, variance, and autocorrelation, remain constant over time

Usage of generative AI models

‡MG	Media generation: Creating entire passages from given content.

Explanations for the usage of generative AI models and its notation: The

bottommost level at which the identification is presented regarding the possible uses of generative AI models are subchapters of the 2nd order (e.g., 1.1.1, which may also appear without numbering), as otherwise, the identification would disrupt the reading flow due to frequent occurrences. Algorithms used for implementing generative AI models are mentioned at least in the text or provided as pseudo-code to facilitate appropriate identification.

1 Introduction

Automated driving has emerged as a promising solution to enhance road safety and reduce the incidence of traffic accidents. The underlying premise is that automated systems, leveraging advanced sensors, artificial intelligence, and real-time data processing, can outperform human drivers in terms of precision, reaction time, and adherence to traffic laws. Despite these advancements, the realization of fully automated driving remains fraught with challenges. One of the most significant hurdles is the failure of the automated system to make safe decisions in all driving situations, especially in unforeseen or complex scenarios. This limitation arises due to conditions that were not anticipated during the system's development and training phases. In such situations, the intervention of a human driver becomes necessary. This intervention can take the form of collaborative control, where the human driver assists the automated system in making decisions and executing driving tasks, or a complete takeover, where the human driver takes full control of the vehicle. The transition between automated and manual control is critical and must be seamless to ensure safety. The interaction between the human driver and the automated system, particularly in scenarios requiring collaboration, plays a central role in maintaining road safety. Effective communication and coordination between the human driver and the automated system are essential to manage the control transition smoothly and mitigate the risks associated with sudden or unexpected takeovers. Understanding and optimizing this interaction is crucial for the development of reliable automated driving technologies that can be trusted to operate safely in diverse and dynamic driving environments.

1.1. Motivation and objectives

The motivation behind this dissertation stems from the critical need for efficient collaboration between automated driving systems and human drivers. For such collaboration to be effective, it is essential that the automated system understands and predicts the behavior of the human driver. To achieve this aim, the system must be equipped with a model that, by knowing the current and past driver states, can predict future changes in their state and subsequent reactions. This predictive capability is not only beneficial but necessary during various driving modes. During automated driving, it can enhance the acceptance of driving automation by reducing the driver's anxiety. During manual driving, it can improve the driver's comfort and situation awareness by sharing information,

leading to a more relaxed and informed driving experience. However, the most critical application of this interaction model is during the transition from automated to manual driving, particularly when the automated system reaches its operational limits.
Although technical advancements in automated driving are progressing, the challenge of effective interaction between the human driver and the automated system persists. This challenge is specifically pronounced in scenarios involving driving automation Level 3 (society of automotive engineers (SAE) L3; Standard J3016). At this level, the automated system controls the vehicle, but the human driver acts as a fallback for scenarios where the automated system cannot make a decision or fails. During these times, the driver might be engaged in a non-driving related task (NDRT), such as playing games, and must suddenly take control of the vehicle due to a takeover request (TOR) issued by the automated system.

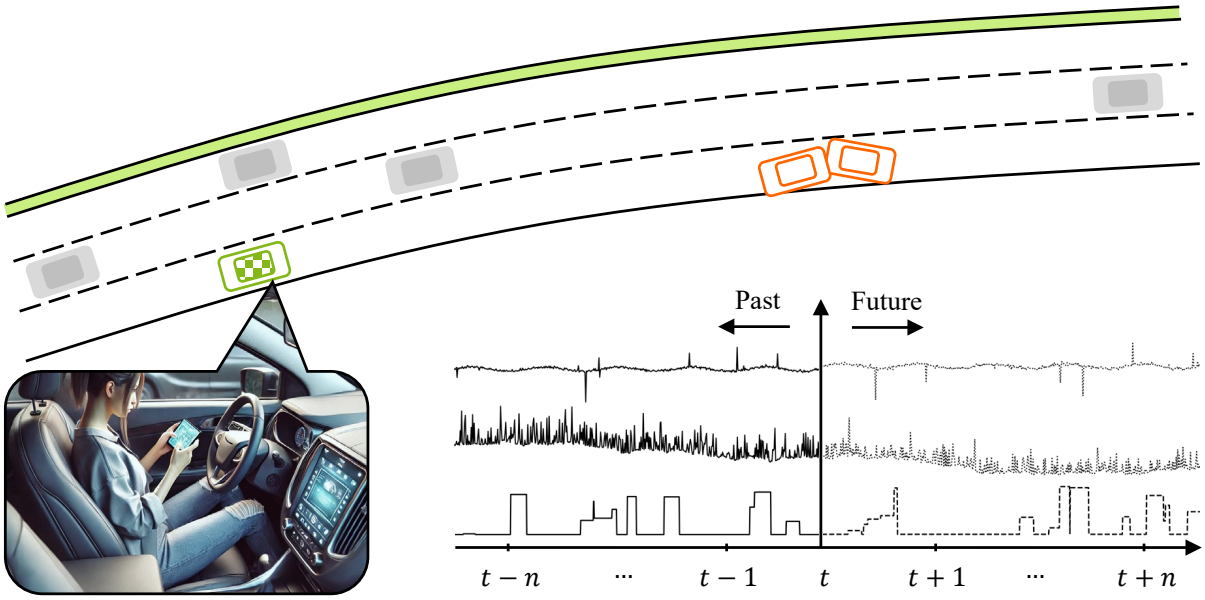


Figure 1.1.: A critical driving situation and a schematic driver state development through time ‡MG

An example of such a situation is depicted in Figure 1.1. Here, the ego-vehicle (checkered roof) is driving on a highway in automated mode at SAE L3, where the human driver is not required for vehicle control and is instead engaged in a game on their cell phone. Suddenly, two vehicles (white roof) crash at high speed directly in front of the ego-vehicle, creating a critical situation. Several factors make both automated and manual driving challenging in this scenario. For the automated system, challenges include the impulsive behavior of other traffic participants and the uncertainty surrounding the progression of the hazard. For the human driver, challenges stem from the high speed, the limited time available for reaction, and the fact that the driver is distracted by the game, rendering them less aware of the driving situation. If the automated system can manage the situation safely, the criticality is mitigated. However, if it cannot determine a safe response, the human driver will be prompted to take over the driving task. In this context, the distracted driver likely first perceives the crash through sound before receiving a TOR issued by the automated system. This TOR, if not tailored to the driver's current state, may induce panic, be

ignored, or lead to a misinterpretation of the situation, resulting in an unsafe reaction. This scenario is safety-critical due to the uncertainty about the driver's ability to respond promptly and safely. Therefore, the automated system must continuously monitor the driver to assess their state and perception, employing a driver model that accounts for their experience and predicts readiness and reaction. Then, by integrating this model into a control loop, the system can ensure that the driver remains informed and attentive to the driving situation when necessary. The primary objective of this dissertation is to develop such a model, utilizing psychological theories and neuroscience to accurately capture the complexities of human behavior and the decision-making process.

1.2. Main contributions and outline

This dissertation makes several contributions to the field of driver modeling and driver-vehicle interaction (DVI).

- **Design and execution of a subject study:** A comprehensive subject study is designed and executed using a driving simulator. This study is accurately crafted to collect synchronized data from individuals across a variety of driving contexts and driver states, encompassing a broad spectrum of emotional and cognitive changes. The driving simulator allows for a controlled and reproducible environment where different scenarios can be tested, and detailed behavioral data can be gathered.

- **Creation of the manD 1.0 dataset:** One of the key outputs of this research is the human dimension in automated driving - version 1.0 (manD 1.0) dataset [DB24a] gathered during the subject study. This dataset, which is open-access and available for further research [DB23b], includes comprehensive data on driver behavior, collected under various conditions and states. manD 1.0 stands as a resource for researchers aiming to explore driver behavior in automated driving systems further, enabling the replication of studies and the development of new models.

- **Analysis and extraction of initial correlations:** The collected data is subjected to thorough analysis to identify and extract initial correlations between driver characteristics, driver state, driving context, and subsequent actions. These correlations provide initial insights into the patterns of driver behavior, offering a foundation for future enhancements and studies.

- **Development of a driver model:** The core contribution of this study is the development of a driver model based on the neuroscientific findings and psychological constructs (latest version of adaptive control of thought-rational (ACT-R) cognitive architecture developed by Anderson [And07] and the theory of constructed emotion proposed by Feldman Barrett [Bar17]). This model simulates the human mind's ability to consider the driver state and the driving context, predicting subsequent changes in both driver state and context. By integrating cognitive and emotional theories, the model aims to provide a more accurate and nuanced prediction of driver reaction.

Overall, this work advances the field of automated driving by presenting a driver model that addresses the complexities of driving scenarios while incorporating psychological constructs and neuroscientific principles. In this way, it offers a more reliable prediction of driver behavior, particularly during critical transitions between automated and manual control. This contribution can enhance the safety and efficiency of automated driving systems and promote their acceptance by addressing the interplay between human drivers and automated systems.

The following provides a brief overview of the chapters contained in the present dissertation, highlighting their central contributions.

Chapter 2 deals with the historical progression of driver modeling objectives over the past decades. It introduces the concept of DVI and its associated terminology, providing a foundational understanding for the subsequent development of the driver model. Additionally, Chapter 2 explains the ACT-R cognitive architecture, which serves as one of the core frameworks for the model developed in this work.

Chapter 3 explores the neural architecture and activation patterns in the human brain and how they are mapped within the driver model. This chapter also introduces similar artificial learning algorithms from the field of machine learning, drawing parallels between biological and artificial systems to enhance the model's predictive capabilities.

Chapter 4 details the integration of emotion into the driver model based on the theory of constructed emotion. It introduces various physiological and behavioral metrics that are used as inputs and outputs within the model, providing a comprehensive approach to considering the emotional state of drivers.

Chapter 5 explains the data collection process for the manD 1.0 dataset, which is essential for training and evaluating the driver model. It details the experimental design, the scenarios used in the driving simulator, and the selected events from the dataset for modeling purposes.

Chapter 6 presents a thorough analysis of the collected data. It provides an initial summary of the data, offering insights into the patterns and trends observed in driver behavior under various conditions.

Chapter 7 outlines the structure of the computational driver model, detailing the selection of layers and activation functions. This technical description lays the groundwork for understanding how the model processes input data simulating the driver's brain and generates predictive outputs.

Chapter 8 evaluates the performance of the driver models that are separately trained on individuals. It compares the trained main model with a trained base model to quantify improvements in performance, demonstrating the model's effectiveness in predicting the driver state. Furthermore, the results of an ablation study are presented to clarify the influence of each component on the model's overall performance.

Chapter 9 discusses the findings of the study, draws conclusions, and suggests directions for future research. It encapsulates the contributions of the work and highlights the potential for further advancements in the field of automated driving.

Further explanations and supplementary materials are provided in the appendices. A list of the appendices with a brief explanation can be found below.

Appendix A provides terminology for different brain regions and views, serving as supplementary material for Chapter 3.

Appendix B gives a brief introduction to human sensations introduced in Chapter 4, which give the human brain information about actual states and help to decide on a response.

Appendix C describes the apparatus utilized in the subject study for data collection (discussed in Chapter 5), ensuring transparency and replicability of the experimental setup.

Appendix D explains the computational processes undertaken during and after the collection of manD 1.0 dataset, providing a comprehensive view of the technical aspects involved with the sensor data.

Appendix E presents the correlation matrix heatmaps for the features utilized in training the driver models. A more detailed discussion of the correlations is provided in Chapter 6.

2

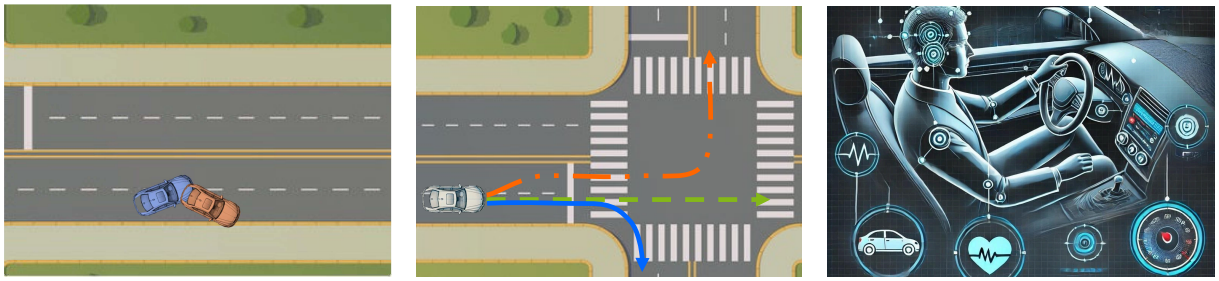
Evolution of driver models: fundamentals and state-of-the-art

This chapter presents a summary of the evolution of driver modeling over the last decades and provides the necessary background for the remainder of this dissertation. In the literature, driver models are used for a variety of purposes, ranging from predicting the number of accidents to forecasting driver state based on the cognitive architectures of the human brain. These models can be broadly categorized into rational and functional types. Rational models are predicated on ideal drivers who make optimal decisions based on their driving goals and environmental information. These models typically address aggregated driving behavior, such as Smeed's formula [Sme49] for estimating road fatalities or Wilde's risk homeostasis theory [Wil82], which suggests that drivers adjust their behavior to maintain a constant level of perceived risk. In contrast, functional models focus on the actual behavior of drivers, explained through cognitive functions and mental world models. These models aim to replicate the exact performance of drivers by considering their real-world behavior.
Michon [Mic89] criticizes rational models for neglecting crucial aspects of human driving behavior and argues that functional models, which link aggregate traffic behavior to individual information processing functions, are more effective for traffic safety. Given that rational models are not relevant to this study, the following section reviews only driver models with functional aspects or combinations of rational and functional aspects. For simplicity, these will be referred to as driver models throughout the remainder of this work.

2.1. Historical development of driver modeling objectives

Driver modeling has been studied in the literature for several decades. The objective of modeling has always been to increase traffic safety by predicting future events, with a focus on risk-taking behaviors. However, the object of prediction has evolved over time according to problem formulation (see Figure 2.1). Initially, models focused on predicting the number of accidents and resulting deaths. In the next generation of models, the primary aim was to predict driving behavior to assess road safety based on risky driving. The third generation of models, which are also aimed at collaborative and automated

driving, primarily focuses on predicting the driver state, including physical and mental conditions, to improve comfort and increase acceptance and trust in the systems alongside the increased safety. The following subsections provide brief overviews of driver modeling with the aforementioned prediction topics.



(a) Accident prediction (b) Driving behavior prediction (c) Driver state prediction ‡MG

Figure 2.1.: Driver modeling objectives

2.1.1. Accident prediction models

The first driver models focus on predicting accidents and driving failures with the aim of increasing road safety. Fuller [FUL84] proposes the threat avoidance model to investigate the dependency between stimuli and response at the functional level. In this model, associative mechanisms lead to avoiding harm and risky situations. It has both functional and rational levels that operate in a separate manner. The functional mechanisms are associations (reinforcements and punishments), and the behavior that results from these rules is rational since it optimizes an objective (subjective utility). The problem with this model is that as soon as the task becomes complex and includes parallel subtasks, the complexity of the rule-based model increases significantly.
Schank [Sch86] uses cognitive psychology to provide a mentalistic model. In a cognitive process, when a driver faces a situation, they recall a similar representation, adjust it to fit the new situation, then decide and react. For a new event, humans go through the recall process more, but for frequently occurring tasks, humans mainly consider the situational state [VH88]. Learning is failure-driven: when the existing representation fails, the brain makes changes, and the reaction to the new situation is learned. The model also has a rational aspect since recalling and tweaking involve changing past behavior to an optimal one for the current situation.
Michon [Mic85] indicates that a rule-based model with both functional and rational levels should be used, where the rules at the two levels do not necessarily correspond. The rules at the functional level cause behavior that is presumed to follow the rational rules as well.
Soar [Lai+86], a rule-based model, is applied to mimic the problem-solving behavior of humans. Since driving is a problem-solving and decision-making task, Soar can also be used to model it. In Soar modeling, a top goal is defined and some information about the problem state is available. The model looks for similar memories and identifies all scenarios in memory that have something in common with the new problem state. If Soar does not find any matching scenarios or finds several conflicting matching scenarios, it defines a new goal in the same direction as the top goal and performs the memory search

again. In all cases, it finds a pattern to solve the present problem. This feature of Soar in its learning mechanism is called chunking. Soar often assumes that behavior is rational and goal-directed, which may not always be the case. Human behavior can be irrational or driven by emotions, habits, or biases, aspects that Soar might not model accurately.
The generic error modeling system [Rea90] combines cognitive information processing with the three-level control hierarchy of Rasmussen [Ras83]. For example, a driver in routine driving performs periodic attentional checks. Whenever this check detects a problem, the cognitive process moves from skill-based to rule-based control. If no relevant rule is found for the new situation, the cognitive process shifts to knowledge-based control. This model focuses on modeling failures in monitoring and decision-making that cause accidents.

2.1.2. Driving behavior prediction

In this study, driving behavior refers to the behavior of the human driver that influences or indicates the driving trajectory. Driving behavior plays a crucial role in predicting the safety state, as a safe trajectory significantly enhances road safety. Numerous models have been developed to understand and forecast driving behavior, focusing on various aspects of the driving task.
One of the pioneering contributions in this field is made by Lee [Lee76], who proposes the concept of τ, defined as $\tau = \theta/\dot{\theta}$, where θ is the angle subtended by an obstacle on the driver's retina, and τ is a close estimation of the time-to-collision (TTC). Lee suggests that θ has a threshold value that, when reached, prompts the driver to brake, regardless of the vehicle's speed. Additionally, τ could be used for the control of braking, providing a fundamental basis for understanding collision avoidance behavior.
Wickens and Hollands [WH00] develop a comprehensive model of human behavior in driving that includes several stages: sensor processing, perception, cognition and memory, response selection, and response execution. According to their model, accidents occur due to failures in one of these stages or deviations from normal behavior. Factors such as tiredness, distraction, age, and driving experience significantly influence the frequency of these failures, highlighting the complexity of predicting and preventing accidents.
Tran et al. [Tra+15] employ a hidden Markov model (HMM) [BP66] to predict driving behavior intentions such as stopping or not stopping, changing lanes to the left or right, turning left or right, and general driving behavior. This approach leverages the probabilistic nature of HMMs to account for the inherent uncertainties in driving actions and decisions, providing a robust framework for predicting driver intentions. Similarly, Liu et al. [Liu+14] use HMMs to predict the trajectory of lane changes. They develop two separate HMMs: one trained on normal lane change data and the other on critical lane change data, which includes crash scenarios. This dual-HMM approach enables the differentiation between regular and potentially hazardous lane changes, thus improving the predictive accuracy for critical driving situations. Furthermore, the review by Elassad et al. [Ela+20] gives an overview of the models using machine learning methods for analyzing and predicting driving behavior.
All these models focus on typical driving behavior and the conditions under which deviations might occur. They aim to understand routine driver actions and the factors influencing them, providing a baseline for identifying abnormal or risky behaviors. Over-

all, the literature indicates that while significant progress has been made in understanding and predicting driving behavior, challenges remain, particularly in integrating diverse data sources and accounting for the wide range of factors influencing driver actions [Mar+12; Ran94].

2.1.3. Driver state prediction

Recent advancements in driver modeling increasingly focus on the mental state of drivers in addition to their driving behavior, as the mental state directly affects driving behavior and, consequently, road safety. Tavakoli and Heydarian [TH22] develop a model of driver state and behavior using naturalistic driving data. The inputs to this model include physiological data such as heart rate (HR) and gaze patterns, along with vehicle kinematic data. The outputs are classified driver state and driving behavior. The modeling method employs Bayesian change point detection [BH93] to identify patterns in the data. The performance of the model is demonstrated through two case studies, showing that it can successfully detect distinct patterns in driving behavior and driver state, such as elevated HR during stressful events. However, the model has limitations, including the need for high-resolution data and the challenges of generalizing findings across different driving contexts and individual driver behaviors. Additionally, reliance on predefined parameter values can restrict its applicability in more varied driving scenarios. In a subsequent study [Tav+23], the driver state is modeled with a focus on stress and cognitive load. The inputs to this model include various physiological measures such as HR, gaze direction, and skin temperature, while the outputs are the estimated latent states of the driver, including stress levels and cognitive load. The modeling method used is a state-space model [Mir19], which incorporates a Kalman filter [Kal60] to handle the time-dependent nature of the data and provide recursive estimates of the driver latent state. The model's performance shows promising results in accurately predicting driver state changes based on the observed variables, demonstrating its utility in real-world driving scenarios. However, the limitations include the complexity of modeling the dynamic interactions between multiple observed variables and the need for a psychological backbone to improve the model's structure. Additionally, a potential need for more extensive datasets to improve generalizability is reported.

Salvucci [Sal06] explores the development of a computational model of driving behavior within the ACT-R cognitive architecture. The model aims to simulate driver actions such as lane keeping, curve negotiation, and lane changing in a multilane highway environment. The inputs to the model include visual cues and environmental data, while the outputs are the driver's steering angles, lateral positions, and gaze distributions. The ACT-R cognitive architecture incorporates human-like limitations and constraints to produce a psychologically plausible representation of driving behavior. The model's performance is evaluated by comparing its outputs with real human driving data, showing a good match in steering profiles, lateral positions, and gaze behaviors. However, the model's reliance on predefined parameter values and its current inability to fully account for complex multitasking scenarios involving significant cognitive load limit the model's applicability.

Chao Deng et al. [Den+19] model the reaction time of drivers reading direction signs on expressways. The model uses the amount of information on the direction signs and the task

conditions (single-task vs. dual-task) as inputs and outputs the predicted reaction times for reading the signs. The employed modeling method is the queueing network-adaptive control of thought rational cognitive architecture, which integrates the queueing network [Liu+06] and ACT-R architectures to simulate human multitasking performance. The performance of the model is validated by comparing its predictions with empirical data from human participants, showing good alignment with a root mean square error (RMSE) of 0.3 s and a mean absolute percentage error (MAPE) of 12 %. However, limitations include the use of only one search strategy (e.g., left-right, top-down search order for tasks such as map reading), which may not account for variations in individual driver behaviors, and a focus on reaction time without modeling errors, which may impact overall prediction accuracy.

Considering driver state in driver modeling generally shows better performance. However, several aspects, such as generalizability across different driving situations and accounting for emotional variation in drivers, remain unaddressed. This dissertation aims to improve these aspects by developing a model structure that aligns with human brain activation patterns and integrates emotion into the model. This model can then enhance driver-vehicle interaction in assisted and automated driving scenarios.

2.2. Driver-vehicle interaction: key concepts and terminology

DVI is a critical component in the development of automated vehicles. The process of interaction involves both obtaining data from the driver and exposing the driver to various stimuli to ensure seamless communication between the human and the automated system. The collaboration between humans and technology necessitates precise product design grounded in psychological and physiological principles to accommodate the user effectively. In automated vehicles, DVI goes beyond mere interface design. It is responsible for information processing and transitions (takeovers) during dynamic, complex driving situations. This interaction is particularly vital in the initial generations of automated vehicles, where the driving task is shared between the human driver and the automated system. Efficient collaboration hinges on appropriate communication to achieve a mutual understanding of the driving situation, the state, and the intentions of both parties.

The distribution of tasks between the driver and the automated system varies with the level of automation, thereby altering the objectives and design of DVI. SAE [SAE18] defines five levels of driving automation: SAE L0 refers to no automation, where the human driver performs all driving tasks. SAE L1, or driver assistance, involves the system assisting with either steering or pedals, while the human driver is responsible for the other tasks. SAE L2, known as partial automation, allows the system to control both steering and pedals, but the human driver must remain engaged and monitor the driving environment. SAE L3, or conditional automation, means the system handles all driving tasks under certain conditions, though the driver must be ready to take over when requested. SAE L4, high automation, enables the system to perform all driving tasks in specific conditions without requiring driver intervention. Finally, SAE L5, full automation, allows the system to perform all driving tasks in all conditions without any human intervention.

At different levels of SAE automation, the goals of DVI vary depending on the level of

automation. Specifically,

- in SAE L1 and L2, the DVI aims to keep the driver attentive to the driving situation throughout the journey,
- in SAE L3, the interaction focuses on re-engaging the driver when the system reaches its operational limits, utilizing a TOR, and
- in SAE L4 and L5, the primary goal of interaction is to provide a comfortable experience for the driver (now more of a passenger) and to enhance trust in the automated system.

Generally, at higher levels of automation, DVI aims to increase driver comfort, while at lower levels, the focus is on safety, necessitating careful design of interfaces.
During automated driving at SAE L3 or higher, drivers are allowed to engage in NDRTs, which can significantly alter the driver state, reducing situation awareness and readiness for action. Flemisch et al. [Fle+11] provide comprehensive guidelines for designing human-machine interface (HMI) that help users form an appropriate mental model of the automated system. They stress the importance of verifying the driver's activity level before issuing a task transition request. The driver state assessment component plays a crucial role in this process, monitoring the driver both directly through sensors and indirectly by recording driver performance. This component is designed to detect inattention caused by distractions and drowsiness [Rau+10]. However, to accurately identify the driver state, it is necessary to consider additional aspects that capture the complexity of human behavior. To create effective interactions, the automated system needs a model that delineates the relationship between the driver state and the interaction signals within the context of automated driving [Lap+16]. Driver state encompasses various factors affecting a driver's abilities and can be subdivided into sensory, decision-making, and motor states [Dar+22a]. The sensory state involves the collection and interpretation of sensory inputs, with visual being the primary sense for drivers, supplemented by auditory and haptic feedback. Decision-making includes elements that influence cognitive processes such as workload, memory, attention span, emotional state, and intention. The motor state pertains to the position of body parts and the physical activities in which the driver is engaged.
This dissertation focuses on driving scenarios that involve takeover situations where the vehicle reaches its operational limits and hands over control to the human driver with warning stimuli. These situations are typical of SAE L3 automation. Marberger et al. [Mar+18] propose a comprehensive model for the transition process in SAE L3, outlining several phases: the automated mode with a compatible driver state, the takeover mode involving a transition of the driver state, and a post-transition mode where the driver intervenes and stabilizes vehicle control. The transition of the driver state involves shifting from an NDRT or any non-attentive state to a fully attentive driver state. Driver intervention refers to the deactivation of the automated mode by the driver, which can be executed in various ways depending on the system design [Gol16]. The control stabilization interval is the additional time required for the driver to achieve precise control and enhance performance to the average driving standards.

The interaction between the driver and the automated system is fundamentally a decision-making task. This decision-making results in an action, such as taking over the driving task, or it simply leads to a change in the driver state. Initially, decision-making is considered a cognitive process involving thinking, computation, and problem-solving. Therefore, the model proposed in this dissertation is developed based on a cognitive architecture. The subsequent section explains and details the foundational cognitive architecture considered for this model.

2.3. Cognitive architecture of the human brain: implications for driver modeling

Understanding human information processing is crucial for modeling driver behavior, particularly in the context of both manual and automated driving. Human information processing can be delineated into four fundamental stages of sensory processing, perception, decision-making, and motor response [Par+00], as depicted in Figure 2.2. Sensory perception involves the collection and interpretation of sensory inputs. In the context of driving, visual and auditory attention are the most relevant. Drivers rely heavily on visual cues to navigate, recognize hazards, and make split-second decisions. Auditory inputs, such as honking horns, warnings, or emergency sirens, also contribute to situation awareness. Effective sensory perception is the foundation upon which subsequent decision-making and motor responses are built. Once sensory information is processed, it must be evaluated and acted upon. Decision-making in driving is influenced by several factors, including emotion, attention, and cognition. Emotions can affect how quickly and thoroughly decisions are made [Hu+13; Jeo+14], while attention and cognition determine a driver's ability to process information and respond to dynamic driving conditions. High workload or divided attention can impair decision-making, leading to delayed or inappropriate responses [Ker+22]. The final stage of information processing is motor response, which involves executing physical actions based on decisions made. This includes activities such as steering, accelerating, braking, and operating other controls.

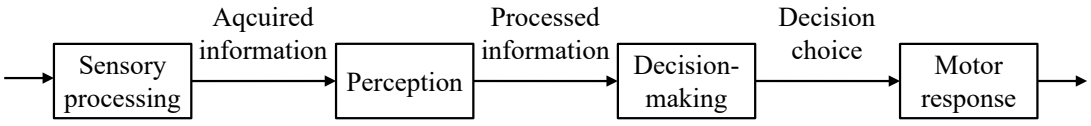


Figure 2.2.: Four stages of human information processing

Over the past decades, numerous architectures and decision-making models have been proposed to understand human behavior, with a focus on the cognitive aspect of decision-making. Some of these models are also used to predict driving behavior. In this dissertation, the ACT-R cognitive architecture is selected as the foundation for modeling the cognitive functions of the human brain in driving scenarios. This choice is grounded in the proven efficacy of the ACT-R model in multitasking and particularly in estimating drivers' steering reactions in modeling lateral vehicle control [Sal06]. The current subsection explains the structure of ACT-R, the communication between its modules, and details relevant to the driver decision-making process.

ACT-R cognitive architecture has been repeatedly implemented in various applications. Initially proposed by Anderson [And93], ACT-R has undergone numerous enhancements, with the most recent versions being continually refined. Ritter et al. [Rit+19] provide a comprehensive review of the development of the ACT-R architecture, highlighting key research contributions that have shaped its evolution.
ACT-R originates from the Human Associative Memory (HAM) model [AB73], which serves as the foundational memory simulation model. Over the years, ACT-R has integrated advancements in cognitive modeling and computational architecture. The progression includes the original ACT theory, which introduced procedural memory alongside HAM's declarative memory. Later versions, such as ACT-Embodied and ACT*, incorporate environmental interaction and a unified theory of cognition, respectively. The ongoing refinement in versions like ACT-R 2.0, 4.0, 6.0, and 7.0 introduces modular enhancements, including perceptual-motor modules and improved memory modeling, reflecting an increasingly sophisticated understanding of cognitive processes, influenced by both internal developments and external contributions from other models such as the executive-process interactive control (EPIC) [MK97].
As illustrated in Figure 2.3, the ACT-R architecture consists of several components, including modules for different cognitive functions and buffers that connect these modules to a central production system. The central production system (procedural memory) coordinates cognitive activities by selecting and applying rules based on the current state of the buffers. These buffers temporarily hold information that modules produce or require, ensuring efficient synchronization and information flow between cognitive functions. Both modules and the procedural memory can update the buffer values, ensuring that cognitive processes are well-coordinated. Further key modules within the ACT-R architecture include

- declarative memory that stores facts and knowledge, retrieving information from long-term memory as needed,
- goal module that manages the current task or goal, including internal states for problem-solving and decision-making, and prioritizes subgoals,
- sensory modules (visual and auditory), which process visual and auditory information, respectively, contributing to data acquisition and situation awareness, and
- motor modules (manual and vocal), which control motor functions, particularly those involving the head, hands, feet, and speech production.

Each module processes inputs from the environment or internal states, placing processed information into buffers. The procedural memory then uses this information to determine actions or further cognitive processing, guiding behavior and internal state changes. Each module contributes uniquely to the cognitive processes. The procedural memory contains rules essential for task decisions, pattern recognition, action selection, conflict resolution, and execution control. These rules evolve and become more efficient with task repetition, enhancing the model's ability to simulate learned behaviors. The goal module manages multiple goals, prioritizes tasks, and handles internal states, ensuring that higher-priority

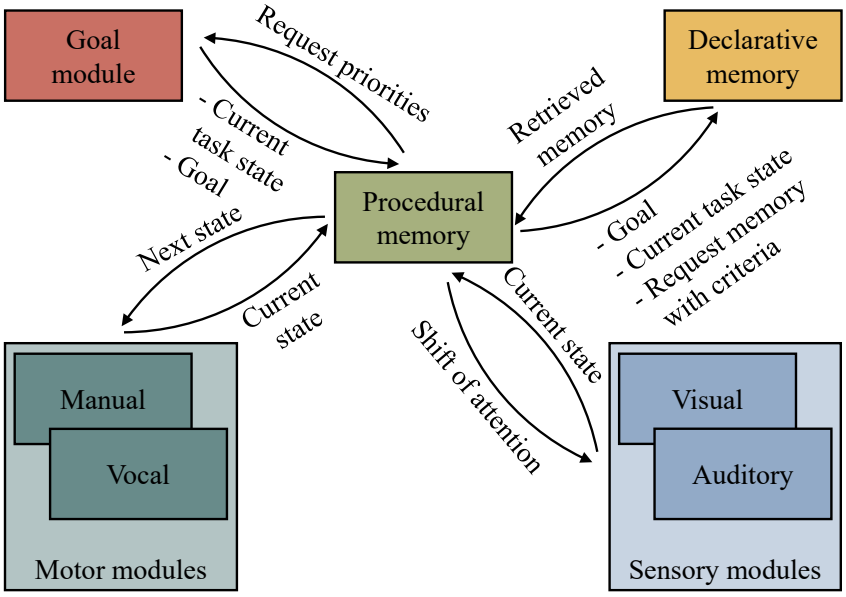


Figure 2.3.: Adaptive control of thought-rational architecture

goals take precedence in the cognitive queue. It maintains a hierarchy of goals, with higher numbers of subgoals potentially increasing the module's latency.

The declarative module is responsible for retrieving and storing facts, boosting the activation of frequently accessed information to reduce retrieval time. This module's efficiency is influenced by factors such as retrieval frequency and individual characteristics. The motor module oversees the control of body parts, such as hands and feet (manual), translating cognitive decisions into physical actions necessary for driving tasks and NDRTs. The sensory module (visual) manages visual attention, object detection, and identification, which are critical for maintaining situation awareness during driving.

The connections and data flow between these modules facilitate seamless cognitive processing. The procedural module requests goal priorities from the goal module, which, in turn, provides the current task state and goals. This information is then relayed to the declarative module. The declarative module supplies retrieved memory to the procedural module, aiding in the decision-making process. The visual sensory module identifies objects and locations and can occasionally shift attention based on directives from the procedural module to adjust gaze direction. The motor module reports its current state to the procedural module, which then directs the next motor actions based on the integrated cognitive state. Sensory and motor modules maintain a queue of tasks, executing them based on priority, ensuring that cognitive and physical actions are well-coordinated to simulate realistic driving behavior.

Salvucci [SAL01] applies computational ACT-R cognitive architecture to model driving behavior during multitasking (driving and dialing). The model comprises three core components: control, monitoring, and decision-making. Inputs to the model include visual angles to near and far points on the road, types of far points (lead car, tangent point, or vanishing point), and time headway to the lead car. Outputs are the driver's steering angles, lane positions, and gaze distributions. The data used for modeling are gathered through an empirical study where drivers navigate a simulated four-lane highway, encountering

moderate traffic and performing tasks such as lane keeping, curve negotiation, and lane changing. The secondary task is dialing with four different cell phone interfaces. The dialing methods are included and compared, namely full manual, speed manual, full voice, and speed voice. Different dialing speeds result in different occupation times and different dialing modalities (manual or speech) execution result in the occupation of different modules in the process (motor execution and visual attention versus speech execution and aural attention). The ACT-R model consists of a declarative memory that stores chunks of facts, a procedural memory that includes pairs of rules, and sensory and motor modules that interact with the environment. The rules depend on the goal and the state of the declarative knowledge. The sensory input affects the state of declarative memory, and based on the updated state, the procedural memory initiates motor actions. The model has additional performance parameters such as time constants for cognitive processing, perception of sensory data, and motor delays. The architecture also has a learning algorithm for adjusting these parameters. All of these parameters have initial values at the beginning, which work for general activities. The results show that the model successfully predicts aspects of lower-level control (like steering and eye movements) and higher-level cognitive tasks (like task management and decision-making). The model performs well in replicating human steering behavior and visual attention patterns during driving tasks, demonstrating strong alignment with human data. However, the model's reliance on predefined parameter values and the complexity of modeling dynamic interactions between multiple variables pose limitations to its applicability in more varied and real-world driving scenarios.

Besides its advantages, the ACT-R model has limitations, including its reliance on predefined parameter values and its challenge in handling complex multitasking scenarios. Additionally, it can only handle one goal at a time, it lacks automatic learning processes for procedural rules and declarative facts, and does not account for the driver's emotional state.

This dissertation aims to address these limitations by developing a driver model that incorporates a more robust multitasking capability, integrates learning processes, and includes emotional state considerations to enhance driver-vehicle interaction in assisted and automated driving contexts.

3

From biological neural networks of drivers to artificial neural network models

During both manual and automated driving, drivers continuously make decisions to ensure safety and efficiency. For effective driver modeling, it is essential to understand the functionality of the human brain and the entanglements of the decision-making process in greater detail. Investigating the neural activities that underpin these cognitive processes enables us to create a more accurate and reliable driver model. The following subsections look at the neuroscientific aspects of decision-making and the activation of different brain regions during various tasks. A detailed mapping between the activation patterns of brain neurons and the ACT-R model is presented. Afterwards, the structures of a selection of artificial neural network (ANN) models are discussed, which are designed to imitate the functions of the brain as closely as possible. By drawing parallels between human neural processes and computational models, a robust framework for simulating driver behavior is developed that can be used to enhance both manual and automated driving systems.

3.1. Neuronal activation patterns and ACT-R mapping

The human brain's anatomy encompasses several critical structures that contribute to its complex functions. These structures are divided into four primary parts: the cerebrum, diencephalon, brainstem, and cerebellum (see Figure 3.1).
The cerebrum, the largest portion, consists of the cerebral cortex and subcortical structures as illustrated in Figure 3.2. The cerebral cortex is the brain's outermost layer, playing a pivotal role in higher-order functions such as perception, cognition, and decision-making. This cortex is traditionally divided into four lobes: frontal, parietal, occipital, and temporal, each responsible for distinct but interconnected functions.
The frontal lobe, particularly the prefrontal cortex, governs executive functions and is crucial for planning, problem-solving [Unn+15], and motor skills. The dorsolateral prefrontal cortex (DLPFC) region is integral to tracking internal states and problem-solving steps, contributing significantly to goal-directed behavior and cognitive control. The ventrolateral prefrontal cortex (VLPFC) plays a role in memory recall and is crucial for cognitive tasks, supporting cognitive functions that depend on the retrieval of relevant knowledge from long-term memory.

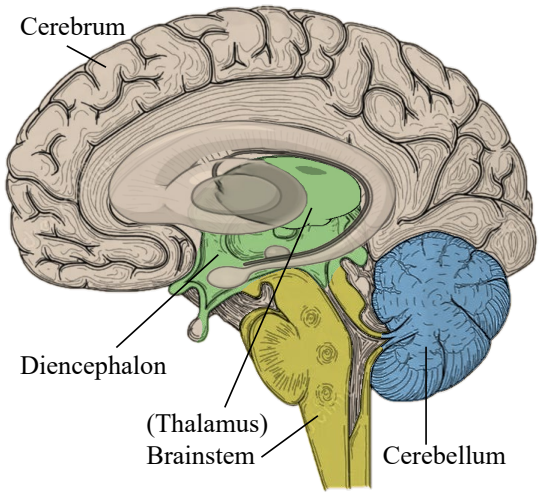


Figure 3.1.: Four major parts of the human brain

The parietal lobe processes sensory information from various parts of the body, enabling spatial awareness and coordination. Visual processing is the primary function of the occipital lobe, while the temporal lobe plays a critical role in auditory perception and memory formation. In addition to these four primary lobes, two other cortical regions are the insula and the cingulate cortex. The insula, located between the temporal lobe and the upper lobes, is deeply involved in interoception [Bud03], the brain's ability to interpret internal bodily signals such as HR, contributing to the conscious awareness of internal states [BK21]. The anterior cingulate cortex, located just behind the prefrontal cortex, is essential in emotion regulation, decision-making, and attention, integrating interoceptive information with emotional and cognitive processes [Has+17]. The neocortex comprises a significant part of the cerebral cortex in the human brain. It is referred to as the neocortex because it is the most recently evolved portion of the cerebral cortex, distinguishing it from the older, more primitive regions of the brain.
Beneath the cortex lie the subcortical structures, which include the basal ganglia and the limbic system. The basal ganglia, comprising the striatum and pallidum, are central to coordinating movement and procedural learning. The striatum, which receives input from various cortical areas, is crucial for movement control and reward processing, while the pallidum serves as a major relay station within the basal ganglia circuitry. The limbic system, another key subcortical structure, is responsible for regulating emotions, behavior, and long-term memory. This system includes the hippocampus, critical for memory formation and retrieval, and the amygdala, which processes emotional experiences and emotional memories.
The amygdala is not a single structural or functional entity but rather a collection of regions interconnected with other brain systems [SP98]. For example, the lateral and basal nuclei are often considered extensions of the cerebral cortex. The lateral nucleus, in particular, receives signals from various sensory systems, including visual, auditory, somatosensory, olfactory, and gustatory pathways. Other parts of the amygdala are connected to different brain regions, allowing it to integrate diverse types of information. While the lateral nucleus acts as the primary sensory gateway to the amygdala, the central nucleus serves as a major output center [Šim+21]. It is particularly important for the expression of

innate emotional responses and associated physiological reactions. The flow of information within amygdala circuits is regulated by multiple neurotransmitter systems. Despite this understanding, the mechanisms by which these chemical systems interact to establish the overall functional state of the amygdala remain poorly understood [LeD23]. The amygdala's role involves detecting salience, processing uncertainty, and contributing to predictions. The amygdala also has a critical function in social affect by recognizing and interpreting facial expressions of emotion [Šim+21].

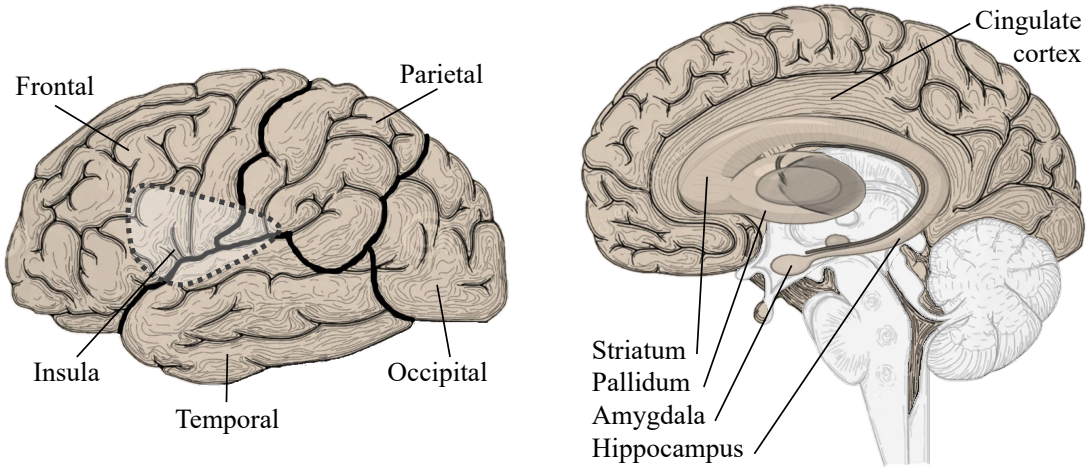


Figure 3.2.: Human brain cerebrum; left: four lobes of cerebral cortex and the insula; right: subcortical structures and cingulate cortex

The diencephalon (see Figure 3.1) is another major component of the brain that encompasses the thalamus. The thalamus serves as a central relay station for transmitting sensory and motor signals to the cortex. Table 3.1 summarizes the brain regions introduced and their roughly assigned functions, as mentioned earlier in this section. These regions represent key components involved in decision-making processes. Figure 3.3 illustrates a serial progression of these regions, from sensory input to decision-making and action. However, it is important to note that, due to the brain's complex structure, these steps do not always occur sequentially; parallel connections and interactions are also possible. Through these interconnected regions, the human brain orchestrates the intricate balance of cognitive, sensory, motor, and emotional functions necessary for daily life and survival. Anderson et al. [And+04] discuss the integration of perceptual-motor, goal, declarative, and procedural modules in the ACT-R cognitive architecture through previously collected knowledge of brain activity and a brain imaging study using functional magnetic resonance imaging (fMRI) with the goal of finding neural correlates for the ACT-R model concepts. Figure 3.4 presents a visual summary of the findings at a glance. Their study involves an artificial algebra task where participants solve symbolic equations through mental transformations [Qin+03]. This process involves multiple steps, such as moving and inverting symbols, performed mentally before keying out the final answer. Participants practice this task over five days, with their performance speed tracked. The fMRI data are collected as participants perform the task, looking at brain regions associated with different ACT-R modules. The fMRI data reveal distinct activation patterns for different cognitive tasks.

Table 3.1.: Overview of introduced brain regions and their roughly assigned functions

Region	Functions
Thalamus	Relay station for sensory and motor signals
Occipital cortex	Visual processing
Temporal cortex	Auditory perception, memory formation
Amygdala	Emotional experience and memory
Prefrontal cortex	Executive functions, planning, problem-solving, motor skills
Hippocampus	Memory formation and retrieval
Anterior cingulate cortex	Emotion regulation, decision-making, attention
Insula	Interoception
VLPFC	Memory retrieval from long-term memory
DLPFC	Tracking internal state, goal-directed problem solving
Basal ganglia	Movement coordination, procedural learning
Parietal cortex	Spatial awareness, coordination

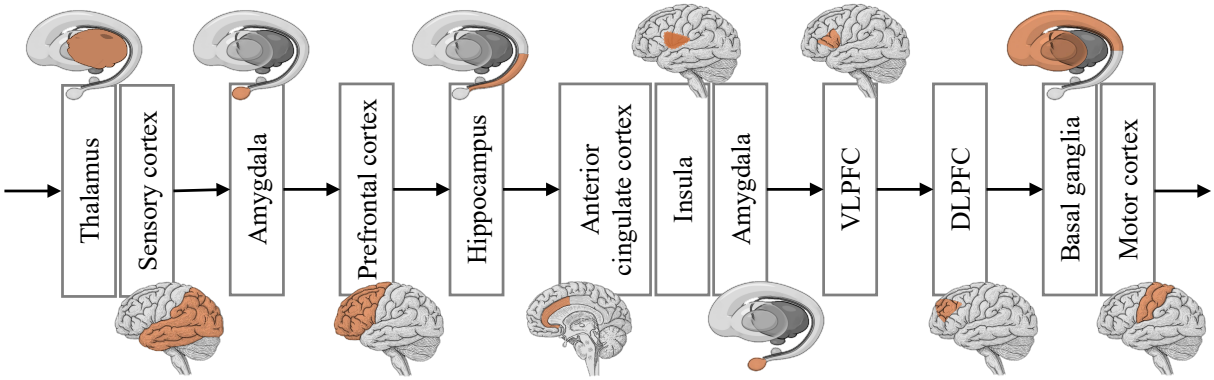


Figure 3.3.: Brain regions involved in decision-making

The DLPFC is associated with the goal module in the ACT-R cognitive architecture [Fin+02]. The DLPFC is implicated in working memory, which is essential for managing subgoals in complex tasks such as the "Tower of London" [Krc17] and the "Tower of Hanoi" [Sch10]. Its role in maintaining and manipulating information over short periods enables individuals to plan, initiate, and execute complex behavioral sequences necessary for achieving specific goals.

The posterior parietal cortex plays a significant role in maintaining the problem state within the ACT-R framework [And+04]. It contributes to the integration of cognitive functions by holding representations of problems, facilitating the manipulation and maintenance of information related to the task at hand. This region supports spatial awareness and the coordination of attention, which are critical for managing complex problem-solving activities.

The basal ganglia are proposed to implement production rules in the ACT-R architecture, performing essential pattern-recognition functions and playing a crucial role in procedural learning [Sto+10]. The basal ganglia's involvement in habit formation and the execution of learned behaviors makes it a key figure in procedural memory. This system allows

for the automation of repetitive tasks, freeing cognitive resources for more complex and novel problem-solving activities. Within the basal ganglia, the striatum is hypothesized to perform a pattern-recognition function [And+04]. It receives projections from cortical areas corresponding to various buffers and is involved in procedural learning and the execution of production rules. The striatum acts as a hub for the integration of cognitive functions and procedural memory, enabling the efficient execution of learned behaviors. The pallidum receives inhibitory projections from the striatum and, in turn, inhibits cells in the thalamus. This inhibitory mechanism is part of a pathway that ultimately influences the selection of actions in the cortex. The pallidum serves as a conflict-resolution function within the basal ganglia loop, contributing to the regulation of motor and cognitive actions. By managing inhibitory signals, the pallidum helps to prioritize and select appropriate responses during cognitive and motor tasks. The thalamus plays a critical role in the relay of information from the pallidum to the cortex, participating in the selection of actions [And+04]. It is a key component of the cortical-striatal-thalamic loop, which underlies the execution of production actions and procedural learning. The thalamus facilitates the flow of information necessary for executing production rules in ACT-R, supporting both cognitive and motor functions by ensuring efficient communication between different brain regions.

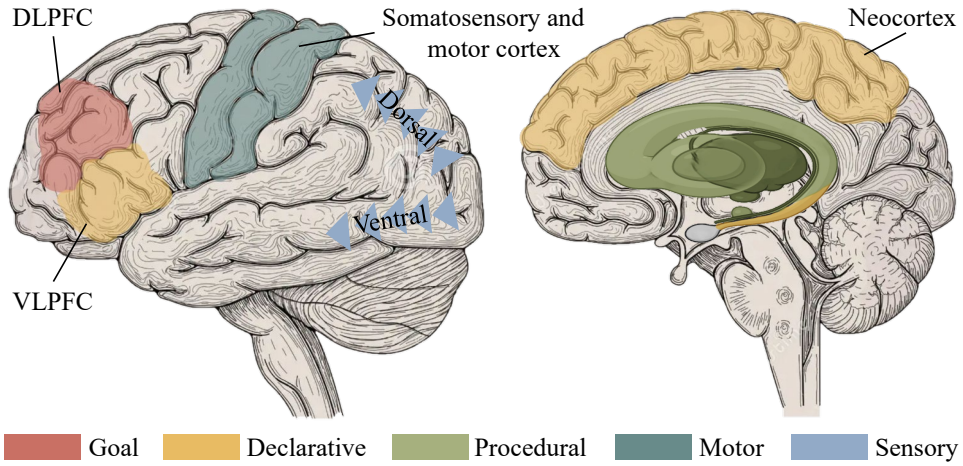


Figure 3.4.: Mapping of the ACT-R architecture onto the human brain

The visual and manual buffers are associated with the dorsal "where" and ventral "what" pathways in the visual system [And+04] and adjacent motor and somatosensory cortical areas for leg, hand, and face control [Pal+18], respectively. These buffers are essential for tasks involving scanning, typing, and mouse use. The integration of sensory information from the visual system with motor commands enables the precise coordination of eye and hand movements necessary for interacting with the environment.
Memory, a central component of the cognitive process, is modeled in the ACT-R architecture through declarative memory (facts and knowledge) and procedural memory. The life cycle of a memory involves three critical stages: encoding, consolidation, and retrieval [GC23]. Memory encoding is the initial phase where incoming information is transformed into a neural representation, often referred to as an engram or a trace in ACT-R terminology. The hippocampus plays a pivotal role in this process, as demonstrated by studies

on amnesic patients [Cor02]. Bilateral damage to the hippocampus results in severe amnesia, indicating its essential role in forming declarative memories, which include both semantic (general knowledge) and episodic (personal experiences) memories. The hippocampus receives organized projections from the cortical mantle as illustrated in Figure 3.5 and consists of densely interconnected projection neurons, enabling it to rapidly encode new information through associative learning mechanisms. Following encoding, memory consolidation stabilizes and stores these representations for long-term retention. This stage involves transferring information from the hippocampus to the neocortex. Evidence from patients with bilateral medial temporal lobe damage, who retain past memories but cannot form new ones [Cor02], supports the notion that while the hippocampus is crucial for the formation of new memories, it is not the ultimate repository for long-term storage. Instead, over time, memories become distributed across the neocortex as patterns of synaptic connections. This distributed storage aligns with the theory of systems consolidation, which proposes that the hippocampus and neocortex work together over extended periods, ranging from months to years, to stabilize memories [Dud12]. During this process, the hippocampus handles rapid, short-term learning and initial encoding, whereas the neocortex is responsible for the slow, long-term consolidation and storage of these memories. The VLPFC is linked to the retrieval buffer in ACT-R and is involved in retrieving information from long-term declarative memory [And+08]. This association aligns with neuroscience findings on memory and cognitive control [BW07], emphasizing the VLPFC's role in accessing and managing memory retrieval processes. This division of labor ensures that memories are efficiently encoded and robustly stored, facilitating their retrieval when needed.

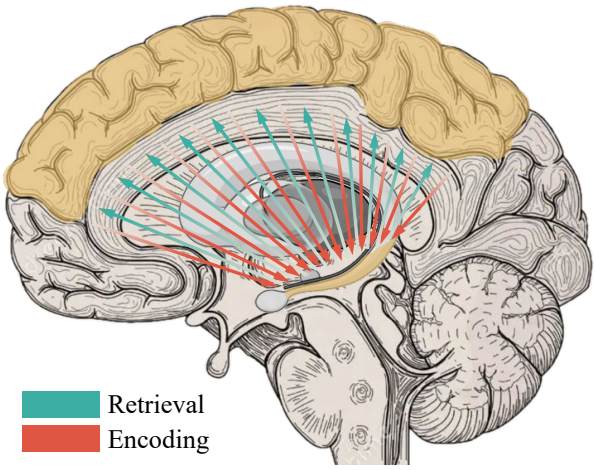


Figure 3.5.: Mapping of memory encoding and retrieval

In ACT-R, each memory m is represented only once in long-term memory, but each time it is recreated, a trace of it is listed and also kept in long-term memory, which means adding a new creation time to the list. The memory activation $a_{m,t}$ [Sto+23] in the declarative module of the ACT-R model is computed using the equation

$$a_{m,t} = b_{m,t} + s_m + p_m, \tag{3.1.1}$$

where $b_{m,t}$ represents base-level activation, s_m denotes spreading activation from contextual

cues, and p_m is the partial matching penalty (used when the retrieval cue only partially matches the memory). However, the partial matching term p_m is often not included in simpler models. The base-level activation $b_{m,t}$ reflects the history of how often and how recently the memory has been used and is computed as

$$b_{m,t} = \log \sum_{i=1}^{n} (t - t_{\text{trace}_i})^{-d_{\text{trace}_i}}. \tag{3.1.2}$$

Here, t represents the ongoing time, t_{trace_i} denotes the encoding time of the i-th trace, d_{trace_i} is the decay rate parameter of the i-th trace, and n is the number of times the memory has been accessed.
Spreading activation s_m models how activation spreads from related concepts or contextual cues to the target memory and is computed as

$$s_m = \frac{1}{N} \sum_{q \in \mathbb{Q}} \mathbf{w}_q s_{q \to m}. \tag{3.1.3}$$

In this equation, $\mathbf{w}_q$ is the attentional weight, N is the number of contextual features, $\mathbb{Q}$ is the set of contextual cues, and $s_{q \to m}$ is the strength of the link between cue q and memory m. The partial matching term p_m accounts for the similarity between the retrieval cue and the memory. It penalizes mismatches to reflect the reduced likelihood of retrieving a memory that only partially matches the cue and is calculated as

$$p_m = MP \sum_{\mathbf{f_r} \in \mathbb{F}_r} \delta(\mathbf{f_r}, \mathbf{f_m}), \tag{3.1.4}$$

where $\mathbf{f_r}$ and $\mathbf{f_m}$ represent features of retrieval and memory, respectively, MP is a mismatch penalty scaling factor, and $\delta(\mathbf{f_r}, \mathbf{f_m})$ shows the dissimilarity between the required feature and the corresponding memory feature.
In determining the $a_{m,t}$ formula, several psychological phenomena are considered. The power law of forgetting, recency, and frequency are three of the main effects that are reflected in the term $(t - t_{\text{trace}_i})^{-d_{\text{trace}_i}}$. The power law of forgetting states that memory retention declines in a predictable manner over time [NR81]. As time t increases, the influence of older traces trace_i diminishes according to a power law, where d_{trace_i} controls the rate of decay. Recency refers to the observation that more recent events are remembered better. When t_{trace_i} is close to t (i.e., the memory is recently encoded), the value of $(t - t_{\text{trace}_i})^{-d_{\text{trace}_i}}$ is higher, thus contributing more to the sum. Frequency effects indicate that memories accessed more often are easier to recall. Each additional retrieval adds another term $(t - t_{\text{trace}_i})^{-d_{\text{trace}_i}}$ to the sum, increasing $b_{m,t}$ and thereby making frequently accessed memories more active.
Another fundamental effect is the spacing effect, which suggests that information learned over spaced intervals is remembered better than information learned in a short period [Cep+08]. In the $b_{m,t}$ formula, the spacing effect is addressed by using trace-specific decay rates d_{trace_i} instead of a single decay rate d. The decay rate d_{trace_i} is computed as

$$d_{\text{trace}_i} = C \exp(b_{m,t_{\text{trace}_i}}) + \alpha, \tag{3.1.5}$$

where C is a constant and α is an additional decay parameter. This modification allows each memory trace to have a decay rate that depends on the base-level activation at the

time of encoding. When learning sessions are spaced out, each new trace is encoded when the memory's activation is lower, leading to a smaller decay rate d_{trace_i}. This effectively captures the spacing effect by ensuring that spaced learning results in more durable memory traces.

Finally, the fan effect describes how increased associations with a memory reduce its activation due to divided attention [AR99]. This is modeled in the s_m term. When a memory m is associated with many cues, the strength $s_{q\to m}$ is spread across these associations, diluting the activation each cue provides. Thus, s_m reflects the combined influence of all associated cues, but each cue's contribution is reduced as the number of associations (fans) increases.

Overall, the base-level activation $b_{m,t}$ models the power law of forgetting, recency, frequency, and spacing effects by summing the contributions of each memory trace, weighted by their age and frequency of access. The spreading activation s_m accounts for the fan effect by distributing activation from multiple contextual cues to the target memory. The partial matching p_m distinguishes similar memories to retrieve. Together, these components provide a robust framework for understanding and predicting memory activation in cognitive tasks.

These computations are based on the principles of Bayesian rational analysis inference, reflecting the probability of memory retrieval given prior usage and current context [And13]. The decay rate α and the strength of contextual links ($s_{q\to m}$) play significant roles in determining the dynamic availability of memories, aligning ACT-R's computational mechanisms with neurobiological processes observed in empirical studies.

Procedural memory in the ACT-R is essential for modeling adaptive and goal-oriented behavior [And+04]. The architecture comprises various modules that operate independently, exchanging information through buffers to maintain coherence. A central production system, which selects actions based on production rules, governs these modules. The selection of these rules is based on their utility, a noisy variable similar to the declarative activation. The utility of a production rule j is defined by the equation [Leb99]

$$U_j = P_j G - C_j, \tag{3.1.6}$$

where P_j is the probability that the current goal will be achieved if production j is selected, G is the value of the current goal, and C_j is the cost (typically measured in time) to achieve that goal. These parameters are learned from experience with the production rule. The probability P_j of selecting a production j among $\mathbb{O}$ matching productions is given by the softmax function $\phi(\cdot)$ as

$$P_j = \phi(U_j/v) = \frac{\exp(U_j/v)}{\sum_{j\in\mathbb{O}} \exp(U_j/v)}, \tag{3.1.7}$$

where v is the noise parameter, introducing variability into the selection process. ACT-R assumes a mixture of parallel and serial processing. Within each module, parallelism is prevalent, such as simultaneous processing in the visual system or parallel memory searches in the declarative system. However, serial bottlenecks exist at two levels: the content of any buffer is limited to a single declarative unit of knowledge (chunk), and only one production rule can fire at each cycle.

The procedural memory mechanism involves recognizing patterns in buffer contents, selecting a production rule, and updating buffers to achieve coherent behavior. This cyclic process takes about 50 ms, aligning with estimates from other cognitive architectures such as Soar and EPIC.

The architecture's production rules enable it to adaptively control thought processes by recognizing patterns and applying learned utility values to guide behavior. Despite its robust structure, ACT-R's procedural memory system faces challenges, including its reliance on predefined production rules and limitations in multitasking, where only one goal can be pursued at a time.

Overall, ACT-R's procedural memory framework provides a sophisticated model for simulating human cognitive processes, balancing parallel processing within modules with serial control through production rules to achieve adaptive and goal-directed behavior.

3.2. Relevant artificial neural network architectures

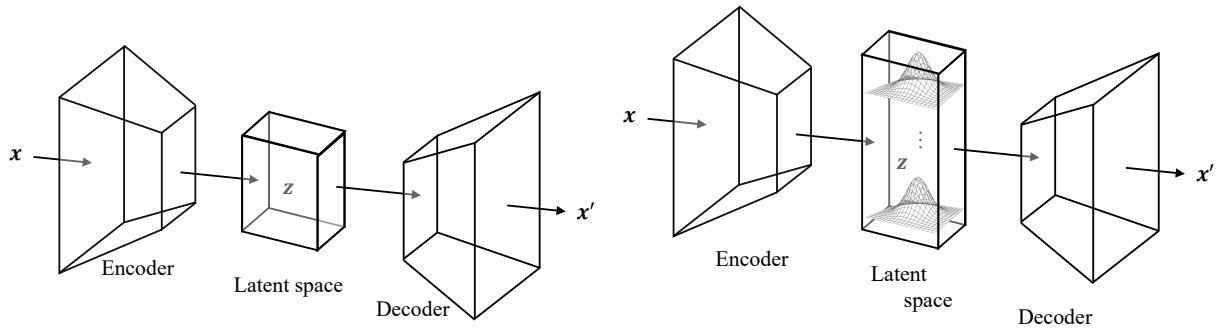
Neural activation patterns in the hippocampus and cortical areas during the encoding and retrieval of memories exhibit a resemblance to the structure and function of an autoencoder [Lio+14] in machine learning. The hippocampus and cortical regions work together to encode new information and later retrieve it, much like the encoder and decoder components. An autoencoder is a type of artificial neural network used for unsupervised learning and efficient coding of input data. The network learns to compress the input into a latent-space representation $\mathbf{z}$ and then reconstruct the output from this representation. An autoencoder consists of three main parts.

- Encoder: This part of the network compresses the input data into a latent-space representation. It reduces the dimensionality of the input, capturing the essential features required for reconstruction.

- Latent space: Also known as the bottleneck, this is the compressed representation of the input data. It is the key part of the autoencoder where the input information is encoded into a lower-dimensional space.

- Decoder: This part reconstructs the input data from the latent-space representation. It aims to produce an output as close as possible to the target data.

The strength of an autoencoder lies in its ability to learn efficient data representations in an unsupervised manner. It can capture the most important features of the data, making it useful for tasks such as dimensionality reduction, anomaly detection, and generative modeling. Figure 3.6a is an illustration of the structure of an autoencoder.



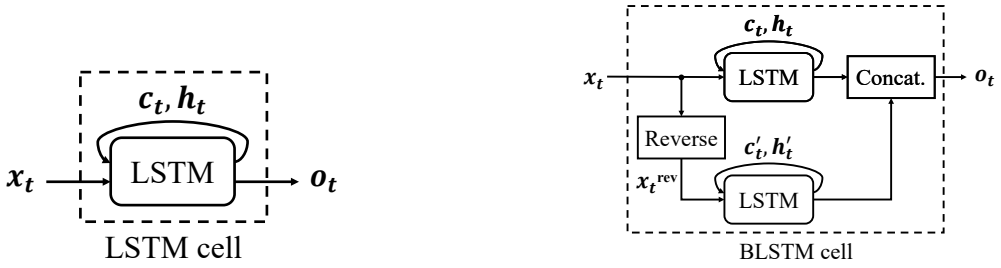
(a) Standard autoencoder (b) Variational autoencoder

Figure 3.6.: Schematic representation of the autoencoder architecture

A variational autoencoder (VAE) [KW19] is an extension of the traditional autoencoder that introduces a probabilistic approach to encoding the input data.
Unlike the standard autoencoder, which directly learns a mapping to a latent space, a VAE learns a distribution over the latent space $\mathbf{z}$, allowing for more robust and flexible data representations. In a VAE, the encoder produces the parameters of a vector of probability distributions (typically Gaussian [PR96]) rather than a single vector in the latent space. These parameters include the means $\boldsymbol{\mu}_z$ and the variances $\boldsymbol{\sigma^2}_z$, which define the distributions over the latent space.
During training, samples are drawn from the distributions. The decoder in a VAE generates outputs from the sampled latent variables, allowing for the generation of new data instances by sampling from the learned latent space. The strength of a VAE lies in its ability to generate new data samples that are similar to the training data, making it highly effective for generative tasks. It also provides a more robust latent representation by capturing the underlying distribution of the data. Figure 3.6b depicts the structure of a VAE.
The units and networks inside the encoder and decoder can be tailored based on the nature of the input data, such as images, numbers, or binary data, and the desired function of the network, whether it be for description or prediction. Examples of such networks include long short-term memory (LSTM) [Hoc97], bidirectional long short-term memory (BLSTM) [Zha+15], and Hopfield networks [Hop82].

LSTM networks are a type of recurrent neural network (RNN) specifically designed to model sequential data and capture long-term dependencies, addressing the limitations of traditional RNNs. LSTMs are particularly effective in tasks where context and sequence order are crucial, such as natural language processing, time series prediction, and speech recognition. LSTM is inspired by certain principles of memory and sequential processing that are important in biological neural systems. By applying various gates, it controls what information is stored, used, or forgotten over time. This mimics the brain's ability to store information temporarily and selectively forget unimportant information. Additionally, the LSTM has non-linear activation functions (like the sigmoid function $\psi(\cdot)$ and hyperbolic tangent function $\tanh(\cdot)$) that simulate the non-linear response of biological neurons to different inputs, which allows for complex and dynamic behavior based on various stimuli. The LSTM network is composed of LSTM cells, each of which contains several components that regulate the flow of information. Figure 3.7a presents the structure of an LSTM cell.

An LSTM cell includes several key elements.



(a) A standard cell (b) A bidirectional cell

Figure 3.7.: Two variants of long short-term memory (LSTM)

- Cell state $\mathbf{c}_t$ is a key feature of LSTMs that runs through the entire chain of cells, carrying information across time steps. It acts as a memory that can retain or discard information over long sequences.
- Forget gate $\mathbf{f}_t$ determines which information from the cell state should be discarded. It takes the current input $\mathbf{x}_t$ and the previous hidden state $\mathbf{h}_{t-1}$ as inputs, and applies $\psi(\cdot)$ to output a value between zero and one for each number in the cell state, where zero means completely forget and one means completely retain.

$$\mathbf{f}_t = \psi(\mathbf{W}_f \cdot [\mathbf{h}_{t-1}, \mathbf{x}_t] + \mathbf{b}_f) \tag{3.2.1}$$

- Input gate $\mathbf{x}_t$ controls the extent to which new information is written to the cell state. It also uses $\psi(\cdot)$ to output values between zero and one, determining which values are updated.

$$\mathbf{x}_t = \psi(\mathbf{W}_x \cdot [\mathbf{h}_{t-1}, \mathbf{x}_t] + \mathbf{b}_x) \tag{3.2.2}$$

- Candidate state $\tilde{\mathbf{c}}_t$ is created by passing the input through $\tanh(\cdot)$ to produce new candidate values, which could be added to the cell state.

$$\tilde{\mathbf{c}}_t = \tanh(\mathbf{W}_c \cdot [\mathbf{h}_{t-1}, \mathbf{x}_t] + \mathbf{b}_c) \tag{3.2.3a}$$
$$\mathbf{c}_t = \mathbf{f}_t \odot \mathbf{c}_{t-1} + \mathbf{x}_t \odot \tilde{\mathbf{c}}_t \tag{3.2.3b}$$

- Output gate ($\mathbf{o}_t$) determines the next hidden state ($\mathbf{h}_t$), which is used for the prediction of the next time step. It combines the cell state and the input to decide the next hidden state.

$$\mathbf{o}_t = \psi(\mathbf{W}_o \cdot [\mathbf{h}_{t-1}, \mathbf{x}_t] + \mathbf{b}_o) \tag{3.2.4a}$$
$$\mathbf{h}_t = \mathbf{o}_t \odot \tanh(\mathbf{c}_t) \tag{3.2.4b}$$

In the above equations, $\mathbf{W}.$ and $\mathbf{b}.$ denote the weight matrices and bias vectors of the LSTM cell, respectively. $\odot(\cdot)$ represents the element-wise multiplication of the vectors. LSTMs excel at capturing long-term dependencies in sequential data, addressing the vanishing gradient problem that hampers traditional RNNs. This capability is crucial for tasks where context from earlier in the sequence significantly influences later stages. The gating mechanisms in LSTMs allow them to selectively retain or forget information, making them robust to noise and irrelevant input.

BLSTM networks are an advanced variant of RNNs that enhance sequence processing capabilities by learning from input data in both forward and backward directions. Unlike traditional LSTM models that process data in a single chronological order, BLSTMs have two separate LSTM layers as depicted in Figure 3.7b: one for processing the input sequence from the beginning to the end (forward direction) and another for processing it from the end to the beginning (backward direction). For input sequence of $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_k)$, the backward LSTM processes $\mathbf{X}^{\text{rev}} = (\mathbf{x}_k, \mathbf{x}_{k-1}, ..., \mathbf{x}_1)$. This bidirectional approach enables BLSTMs to capture context from both past and future inputs around a particular input part, leading to a richer and more comprehensive understanding of the sequence. This performance is a gentle reminder of the brain cognition that uses the whole context in decision-making process. In BLSTM, the output is determined by combining the outputs from both LSTM layers, allowing the model to capture long-distance dependencies and context more effectively. The flexibility of BLSTM makes it especially suitable for tasks where important information can appear at any position within the input sequence. The BLSTM network uses the standard LSTM cell architecture with equations defining the gates and cell states, applied in both forward and backward directions.

The attention mechanism is another architecture inspired by the human brain's ability to focus selectively on information while processing complex stimuli. The attention mechanism and the amygdala share similarities in their role of detecting salience, as both prioritize processing relevant or significant information while filtering out less important inputs. The amygdala focuses on emotionally or socially salient stimuli, such as threats or facial expressions. Similarly, the attention mechanism enables models to dynamically weigh the importance of different parts of the input data, allowing them to capture long-range dependencies and intricate relationships within sequences more effectively than traditional architectures.

The attention mechanism operates by computing a weighted sum of input features, where the weights reflect the relevance of each feature to the current context. Given a query vector $\mathbf{q}_{att}$, a set of key vectors $\mathbf{K} = (\mathbf{k}_1, \mathbf{k}_2, ..., \mathbf{k}_n)$, and corresponding value vectors $\mathbf{V} = (\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_n)$, the attention output is calculated as

$$\text{Attention}(\mathbf{q}_{att}, \mathbf{K}, \mathbf{V}) = \sum_{i=1}^{n} \alpha_i \mathbf{v}_i, \tag{3.2.5}$$

where the attention weights α_i are obtained using the softmax function applied to a compatibility score between the query and each key

$$\alpha_i = \phi(\text{score}(\mathbf{q}_{att}, \mathbf{k}_i)). \tag{3.2.6}$$

The compatibility function score($\mathbf{q}_{att}, \mathbf{k}_i$) measures how well the query aligns with the key, commonly implemented as a dot product score$(q, k_i) = \mathbf{q}_{att}^\top \mathbf{k}_i$ or a scaled dot product score$(q, k_i) = \frac{\mathbf{q}_{att}^\top \mathbf{k}_i}{\sqrt{d_k}}$ to maintain stable gradients. d_k denotes the square root of the dimension of the key vectors, used for scaling and stabilizing gradients. In addition to these standard methods, alternative compatibility functions can include additive attention mechanisms, neural network-based scoring, or more complex similarity metrics tailored to specific tasks.

The classic Hopfield networks are a form of RNN that serve as associative memory [Suz08] systems, capable of storing and retrieving patterns. They are introduced by John Hopfield [Hop82] and have since been fundamental in understanding neural network dynamics and associative memory. The Hopfield network consists of a set of binary neurons that are fully interconnected. The network's architecture includes neurons, weights, and energy function.

- Neurons: In a Hopfield network, neurons are binary units that can be in one of two states, typically represented as +1 or -1 (or alternatively, 1 and 0). Each neuron's state is updated asynchronously based on the states of other neurons.
- Weights: Each connection between neurons i and j has a weight w_{ij}, which determines the strength and type (excitatory or inhibitory) of the connection. These weights are symmetric ($w_{ij} = w_{ji}$).
- Energy Function: The network's dynamics are governed by an energy function E, defined as

$$E = -\frac{1}{2} \sum_i \sum_j w_{ij} \xi_i \xi_j + \sum_i \theta_i \xi_i, \tag{3.2.7}$$

  where ξ_i and ξ_j are states of neurons i and j, respectively, and θ_i is the threshold for neuron i.

The network evolves to minimize this energy function, moving towards stable states (local minima), which correspond to stored patterns. The weights in a Hopfield network are typically determined using the Hebbian learning rule [CD08]

$$w_{ij} = \frac{1}{N_n} \sum_{\mu=1}^{N_p} \xi_i^\mu \xi_j^\mu, \tag{3.2.8}$$

where N_n is the number of neurons, N_p is the number of stored patterns, and ξ_i^μ represents the state of neuron i in the μ-th pattern. This learning rule ensures that the network can store and recall patterns effectively.
Unlike the classical version, which primarily operates with binary states and limited storage capacity, the modern Hopfield network [Ram+20] leverages continuous states and an exponential storage capacity relative to the dimensionality of its associative space. This allows it to store and retrieve patterns with high efficiency and minimal error.

Structurally, the modern Hopfield network includes a continuous update rule, characterized by a softmax function that aligns it closely with the attention mechanism. The update rule for continuous states is expressed as

$$\boldsymbol{\xi}_{\text{new}} = \mathbf{Y}\,\phi(\beta \mathbf{Y}^\top \boldsymbol{\xi}), \tag{3.2.9}$$

where β is a parameter that controls the sharpness of the softmax function and $\mathbf{Y}$ represents the stored patterns. This differs from the classical update rule, which relies on binary thresholding rather than continuous softmax adjustments.
The energy function for the modern Hopfield network incorporates an LSE$(\cdot)$ term to manage continuous states and is defined as

$$E = -\text{LSE}(\beta, \mathbf{Y}^T\boldsymbol{\xi}) + \frac{1}{2}\boldsymbol{\xi}^\top\boldsymbol{\xi} + \frac{1}{2}M^2 + \beta^{-1}\ln N_p, \tag{3.2.10}$$

where the LSE$(\cdot)$ is the log sum exponential function defined as

$$\text{LSE}(\beta, \mathbf{Y}^\top\boldsymbol{\xi}) = \beta^{-1}\ln\left(\sum_{i=1}^{N_p}\exp(\beta \mathbf{y}_i^\top\boldsymbol{\xi})\right) \tag{3.2.11}$$

and M is the maximum norm of the patterns. This novel formulation allows for exponentially many patterns to be stored and retrieved with a single update, making the modern Hopfield network a powerful and flexible addition to deep learning architectures. This network shares some structural similarities with the attention mechanism, but operates differently in terms of purpose and design. In modern Hopfield networks, a state vector (query) is updated based on stored patterns (keys) using an energy function that often involves an exponential interaction. The update rule resembles the attention mechanism mathematically because it also relies on a softmax-weighted sum of similarities between the query and keys. However, its purpose is to converge to a stored memory pattern (associative memory) rather than selectively emphasizing relationships within an input sequence.
Hopfield networks are well-known for their ability to store and retrieve patterns. When presented with a noisy or incomplete version of a stored pattern, the network can converge to the closest stored pattern, demonstrating robust associative memory capabilities. The concept of an energy function provides a clear and intuitive way to understand the network's dynamics. The network's tendency to minimize this energy function ensures stability and robustness in pattern retrieval. The simplicity of the Hopfield network's architecture, combined with the mathematical rigor of its underlying principles, makes it a useful model for studying neural dynamics and associative memory.
These networks can quickly store new patterns, reflecting the hippocampus's role in one-shot learning and its adaptability through Hebbian learning rules. Weber et al. [Web+17] utilize an extension of the Hopfield network model that serves as a model for understanding associative memory. The study models memory deterioration in participants with neurodegenerative diseases by examining two types of memory-recall tasks, face recognition task and discrimination task. The face recognition task involves highly correlated memories, where the network is trained to recognize and recall human facial images. The discrimination task uses random, uncorrelated memories near the capacity limit of the

Hopfield network to test memory recall under different levels of injury and noise. The performance of the model is evaluated by measuring the recognition scores for both tasks. The model successfully demonstrated a decrease in recognition ability with increasing levels of injury, quantified through the deterioration rates of memory recall as a function of injury. For instance, in the face-recognition task, the healthy network could correctly identify faces with 90 % accuracy, whereas this accuracy significantly dropped in injured networks. However, several limitations of the model are noted. The computational model is a simplified representation of the actual neural circuitry involved in memory tasks. It does not account for the full complexity of neuronal interactions and pathological effects. Moreover, the study primarily focuses on associative memory and does not consider other forms of memory impairment that may result from neurodegenerative diseases. Nevertheless, the study highlights the potential of using extended Hopfield models to estimate cognitive decline in neurodegeneration, despite the acknowledged limitations and simplifications of the approach.

The choice of networks inside the encoder and decoder can be optimized based on the specific nature of the input data and the expected function of the network. LSTMs and BLSTMs are ideal for sequential data such as time series or natural language, owing to their ability to capture long-term dependencies and maintain information over extended sequences. Hopfield networks are effective for associative memory tasks where the network needs to retrieve stored patterns that most closely match the input data.

4

Enhancing driver model: emotional aspect of decision-making

The ACT-R model has demonstrated significant efficacy in simulating the reaction time of drivers. The model excels in reproducing various cognitive processes involved in driving tasks, providing valuable insights into driver behavior under different conditions. However, a notable limitation of the ACT-R model is its lack of consideration for the emotional states of drivers.

To address the limitations of the ACT-R model, it is crucial to understand the concepts of emotion and affect and their roles in decision-making processes. Affect refers to the fundamental qualities of experience that are typically characterized by pleasure and arousal levels [Qui+21]. Emotions, on the other hand, are events that emerge from the brain's process of interpreting sensory input, including interoception, and are identified in specific contexts using emotion labels like anger or happiness [Qui+21]. These labels help classify strong affective experiences and are also shaped by external, situational cues. While emotions are acute responses to particular stimuli, affect also includes longer-lasting states such as mood. Mood, unlike emotion, is a more diffuse and enduring affective state without a clear trigger. The primary difference between emotion and affect lies in their duration and specificity. The next section explores the integration of emotional state in the driver model.

4.1. Emotion integration in driver model

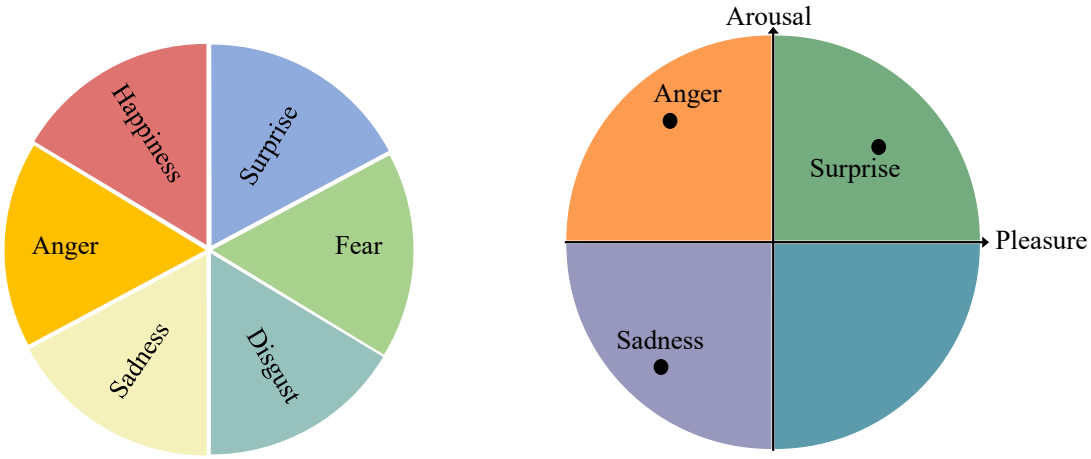
The interaction between a driver and their vehicle can be conceptualized as a decision-making task. This interaction requires the driver to make continuous decisions that can lead to specific actions, such as overtaking another vehicle, or changes in their internal state, such as heightened alertness or frustration. Traditionally, decision-making has been viewed as a purely cognitive process involving logical reasoning, computation, and problem-solving skills.

Recent advancements in psychological theories suggest that emotions play a critical role in decision-making processes [RB10]. Emotions can influence the strategies drivers use to process information and make decisions. For instance, the degree of pleasure associated with an emotion can determine whether a driver relies on top-down or bottom-up processing

strategies. When a driver is in a happy mood, they are more likely to engage in top-down processing, relying on pre-existing knowledge and expectations. Conversely, a sad mood tends to promote bottom-up processing, with increased attention to current details and stimuli [Sch00]. Lerner and Keltner [LK00] further elucidate the impact of emotional state on decision-making. They propose that information processing and decision-making are not only influenced by the degree of pleasure associated with an emotion but also by the underlying appraisal tendencies of different emotions. This means that even emotional states with similar levels of pleasure can lead to different decision-making outcomes based on their appraisal tendencies.

There are different approaches to understanding and categorizing human emotions. Among the most prominent approaches are the discrete emotion models, also known as categorical or basic emotion models [Iza77; Tom62]. These models propose that human emotions can be categorized into a finite set of distinct, universally experienced emotions. These models assert that each basic emotion is biologically determined, serving evolutionary functions essential for survival and social communication. Prominent proponents of this approach include Paul Ekman et al. [Ekm+87] (see Figure 4.1a), Carroll Izard [Iza77], and Silvan Tomkins [Tom14]. Discrete emotion models suggest that basic emotions are innate, meaning they are hardwired into the human brain from birth. This innateness implies that these emotions are not learned but are part of the human evolutionary heritage. These models argue that basic emotions are universally experienced across all human cultures. Despite cultural differences in emotional expression, the underlying emotions themselves are consistent worldwide. Each basic emotion is considered qualitatively different from the others. For example, happiness, sadness, anger, fear, disgust, and surprise are viewed as separate entities, each with its own unique set of physiological responses, facial expressions [Zha+21], and behavioral tendencies. Discrete emotions are believed to be linked to specific neural circuits and physiological processes. For instance, the fight-or-flight response associated with fear involves distinct autonomic nervous system activation patterns.

One of the most significant criticisms of discrete emotion models is their claim of universality. While basic emotions like happiness, sadness, and anger are widely recognized across cultures, the expression, interpretation, and contextual use of these emotions can vary considerably. Cross-cultural studies have shown that cultural norms and values significantly influence emotional expression and perception, suggesting that emotions may not be as universal as discrete emotion models propose [Har+15]. Besides, discrete emotion models typically focus on a small set of basic emotions, often excluding the vast array of nuanced and complex emotional states that humans experience. Emotions such as jealousy, pride, guilt, and nostalgia, which are common in human experience, do not fit neatly into the categories defined by basic emotion models. This limitation raises questions about the comprehensiveness of the discrete emotion approach. Furthermore, by categorizing emotions into discrete types, these models may oversimplify the complex and dynamic nature of emotional experiences. Human emotions are often multifaceted and co-occur in mixed or blended forms. For instance, one can feel both happiness and sadness simultaneously, a phenomenon that discrete emotion models struggle to account for adequately.



(a) Six basic emotions of Ekman et al. [Ekm+87] (b) Russell's 2D emotion model [Rus80]

Figure 4.1.: Examples of discrete (a) and dimensional (b) emotion models

Dimensional emotion models offer a different perspective from that of the discrete emotion models by proposing that emotions can be described along continuous dimensions rather than as distinct categories (see Figure 4.1b). This approach suggests that emotional experiences are complex and can be mapped within a multidimensional space, typically characterized by the dimensions of pleasure and arousal. Pleasure refers to the positivity or negativity of an emotion [Rus80]. Emotions can range from pleasant (positive pleasure) to unpleasant (negative pleasure). Examples include joy (positive pleasure) and sadness (negative pleasure). Arousal refers to the intensity or activation level of an emotion [Rus80]. Emotions can range from high arousal (excited, energetic) to low arousal (calm, relaxed). Examples include excitement (high arousal) and serenity (low arousal). Dimensional emotion models have their roots in the work of early psychologists such as Wilhelm Wundt [Wun97], who proposed a three-dimensional model (pleasure-arousal-tension). However, modern dimensional models often focus on the two primary dimensions of pleasure and arousal, as popularized by researchers like Russell (pleasure-arousal) [Rus80] and Mehrabian (pleasure-arousal-dominance) [Meh80].
Despite their utility, dimensional emotion models have several limitations and weak points that demand critical examination. One major limitation of dimensional emotion models is the potential loss of specificity. While these models effectively describe the general feeling (e.g., pleasant or unpleasant, activated or deactivated), they may not capture the qualitative differences between distinct emotions. For instance, the emotions of anger and fear may both be characterized by high arousal and negative pleasure, but they are qualitatively different experiences with unique triggers and behavioral outcomes. By reducing emotions to just a few dimensions, these models may oversimplify the richness and complexity of human emotional experiences. Emotions are often multifaceted and context-dependent, involving intricate interactions between cognitive, physiological, and social factors. Dimensional models may overlook these complexities by focusing primarily on pleasure and arousal. Dimensional models can sometimes fall short in explaining why certain stimuli evoke specific emotional responses. They describe the resultant emotional state but do not provide a detailed account of the processes leading to that state. This limitation can hinder a deeper understanding of the underlying mechanisms of emotion

regulation and expression. While pleasure and arousal are the most commonly used dimensions, they may not be sufficient to fully capture all aspects of emotional experience. Other dimensions, such as dominance or control (the degree to which one feels in control of the situation), might also play significant roles in shaping emotional responses. However, these additional dimensions are often not included in traditional dimensional models, potentially limiting their comprehensiveness.

Brain imaging studies have revolutionized our understanding of the neural underpinnings of psychological phenomena. However, the validity and reliability of these studies are contingent upon several foundational assumptions that have been called into question by recent research [Wes+23]. These assumptions, including the localization, one-to-one mapping, and independence assumptions, may hinder the accurate mapping of the brain's functions. The localization assumption posits that specific psychological events are caused by distinct neural ensembles localized in particular brain regions [Pos+88]. The one-to-one assumption suggests a unique correspondence between each neural ensemble and a specific psychological category [EG01]. The independence assumption presumes that neural ensembles function independently of their broader context, both within the brain and in relation to external factors [HB19].

These assumptions have provided a framework for interpreting brain imaging data, but have also led to issues with replicating findings across different contexts and studies. Contrary to the localization assumption, recent research indicates that mental events (any instance of a psychological category, such as an instance of behavior or subjective experience, e.g., memory recall, focused attention, emotional experiences, or performing an action) arise from distributed activity across the entire brain [BS13]. Studies have demonstrated that localized neural computations contribute to a variety of phenomena, and whole-brain models are more effective in capturing the complexity of mental events. Evidence from human neuroimaging and non-human animal studies supports this distributed approach, showing that a wide array of neural ensembles contributes to single psychological events.

The one-to-one mapping assumption is increasingly questioned by findings related to degeneracy, the phenomenon where multiple neural mechanisms can produce the same psychological outcome. Degeneracy implies a many-to-one mapping, where different patterns of brain activity can lead to similar behavioral and psychological states. Lindquist et al. [Lin+12] find that different individuals experiencing the same emotion (e.g., sadness) exhibited diverse neural activation patterns, suggesting that the same psychological state can emerge from different neural configurations. Similarly, during cognitive tasks, functional connectivity patterns vary yet lead to the same cognitive performance, highlighting the flexibility and redundancy in the brain's functional architecture.

The independence assumption is challenged by the brain complexity hypothesis, which posits that mental events emerge from dynamic interactions between the brain, body, and environment. Neurons do not function in isolation; their activity is influenced by their neural context and external factors. Hutchison and Morton [HM15] demonstrate that neural activity is modulated by physiological states (e.g., HR, breathing) and environmental contexts (e.g., ambient noise, social interactions). This interdependence suggests that brain-behavior relationships cannot be fully understood without considering the broader, interconnected systems that contribute to psychological phenomena.

To address the limitations posed by these assumptions, it is essential to sample relevant

signals from the brain, body, and environment together to capture the complexity of mental events comprehensively. An emerging framework, the theory of constructed emotion, proposes a novel understanding of emotional experiences, offering advancements over previous models. The theory of constructed emotion [Bar17] posits that emotions are not innate, biologically hardwired phenomena, but rather dynamic constructions of the brain. According to this theory, emotions are not universal, predefined categories but are constructed in real-time by interoception, prior experiences, and contextual information. This framework is rooted in the concept of predictive coding, where the brain continuously generates and updates hypotheses about sensory inputs and their causes [HR11]. The central concepts of the theory of constructed emotion include predictive coding, conceptual knowledge, core affect [Bar06], and contextual influence.

- Predictive coding: The brain predicts sensory input and adjusts its predictions based on incoming information. Emotions are constructed from these predictions and the resultant sensations.
- Conceptual knowledge: Emotions are shaped by an individual's prior experiences and knowledge. The brain uses this conceptual knowledge to categorize and interpret sensory inputs.
- Contextual influence: The context in which a sensation occurs significantly influences the construction of emotions. This includes environmental factors, social cues, and situational contexts.
- Core affect: This refers to the basic, continuous stream of feelings that ranges from pleasant to unpleasant and from activated to deactivated states. Emotions emerge when core affect is interpreted through the lens of contextual and conceptual knowledge.

The theory of constructed emotion offers several advancements over traditional discrete and dimensional models of emotion. Discrete models suggest a limited set of basic emotions, each with distinct biological signatures. In contrast, the theory of constructed emotion accommodates the vast variability in emotional experiences, acknowledging that the same physiological state can lead to different emotions depending on context and prior knowledge (flexibility and variability).
Unlike dimensional models that plot emotions on axes of pleasure and arousal, the constructed theory integrates the crucial role of context in shaping emotions (contextual sensitivity). This approach recognizes that emotions are not static points in a two-dimensional space but dynamic constructions influenced by ongoing contextual factors.
The theory aligns with contemporary neuroscientific findings that emphasize the brain's predictive nature and the importance of context in perception and cognition. It provides a more accurate framework for understanding how emotions are neurologically and psychologically constructed.
By highlighting the role of individual experiences and conceptual knowledge, the theory of constructed emotion accounts for the vast individual differences in emotional experiences (individualization). This contrasts with the more uniform predictions of discrete and dimensional models.

Similar to cognition, the brain constructs emotions at each moment based on bodily sensations, contextual knowledge from the environment, and prior experiences stored in memory. Emotions, therefore, can be directly integrated into the decision-making model, following the same pathways as cognitive processes. This suggests that emotions are stored alongside memories within the model, enabling their simultaneous retrieval and prediction. Consequently, the inputs to the model can be defined by the data necessary for constructing emotions, such as bodily sensations and contextual knowledge, which are collectively referred to as perception in Figure 4.2. This research hypothesizes that driver-vehicle interaction involves both emotional and cognitive decision-making processes. By integrating emotional states into the driver model, a more comprehensive and accurate representation of driver behavior can be developed. This integration acknowledges that drivers' decisions are not solely based on cognitive evaluations but are also significantly influenced by their emotional states. Thus, in this study, the theory of constructed emotion serves as the foundational framework for examining how emotions are constructed in human drivers. Driving is a complex task that involves continuous interaction with a dynamic environment, requiring real-time emotional regulation and decision-making. Traditional emotion models may fall short in capturing the nuanced and context-dependent nature of emotions experienced by drivers.

In the modeling approach, sensory perception and motor response, depicted in Figure 4.2, are assumed to be ideal for the sake of simplicity. This assumption implies that the environment state, internal state, and body schema are perceived by the driver with complete accuracy, exactly as they exist in reality. Consequently, the inputs to the driver model are taken directly from real-world signals rather than from perceived sensory signals. Similarly, the motor response is assumed to be perfect, meaning that the driver performs actions precisely as intended, with negligible motor response time. Under these assumptions, the focus of the modeling is on the decision-making component of human information processing, analogous to the procedural and declarative modules defined within the ACT-R architecture.

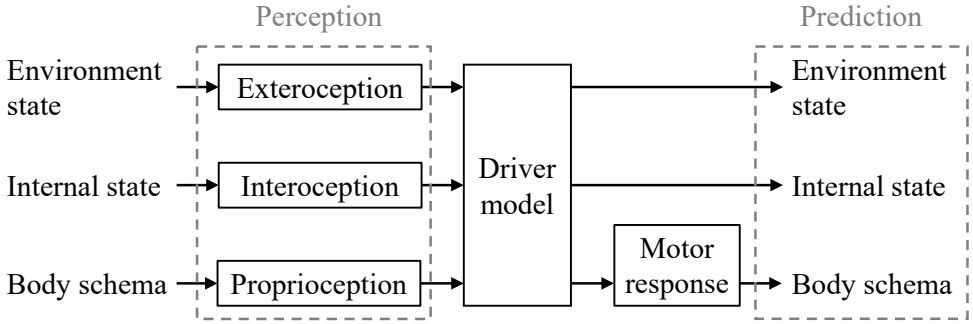


Figure 4.2.: Information processing of a driver with driver model as the decision-making component

In the future, models based on the present work can integrate the sensory perception and motor modules, including visual, auditory, vocal, and manual functionalities. It is essential to account for the consistent delay of approximately 0.05 s associated with each perception and manual reaction, which has been incorporated into the current driver model during the training phase. Therefore, when adding perceptual and motor components to this

model, it is crucial to avoid duplicating these constant delays to maintain accuracy and performance integrity.
In this research, a subject study is executed with synchronized sampling of relevant data from both the driver and the environment. Chapter 5 provides a detailed explanation of the data collection process. The data are gathered from individuals across several driving scenarios, ensuring sufficient within-subject data. This approach enhances the robustness and replicability of findings by accounting for individual variability. Following data collection, individualized models are created using machine learning methods. These methods offer deeper insights into the distributed nature of neural activity, providing a more accurate representation of brain-behavior relationships. The next section examines the factors driver state factor and features driver state feature that can be applied to the driving context and enhance the driver model, considering cognitive structure and integration of emotional state.

4.2. Key factors for driver modeling

The human brain has limited access to the outside, relying on three main streams of sensation consisting of exteroception, interoception, and proprioception. Exteroception is the perceptual inference on the environment state [Pet+17] and relies on the external senses such as vision and audition to perceive the surroundings. In the context of driving, exteroception encompasses all data collected from the driving environment and in-vehicle communication systems, such as traffic signals, road conditions, vehicle state, and auditory cues from navigation systems or other passengers.
In addition to external sensory inputs, the brain also receives information about the internal state of the body through the interoception process [BK21].
The third information source of the brain is the proprioception process, which is the sense of the relative positioning of body parts [TA18]. It allows us to perceive the position and movement of our limbs without having to look at them. This is accomplished through specialized sensory receptors located in the muscles, tendons, and joints. This awareness allows the brain to maintain homeostasis and respond appropriately to changes. More information about human sensations is given in Appendix B.
By integrating information from all sources and leveraging previous experiences, the brain predicts future events and makes informed decisions. This predictive capability is crucial in dynamic and complex activities such as driving, where timely and accurate decisions are essential for safety and performance. Table 4.1 presents an overview of some of the suitable factors and features for the driver state to be considered as inputs to the driver model. The following subsections discuss the mentioned factors. These factors encompass representations of the driving context, incorporating sensory inputs from the external environment, as well as internal sensory signals that reflect the internal state and body schema.

4.2.1. Factors influencing exteroception

The context in which drivers find themselves significantly affects their perception, estimation of situations, emotional state, prediction capability, and consequently, their

decision-making process. Understanding and modeling driver behavior necessitates considering the multifaceted elements that define the driving situation, surrounding events, and all stimuli perceived by the driver. This information, primarily received through the exteroceptors, is essential for accurate situation awareness and response. During both manual and automated driving, the primary information collectors for a human driver are vision and audition. Although tactition (the sense of touch) can also contribute, it is here neglected to simplify the model due to its relatively minor role compared to vision and audition.

Table 4.1.: Overview of driver state factors and features for driver modeling

Driver state	**Factors**	**Features**
Exteroception	Daytime and weather conditions	Daytime Visibility distance Road traction
	Ambient sounds and surrounding scenes	Visual distraction class Auditory distraction class
	Distance to others	Spatial distance to traffic members Spatial distance to objects
	Relative speed and acceleration	Relative speed Relative acceleration
	Criticality of the situation	Time headway TTC
	Visual engagement	Gaze direction Visual stimuli class Visual activity class
	Auditory engagement	Auditory stimuli class Auditory activity class
Interoception	Heart activity	Mean HR Standard deviation of normal-to-normal interbeat intervals Root mean square of successive differences between heartbeats Peak-to-peak interval Pulse amplitude Heart rate variability
	Respiration	Mean respiration rate Standard deviation Depth of inhale Variability of the rate

Continued on next page

Driver state	Factors	Features
	Electrodermal activity	Mean Frequency Amplitude of peaks in skin conductance response
	Skin temperature	Mean Variability
	Pupil dilation	Mean Variability Response amplitude to specific stimuli
Proprioception	Hand acceleration	Mean acceleration Peak amplitudes Standard deviation Variability
	Steering wheel engagement	Mean torque Maximum torque Frequency of engagement Engagement duration Angle variability
	Pedals engagement	Frequency Duration
	Driver activity	Activity classes
	Seat pressure readings	Mean pressure Maximum pressure Pressure distribution pattern Center of mass Pressure distribution variability Posture class Activity class

Factors defining the environment and traffic conditions

Several factors characterizing the environment and the vehicle's surroundings are directly perceived by drivers, shaping their exteroception of the external context.

Daytime and weather conditions: These conditions profoundly influence visibility and road traction, affecting driver perception and reaction times. For instance, driving at night or in heavy rain significantly alters how a driver perceives their surroundings and can increase the perceived risk of certain maneuvers.

Ambient sounds and surrounding scenes: Events such as nearby accidents, sudden lightning, or recreational activities (e.g., people camping) can distract drivers and draw their attention away from the road. These stimuli can lead to abrupt changes in driving behavior or driver state as drivers react to unexpected sights and sounds.

Proximity to other traffic participants and static obstacles: The spatial relationship between the ego-vehicle and surrounding vehicles or obstacles is crucial. Drivers continuously assess the distance to other traffic participants and any potential obstructions to adjust their perceived safety and navigate safely and efficiently.

Relative speed and acceleration of traffic participants: Understanding the relative velocity and acceleration of surrounding traffic helps to predict future positions and make informed decisions. Drivers use this information to maintain safe distances and anticipate necessary adjustments in speed or direction.

Criticality of the situation: Time headway and TTC are two extensively used metrics in the literature to measure situation criticality [Ram+21]. Time headway measures the time gap between two successive vehicles, providing a buffer time for the driver to react to the leading vehicle's movements [Eva91]. TTC estimates the time remaining before a collision would occur if the current speeds and trajectories are maintained [Hay72]. Both factors help to assess the urgency of driver responses and prevent accidents.

Factors defining in-vehicle conditions

Similar to the factors defining the environment, various factors within the vehicle contribute to shaping the driver's exteroception of the in-vehicle conditions.

Visual engagement of the driver: Drivers are exposed to various visual stimuli from the automated system, including information signals, warnings, and takeover requests. Any activity that visually engages the driver, such as reading or interacting with in-vehicle systems, also affects their situation awareness. Monitoring the driver's gaze direction can provide valuable insights into their focus and attention distribution.

Auditory engagement of the driver: Similar to visual stimuli, auditory signals from the automated system, including warnings and takeover requests, require the driver's attention. Other in-vehicle activities, such as listening to music or having conversations with other passengers, also engage the driver's auditory sense and can impact their ability to process external auditory information.

When using these factors to define the driving context, it is assumed that the driver perceives all this information through their external sensors. This assumption simplifies the modeling process but may not fully capture the nuances of human perception. To provide a more precise statement about a driver's perception, a deeper investigation into the driver's state, including cognitive and emotional factors, is essential.

4.2.2. Factors related to interoception

In the context of automotive environments, understanding the interoceptive states of drivers (future passengers) can enhance vehicle safety, comfort, and user experience. Interoception encompasses a variety of physiological signals that reflect the internal state of

the body. This subsection explores the physiological signals that can be measured to assess interoception, the methods for acquiring these signals, and the typical feature extraction techniques applied in this domain.
HR is the number of heartbeats per minute, while heart rate variability (HRV) refers to the variation in time between successive heartbeats. HRV is a key indicator of autonomic nervous system activity, reflecting the balance between the sympathetic and parasympathetic nervous systems. HR and HRV are commonly measured using electrocardiogram (ECG) sensors, which can be integrated into seat belts, steering wheels, or wearable devices like chest straps and smartwatches. Typical features include mean HR, standard deviation of normal-to-normal intervals (SDNN), root mean square of successive differences (RMSSD), and frequency domain measures such as low-frequency and high-frequency power.
The blood volume pulse (BVP) measures blood volume changes in the microvascular bed of tissue, reflecting cardiovascular dynamics. The photoplethysmogram (PPG) sensors, often embedded in wearable devices like smartwatches or integrated into vehicle steering wheels, use light-based technology to detect blood volume changes. Common features include the peak-to-peak interval (PPI), pulse amplitude, pulse rate, and derived HRV metrics.
Respiratory rate is the number of breaths taken per minute, providing insights into respiratory function and autonomic regulation. Respiratory rate can be measured using respiratory belts, impedance pneumography, or piezoelectric sensors embedded in seats or seat belts. Common features include mean and standard deviation of respiratory rate, and measures of respiratory depth and variability.
The electrodermal activity (EDA), the electrical conductance of the skin, varies with sweat gland activity and is influenced by sympathetic nervous system arousal. EDA is typically measured using sensors placed on the fingertips, palms, or wrists. Features include mean EDA level, skin conductance response (SCR) frequency, and amplitude of SCR peaks.
Skin temperature reflects peripheral blood flow and thermoregulatory responses. Infrared thermometers or thermocouples can be used to measure skin temperature. Mean skin temperature and temperature variability are two common features.
Pupil dilation is a measure of autonomic nervous system activity, particularly sympathetic arousal. Eye-tracking cameras can measure pupil size and response to stimuli. Features include mean pupil diameter, pupil diameter variability, and response amplitude to specific stimuli.
The interoceptive state of drivers in a vehicle can be assessed through various physiological signals, each providing unique insights into the internal state of the body. By leveraging advanced sensor technologies and feature extraction methods, the aim is to create a comprehensive picture of interoceptive processes.

4.2.3. Factors associated with proprioception

Proprioception refers to the body's ability to sense its position, motion, and equilibrium. It plays a crucial role in driving, where awareness of limb positions and movements is essential for safe and effective vehicle control. Proprioception involves various signals that reflect the body's awareness of its position and movement. In the following, the proprioceptive factors that can be measured from drivers in a vehicle are introduced, detailing the relevant

signals, measurement methods, and common feature extraction methods.
The driver's hands are among the most critical parts to observe, as they have the greatest freedom of movement in a seated position and are directly involved in controlling the vehicle. Accelerometers can be embedded in gloves, steering wheels, or wearable devices to measure hand movements and acceleration. Typical features include mean acceleration, peak acceleration, standard deviation, and jerk (rate of change of acceleration).
The engagement with the steering wheel and pedals provides insights into the driver's control inputs and limb positioning. Position, angular velocity, and torque on the steering wheel, as well as pressure on the pedals, can be considered. Common features include mean and maximum torque, frequency of engagement, engagement duration, and steering wheel angle variability.
Monitoring the overall activity of the driver helps in assessing the interaction with the vehicle controls and the driving environment. Cameras and inertial measurement units can be used to track the driver's movements and activities. Then, machine learning algorithms can process these data to identify specific activities.
Furthermore, pressure sensors or mats embedded in the driver's seat can capture the distribution of pressure across different areas of the seat. Seat pressure sensor readings provide information about the distribution of pressure on the driver's seat, indicating posture [Dar+22c], movement [DB22b], and position. Possible features to consider include mean pressure, maximum pressure, pressure distribution patterns, center of pressure, and changes in pressure distribution over time.

5

Subject study and data collection in a driving simulator

One main challenge in analyzing human states through machine learning, especially in driving contexts, lies in obtaining a dataset that captures the full scope of the driving environment and driver behavior [Vin+09]. However, developing and evaluating driver models effectively requires a comprehensive dataset that includes both internal driver states and external environmental factors. Such datasets are vital for producing accurate and reliable driver state assessments. An ideal dataset must be sufficiently large to support statistically robust conclusions, include a wide range of relevant information about both the driver and the driving environment, maintain high precision, and exhibit strong reliability.

Many available datasets focus on specific elements, like detecting drowsiness; however, driver behavior is shaped by multiple interconnected factors such as intention, cognition, and emotion. A more holistic dataset should reflect variations across these multiple factors. Additionally, driving dynamics and environmental conditions (e.g., traffic and weather) should be aligned with other collected data to enable a richer interpretation of driver behavior. Including these elements is crucial for a meaningful understanding of driving behavior. Moreover, data collection should encompass a diverse set of participants and driving scenarios to ensure that the analyses and findings are representative and comprehensive.

The present research introduces the manD 1.0 dataset [DB24a], developed through a controlled driving simulator experiment. manD stands for human dimension in automated driving. The static simulator setup and sensor configuration are depicted in Figure 5.1, and more detailed information on the apparatus is given in Appendix C. The controlled nature of the simulator allows for consistent reproduction of driving scenarios among participants, facilitating the study of critical situations and emotional states that may impair driver performance and increase accident risks. The manD 1.0 dataset is designed to be comprehensive, incorporating all of the aforementioned considerations. It integrates data from various sensors to measure synchronized psychophysiological indicators, along with vehicle and environmental state data, thereby providing an extensive view of driving conditions.

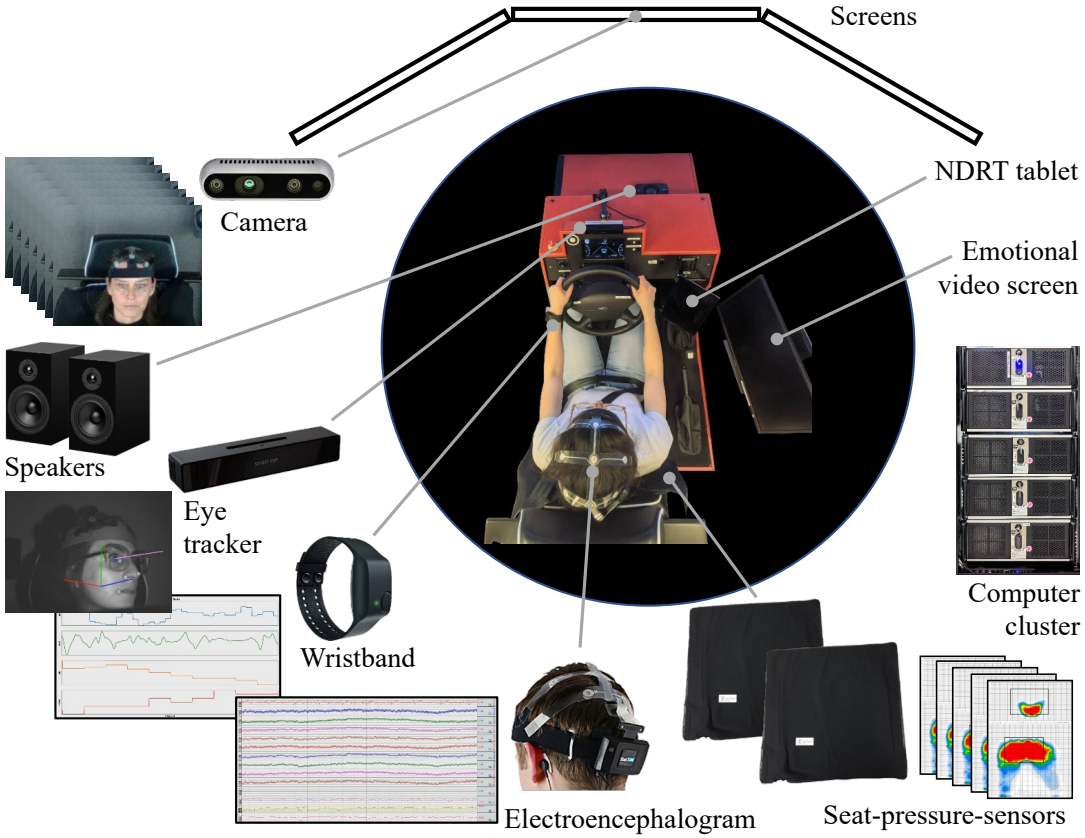


Figure 5.1.: Experimental equipment with a static driving simulator and driver monitoring sensors [DB24a]

The manD 1.0 dataset is openly accessible to the research community through a dedicated repository [DB23b]. Researchers have the flexibility to download the entire dataset or specific parts, depending on their research requirements. This versatility allows the dataset to support a wide range of research applications. For instance, it can be employed to develop models for understanding driving behavior based on data from manual driving sessions. Additionally, it serves as a valuable resource for predicting drivers' responses and actions in high-pressure scenarios, such as instances requiring the driver to take over control of the vehicle. The dataset also provides data crucial for building and refining cognitive frameworks and mental state models. It contains various interaction signals within different scenarios, enabling in-depth analysis of how these signals influence driver behavior. This paves the way for the development of innovative interaction strategies. Dargahi Nobari and Bertram [DB24a] provide a comprehensive description of the dataset. The following sections summarize the dataset and highlight the events in the dataset specifically utilized for driver modeling in this study.

5.1. Overview of the experimental procedure

Each participant in the study spends approximately two hours undergoing the experiment. Figure 5.2 presents an overview of the experimental procedure. At the beginning, participants receive detailed information about the purpose of the study, the procedures to be

followed, and the data collection process involving various sensors. After obtaining their consent through signed forms, the sensors are attached and calibrated individually for each participant. Following this, participants are allowed to practice in the simulator, becoming familiar with both manual driving and various levels of automated driving, from SAE L1 to SAE L3, while also getting accustomed to the simulated environment. They are free to practice multiple times until they are comfortable operating the simulator. During this phase, participants experience TOR and become acquainted with the NDRTs available. The TOR events involve auditory and visual cues. The auditory cues are composed of warning beeps and a spoken instruction indicating "Attention! Continue driving yourself!". The visual cues are displayed as text on the main screen and through color effects on the dashboard (refer to Figure 5.3). In certain takeover situations, alongside the text on the main screen, an additional reaction cue is provided in the form of arrows pointing left or right, or the text "Brake" serving as a recommendation for the driver. Three types of NDRTs are utilized: the auditory digit-span task [DM96], the n-back game [Kir58], and the Subway Surfers game (co-developed by Kiloo and SYBO Games, released in 2012) [GK12]. These tasks are accessible to the driver via a fixed tablet, which the automated system prompts them to use during specific moments. Drivers are encouraged to treat the simulation as if it were a real driving scenario, aiming to drive as naturally as possible.

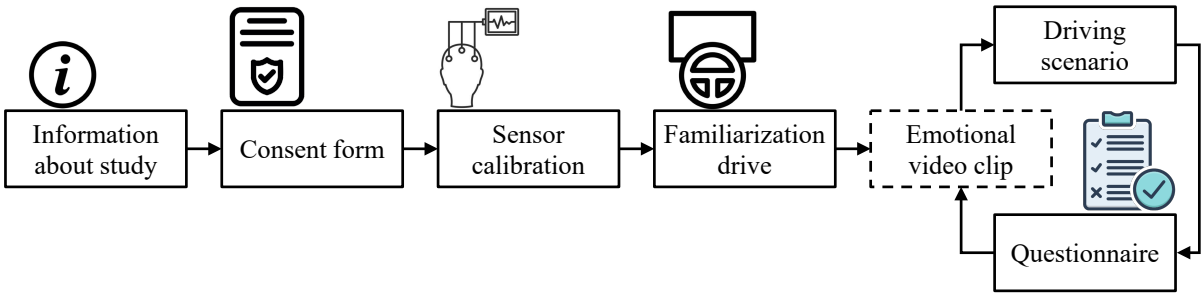
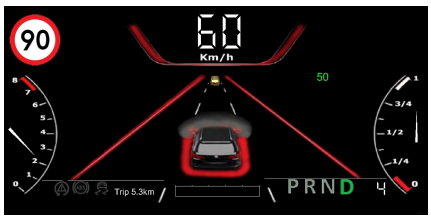


Figure 5.2.: Experimental procedure in the subject study

Participants then engage in five predetermined driving scenarios, presented in a varied sequence with rest periods between scenarios. Before and after these scenarios, participants complete the Differential Emotions Scale (DES) questionnaire [Iza+74], which helps them to provide a subjective evaluation of their emotions and to rate their feelings. Additionally, during the scenarios, the automated system periodically asks participants about their emotional state to confirm that any previously induced emotions have faded. To induce specific emotional responses, certain scenarios begin with the presentation of a video clip to the drivers.

During the scenarios, participants receive prompts or guidance from the automated system about the current driving context and available functionalities. However, they have the autonomy to follow these prompts or make their own decisions regarding responses and activities. This flexibility allows for the accumulation of comprehensive data on driver behavior and state across different driving conditions and scenarios.

(a) On the main screen



(b) On dashboard

(c) On the main screen with reaction clues

Figure 5.3.: Visual modalities of TOR [DB24a] in different speeds and speed limits

5.2. Statistical considerations in design of experiments

The aim of this study is to create a robust, multimodal dataset that captures various aspects of driver behavior. A synchronized multi-sensor system is utilized to monitor participants, gathering extensive data on a wide array of factors that influence driver state under diverse driving scenarios. In addition to capturing information about the driver, data on vehicle dynamics and environmental factors are collected. This approach ensures a holistic view of the interactions between the driver, vehicle, and surrounding environment, thereby enhancing the reliability and validity of the dataset.

The experimental design of this study focuses on capturing a diverse range of driver states and their variations to achieve a comprehensive and widely distributed dataset. Figure 5.4 visualizes the variation of several driver state factors across different events planned within the driving scenarios. In the figure, each (partial) ring represents a specific driver state factor, with a total of six factors encompassing audition, motor engagement of hands, motor engagement of feet, vision, cognition, and emotion. Each color corresponds to a distinct class of event, illustrating diverse combinations of state factors for events and, consequently, different combinations of these factors within the dataset. Factors aligned along the same radial axis are associated with the same event and situation, indicating that a single event can influence multiple driver state factors simultaneously. The angular width of the circular sectors represents the relative frequency of each planned event (i.e., state factor combination) for each participant.

The empty sections of the rings indicate areas where factors vary freely, suggesting that the corresponding event does not impose strict constraints on the state factor. Notably, in certain small circular sectors, all depicted state factors are empty, signifying that during these events, no specific manipulations are applied to the state factors. This portion of the dataset can serve as a control baseline and reference point, representing scenarios where no emotional induction or cognitive tasks are assigned to participants.

Additionally, the figure reveals instances where only one factor is affected or altered, providing an opportunity for in-depth analysis of individual factors. Conversely, there are events

where two or more factors vary simultaneously, offering insights into the potential interactions between these factors. This variation enables a comprehensive investigation of both isolated and combined effects of different driver state factors, enriching the understanding of their dynamics under various driving conditions.

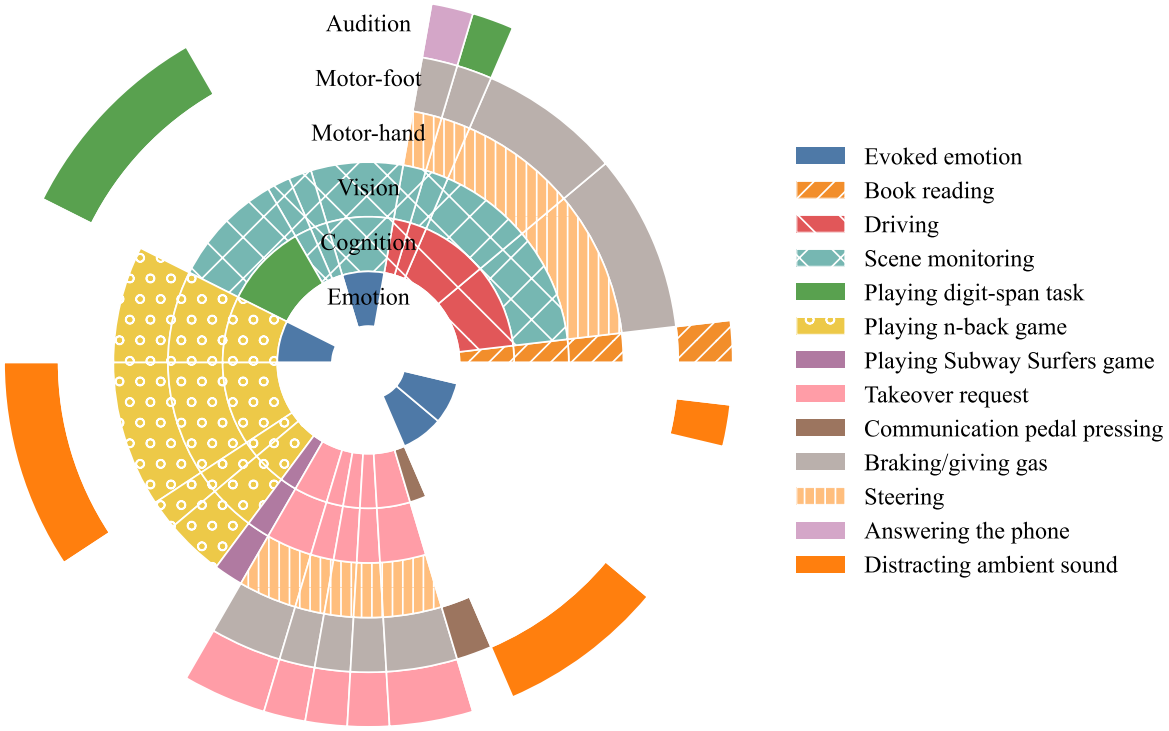


Figure 5.4.: Planned diversity of driver state factors during the experiment for each participant

Participants are informed about the overall goals and procedures of the study; however, specific details and focus areas are not revealed until the experiment's completion to ensure the validity of the data collected. The study's research framework takes into account various elements, including factors related to the driver, vehicle, and surrounding environment. These elements are represented through both qualitative (ordinal and nominal) and quantitative (continuous and discrete) variables. Potential confounders and covariates are carefully considered throughout the experimental setup. Characteristics like age and gender are treated as covariates and managed using stratified sampling [Ker+99]. Other participant attributes, such as height and body mass index (BMI), are documented in the dataset but were not included as covariates in the experimental planning because they were not predetermined before the experiment. Laboratory settings, such as lighting, noise, and temperature, are consistently maintained across all participants. It is assumed that the variables under study may have interdependencies that can be explored further with a sufficiently large dataset.

The study employs a repeated measures design to collect comprehensive data from participants. Each individual is tasked with driving through five different scenarios, allowing for the capture of varied data under diverse conditions from the same person. This approach provides insights into within-subject and between-subject variations. The study's sample

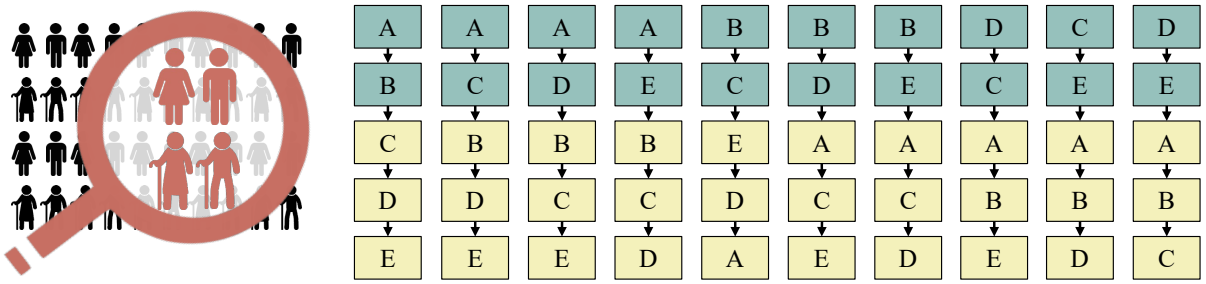


Figure 5.5.: Left: stratified sampling for age and gender covariates; right: partial counterbalancing of five driving scenarios, with each scenario appearing in one of two possible positions: either among the first two scenarios driven or among the last three scenarios driven, A: "no emotion", B: "anger", C: "surprise", D: "sadness", E: "fear"

comprises licensed drivers in Germany. From the original group of 50 participants, those who experienced motion sickness are removed from the core dataset, resulting in a final sample size of 39 participants. In small-scale studies, it is essential to maintain a balanced sample size and control for covariates. For age and gender, stratification is applied, with participants divided into four groups: female or male, and under or over 30 years of age (see Figure 5.5). To reduce potential biases like learning, practice, fatigue, and contextual effects, a counterbalancing method is employed across these groups. Due to the large number of participants needed for complete counterbalancing, a partial counterbalancing strategy is used. This ensures that each scenario is equally represented in both the early and later parts of the sequence in each group. Each participant completes the experiment once, without any repetitions.

5.3. Description of driving scenarios

The experiment includes one familiarization session and five distinct driving scenarios. The familiarization session aims to help participants get comfortable with driving in the simulator and adapt to the virtual setting before moving on to the core part of the study. The primary goal of the five driving scenarios is to induce various driver states and elicit specific emotional responses. Each scenario comprises a series of events, each slightly varied to reduce the learning effect, thereby enabling a thorough analysis of different driver state factors.

The length of each event is not fixed and varies for each participant based on factors such as their driving speed and when they switch between manual and automated modes. The duration for the takeover event is limited to 10 s, the maximum time allocated for the overtaking. If a participant fails to respond within this period, an accident is simulated, and the scenario is terminated. The n-back cognitive tasks are set to last around 30 s per round; however, this can be modified depending on the participant's level of engagement. Some events may be absent from certain participants' data if specific conditions are not met. For example, if a participant does not switch to automated mode, the n-back game will not be triggered, resulting in missing data for those events. With the exception of takeover situations, participants are not compelled to complete specific tasks, allowing

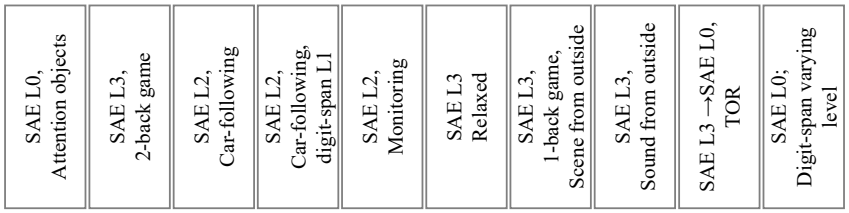
for a more authentic driving experience and resulting in different task completion times among individuals.

Each scenario has been uniquely crafted to have distinct characteristics, yet with enough commonalities to allow for a comparative analysis of the collected data. The "no emotion" scenario serves as an emotional baseline, providing a control condition without inducing any particular emotional state. In contrast, the other four scenarios are designed to evoke different emotional states using methods like emotionally charged video clips shown before driving and either monotonous or event-driven methods to provoke the targeted emotional responses [Dar+21a]. The design specifics of each scenario are outlined below and in Figure 5.6. Each scenario is named according to the emotion that is intended to be elicited by its emotional triggers. All scenarios, except the emotional baseline scenario, are divided into two parts. In the first part, the target emotion (whether anger, surprise, sadness, or fear) is induced and sustained. Each scenario starts with manual driving and incorporates elements designed to capture attention, including a pedestrian on the side of the road, an oncoming car and bicycle on the left side, and a person crossing the road despite the approaching car.

In the first segment of the emotionally driven scenarios, planned events include a car-following activity at SAE L2 with a slow-moving vehicle ahead, automated driving at SAE L3 without any specific tasks assigned to the driver, driving at SAE L3 while engaging in a 2-back game on the NDRT tablet, and auditory cues coming from outside the vehicle to capture the driver's attention. The arrangement of these events is based on the layout of the roads in the simulation environment. Apart from the initial attention objects event, the sequence of events is different across the scenarios. The second segment of each scenario is designed to establish a neutral emotional context without any deliberate emotional triggers. This part features additional tasks such as TOR, an auditory digit-span task, and speaking on the phone via loudspeakers.

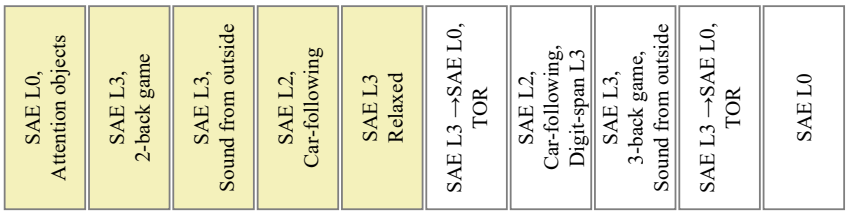
The familiarization drive takes place on a calm circuit of country roads, with no traffic signals or other vehicles, ensuring a safe environment for participants to practice without the risk of accidents. This session is set during midday to provide clear visibility, allowing participants to concentrate on getting accustomed to the simulator.

In the "no emotion" scenario, participants are not exposed to any particular emotional stimuli. This scenario also takes place during midday, ensuring an emotionally neutral environment. First, participants are engaged in a 2-back cognitive task, which then transitions to a 1-back task, offering a lower cognitive load. Then, a TOR is presented during the scenario, supplemented by a visual signal (the word "Brake" in red appears on the lower-left corner of the main display, near the navigation information, to prompt the driver further). Towards the scenario's conclusion, participants are introduced to an auditory digit-span task with varying difficulty levels ranging from one to five.



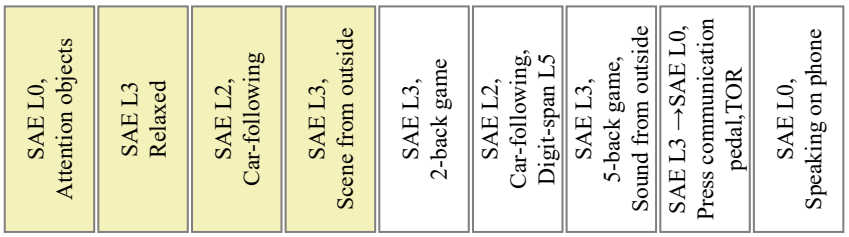
(a) "No emotion" scenario

(b) "Anger" scenario



(c) "Surprise" scenario

(d) "Sadness" scenario



(e) "Fear" scenario

Figure 5.6.: Sequence of the events in the driving scenarios. Colored blocks refer to adjacent emotional state.

The "anger" scenario is specifically designed to provoke feelings of anger in participants. To trigger this emotional state, participants watch a brief clip from the film Seven, sourced from the FilmStim database [Sch+10], prior to starting the driving session. This particular clip is known for its effectiveness in inducing anger. The scenario unfolds during midday, and throughout the drive, participants engage in both an n-back game and an auditory digit-span task, both set at a moderate difficulty level of 2. The scenario includes two TOR events. The first TOR is accompanied by a visual prompt (a red arrow pointing right). During the second TOR, which appears in the final part of the session, participants

are required to play the Subway Surfers game when a simple TOR is issued.
The "surprise" scenario is designed to generate surprise at the beginning. This is achieved through a series of unexpected events, such as a sudden snowfall during sunny weather, a pedestrian unexpectedly crossing the road and quickly retracing their steps, and then an abrupt end of the snowfall. Participants begin with an n-back task set at difficulty level 2, which later increases to level 3, while the auditory digit-span task is maintained at level 3. In the final part of this scenario, a TOR is presented with a visual cue: a red arrow pointing left, placed on the lower left of the main screen near the navigation details.
The "sadness" scenario is crafted to invoke and sustain sad feelings. To initiate this emotional state, a short clip from the movie City of Angels, also chosen from the FilmStim database, is shown before the driving session. This clip is recognized for its ability to evoke feelings of sadness. To reinforce the melancholic atmosphere, the music piece "Adagio for Organ and Strings in G Minor" (Tomaso Albinoni) plays during the initial part of the drive, up until the end of the first car-following event. Additionally, the rain and dark clouds further intensify the gloomy mood. Participants start with an n-back task at level 2 and advance to level 4, while the auditory digit-span task is also set at level 4. At a certain point, participants are asked to read aloud from a book, which keeps their hands occupied. Then a TOR appears as a group of horses crosses the road.
The "fear" scenario is intended to elicit fear. This emotion is triggered by showing a short clip from The Shining, selected from the FilmStim database, before the driving session begins. This particular clip is well-known for its effectiveness in inducing fear. To maintain a sense of dread throughout the scenario, the piece "A Night on the Bare Mountain" (Modest Mussorgsky) is played from the start until the beginning of the second car-following event. The driving environment is set to dark daytime conditions, requiring participants to use the vehicle's headlights. The session begins with an n-back game at difficulty level 2, which escalates to level 5, and an auditory digit-span task, which is also set at level 5, creating a high cognitive demand. Participants must perform a specific task of pressing a communication pedal whenever they see a deer, and a simple TOR is triggered when a deer is detected on the road in front of the vehicle.

5.4. Selected events from the dataset for modeling purpose

This study focuses on understanding driver behavior during takeover situations where the driver must assume control of the vehicle from an automated driving system. To develop a predictive model for a driver's reaction during these takeovers, segments from the manD 1.0 dataset that contain takeover situations with a TOR, are utilized for training and testing. Each scenario in this dataset includes one or two takeover events, providing a diverse range of driver states that contribute to the robustness of the predictive models. The variability in driver states, influenced by different cognitive and situational contexts, allows the model to learn from a wide range of driver reactions, enhancing its generalizability across various real-world scenarios.
The "no emotion" scenario features one takeover event. Before the TOR is initiated, the vehicle is in automated mode, and the driver is occupied with a 1-back cognitive task on the tablet. The TOR is triggered when a cat appears on a country road during midday

under sunny weather conditions. Drivers are given 10 s to react and take control of the vehicle after the TOR is issued. This scenario captures a relatively calm state, allowing the model to learn from a baseline with no extra cognitive load.
The "anger" scenario includes two takeover situations. In the first instance, the ego-vehicle operates in automated mode while the driver is relaxed. The TOR is triggered by the detection of a cow on a low-traffic highway, requiring an immediate driver response. In the second situation, the driver is engaged in the 2-back game on a tablet, which involves visual-motor processing. A TOR is issued due to a construction site that the automated system cannot safely navigate. The variation in driver engagement provides the model with data on different levels of cognitive load, enhancing its ability to predict motor reactions in higher-stress situations.
The "surprise" scenario also contains two takeover events. The first TOR is similar to the first TOR in the "anger" scenario, where the driver is relaxed in an automated mode. However, this time, the TOR is due to lane marking recognition errors, which pose a potential crash risk with roadway structures. In the second event, the driver is engaged in a 3-back game, a more complex cognitive task than the 2-back game. A TOR is initiated following an accident that happens in front of the ego-vehicle involving two other cars. This scenario emphasizes how unexpected changes in traffic conditions and cognitive load affect driver response times and actions, further diversifying the data available for model training.
The "sadness" scenario includes a single takeover event on a narrow country road. Here, the vehicle is in automated mode, and the driver is reading a book, thus occupying visual, cognitive, and motor resources. A TOR is triggered when a group of horses crosses the road, creating a dynamic situation that challenges the automated system's capacity to respond safely. This scenario provides insights into how motor and prolonged cognitive engagement in NDRT affects takeover readiness.
Lastly, the "fear" scenario takes place at night on a country road surrounded by woods. The driver is tasked with pressing the communication pedal near the gas pedal whenever a deer is spotted, involving a continuous visual-motor occupation. A TOR occurs when a deer unexpectedly enters the driving lane, necessitating immediate driver intervention. This scenario illustrates how nocturnal driving conditions and a visual-motor task influence driver responses during sudden, high-risk events.
Although the experimental design intended for all participants to undergo all driving scenarios and experience all planned takeover events, this was not fully realized due to various constraints. Time limitations prevented some drivers from completing all the scenarios, and technical issues such as malfunctioning sensors or unexpected software problems led to certain scenarios not running as intended. Additionally, the occurrence of specific events was contingent on the reactions and performance of individual drivers, resulting in some planned takeover situations not materializing. Consequently, the number of takeover situations experienced varied among participants. Table 5.1 provides an overview of the number of takeover situations available in the dataset for each participant, with bold values being the highest number of takeover situations per person. Given that the driver model needs to be trained individually for each subject and that a sufficient amount of data is required for effective model training, three separate models are trained, each corresponding to one of the three participants who have the largest amount of data

available in the dataset (six TORs).

Table 5.1.: Available takeover situations in the manD 1.0 dataset for each subject

Subject	**No emotion**	**Anger**	**Surprise**	**Sadness**	**Fear**	**Sum**	**Subject**	**No emotion**	**Anger**	**Surprise**	**Sadness**	**Fear**	**Sum**
1		++	++			4	19			+		+	2
2			+			1	20		+	+	+	+	4
3		+			+	2	21		++			+	3
5		++	++	+		5	22	+	+	+		+	4
6		+	++	+	+	5	23	+	+	++	+	+	**6**
9		+	++		+	4	39	+	++	+	+	+	**6**
10		+	+	+		3	41		+	++			3
11		++	++	+	+	**6**	42	+		++	+		4
12		+			+	2	43		+	++	+	+	5
13				+	+	2	44		+	+			2
14		+				1	46		+	+	+		3
15		+			+	2	47	+	+	++			4
16		++	+	+	+	5	48	+	++		+	+	5
17		+		+	+	3	50		+	+		+	3

6

Statistical analysis of the collected data

The selected subset of data from the manD 1.0 dataset is employed to train individualized driver models for each participant. Each driver model undergoes a separate training process using the data collected exclusively from that driver. The dataset encompasses comprehensive information from the driving context, interoception, and proprioception of drivers. For the preparation and preprocessing of the selected subset from the manD 1.0 dataset, a thorough statistical data analysis is first conducted. This comprehensive data analysis provides insights into the underlying distribution and patterns within the dataset. Driver modeling often requires the extraction of meaningful features from raw data. Data analysis facilitates the identification of key features that significantly impact driving behavior. Effective feature extraction and engineering enhance the model's ability to capture relevant driving behaviors and physiological responses. Features such as sudden changes in TTC or variations in BVP can be indicative of critical driving events or stress levels.

In datasets with multiple interrelated variables, such as those from drivers in a driving context, multicollinearity can be a significant issue. Data analysis allows for the examination of correlations between variables, helping to identify and address multicollinearity. For instance, if high correlations are found between pedal engagement and steering wheel movements, these features might need to be transformed or combined to reduce redundancy. Data analysis enables the identification of individual differences in driving behavior and physiological responses. This is particularly important for personalized driver models that adapt to individual drivers' unique characteristics. By analyzing data from interoception (e.g., BVP, EDA, skin temperature) and proprioception (e.g., hand movements, pedal and steering wheel engagement, seat-pressure distribution), it is possible to tailor the models to better fit individual drivers. This personalized approach ensures that the models accurately reflect the unique driving styles and physiological states of each driver. Finally, data analysis provides a foundation for the validation and verification of driver models. By understanding the underlying data characteristics and ensuring their quality, the models can be validated more effectively against real-world scenarios, verifying that they perform reliably under various conditions. This validation process is crucial for developing robust and reliable driver models that can be confidently deployed in real-world applications. The subsequent sections commence with a comprehensive analysis of the complete manD 1.0 dataset, providing an overview of participant characteristics, as well as exteroceptive, interoceptive, and proprioceptive data, alongside the subjective

ratings reported by participants. This analysis also outlines the measures implemented to ensure data quality and reliability for modeling purposes. Following this, a targeted analysis focuses on the subset of data chosen for this study, including the takeover events recorded from three selected participants.

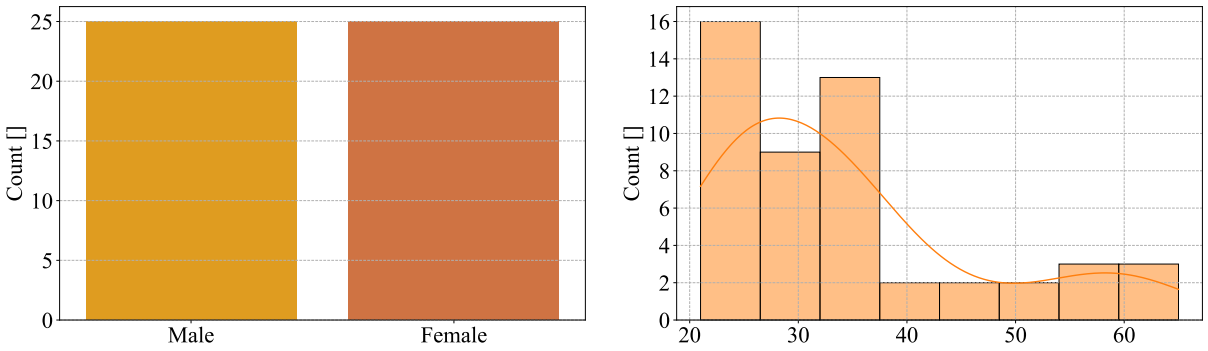
6.1. Analysis of participants' demographics and characteristics

This section provides an analysis of the characteristics of the participants included in the manD 1.0 dataset. The participants drove up to five driving scenarios in both manual and automated modes. The characteristic data collected include demographic details, driving experience, and other relevant metrics. This analysis provides insights into the diversity and background of the participants included in the dataset. Additionally, it sheds light on the demographic groups that exhibit greater interest in automated driving technologies. A summary of the participants' numerical characteristics is presented in Table 6.1.

Table 6.1.: Summary of manD 1.0 participants' numerical characteristics

Characteristics	m_{stat}	σ_{stat}	med_{stat}	min.	max.
Age in years	34.40	12.12	31.50	21.00	65.00
Driving experience in years	16.14	12.33	13.50	3.00	47.00
Height [m]	1.77	0.09	1.78	1.55	1.95
BMI	25.96	5.67	24.92	16.18	41.18

m_{stat}: mean, σ_{stat}: standard deviation, med_{stat}: median



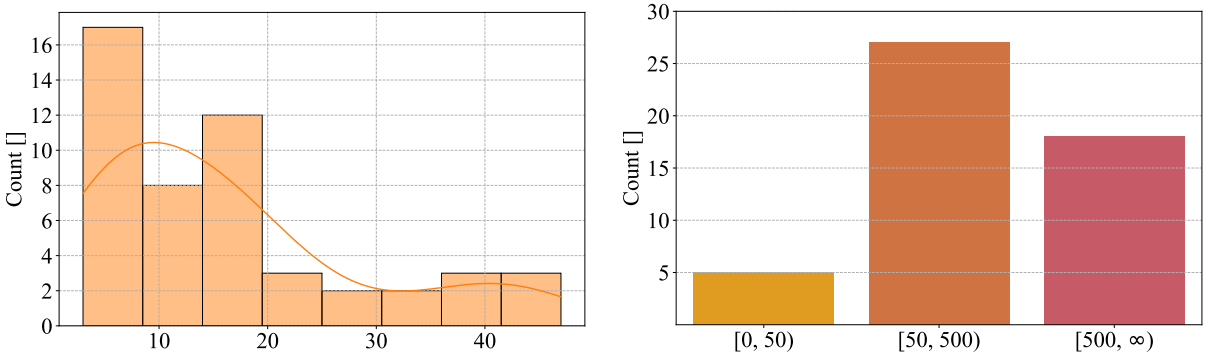
(a) Gender distribution

(b) Age distribution in years

Figure 6.1.: Demographic distribution of drivers included in the manD 1.0 dataset

The gender distribution shown in Figure 6.1a has an equal representation of genders, with 25 male and 25 female participants. The ages of the participants, as illustrated in Figure 6.1b, span a broad range, from 21 to 65 years, with a mean age of 34.4 years ($\sigma_{\text{stat}} = 12.12$). A noticeable tendency toward younger age groups is observed, as indicated by the

lower quartile age of 25 and a median of 31.5 years, suggesting that the majority of the participants belong to the younger segment of the population. This demographic diversity shows the inclusion of perspectives from both younger and older drivers, which is crucial for studying age-related variability in driving behaviors.
The driving experience of the participants, measured by two indicators, reflects a diverse range of experience levels (see Figure 6.2). The years since the issuance of their driver's license span from 3 to 47 years, with a mean of 16.14 years ($\sigma_{\text{stat}} = 12.33$), highlighting the inclusion of both novice and highly experienced drivers. The average monthly driving distance is categorized into three levels: less than 50 km per month, 50 km to 500 km per month, and more than 500 km per month. Most participants fall into the higher driving categories, indicating that the majority drive more than 50 km per month, and a considerable proportion exceed 500 km.



(a) Possession period of the driver's license in years (b) Average distance traveled per month [km]

Figure 6.2.: Distribution of the subjects' driving experience in the manD 1.0 dataset

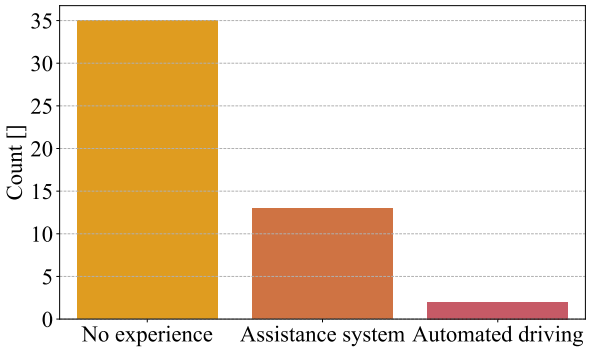
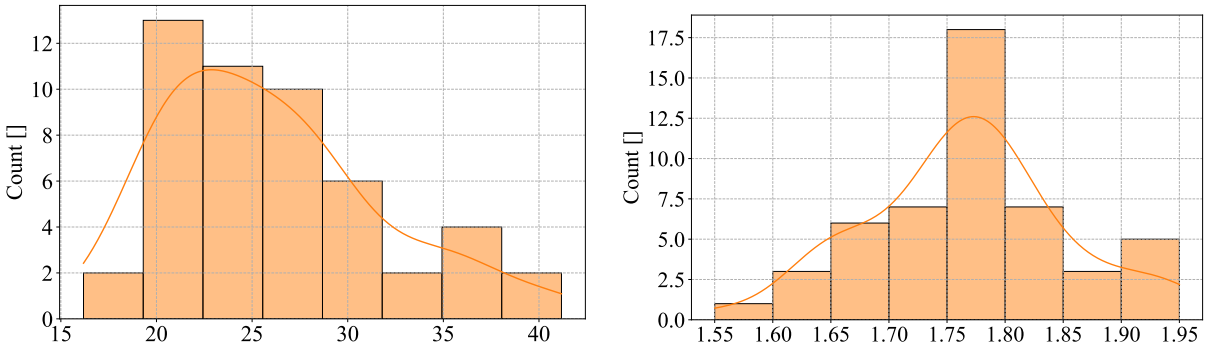


Figure 6.3.: Prior experience with driving assistance systems

The dataset classifies participants' experiences with the driving assistance systems into three levels, as depicted in Figure 6.3. A majority of participants (70 %) have no prior experience with assistance systems. About 27 % of participants have experience with driver assistance systems such as cruise control, indicating some familiarity with partial automation (SAE L2). Only 4 % of the participants report prior experience with fully automated driving, either in a simulator or a vehicle. This distribution highlights a predominantly low level of exposure to automated driving technologies within the participant

pool, providing an opportunity to assess their adaptability and responses to such systems during the study.
The height and BMI of the participants exhibit notable variability, factors that could influence the proprioception of the drivers. Heights range from 1.55 m to 1.95 m, with a mean of 1.77 m ($\sigma_{\text{stat}} = 0.09$), indicating a diverse physical stature among the drivers. Similarly, the BMI values span from 16.18 to 41.18, with a mean of 25.96 ($\sigma_{\text{stat}} = 5.67$). This distribution encompasses a broad spectrum of body types, from underweight to significantly overweight individuals, ensuring the inclusion of diverse physical characteristics among the proprioceptive signals.



(a) Body mass index distribution

(b) Height distribution

Figure 6.4.: Participants' physical characteristics

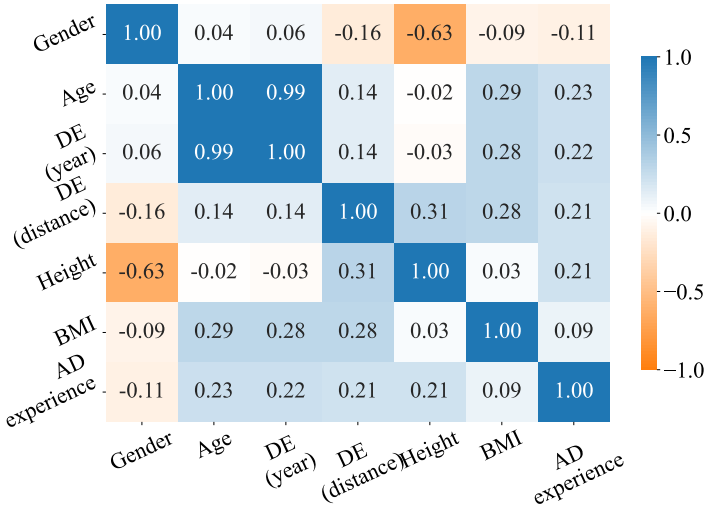


Figure 6.5.: Correlation matrix of participant characteristics; DE: driver experience

The Pearson correlation analysis (see Figure 6.5) reveals a strong positive relationship between age and the years of driving experience ($r = 0.99$, $P = 0.00$), where r denotes correlation coefficient and P gives the likelihood of obtaining the observed data under the null hypothesis. This result is intuitive, reflecting that older participants have typically held their driver's licenses for longer periods, accumulating more years of driving experience. A significant negative correlation is found between height and gender ($r = -0.63$, $P = 0.00$). Further significant but weak correlations found in the data are between BMI and age

($r = 0.29$, $P = 0.04$), BMI and driving experience in years ($r = 0.28$, $P = 0.04$), BMI and distance driven per month ($r = 0.28$, $P = 0.04$), and height and distance driven per month ($r = 0.31$, $P = 0.03$).
No other significant correlations ($P < 0.05$) are identified between the remaining characteristics. This suggests that variables like BMI, height, or automated driving experience are largely independent of each other and age, providing a diverse set of participant profiles without substantial confounding effects between these factors.

6.2. Driving context and exteroceptive perception

The context in which a human operates significantly impacts their emotions, decision-making processes, and overall performance. In automated driving, these contextual factors are crucial in shaping how drivers interact with their environment and vehicle. The manD 1.0 dataset provides a comprehensive collection of data encompassing both environment and vehicle states, offering a robust foundation for understanding these interactions. This study incorporates key environmental and vehicle factors as contextual identifiers into the driver model to improve its accuracy in predicting driver behavior, particularly in takeover driving scenarios.

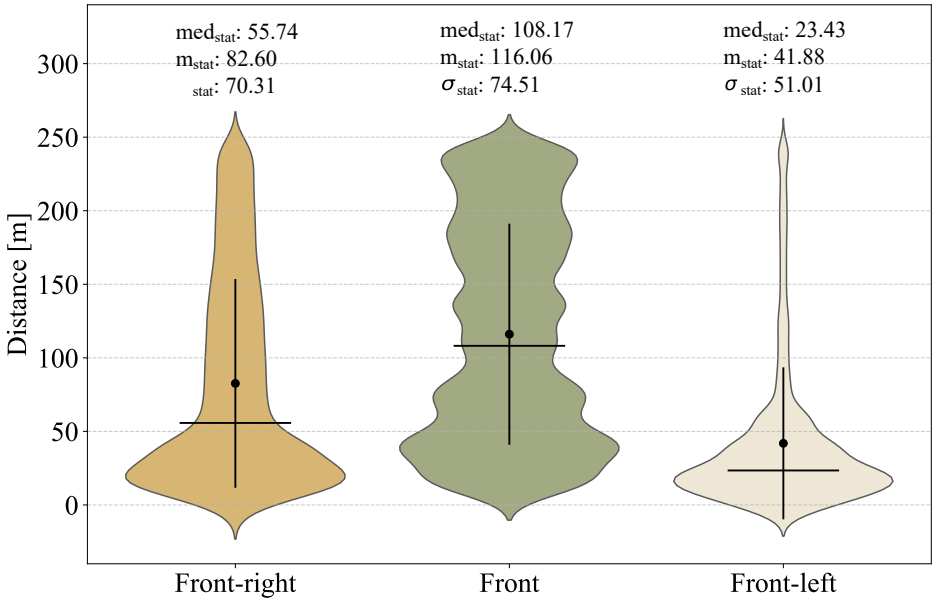
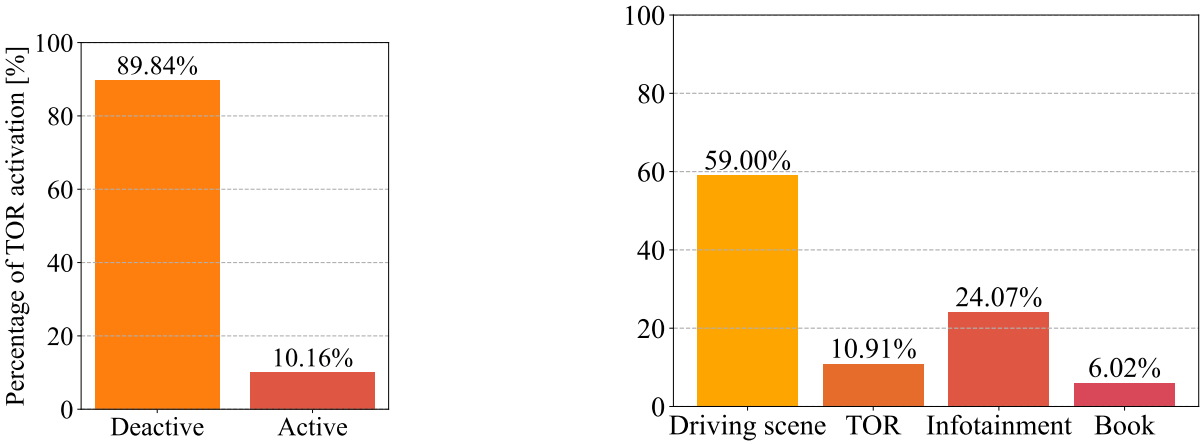


Figure 6.6.: Violin plots of distributions across the entire data section selected from manD 1.0 dataset

The environmental factors included in the data, such as the distances to other traffic users and static obstacles in the front half of the vehicle, are critical for influencing driving style, triggering takeover situations, or potentially causing accidents. Figure 6.6 illustrates the distribution of distances to front vehicles across the entire dataset, where

sensor range is set to 250 m . The sensation of the driver from the environment, including auditory and visual information, is also considered in the model. Auditory information encompasses the auditory modality. Figure 6.7 depicts statistics the auditory modelity during TOR scenarios, with Figure 6.7a showing the percentage of data with activated auditory warnings versus those without.



(a) TOR activation [%]

(b) Gaze direction class [%]

Figure 6.7.: Percentage of available data across participants in manD 1.0 dataset

Visual information is represented through various gaze direction classes, as shown in Figure 6.7b. These comprehensive data inputs enable a nuanced understanding and modeling of driver behavior in complex driving scenarios. However, the unbalanced distributions should be adjusted during modeling.

6.3. Unconscious interoception

Interoception, the sensing of internal bodily signals, plays a crucial role in maintaining homeostasis and influencing emotions and behavior. Various methods have been developed to measure interoception, which can be categorized into several broad approaches: self-report measures, behavioral tasks, and physiological measures. Each of these methods provides unique insights into the complex processes underlying interoceptive awareness and sensitivity. Self-report questionnaires are commonly used to assess interoceptive sensibility, which refers to the self-perceived tendency to focus on internal bodily sensations [Bud03]. Behavioral tasks are commonly used to measure interoceptive accuracy, such as counting one's own heartbeat without feeling the pulse [Bud03; HP12].
In this study, objective physiological measures are used as indicators of the interoception. These measures include various peripheral physiological measurement methods. Understanding these physiological processes is crucial for comprehending how the body maintains homeostasis and how these internal states influence behavior and emotional regulation.
BVP is a measure of the blood flow in peripheral vessels, typically recorded using PPG sensors. BVP is relevant to interoception as it provides data on cardiovascular dynamics, which are integral to the body's internal sensing mechanisms. Changes in BVP can reflect variations in emotional states and stress levels, contributing to the understanding of how

interoceptive signals influence psychological well-being [Bud03]. HR and HRV extracted from BVP provide insights into autonomic nervous system functions and cardiac interoceptive accuracy. The number of peaks in BVP indicates the frequency of physiological arousal events and can be used to measure the sensitivity to internal physiological changes and responses to stress.

EDA measures the skin's electrical conductance, which varies with sweat gland activity. EDA is relevant to interoception as it is a direct indicator of sympathetic nervous system activity, often associated with emotional and arousal states. Increased EDA is linked to heightened interoceptive awareness, particularly in response to stress or emotional stimuli [Bud03]. The skin conductance level (SCL) reflects baseline sympathetic nervous system activity and overall arousal state, which is important for understanding tonic arousal and baseline interoceptive awareness. Figure 6.8 depicts the spread of average SCL among the participants and driving scenarios in the data.

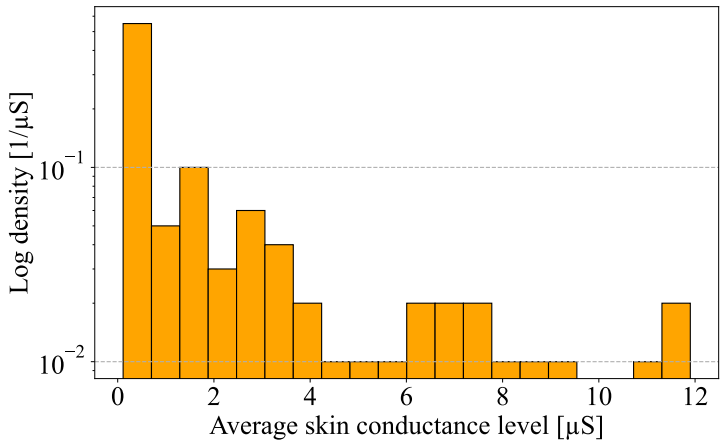


Figure 6.8.: Spread of average SCL among the participants and driving scenarios

SCR indicates transient changes due to specific stimuli, providing insights into momentary arousal and emotional responses. SCR is used to measure phasic interoceptive responses to internal and external stimuli. Figure 6.9 illustrates the distribution of the number of SCR peaks for each scenario.

Body temperature regulation is another critical aspect of interoception, involving complex feedback mechanisms that maintain homeostasis. The sensation of temperature is processed by thermoreceptors in the skin and internal tissues, providing the brain with information about the body's thermal state. Studies have shown that interoceptive awareness of body temperature is linked to emotional states and can influence behaviors aimed at thermoregulation, such as seeking warmth or cooling [Bud03]. Figure 6.10 depicts the mean and variance of skin temperature collected from participants during the driving scenarios, reflecting baseline thermal homeostasis and the stability of thermal regulation.

These features provide a comprehensive understanding of how individuals perceive and respond to their internal bodily states, contributing valuable insights into interoceptive processes and their implications for emotional and cognitive functions.

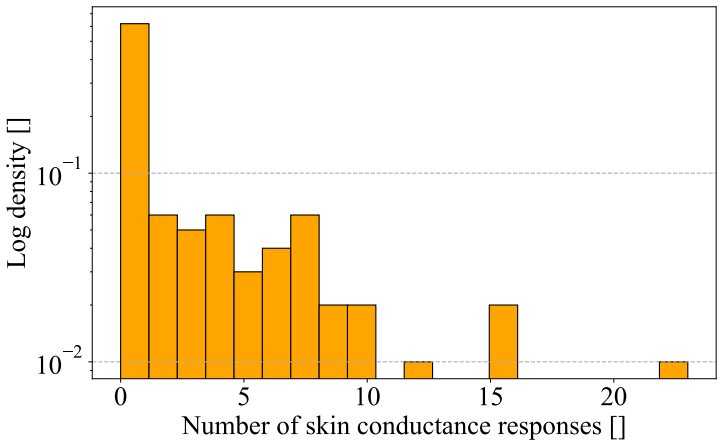


Figure 6.9.: Distribution of the number of SCR peaks

6.4. Proprioceptive responses

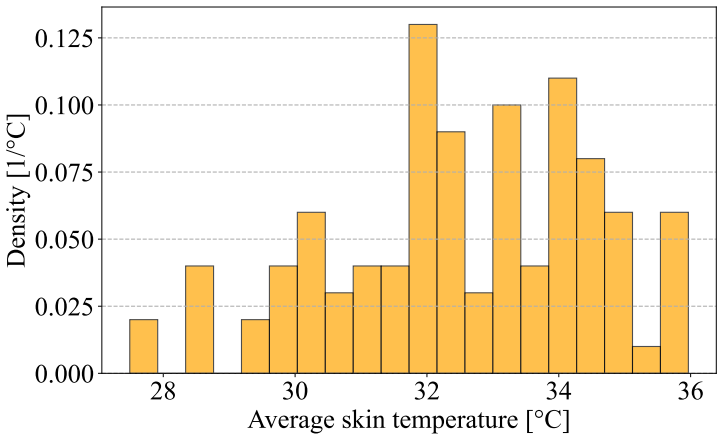


Figure 6.10.: Collected skin temperature from participants

Proprioception is the body's ability to sense its position, movement, and action of its limbs in space. The term is first introduced by Charles Sherrington, who described it as the perception of joint and body movement as well as the position of the body or body segments in space [She11; Ash+01]. This sense is crucial for motor control and coordination, and involves the integration of sensory signals from mechanoreceptors located in muscles, joints, and skin. Proprioception is not solely a physiological process but involves both physiological and psychological aspects. Mechanoreceptors such as muscle spindles, Golgi tendon organs, and joint receptors play a key role in this sensory system. Muscle spindles, in particular, detect changes in muscle length and the velocity of contraction, providing critical information for proprioceptive feedback. This sensory feedback system operates through both conscious and subconscious pathways, with proprioceptive information being processed at multiple levels within the central nervous system, including the spinal cord, brain stem, cerebellum, and cortex.
During this study, all participants remain seated in the driver's seat throughout the experiment. To measure proprioception, the pressure distribution on the driver's seat is

monitored using two seat-pressure-sensor mats placed on the seat and backrest. These pressure readings enable the detection of body posture and movement. The drivers' center of mass is also calculated from the pressure readings, which enables an initial assessment of body movement.
Furthermore, during automated driving, drivers can engage in NDRTs, which can involve the physical engagement of the driver's hands and feet. To capture this, the acceleration of the driver's hand is monitored using the accelerometer of the Empatica E4 wristband. The dataset includes various NDRTs, such as playing a game on the NDRT tablet, reading a book held by hand, and pressing a pedal with the right foot when spotting a deer on the road. Consequently, the data also contain columns for hand and foot activities, categorizing the specific tasks the driver is engaged in. Figure 6.11 illustrates the distribution of activity classes within the dataset.

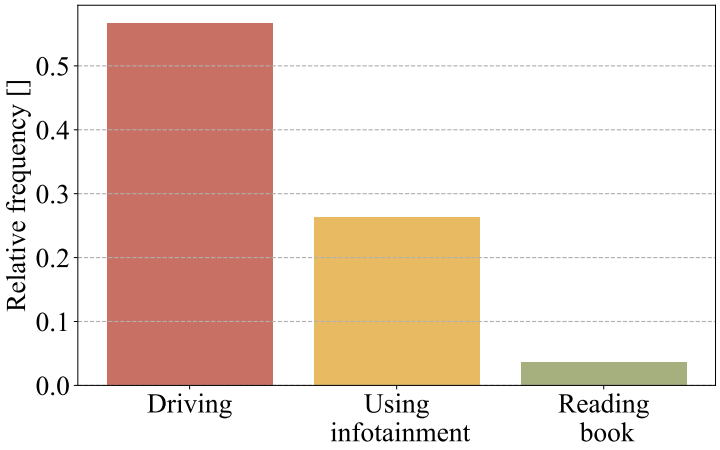


Figure 6.11.: Distribution of drivers' activity classes

As an additional piece of information, the data selected for this study include takeover situations, where the drivers' reaction times and types in response to TORs are of particular interest. Figure 6.12 shows the distribution of drivers' takeover reactions in the data, highlighting the variability in how drivers responded to TORs under various conditions. This aspect of the study is crucial for understanding the drivers' ability to transition from automated to manual driving and the factors influencing their reaction times and behaviors.
Moreover, since the movement of the right foot and hands directly affects the pedals and steering wheel during driving, engagement of the pedals and steering wheel is also recorded. Figure 6.13 presents the distribution of pedal and steering wheel usage among the drivers, offering a detailed view of how these controls are utilized across different driving scenarios directly after takeover while regaining the vehicle control.

6.5. Analysis of the subjective ratings

To introduce variations in driver states, various elements and events are integrated into the driving scenarios. A specific approach to invoke distinct emotional responses in participants involves presenting them with emotionally charged video clips, designed to elicit feelings

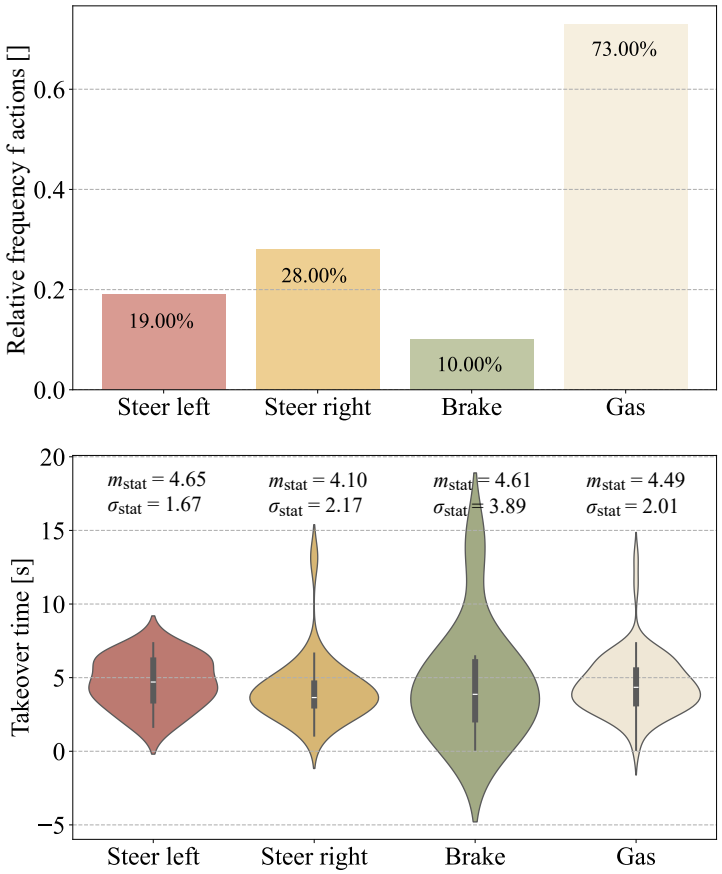


Figure 6.12.: Topp: relative frequency of four reaction types to TOR; down: distribution of takeover time for each reaction type

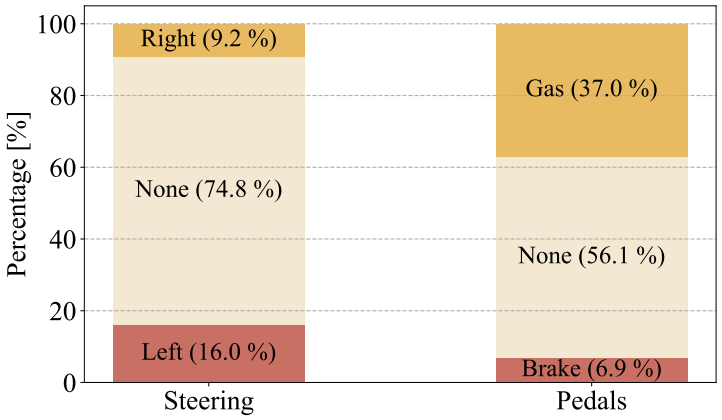


Figure 6.13.: Distribution of available data from pedal and steering wheel usage among the drivers [%]

of anger, sadness, and fear, prior to the start of each of the three scenarios. Participants are instructed to view these video clips, subsequently complete the DES, and then proceed with the driving task. Results of the collected DES are shown in Figure 6.14.
Analysis using a one-way analysis of variance (ANOVA) [Gir92] reveals statistically significant differences across the scenarios in terms of the average ratings for several emotions: enjoyment ($F(4, 145) = 11.008, P = 7.9 \times 10^{-8}$), surprise ($F(4, 145) = 4.88, P = 1.0 \times 10^{-3}$), sadness ($F(4, 145) = 13.86, P = 1.3 \times 10^{-9}$), anger ($F(4, 145) = 9.17, P = 1.2 \times 10^{-6}$), disgust ($F(4, 145) = 18.93, P = 1.5 \times 10^{-12}$), contempt ($F(4, 145) = 4.44, P = 2.1 \times 10^{-3}$), and fear ($F(4, 145) = 5.16, P = 6.4 \times 10^{-4}$). Here, F represents the ratio of between-group variance to within-group variance, and P indicates the likelihood of obtaining an F-ratio as extreme or more under the assumption of no significant group mean differences.

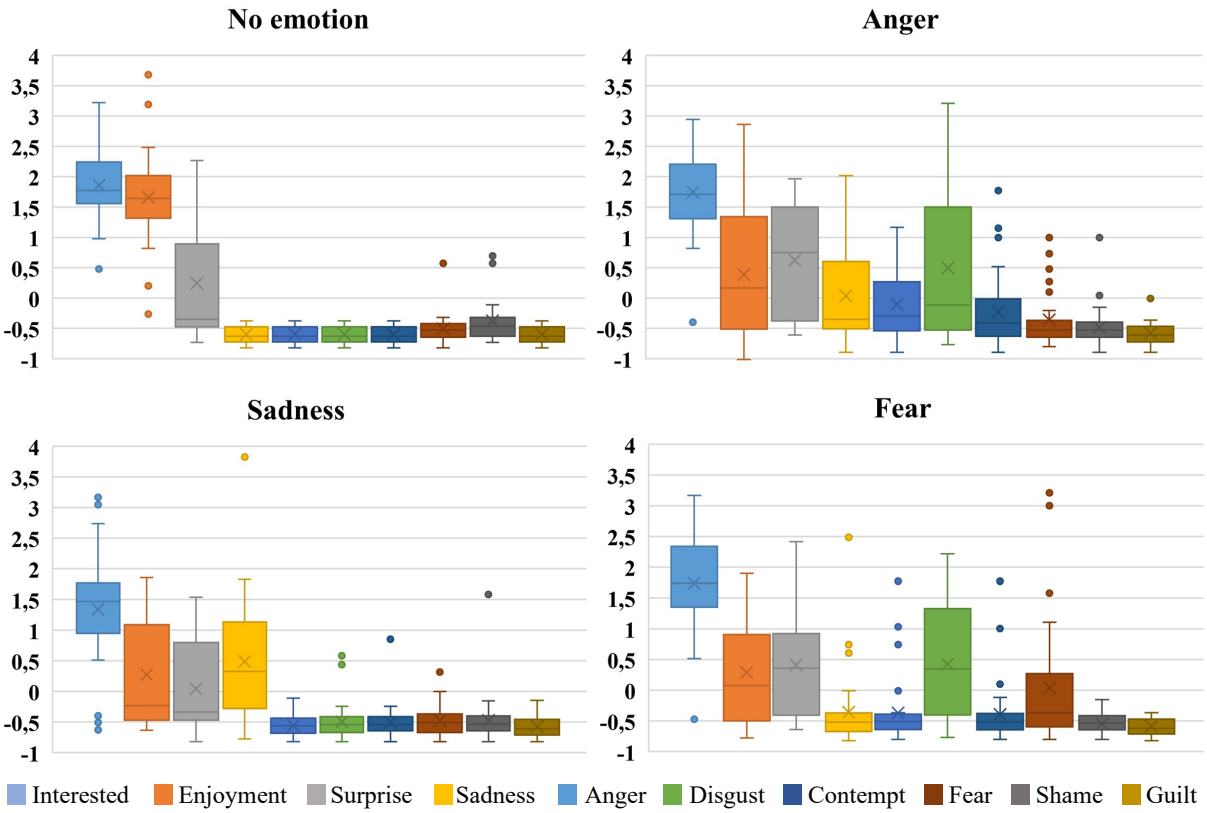


Figure 6.14.: Z-score [ ] of participants' subjectively rated emotions (DES) immediately before the start of the driving scenarios noemotion, anger, sadness, and fear [DB24a]; cross :mean, line: median

Post hoc t-tests [Stu08] confirm that each video clip effectively triggers the intended primary emotion along with additional emotional responses. Table 6.2 summarizes the significant emotions ($P < 0.05$) associated with each video clip.

6.6. Analysis of the selected data section for modeling

In this section, an overview of the selected dataset used for training the driver models is presented. The focus of this study is on the drivers' takeover reactions, making events

Table 6.2.: Post-hoc analysis of significant emotions evoked by video clips [DB24a]

Video	**Evoked emotions**											
	Enjoyment		**Sadness**		**Anger**		**Disgust**		**Contempt**		**Fear**	
(No emotion as reference)	m_{stat}	1.66	m_{stat}	-0.60	m_{stat}	-0.60	m_{stat}	-0.60	m_{stat}	-0.60	m_{stat}	-0.51
	σ_{stat}	0.96	σ_{stat}	0.21	σ_{stat}	0.35	σ_{stat}	0.93	σ_{stat}	0.20	σ_{stat}	0.29
Seven	m_{stat}	0.39	m_{stat}	0.04	m_{stat}	-0.10	m_{stat}	0.50	m_{stat}	-0.23		
	σ_{stat}	0.99	σ_{stat}	0.74	σ_{stat}	0.62	σ_{stat}	1.11	σ_{stat}	0.57		
	$P:$	7.1E-06	$P:$	1.8E-04	$P:$	2.5E-04	$P:$	3.8E-05	$P:$	5.8E-03		
City of Angels	m_{stat}	0.28	m_{stat}	0.49								
	σ_{stat}	0.84	σ_{stat}	0.96								
	$P:$	3.0E-07	$P:$	2.0E-06								
The Shining	m_{stat}	0.29					m_{stat}	0.42			m_{stat}	0.04
	σ_{stat}	0.88					σ_{stat}	0.89			σ_{stat}	0.98
	$P:$	4.7E-07					$P:$	1.3E-06			$P:$	1.4E-02

involving TOR the primary subject of analysis. From the manD 1.0 dataset, the three participants with the highest number of driven scenarios, namely P11, P23, and P39, are chosen for modeling. Each of these participants completed six driving scenarios, resulting in distinct datasets used for training and testing individualized driver models created for each participant. Table 6.3 provides the length of the events collected from each participant.

Table 6.3.: Length (number of steps) of the selected takeover events from the manD 1.0 dataset for each chosen participant

ID	**Length of the event**						**Sum**
	TOR 1	**TOR 2**	**TOR 3**	**TOR 4**	**TOR 5**	**TOR 6**	
P11	1180	1371	822	582	357	2104	6416
P23	1196	818	686	555	343	2211	5809
P39	1169	1321	816	707	579	353	4945

Figure 6.15 illustrates the distribution of reaction types among the three drivers. The analysis reveals that braking and steering left are two dominant reaction categories, indicating an imbalance in the dataset.

To gain an initial understanding of the data selected for driver modeling, a correlation analysis is conducted on the feature signals. This analysis provides insight into the relationships between various features, revealing potential linear dependencies relevant for driver modeling. The correlation matrix heatmaps for the three participants are presented in Appendix E, offering a visual representation of these relationships. The results highlight distinct patterns for each participant, reflecting individual differences in their responses and behavior during TOR events.

For P11, several significant correlations are identified. Among the exteroceptive signals a strong negative correlation (r = - 0.72) is observed between the activation of the auditory modality of the TOR and the TTC, suggesting that the driver reacts to the auditory modality of TOR and takes over the control of the vehicle in critical situations, so the TOR goes off. The activation of the visual modality of the TOR exhibited a positive correlation with the driver's activity class ($r = 0.80$) and hand acceleration in the y direction ($r =$

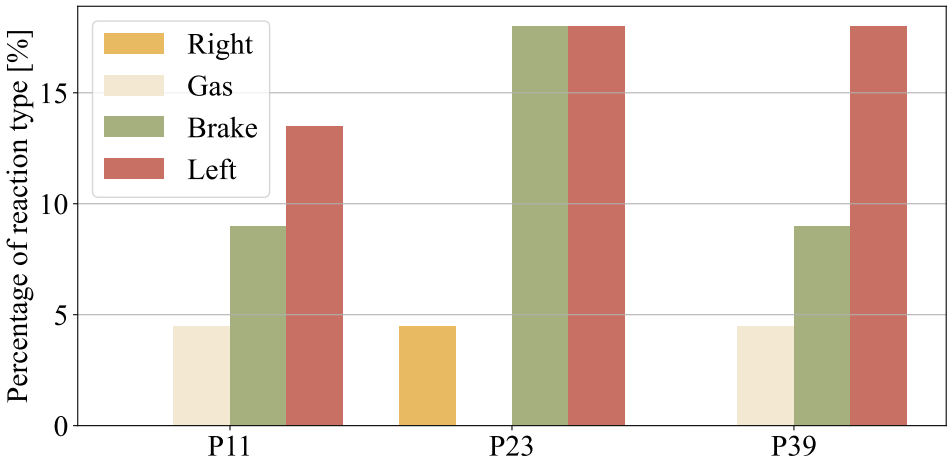


Figure 6.15.: Percentage of reaction types for the three selected participants from the manD 1.0 dataset across the available TOR events

0.65), indicating as well an association between visual alerts and the driver's immediate response behaviors. A strong negative correlation (r = -0.75) is found between the visual modality activation and the center of mass of the driver's body on the backrest in the x direction, implying a positional adjustment by the driver, such as pressing pedals, in response to visual cues. Similarly, a positive correlation ($r = 0.61$) is observed with the center of mass in the y direction, suggesting a nuanced movement pattern during visual engagement.

Hand acceleration metrics also revealed important relationships. A strong positive correlation ($r = 0.72$) exists between hand acceleration in the x and z directions, which is an expected result. The y direction hand acceleration is negatively correlated (r = -0.52) with the x direction center of mass but positively correlated ($r = 0.54$) with the y direction center of mass on the backrest, suggesting a link between hand dynamics and driver posture. Additionally, the driver's activity class correlated negatively (r = -0.52) with the x direction center of mass, while a moderate negative correlation (r = -0.56) is found between the x and y directions of the center of mass on the backrest.

In correlation analysis of the P23, again a moderate negative correlation ($r = -0.50$) between the activation of the auditory modality of the TOR and TTC is detectable. Similarly, a strong negative correlation ($r = -0.73$) between the distance to the vehicle in the front-right of the ego vehicle and the steering angle (steering to left) suggests that closer proximity to surrounding vehicles influences the driver's steering behavior or vice versa. This proximity is also moderately negatively correlated with EDA ($r = -0.56$), suggesting heightened physiological arousal when the ego-vehicle is near other vehicles.

The activation of the visual TOR shows strong positive correlations with key driver responses. A significant correlation with hand acceleration in the y direction ($r = 0.59$) and the driver's activity class ($r = 0.81$) highlights the role of visual alerts in eliciting physical and behavioral responses from the driver. Additionally, the positive correlation between the visual TOR and the center of mass of the driver's body on the backrest in the y direction ($r = 0.54$) indicates positional adjustments in response to visual stimuli.

EDA and steering angle exhibit a moderate positive correlation ($r = 0.56$), suggesting

that physiological arousal may be linked to steering adjustments during driving. EDA negatively correlates with the center of mass of the driver's body on the backrest in the y direction ($r = -0.56$), indicating that arousal and body posture could also be related. Notably, skin temperature shows a strong positive correlation ($r = 0.88$) with the center of mass in the x direction, pointing to potential thermoregulation effects on driver posture. Furthermore, hand accelerations in the x and y directions are strongly correlated ($r = 0.74$) as an expected result.

A moderate negative correlation ($r = -0.60$) between the activation of the auditory modality of the TOR and TTC is observed for P39, as well. The distance to the vehicle on the front-right side of the ego-vehicle shows significant correlations with other proximity-related features, including a moderate positive correlation with the distance directly ahead ($r = 0.52$) and a stronger correlation with the distance to the front-left vehicle ($r = 0.60$), indicating a relatively dense driving environment.

The distance to the front-right vehicle also exhibits a moderate negative correlation with skin temperature ($r = -0.57$), suggesting a physiological response linked to proximity, potentially reflecting stress or thermal regulation in reaction to closer vehicles. Additionally, the activation of the visual modality of the TOR correlates strongly with driver activity ($r = 0.76$), highlighting the direct influence of visual alerts on eliciting observable driver responses during critical events.

EDA positively correlates with the center of mass of the driver's body on the seat in the y direction ($r = 0.58$), suggesting that heightened arousal is associated with positional adjustments along this axis. Hand accelerations in the x and y directions display a strong negative correlation ($r = -0.73$), reflecting complementary movement patterns that may indicate precise manual control during steering or other tasks. The center of mass in the x direction on the seat correlates strongly ($r = 0.80$) with the center of mass in the y direction, revealing synchronized positional adjustments of the driver's posture.

These correlations emphasize the diverse nature of driver behavior, where environmental factors, physiological responses, and physical movements interact dynamically.

7

Computational modeling for driver state prediction: a data-driven approach

The driver model introduced here aims to predict the driver state using a data-driven approach, specifically trained on the manD 1.0 driver monitoring dataset. The structure of the model is inspired by the functional mechanisms of the human brain, as observed in various studies examining brain activation during different decision-making tasks and cognition phases. It is important to note that the goal of modeling is not to replicate the brain's activation patterns directly within artificial neural layers. Instead, the aim is to design a realistic network of layers that emulate the decision-relevant functions of the driver's brain. These modules are intended to represent functional roles rather than a strict sequential ordering. Additionally, it is crucial to acknowledge that the mechanisms underlying learning in artificial neural networks differ fundamentally from those in biological systems.

By emulating the brain's processing of exteroceptive, interoceptive, and proprioceptive data, the model predicts subsequent sensations of the brain based on the provided training data. Specifically, it takes a sequence of sensations as input and forecasts the next sequence. This predictive capability mirrors the brain's ability to anticipate and respond to sensory inputs.

In modeling, it is assumed that the drivers' sensations are seamless; thus, exteroceptive, interoceptive, and proprioceptive senses are represented using sensor data collected from the environment and the driver. For instance, exteroceptive data include signals describing the driving scene and in-vehicle situations, under the assumption that drivers accurately perceive all relevant visual and auditory cues. Similarly, proprioception data involve actual hand acceleration values, presuming that drivers sense these values correctly. The model processes sequential numerical features, necessitating components capable of handling temporal sequences. To achieve this, the model integrates BLSTM layers alongside a Hopfield layer within a VAE framework. The following sections provide detailed explanations of the model, its components, and the training process employed to optimize the predictive performance.

7.1. Input design and model architecture

This section provides a detailed overview of the input data used to represent the driver state during both automated and manual driving. Then, the structure of the model is introduced, which includes several parts that process these data and encode them into a lower-dimensional latent representation.

7.1.1. Input data representation

The model presented in this study utilizes a set of 20 diverse inputs that include interoceptive, exteroceptive, and proprioceptive signals to simulate a driver's reaction in an automated driving scenario. Figure 7.1 depicts an overview of the selected input features for the driver model. These inputs are chosen to capture the comprehensive dynamics of a driver's physiological, perceptual (from the environment), and behavioral state, which are essential for predicting reactions during automated driving. The data are sampled at a rate of 16 Hz. Furthermore, all input data are normalized to the scale of $[0, 1]$ to ensure consistency and improve the performance of the deep learning algorithms used in the model.

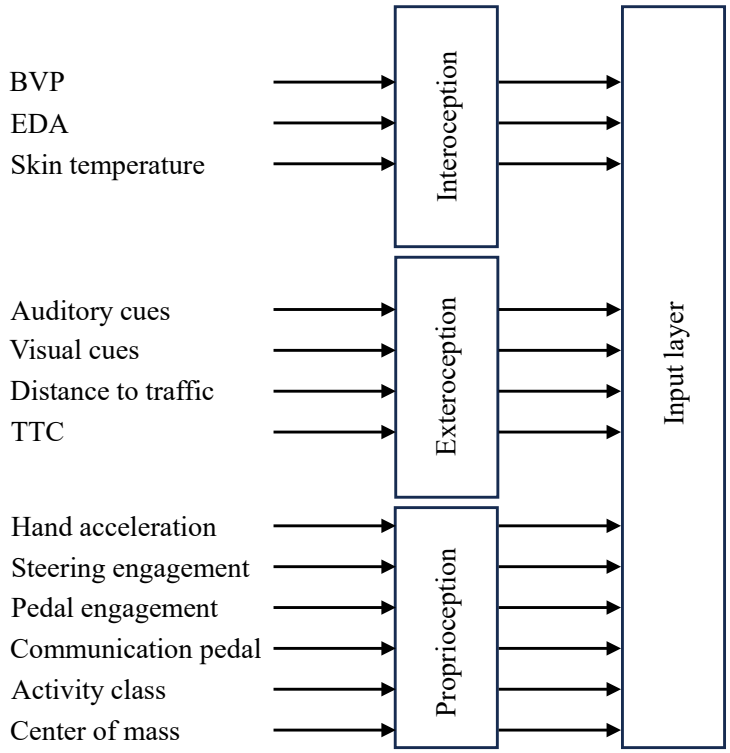


Figure 7.1.: Overview of the model inputs

The interoceptive inputs, which represent internal physiological states, are derived from the Empatica E4 wristband (Empatica Inc., Boston, Massachusetts, U.S.). These inputs include BVP and EDA. Both BVP and EDA signals are preprocessed using a moving window of one second to compute averaged values, obtaining smoother and more stable signals (Appendix D). In addition, the skin temperature is also collected from the wristband.

These interoceptive inputs help in understanding the driver's emotional and physiological state, which could influence reactions in different driving contexts.
Exteroceptive inputs pertain to the external environment that affects the driver. In this model, they include auditory warnings issued by the automated system as part of the TOR, the driver's visual attention class, including driving scene, tablet, book, or none; the distances to front vehicles in the same or adjacent lanes, and the TTC with vehicles in close proximity ahead of the ego-vehicle. The TTC considers the relative speed and distance between the ego-vehicle and the vehicle ahead and in this study is included as a measure of perceived situation criticality. These inputs can be helpful in modeling the situation awareness of the driver and how they might respond to various external stimuli during driving.
Proprioceptive inputs provide information about the driver's body movements and vehicle control actions. These include the acceleration of the driver's dominant hand, recorded by the Empatica E4 wristband. Furthermore, the driver's movement and body posture are tracked and defined by several inputs. One of these inputs is the engagement with the steering wheel, categorized into three classes: steering right, steering left, and driving straight. Another input captures the use of the gas or brake pedals, providing information about the foot's position. The engagement with the communication pedal, used to signal the automated system, is also included as a binary variable. Additionally, there is an activity class that indicates whether the driver's hands are steering, playing a game on a tablet, holding a book, or not engaged in any of these activities. Another critical proprioceptive input involves the center of mass calculated from the pressure sensor mats placed on the driver's seat and backrest. These mats contain 1024 pressure sensors each, and the center of mass is computed using the `center_of_mass` function from `SciPy.ndimage` [Vir+20], providing two 2D coordinates that represent the driver's center of mass on the driver seat and backrest. These proprioceptive inputs are key to understanding the driver's physical state and interaction with the vehicle controls.
To implement this model, several Python packages are employed. These include `os` for file handling, `NumPy` [Har+20] for numerical computations, `Pandas` [McK+10] for data manipulation, and `PyTorch` [Pas+19] for building and training deep learning models.

7.1.2. Driver model architecture

The model combines a VAE framework with a BLSTM network, enhanced by Hopfield networks and an attention mechanism as depicted in Figure 7.2. This hybrid architecture aims to capture both the temporal dependencies and the underlying latent representations of the input sequences, providing a robust approach to sequence prediction and reconstruction. The model comprises three primary components of

- encoder, which processes the input sequence to extract meaningful representations,
- latent space, which encodes the sequence into a latent vector using the VAE framework, and
- decoder, which reconstructs the output sequence from the latent vector, utilizing attention over the encoder outputs.

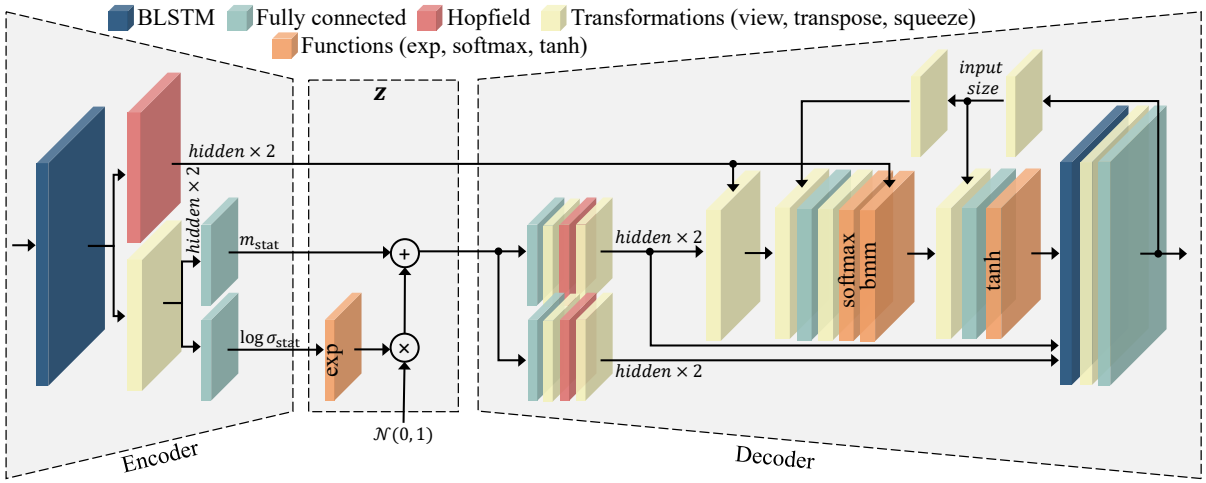


Figure 7.2.: Architecture of the driver model implemented in this study, simulating driver's brain; bmm: matrix multiplication

Encoder begins with a BLSTM layer. The BLSTM network, being a variant of the LSTM network, includes several gates that collectively emulate the concept of human brain memory. In the human brain, memory is formed and processed through stages of encoding, consolidation, and retrieval, during which certain pieces of information are selectively retained or discarded. Similarly, the gates in the BLSTM network allow the model to forget irrelevant parts of the input data during the encoding and consolidation phases. When the encoded information is later retrieved, the memory can be manipulated, akin to how the human brain adjusts memories based on new experiences or insights. This gate-based mechanism enables the network to maintain a dynamic memory state that evolves as new data are processed.

Furthermore, the BLSTM network's design allows it to modify its internal state based on new inputs, paralleling the way the human brain continuously updates its memory pool as new information is acquired. When the brain receives new stimuli, it does not just store them passively; rather, it integrates this new information with existing memories, potentially altering the current memory state. Similarly, in a BLSTM network, each new input can affect the state of the LSTM cell, influencing the model's future predictions and decision-making processes. This dynamic state adaptation is crucial for applications where the sequence of events and their interdependencies must be learned, such as in modeling driver reactions to TOR.

Another notable feature of the LSTM, and by extension, the BLSTM, is its predictive capability, which mirrors the human brain's capacity to anticipate future events. The human brain is constantly engaged in predicting the next sensations based on current and past inputs, allowing it to prepare appropriate responses. In a similar manner, the LSTM network, with its recurrent connections and memory cell states, makes predictions about the next data points in a sequence. This autoregressive nature is particularly valuable in time-series data modeling, where understanding the sequence and timing of events is key to making accurate forecasts.

The BLSTM processes the input sequence $\mathbf{X} = (\mathbf{x_1}, \mathbf{x_2}, ..., \mathbf{x_T})$, where $\mathbf{x_t} \in \mathbb{R}^{d_{\text{in}}}$ represents the input vector at time step t and T is the sequence length.

In the BLSTM, the forward LSTM computes hidden states $\mathbf{h}_t$ from $t = 1, ..., T$. The backward LSTM computes hidden states $\mathbf{h}'_t$ from $t = T, ..., 1$. The outputs are concatenated at each time step to form the encoder outputs $\mathbf{h}_t^{\text{BLSTM}} = [\mathbf{h}_t \,||\, \mathbf{h}'_t] \in \mathbb{R}^{2d_{\text{hid}}}$, where d_{hid} is the hidden size of the LSTM.
To capture associative memory and enhance pattern retrieval capabilities, the encoder outputs are passed through a Hopfield network, implemented using the `Hopfield` module from the `hflayers` package [Ram+20]. The Hopfield layer operates on the sequence of hidden states $\mathbf{h}^{\text{BLSTM}}$ to produce updated representations

$$\mathbf{h}^{\text{Hop}} = \text{Hopfield}(\mathbf{h}^{\text{BLSTM}}, \mathbf{H}), \tag{7.1.1}$$

where $\mathbf{H}$ is the set of all hidden states, and the Hopfield network updates each $\mathbf{h}^{\text{BLSTM}}$ based on its similarity to other states in $\mathbf{H}$.

Latent space obtains the encoder's final hidden state $\mathbf{h}_{\text{final}}$ by concatenating the last hidden states from both directions $\mathbf{h}_{\text{final}} = [\mathbf{h}_T \,||\, \mathbf{h}_1] \in \mathbb{R}^{2d_{\text{hid}}}$. Two linear layers map $\mathbf{h}_{\text{final}}$ to the mean $\boldsymbol{\mu}$ and log variance $\log\boldsymbol{\sigma}^2$ of the latent space as

$$\boldsymbol{\mu} = \mathbf{W}_\mu \mathbf{h}_{\text{final}} + \mathbf{b}_\mu, \tag{7.1.2}$$

$$\log \boldsymbol{\sigma}^2 = \mathbf{W}_\sigma \mathbf{h}_{\text{final}} + \mathbf{b}_\sigma, \tag{7.1.3}$$

where $\mathbf{W}_\mu, \mathbf{W}_\sigma \in \mathbb{R}^{d_{\text{lat}} \times 2d_{\text{hid}}}$ and $\mathbf{b}_\mu, \mathbf{b}_\sigma \in \mathbb{R}^{d_{\text{lat}}}$. To enable backpropagation through the stochastic sampling process, the reparameterization method [Kin13] is used. A latent vector $\mathbf{z}$ is sampled from the latent space

$$\mathbf{z} = \boldsymbol{\mu} + \boldsymbol{\sigma} \odot \boldsymbol{\epsilon}, \tag{7.1.4}$$

where $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and $\odot$ denotes element-wise multiplication.

Decoder in the proposed model architecture is designed to generate output sequences by reconstructing or predicting future data points based on encoded representations. The decoder in the proposed model comprises three primary components: a Hopfield layer (analogous to the hippocampus), a decoder BLSTM (analogous to VLPFC), and an attention mechanism (analogous to the amygdala). Each component plays a critical role in transforming encoded information into a meaningful sequence, mirroring the human cognitive process of recalling and reconstructing memories. However, it is important to clarify that the model does not claim to replicate the temporal order of neural firing. Instead, each module performs a function that loosely corresponds to the role of the associated brain region. For instance, while the amygdala is often associated with salience detection and the hippocampus with memory encoding and retrieval, these processes in the brain can occur in parallel or involve rapid bidirectional interactions, rather than following a strictly linear sequence. Similarly, in sequence-to-sequence models, pattern completion modules like Hopfield networks are typically placed in the encoder to capture essential patterns before passing them on, while the attention mechanism (salience detection) resides in the decoder, reflecting a functional rather than anatomical correspondence to the brain.

The decoder performs a linear transformation on the latent vector $\mathbf{z}$ to initialize hidden and cell states for the decoder LSTM by

$$\mathbf{h}_{\text{init}} = \mathbf{W}_h \mathbf{z} + \mathbf{b}_h, \tag{7.1.5}$$

$$\mathbf{c}_{\text{init}} = \mathbf{W}_c \mathbf{z} + \mathbf{b}_c, \tag{7.1.6}$$

where $\mathbf{W}_h, \mathbf{W}_c \in \mathbb{R}^{(n_{\text{layers}} \times 2d_{\text{hid}}) \times d_{\text{lat}}}$ and $\mathbf{b}_h, \mathbf{b}_c \in \mathbb{R}^{n_{\text{layers}} \times 2d_{\text{hid}}}$. The initial hidden and cell states are further processed by a Hopfield layer to enhance memory retrieval as

$$\mathbf{h}'_{\text{init}} = \text{Hopfield}(\mathbf{h}_{\text{init}}, \mathbf{H}_{\text{init}}), \tag{7.1.7}$$

$$\mathbf{c}'_{\text{init}} = \text{Hopfield}(\mathbf{c}_{\text{init}}, \mathbf{C}_{\text{init}}), \tag{7.1.8}$$

where $\mathbf{H}_{\text{init}}$ and $\mathbf{C}_{\text{init}}$ are the sets of initial hidden and cell states. At each decoding time step t, the model employs an attention mechanism to focus on the most relevant parts of the input sequence by computing attention over the encoder outputs. Specifically, the Bahdanau (additive) attention mechanism [Bah+15] is utilized. Unlike the dot-product approach used in multiplicative attention, this method applies a feed-forward network to compute attention scores. While the original Bahdanau attention mechanism employs a small feed-forward network with a $\tanh(\cdot)$ activation, the model here uses a linear layer directly. The core idea remains the same: learn attention weights in an additive manner rather than relying on direct vector dot products. These scores are normalized using a softmax function to produce attention weights. The mechanism involves concatenating the last decoder hidden state $\mathbf{h}_{t-1}$, the encoder state, and the decoder input (denoted as $\mathbf{H}$), which is then fed into a learnable linear layer to compute attention scores

$$\mathbf{e}_t = \text{softmax}(\mathbf{W}_{\text{attn}}[\mathbf{H}_{\text{attn}}; \mathbf{x}_{t-1}]), \tag{7.1.9}$$

where $\mathbf{x}_{t-1}$ is the decoder input at time $t-1$, $\mathbf{H}_{\text{attn}}$ represents the concatenation of the input elements, and $\mathbf{W}_{\text{attn}}$ is a learnable weight matrix.
Although the Bahdanau mechanism effectively captures salience in sequence-to-sequence tasks, it simplifies the complexity of real-world neural processes. The amygdala's functioning is more dynamic and context-dependent, involving complex interactions with neurotransmitter systems and other brain regions, which may not be fully captured by the relatively simplified additive attention model.

The choice of the attention mechanism also meets the memory activation functions utilized in the ACT-R cognitive architectures. In ACT-R, the activation of a memory m has a term called base-level $b_{m,t}$. This term is determined by Equation 3.1.2. The decay rate d_{trace_i}, represents the rate at which memory traces fade over time, accounting for the spacing effect. The equation for d_{trace_i} (Equation 3.1.5) in ACT-R is formulated as an exponential decay term that diminishes with time.
The softmax function in the attention mechanism serves a similar purpose. The softmax function, defined as

$$\phi(\mathbf{h}_t) = \frac{\exp(\mathbf{h}_t)}{\sum_{j=1}^{n} \exp(\mathbf{h}_{t_j})}, \tag{7.1.10}$$

normalizes the outputs into a probability distribution. In the context of memory activation, the softmax function can be seen as a resembles to the d_{trace_i} decay rate in the ACT-R model. It emphasizes certain outputs while diminishing others, effectively mimicking the process by which the brain prioritizes certain memories based on their relevance or recentness.
The context vector $\mathbf{c}_t$ is computed as a weighted sum of the encoder outputs (shown as bmm in Figure 7.2)

$$\mathbf{c}_t = \sum_{i=1}^{T} e_{t,i}\mathbf{h}_i. \tag{7.1.11}$$

The decoder LSTM generates the output sequence by processing the combined input

$$\mathbf{x}'_t = \tanh(\mathbf{W}_{\text{combine}}[\mathbf{x}_{t-1}; \mathbf{c}_t]), \tag{7.1.12}$$
$$\mathbf{h}_t, \mathbf{c}_t = \text{LSTM}(\mathbf{x}'_t, (\mathbf{h}_{t-1}, \mathbf{c}_{t-1})), \tag{7.1.13}$$

where $\mathbf{W}_{\text{combine}}$ is a linear layer that combines the context vector and the decoder input. Finally, the decoder's output at each time step is generated using a linear layer

$$\mathbf{y}_t = \mathbf{W}_{\text{out}}\mathbf{h}_t + \mathbf{b}_{\text{out}}, \tag{7.1.14}$$

where $\mathbf{W}_{\text{out}} \in \mathbb{R}^{d_{\text{out}} \times 2d_{\text{hid}}}$ and $\mathbf{b}_{\text{out}} \in \mathbb{R}^{d_{\text{out}}}$.

7.2. Model training and hyperparameter optimization

To set the remaining hyperparameters of the model to near-optimal values, a hyperparameter optimization loop is developed. This loop involves iteratively training the model while varying the hyperparameters. The number of iterations is limited, and the model with the best evaluation metrics is selected as the final driver model. Since each participant has an individualized model, hyperparameter optimization is conducted separately for each participant's data. As a result, the selected hyperparameters for each driver may differ. The following subsections provide a more detailed discussion of the training and optimization processes.

7.2.1. Training procedure

In the training process of the proposed model for predicting driver behavior in automated driving scenarios, several Python [VD09] packages and methods are utilized to handle the data and optimize the model's learning process.
The data used in this study consist of multiple sessions of automated driving, each session containing a unique takeover situation where the driver needs to resume control of the vehicle. These sessions are rich in temporal dynamics and variability, making them suitable for training deep learning models that aim to capture complex driver behavior. In the training process, a leave-one-group-out cross-validation (LOGO-CV) method is applied. Specifically, in each loop of this method, one session is designated as the test data, while

the remaining sessions are used for training the model. This approach allows the model to generalize well by being tested on unseen data, ensuring robust performance across different driving scenarios.
Given the temporal nature of the sequences, data preparation is a crucial step in this process. The original continuous data from each session are divided into several overlapping sequences. These overlapping sequences help the model learn transitions and dependencies across time steps, which is essential for understanding events leading to the takeover situation. Once these sequences are generated, they are further divided into batches for efficient training. Shuffling is deliberately avoided here within the sequences. This is because the data has a temporal aspect, and shuffling would disrupt the continuity and temporal dependencies that the model needs to learn.
The focus of this study is on the takeover performance of drivers, which is a critical aspect of automated driving safety. To emphasize the importance of accurate prediction of the driver's reaction during these takeovers, a weighted mean squared error (WMSE) loss function is defined for training. This loss function assigns higher weights to the driver's reactions. Thus, the model becomes more sensitive to driving performance, especially during crucial takeover scenarios.
For the optimization of the model's parameters during training, the Adam optimizer [KB15] is employed as it converges faster and more reliably than stochastic gradient descent.

7.2.2. Hyperparameter tuning

Hyperparameter optimization helps in fine-tuning the model to achieve the best performance for predicting driver behavior during driving scenarios and takeover situations. The optimization process is performed using the `optuna` package [Aki+19], a tool for automated hyperparameter tuning. Optuna leverages Bayesian optimization, which is particularly effective for optimizing complex models with many hyperparameters.
The model takes an input sequence of 10 s, which corresponds to a series of time steps representing varying physiological and environmental data points. Based on this input, the model predicts the next driver behavior and system states. The length of the prediction horizon is determined by a hyperparameter, which is selected during the optimization process. This variability in the selection of the prediction horizon for each individual partially compensates and resembles the rate of forgetting (α) in the ACT-R architecture, which is defined individually for each person. Moreover, the flexibility in choosing the prediction horizon allows the model to adapt to different temporal resolutions and prediction requirements.
The hyperparameters optimized in this study are listed below.

- **Prediction horizon:** The number of time steps the model should predict into the future. It is selected from a categorical range of varying lengths from 5 s to 10 s.
- **Size of latent space:** This hyperparameter determines the dimensionality of the latent space in the BLSTM network. It is chosen among 32 and 64, and it directly influences the model's capacity to learn complex patterns. A larger hidden size can potentially capture more intricate features but may also increase computational complexity and risk of overfitting.

- **Size of hidden state** This hyperparameter determines the number of units in the hidden layers of the encoder BLSTM, which influences the model's capacity to learn and represent complex patterns within the data. Size of the hidden state is selected from a set of values $\{32, 64, 128\}$ to allow the optimization process to explore a range of model complexities. Smaller values lead to a model with lower computational demands and faster training, though potentially at the cost of capturing fine-grained features. Conversely, larger values increase the model's representational power, potentially enhancing performance on more complex data, albeit with a higher computational load and risk of overfitting.

- **Batch size:** The number of samples processed before the model's internal parameters are updated. It is selected from 32 to 128. Smaller batch sizes can make training noisier but can lead to faster convergence and improved generalization.

- **Learning rate:** This hyperparameter controls the step size during optimization. It is sampled from a log-uniform distribution within the range 10^{-5} to 10^{-2}. A smaller learning rate allows the model to converge to a more precise minimum but requires more training iterations.

Table 7.1 provides an overview of the defined hyperparameter ranges used for optimization. For hyperparameter optimization, the Bayesian optimization algorithm [Pel+99] is utilized. Bayesian optimization is a probabilistic model-based approach that builds a surrogate model to approximate the objective function and uses this model to select the most promising hyperparameters to evaluate next. The key idea is to balance exploration (trying out hyperparameters that may not perform well but provide more information about the objective function) and exploitation (choosing hyperparameters that are likely to perform well based on the current surrogate model).

Table 7.1.: Overview of the defined hyperparameter ranges

Hyperparameter	**Range**
Prediction horizon	(categorical) {80, 96, 112, 128, 146, 160}
Size of latent space	(categorical) {32, 64}
Size of hidden state	(categorical) {32, 64, 128}
Batch size	(categorical) {32, 64, 128}
Learning rate	(log-uniform) $U(10^{-5}, 10^{-2})$

Bayesian optimization involves the following steps.

1. **Initialization:** First hyperparameter sets are selected randomly and evaluated to initialize the surrogate model. The objective function's value (e.g., validation loss) is computed for these initial sets.

2. **Surrogate model construction:** A probabilistic model, such as a Gaussian process [See04], is built to approximate the objective function. This model estimates both the mean and uncertainty (variance) of the objective function for any given set of hyperparameters.

3. **Acquisition function optimization:** An acquisition function is used to determine the next set of hyperparameters to evaluate. The acquisition function balances the exploration-exploitation trade-off by favoring hyperparameters that have a high probability of improving the current best result.

4. **Evaluation of objective function:** The objective function is evaluated for the new set of hyperparameters chosen by the acquisition function, and the surrogate model is updated based on the new data.

In this study, the number of trials for Bayesian optimization is set to ten, considering the computational resources available. During training and hyperparameter optimization, various evaluation metrics are computed to assess the model's performance, including WMSE, mean absolute error (MAE), and RMSE. These metrics provide an understanding of the model's accuracy, variance explained, and prediction error, respectively.
Algorithm 7.2.1 illustrates the pseudocode of the Bayesian optimization procedure for hyperparameter tuning, outlining the steps from initialization to the final selection of optimized hyperparameters.

Algorithm 7.2.1.: Pseudocode of Bayesian optimization procedure for hyperparameter tuning

Require: Data: Dataset, L: Number of trials, $\mathbf{w}$: WMSE weights, $Theta$: Hyperparameter distributions

study $\leftarrow$ create_study($direction =' \min'$) ▷ Initialize study
for $l = 1$ to L **do** ▷ Iteration on trial
  $\theta(l) \leftarrow \arg\max_{\theta \in \Theta} \xi(\theta \,|\, TBE(\mathcal{D}_{l-1}))$ ▷ Sample hyperparameters θ
  for $k = 1$ to $|\text{Data}|$ **do** ▷ Iteration on fold
    Split data: $\text{Train}_k \leftarrow \cup_{j \neq k} \text{Data}_j$, $\text{Test}_k \leftarrow \text{Data}_k$
    M $\leftarrow$ M(θ) ▷ Initialize model with sampled hyperparameters
    for $r = 1$ to R **do** ▷ Iteration on epochs
      Train $M(\theta)$ on Train_k using WMSE:

$$WMSE = \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{d} w_j (y_{ij} - \hat{y}_{ij})^2$$

    $\text{Test\_Loss}_{k,l} \leftarrow$ Evaluate $M(\theta)$ on Test_k
    Append $\text{Test\_Loss}_{k,l}$ to All_Fold_Losses
  Compute average test loss:

$$\text{AvgTestLoss}_l \leftarrow \frac{1}{|\text{Data}|} \sum_{k=1}^{|\text{Data}|} \text{Test\_Loss}_{k,l}$$

  Return AvgTestLoss_l ▷ Objective function value for trial l
return $\theta^* = \arg\min_\theta \text{AvgTestLoss}_l$ ▷ Find optimal hyperparameters

The trained driver models are individualized, which means the models are trained for each participant separately using the data gathered from that participant. In total, three

participants (P11, P23, P39) are selected from the manD 1.0 dataset for this purpose, since the number of takeover sessions available in the dataset from other drivers is not sufficient for the training process. The next chapter reports the performance of the trained models for these participants.

8

Evaluation of the driver model in light of the benchmark dataset

In the present dissertation, a variational autoencoder with BLSTM and Hopfield layers in encoder and decoder structure and attention mechanism (VAE-BLSTM-HA) model is proposed to serve as a driver model. The purpose of this model is to predict the next driver state based on the input of the sequence of current and previous driver states. The VAE-BLSTM-HA model architecture is designed specifically for handling time series data, as it captures temporal dependencies and dynamic patterns in driver behavior. The model processes a sequence of driver states spanning the last 10 s (160 steps) and predicts the subsequent driver state sequence, emphasizing modeling of the evolving nature of driving behavior.

The model is trained individually for each driver, utilizing personalized data gathered from the respective driver. For this study, data from three drivers are selected from the manD 1.0 dataset. The LOGO-CV method is employed for training, where in each iteration, one driving session is held out for testing while the remaining sessions are used for training. This approach evaluates the model's generalization capabilities across different driving sessions. The performance evaluation of the models trained for the three drivers is presented in the following sections of this chapter, offering insight into the effectiveness and accuracy of the proposed VAE-BLSTM-HA model in predicting driver state.

The model is trained using a loss function that combines the reconstruction loss and the Kullback-Leibler (KL) divergence [KL51], following the VAE framework. The reconstruction loss measures the discrepancy between the predicted output $\hat{\mathbf{y}}_i$ and the true output $\mathbf{y}_i$ using a WMSE defined as

$$\text{WMSE} = \frac{1}{k} \sum_{i=1}^{k} w_i \left(\mathbf{y}_i - \hat{\mathbf{y}}_i\right)^2, \tag{8.0.1}$$

where k gives the number of samples and w_i represents the weight for each instance. Since this study emphasizes the takeover situations and the drivers' reactions during these events, the engagement of the steering and pedals is given higher weight in the loss function. This adjustment ensures greater accuracy in predicting these critical aspects of driver behavior. The KL divergence encourages the latent space to follow a standard normal distribution by

$$\mathrm{KL} = -\frac{1}{2}\sum_{i=1}^{d_{\mathrm{lat}}}(1 + \log\sigma_i^2 - \mu_i^2 - \sigma_i^2). \tag{8.0.2}$$

The total loss then combines both terms

$$\mathcal{L}_{\mathrm{total}} = \mathrm{WMSE} + \mathrm{KL}. \tag{8.0.3}$$

8.1. Evaluation metrics

For the evaluation of the trained models, several metrics are incorporated to ensure a comprehensive assessment of their performance. The RMSE is calculated through

$$\mathrm{RMSE} = \sqrt{\frac{1}{k}\sum_{i=1}^{k}(\mathbf{y}_i - \hat{\mathbf{y}}_i)^2}, \tag{8.1.1}$$

to provide a clearer sense of prediction error in the same units as the original data. RMSE is particularly advantageous in this modeling case as it penalizes larger errors more significantly, making it useful for detecting substantial deviations in predicted driver states. This is crucial when modeling sensitive behaviors such as steering and pedal engagement, where even small prediction errors can lead to substantial misinterpretations of driver reactions.
The MAE as an evaluation metric is calculated as

$$\mathrm{MAE} = \frac{1}{k}\sum_{i=1}^{k}\left|\mathbf{y}_i - \hat{\mathbf{y}}_i\right|. \tag{8.1.2}$$

The MAE measures the average magnitude of the errors in the model's predictions. Its strength lies in its simplicity and interpretability, as it directly reflects the average absolute difference between the predicted and actual reaction times.
In this study, reaction time is defined as the interval between the initiation of the TOR and the deactivation of the automated mode. Deactivation can be achieved through steering, pressing the pedals, or using the deactivation buttons on the steering wheel.
The MAE is also separately employed to evaluate the takeover reaction time of drivers. Its simple interpretability is advantageous for assessing takeover reaction time because it treats all errors equally, offering a clear indication of the model's prediction accuracy in capturing the precise timing of driver responses. Unlike metrics that square the errors, MAE ensures that both small and large deviations from the actual reaction times are treated proportionally, making it an effective metric for evaluating driver reaction in the manD 1.0 dataset, where large outliers in behavior are less likely.
Furthermore, match scores are calculated by comparing drivers' predicted and actual reaction type (steering or pressing pedals), evaluating the model's ability to predict specific behavioral responses. The match score is a binary measure indicating whether there is alignment between the actual and predicted steering and pedal actions. Let s_{true}, s_{pred}, p_{true}, $p_{\mathrm{pred}} \in \{0, 1\}$ be the binary values representing the true and predicted steering and pedal actions, respectively. The match score can then be defined as

$$\text{Match score} = \begin{cases} 1 & \text{if } (s_{\text{true}} = s_{\text{pred}}) \text{ or } (p_{\text{true}} = p_{\text{pred}}) \\ 0 & \text{otherwise} \end{cases} \tag{8.1.3}$$

The next section provides an overview of the hyperparameters selected in the model optimization process, detailing the specific configurations that contributed to the performance of the models. Following this, a baseline model is introduced, trained using the same dataset as the VAE-BLSTM-HA model, serving as a reference point for performance comparison. The evaluation of the VAE-BLSTM-HA model's ability to predict driver state is then presented and compared to the performance of the baseline model, highlighting any improvements or differences. Then, the model's performance in predicting driver reactions during takeover situations is specifically reported, with a focus on both reaction time and reaction type. This detailed analysis offers insight into the model's effectiveness in critical driving scenarios and is discussed in relation to the reference model. Finally, an ablation study is conducted on the VAE-BLSTM-HA model for a randomly selected participant to assess the contribution of each component to the model's overall performance.

8.2. Hyperparameter optimization results

The model is trained on three participants with distinct characteristics. The model optimization through Bayesian hyperparameter optimization involves systematically exploring the parameter space to find the optimal configurations. Table 8.1 presents the hyperparameters optimized for driver models. Each hyperparameter plays a role in tuning the model's capacity and learning process.

Table 8.1.: Optimized hyperparameters for each participant's model and the resulting number of trainable parameters

ID	Learning rate	Batch size	Hidden size	Latent size	Prediction horizon	Trainable parameters
P11	$5.91 \cdot 10^{-5}$	128	128	32	80 (5 s)	$1.01 \cdot 10^{6}$
P23	$3.36 \cdot 10^{-4}$	32	32	64	96 (6 s)	$9.00 \cdot 10^{4}$
P39	$8.51 \cdot 10^{-4}$	64	128	64	96 (6 s)	$1.05 \cdot 10^{6}$

Learning rate and batch size are among the training parameters. The learning rate controls how fast the model adapts to the data. For instance, P11 has the smallest learning rate ($5.91 \cdot 10^{-5}$), suggesting that the model required more fine-tuning for gradual learning due to the complexity of his driving patterns. In contrast, P39 has a notably higher learning rate ($8.51 \cdot 10^{-4}$), likely indicating that rapid adjustments are needed, perhaps due to less intricate data patterns or less variability.
The batch size determines how many samples are processed before the model's weights are updated. A higher batch size, such as 128 for P11, may indicate that the model benefits from processing more data at once, possibly due to stable patterns in driving and physiological behaviors. Conversely, smaller batch sizes such as 32 for P23 may reflect

cases where incremental updates benefit learning, potentially due to more nuanced or variable behavior in these participants' data.
The hidden state size is critical for capturing the complexity of the input sequence. For participants like P11 and P39, where a larger hidden size of 128 is selected, this suggests a need to capture a broader range of temporal dependencies. For participants with smaller hidden sizes like P23 (hidden size of 32), the model could be interpreting simpler or more structured data, with fewer dependencies between sequential observations.
The size of the latent space in a VAE determines how well the model can capture the underlying patterns in the data. A larger latent size allows the model to encode more complex and nuanced features of the driver's behavior, enabling it to represent subtle variations and anticipate future actions with greater accuracy. This is particularly beneficial for experienced drivers like P23 and P39, whose driving patterns may involve sophisticated strategies and anticipatory maneuvers that require a richer representation to model effectively. Conversely, a smaller latent size forces the model to focus on the most salient features of the data, promoting simplicity and reducing the risk of overfitting. For a driver like P11, who may exhibit less consistent driving behavior due to less experience, a smaller latent space helps the model generalize better by capturing essential patterns without becoming overly complex. Thus, adjusting the latent size allows the model to balance complexity and generalization based on the driver's individual characteristics and the intricacies of their driving style.
Prediction horizons range from 5 s to 10 s in the trials. The prediction horizon reflects the number of future data points the model can predict based on past sequences. In general, participants with longer horizons may have exhibited more consistent or predictable behaviors, thus making longer temporal predictions possible and beneficial. Conversely, participants with more variability in driving patterns may benefit from shorter horizons like P11 (5 s), potentially capturing short-term changes more effectively.
During the hyperparameter optimization process, all parameters not under optimization are held constant across models to have a controlled comparison. The initial weights of the models are randomly initialized and set with a constant seed of 42. Both the encoder and decoder in the BLSTM are configured with a single layer. This design is selected considering limited training data and aims to maintain model simplicity by minimizing the number of trainable parameters, thereby reducing the risk of overfitting. Training is regularized through an early stopping mechanism to avoid unnecessary epochs that do not lead to improvement, optimizing computational efficiency and generalization. The early stopping approach monitors the average test loss during training and halts the process when no significant improvement is observed over consecutive epochs.

From the demographic and behavioral data of the participants, summarized in Table 8.2, some trends in hyperparameter selection can be hypothesized.

- For P11, a 25-year-old male with 6 years of driving experience and no prior exposure to automated driving, the optimal hyperparameters are a small learning rate of $5.91 \cdot 10^{-5}$, a large batch size of 128, a large hidden size of 128, a prediction horizon of 5 s, and a latent size of 32. The small learning rate suggests that the model needed to make finer adjustments during training to converge properly, potentially due to

Table 8.2.: Overview of the characteristics of the selected participants

ID	Gender	Age in years	BMI	License duration in years	Avg. driving distance per month [km]	Automated driving experience
P11	Male	25	21.3	5	Between 50 and 500	Never before
P23	Male	36	22.5	19	Between 50 and 500	Assistance system
P39	Female	37	20.2	18	Between 50 and 500	Never before

higher variability or unpredictability in P11's driving behavior stemming from his relative inexperience. The large batch size and hidden size indicate a necessity to process more data per iteration and to capture complex patterns within the driving data, which may reflect a less consistent driving style that requires more extensive modeling to understand.

- In contrast, P23, a 36-year-old male with 20 years of driving experience and familiarity with driving assistance systems, had optimal hyperparameters of a higher learning rate of $3.36 \cdot 10^{-4}$, a small batch size of 32, a small hidden size of 32, a prediction horizon of 6 s, and a latent size of 64. The higher learning rate implies that the model could learn more quickly from P23's data, possibly because his extensive experience and prior use of assistance systems result in more consistent and predictable driving patterns. The smaller batch and hidden sizes suggest that the model required less data per iteration and fewer parameters to capture the essential features of his driving behavior, indicating a simpler underlying structure due to his experience. The longer prediction horizon and larger latent size might be necessary to model the anticipatory aspects of his driving, reflecting his ability to plan further ahead on the road.

- P39, a 37-year-old female with 19 years of driving experience but no prior experience with automated driving, showed optimal hyperparameters of the highest learning rate at $8.51 \cdot 10^{-4}$, a batch size of 64, a large hidden size of 128, a prediction horizon of 6 s, and a latent size of 64. The highest learning rate indicates that the model could rapidly adjust to her driving data, suggesting consistent driving patterns due to her extensive experience. However, the large hidden size similar to P11's might reflect a need to capture more complex patterns, potentially because, despite her experience, the lack of familiarity with automated driving introduced variability in her behavior that the model needed to account for.

However, it is important to approach these findings with caution due to the small sample size and the exploratory nature of the analysis. Further studies with larger and more diverse participant groups are necessary to validate these potential correlations and to uncover underlying causal mechanisms.
The Bayesian hyperparameter optimization effectively identifies individualized model configurations that likely reflect both the behavioral and physical characteristics of the drivers.

The observed differences highlight the importance of individualized models in predictive tasks related to human behavior, such as driving. Each model's architecture is fine-tuned not only to the behavioral patterns but also to potential demographic factors.

8.3. Vector autoregressive model as a baseline for multivariate time series prediction

A vector autoregressive (VAR) model [ZW03] is selected as a baseline to evaluate the performance of the developed and trained autoencoder (AE) model for multivariate time series prediction. The VAR model is chosen as a baseline to provide a simplified approach for modeling linear interdependencies among multiple time series variables, as required in the current study. By comparing the predictions of the AE model with those of the baseline, the benefits of incorporating nonlinear dynamics into time series prediction can be effectively evaluated.
The VAR model is a multivariate time series model that generalizes the univariate autoregressive model [BP70] to capture the linear interdependencies among multiple variables. It models each variable as a linear function of its own past values and the past values of all other variables in the system. This approach allows the VAR model to effectively capture the dynamics of systems where variables influence each other over time. The mathematical formulation of a VAR model of order p (denoted as VAR(p)) is given by

$$\mathbf{y}_t = \mathbf{c}_v + \sum_{i=1}^{p} \mathbf{A}_i \mathbf{y}_{t-i} + \mathbf{u}_t. \tag{8.3.1}$$

In this equation, $\mathbf{y}_t \in \mathbb{R}^{D\times 1}$ is a vector of endogenous variables at time t, where D is the number of variables. The vector $\mathbf{c}_v \in \mathbb{R}^{D\times 1}$ is a vector of constants (intercepts). The matrices $\mathbf{A}_i \in \mathbb{R}^{D\times D}$ are coefficient matrices for each lag i, representing the influence of lagged values on the current value. The term $\mathbf{u}_t \in \mathbf{R}^{D\times 1}$ is a vector of error terms (innovations), assumed to be white noise with zero mean. Each variable in the VAR model is thus a linear combination of its own lagged values and the lagged values of all other variables in the system up to lag order p.
The parameters of the VAR model, including the intercept vector $\mathbf{c}_v$ and the coefficient matrices $\mathbf{A}_i$, are typically estimated using ordinary least squares (OLS) [Zda14] regression. This method involves stacking the lagged observations and solving the resulting system of linear equations. OLS is preferred due to its computational efficiency and the desirable statistical properties of the estimators under the assumption that the error terms are uncorrelated and homoscedastic.
VAR models offer several advantages that make them suitable as a baseline for multivariate time series prediction. The linear structure of VAR models allows for straightforward interpretation of the influence of lagged variables, aiding in understanding the dynamics of the system. Additionally, VAR models are well-established in the field of time series analysis, with extensive literature and tools available for estimation, diagnostics, and forecasting. As a linear model, the VAR model provides a benchmark to assess the performance gains achieved by more complex models that capture nonlinearities and higher-order interactions.

Algorithm 8.3.1.: Training process of the VAR model

Require: $\{\mathbf{y}_t\}_{t=1}^{T}$: multivariate time series data, where $\mathbf{y}_t \in \mathbb{R}^D$, p: lag order, h_v: prediction horizon.

Data preprocessing

for $k = 1$ to D **do** ▷ Iteration on feature

  $d_k \leftarrow 0$ ▷ Initialize differencing order

  $z_{t,k}^{(0)} \leftarrow y_{t,k}$

  while True **do**

    Perform augmented Dickey-Fuller (ADF) test on $z_{t,k}^{(d_k)}$ and obtain p-value $p_{\text{ADF},k}$

    if $p_{\text{ADF},k} < 0.05$ **then** ▷ Series is stationary

      break

    else

      $z_{t,k}^{(d_k+1)} = \Delta z_{t,k}^{(d_k)} = z_{t,k}^{(d_k)} - z_{t-1,k}^{(d_k)}$

      $d_k \leftarrow d_k + 1$

$a_{max} \leftarrow \max\{d_1, d_2, \ldots, d_K\}$ ▷ Maximum differencing order

for $k = 1$ to D **do** ▷ Iteration on feature

  $y_{t,k}^{(a_{max})} \leftarrow \Delta^{a_{max}} y_{t,k}$ ▷ Apply differencing

Form differenced time series $\{\mathbf{y}_t^{(a_{max})}\}_{t=a_{max}+1}^{T}$

for $n = 1$ to N **do** ▷ Fit and forecast on folds

  $\mathcal{D}_{\text{train}} \leftarrow \{\mathbf{y}_t^{(a_{max})}\}_{t \notin \mathcal{T}_n}$ ▷ Training set

  $\mathcal{D}_{\text{test}} \leftarrow \{\mathbf{y}_t^{(a_{max})}\}_{t \in \mathcal{T}_n}$ ▷ Test set

Fitting

Fit VAR model of order p using $\mathcal{D}_{\text{train}}$:

$$\mathbf{y}_t^{(a_{max})} = \mathbf{c}_v + \sum_{i=1}^{p} \mathbf{A}_i \mathbf{y}_{t-i}^{(a_{max})} + \mathbf{u}_t$$

Estimate parameters $\mathbf{c}$ and $\{\mathbf{A}_i\}_{i=1}^{p}$ using OLS

Forecasting

for each time t in $\mathcal{D}_{\text{test}}$ **do**

  Use past p observations to predict h steps ahead:

$$\hat{\mathbf{y}}_{t+h_v}^{(a_{max})} = \mathbf{c} + \sum_{i=1}^{p} \mathbf{A}_i \hat{\mathbf{y}}_{t+h_v-i}^{(a_{max})}$$

  Apply inverse differencing to obtain forecasts on original scale:

$$\hat{\mathbf{y}}_{t+h_v} = \hat{\mathbf{y}}_{t+h_v}^{(a_{max})} + \sum_{k=1}^{a_{max}} \Delta^{-k} \mathbf{y}_{t+h_v-k}$$

Evaluation

Compute error metrics over test set: RMSE, r-squared (R^2), Pearson correlation coefficient, and MAE

Compute aggregated evaluation metrics for all folds

The implementation of the VAR model in this study involves data preprocessing, model fitting, forecasting, and evaluation. Algorithm 8.3.1 gives the pseudocode for the implementation of the VAR model. Each step is carefully designed to ensure the model is properly configured and that the results are meaningful for comparison with those of the AE model. In addition to the original data preprocessing, stationarity of the data, a property of a time series where its statistical characteristics, such as mean, variance, and autocorrelation, remain constant over time, is also tested. Stationarity is a critical assumption for VAR models, as non-stationary data can lead to spurious regression results. Each feature is tested for stationarity using the ADF test [CL95]. If a feature is found to be non-stationary, differencing is applied until stationarity is achieved. To ensure consistency in the dataset, the maximum differencing order across all features is applied.
A LOGO-CV approach is employed, where each session served as the test set once, while the remaining sessions constituted the training set. This method allows evaluation of the model's generalization performance across different sessions. For each fold, the VAR model is trained using the differenced series from the training sessions. The lag order of the model is set equal to the sequence input length, which is 10 s. The model parameters are estimated using OLS regression, fitting the model to capture the linear relationships among the variables.
Rolling forecasts are then performed on the test data using a sliding window approach to predict the next horizon. For each position in the test data, a window of past observations equal to the lag order is used to predict the next sequence of future values (prediction horizon of 5 s). Since the data are differenced during preprocessing to achieve stationarity, it is necessary to convert the model's predictions back to the original scale. This is achieved by cumulatively summing the differenced forecasts and adding them to the last observed value before the forecast period.

8.4. Performance evaluation of the VAE-BLSTM-HA driver model

In this section, the performance of the VAE-BLSTM-HA models trained on data collected from participants P11, P23, and P39 is evaluated. For each participant, a separate model with the data gathered from that participant is trained and tested.
During the model training, a LOGO-CV strategy is employed across six folds, with each fold representing data from a distinct driving scenario. The training process for each fold is conducted over a maximum of 50 epochs, with an early stopping approach, to monitor the learning progression over time. At the end of each epoch, the trained model is evaluated using the test data, which is held out during training, to assess the model's development and performance throughout the training process.
Figures 8.1 to 8.3 present a detailed view of the training and validation losses across the six folds during cross-validation for the participants. Each subplot represents a specific fold and illustrates six key loss metrics recorded per epoch, including the total training and validation losses as well as their components: reconstruction losses and KL divergence losses.
For all three drivers, the training loss exhibits a steady decline across epochs, indicating

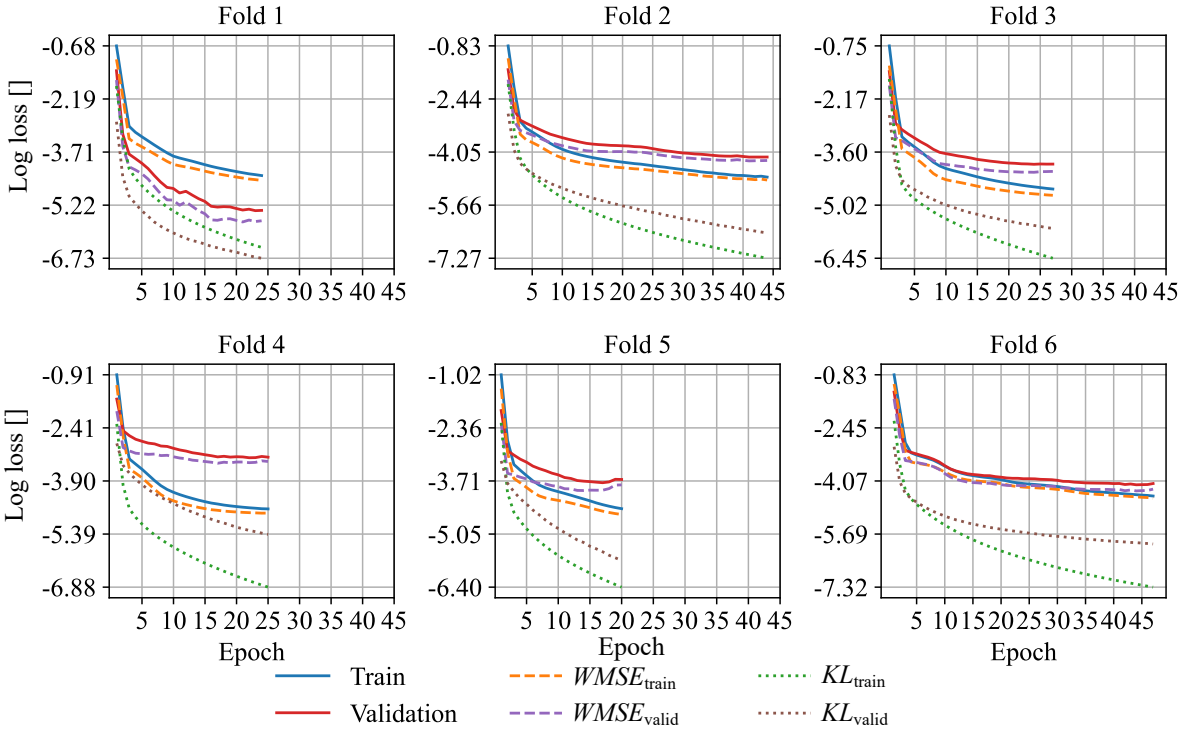


Figure 8.1.: Evolution of training and validation loss for each fold (group) in VAE-BLSTM-HA model of P11

effective optimization of the model parameters. The reconstruction loss $WMSE_{\text{train}}$ decreases more significantly compared to the KL_{train} loss, suggesting that the model prioritizes minimizing reconstruction error while regularizing the latent space with KL divergence. This behavior may be attributed to the unnormalized model outputs during training. The KL_{train} loss decreases initially and then stabilizes, which is consistent with the expected behavior of VAEs, as the latent space distribution aligns with the prior distribution.
The validation losses, evaluated using held-out test data, show a consistent reduction across epochs for all drivers. This suggests that the model generalizes well to unseen data. However, the magnitude and rate of validation loss reduction vary among drivers.

- For P11, the validation loss decreases rapidly in the initial epochs and levels off after approximately five epochs. This pattern indicates that the model quickly learns the underlying data structure but reaches a point of diminishing returns. The reconstruction loss $WMSE_{\text{valid}}$ dominates the total validation loss.

- The validation loss for P23 demonstrates a slower and more gradual decline compared to P11. While the reconstruction loss decreases consistently, the KL_{valid} loss shows fluctuations before stabilizing, suggesting challenges in achieving a smooth latent space distribution. Despite these fluctuations, the overall trend indicates generalization.

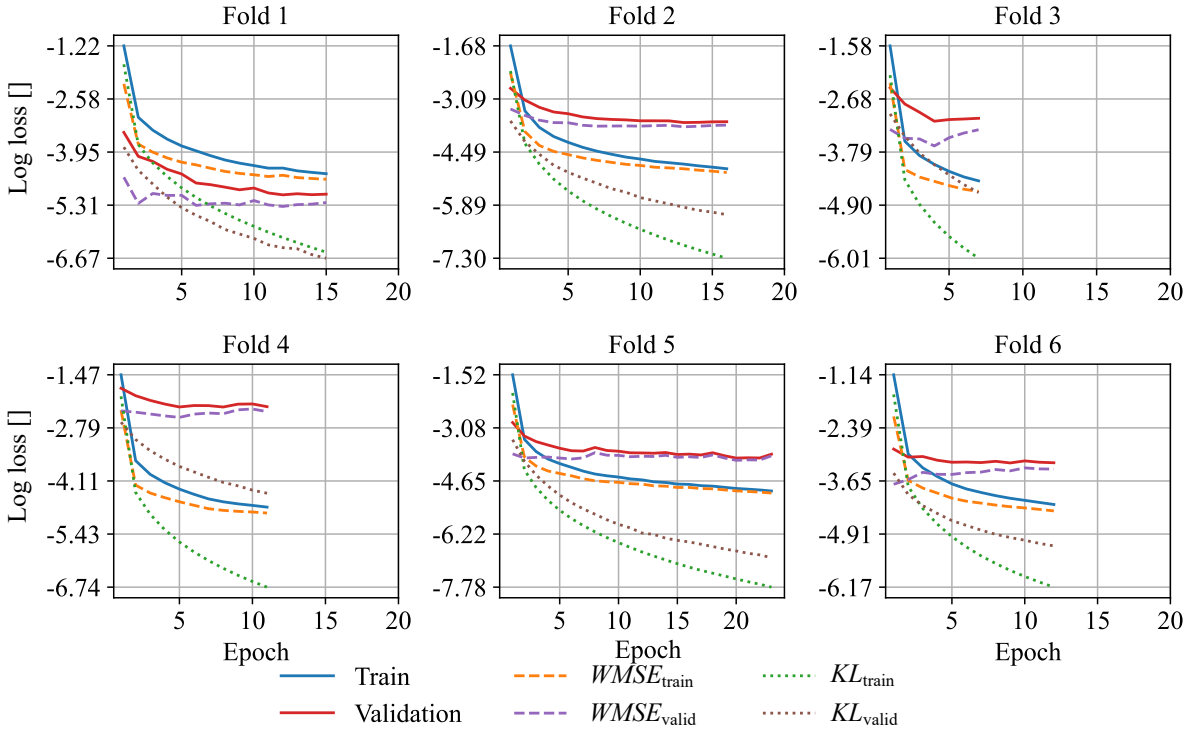


Figure 8.2.: Evolution of training and validation loss for each fold (group) in VAE-BLSTM-HA model of P23

- The validation loss for P39 exhibits a behavior similar to P11, with a rapid initial decrease followed by a plateau. The reconstruction loss $WMSE_{\text{valid}}$ and KL_{valid} loss both contribute to the overall validation loss, with the reconstruction loss showing a more significant reduction. This suggests that the model effectively balances reconstruction accuracy and latent space regularization for this driver.

Across all drivers, the difference between training and validation losses remains small, indicating that the model does not suffer from significant overfitting. The consistent reduction in reconstruction loss highlights the model's ability to capture input and construct the predicted data accurately. The stabilization of KL loss further demonstrates that the latent space achieves the desired regularization, ensuring robust representations.
In summary, the VAE-BLSTM-HA performs effectively in learning driver-specific behaviors. The results demonstrate that the model achieves a balance between reconstruction accuracy and latent space regularization, generalizing well to unseen data. These observations underline the potential of the proposed architecture for modeling driver-specific patterns.
The performance evaluation of the proposed VAE-BLSTM-HA model is conducted in comparison with a baseline VAR model, as mentioned in Section 8.3. During data preprocessing for the VAR model, a differencing level of one is uniformly applied to all datasets based on the results of the ADF test. This step is essential to achieve stationarity in the time series data, which is a critical assumption for the VAR model's validity.
Table 8.3 presents a comparison of the evaluation metrics, RMSE and MAE, between the two models. The reported values are averaged across all features. The comparison

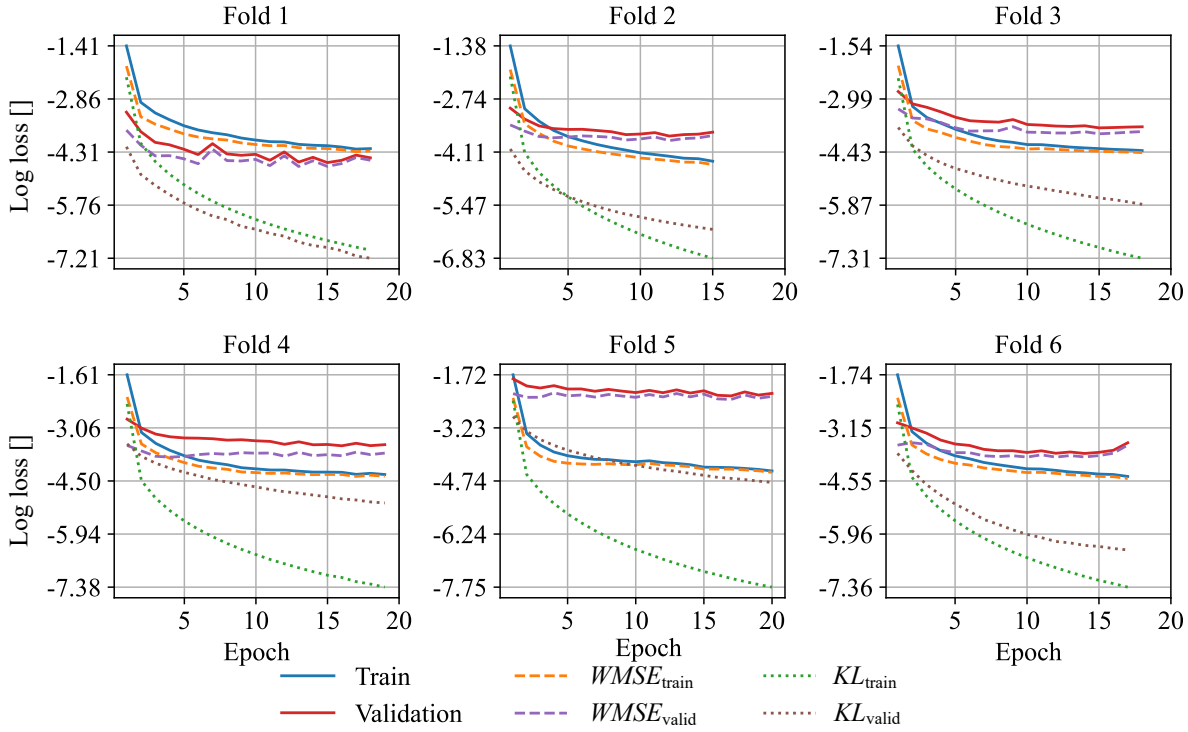


Figure 8.3.: Evolution of training and validation loss for each fold (group) in VAE-BLSTM-HA model of P39

reveals performance improvements with the VAE-BLSTM-HA model over the baseline VAR model for the three participants. Specifically, the VAE-BLSTM-HA model achieves RMSE values of 0.12, 0.16, and 0.15 for the three selected drivers, significantly lower than the VAR models' RMSE of 0.30, 0.65, and 0.87. This indicates that the VAE-BLSTM-HA model has a considerably reduced average error in its predictions, capturing the underlying data patterns more effectively. Similarly, the models' MAE values are 0.05, 0.08, and 0.07, whereas the VAR models' MAE stands at 0.19, 0.30, and 0.41. The lower MAE further confirms the enhanced accuracy of the VAE-BLSTM-HA model in terms of absolute error, reducing the overall prediction discrepancies compared to the baseline.

The improvement in these metrics underscores the benefits of the advanced architecture employed by the VAE-BLSTM-HA model. Incorporating a VAE allows the model to learn latent representations of the data, effectively capturing the inherent variability in driving behaviors. The BLSTM networks facilitate the modeling of temporal dependencies in both past and future directions, enhancing the sequence prediction capabilities. The Hopfield layer and attention mechanism further refine the model's focus on relevant features and patterns within the data, enabling it to handle complex, nonlinear relationships that the VAR model cannot capture.

In contrast, the VAR model, despite being a valuable tool for linear time series analysis, lacks the complexity to model the nuanced and nonlinear dynamics present in driver behavior data. Its reliance on stationarity and linearity limits its ability to adapt to the intricate patterns inherent in human driving actions, leading to higher prediction errors

Table 8.3.: Comparison of prediction performance of the baseline VAR model and the trained VAE-BLSTM-HA on test data for participants P11, P23, and P39

ID	Model	Prediction horizon	Total RMSE	Total MAE	min. RMSE	max. RMSE
P11	VAR	80 (5 s)	0.30	0.19	0.18	0.45
	VAE-BLSTM-HA	80 (5 s)	0.12	0.05	0.05	0.18
P23	VAR	80 (5 s)	0.65	0.30	0.23	2.11
	VAE-BLSTM-HA	96 (6 s)	0.16	0.08	0.05	0.29
P39	VAR	80 (5 s)	0.87	0.41	0.27	3.54
	VAE-BLSTM-HA	96 (6 s)	0.15	0.07	0.06	0.28

as evidenced by the elevated RMSE and MAE values.

8.5. Analysis of takeover reaction prediction accuracy

In the context of automated driving, a critical challenge lies in accurately predicting a human driver's response when prompted to take over control from an automated system. The focus of this study is to evaluate the efficacy of the proposed VAE-BLSTM-HA model in predicting driver reactions during takeover situations, compared to a baseline VAR model. Both models are trained and tested on datasets specifically comprising scenarios where the driver receives a TOR to resume control, capturing the dynamics of these transitional events.

To assess the performance of the models in predicting driver reactions, several evaluation metrics are utilized, such as RMSE and MAE of reaction time and the percentage of match between predicted reaction type and actual reaction. The reaction type match is further refined to steering match and pedal match. Steering match is the percentage of correct predictions regarding steering engagements, including the direction (right or left) of the steering action. Similarly, pedal match is the percentage of accurate predictions of pedal engagements, distinguishing between brake and accelerator pedals. In cases where the driver reacts with both simultaneously, both steering and pedal actions are labeled.

Table 8.4 presents a comparative overview of the performance metrics for both the VAE-BLSTM-HA model and the VAR baseline model.

For all participants, the VAE-BLSTM-HA model significantly outperformed the VAR model in predicting reaction times. Specifically, for Participant P11, the RMSE decreased from 4.74 s (VAR) to 1.71 s (VAE-BLSTM-HA), and the MAE decreased from 3.27 s to 1.20 s. For Participant P23, the RMSE reduced from 9.03 s to 1.83 s, and the MAE from 6.47 s to 1.44 s. For Participant P39, the RMSE decreased from 5.94 s to 1.81 s, and the MAE from 3.50 s to 1.61 s. These results indicate that the VAE-BLSTM-HA model provides a more accurate prediction of driver reaction times, with reductions in RMSE ranging from approximately 3 s to over 7 s, and reductions in MAE of similar magnitudes.

The performance in predicting the type of reaction, as measured by the match scores, presents a more nuanced picture. For the overall match score, the VAE-BLSTM-HA model

Table 8.4.: Comparative performance metrics for takeover reaction prediction in test data for participants P11, P23, and P39

ID	Model	Reaction time RMSE [s]	Reaction time MAE [s]	Match score [%]	Steering Match [%]	Pedal Match [%]
P11	VAR	4.74	3.27	33.33	33.33	33.33
	VAE-BLSTM-HA	1.71	1.20	50.00	66.67	33.33
P23	VAR	9.03	6.47	66.67	40.00	50.00
	VAE-BLSTM-HA	1.83	1.44	50.00	0.00	60.00
P39	VAR	5.94	3.50	50.00	75.00	0.00
	VAE-BLSTM-HA	1.81	1.61	50.00	75.00	33.33

improved from 33.33 % to 50.00 % for P11, while for P23, the VAR model had a higher match score (66.67 %) compared to the VAE-BLSTM-HA model (50.00 %). For P39, both models achieved the same overall match score of 50.00 %. A plausible explanation for these findings may stem from the model's inability to directly incorporate drivers' prior experience, which appears to play a role in shaping reaction types. Specifically, the VAE-BLSTM-HA model performed better for P11, a driver with less experience, suggesting that not accounting for prior experience may have less impact on novice drivers. In contrast, P23 and P39, who both have more extensive driving histories, likely exhibit behavior patterns that the current model does not capture, resulting in less accurate predictions for these participants.

Regarding the steering match, the VAE-BLSTM-HA model shows an improvement for P11, increasing from 33.33 % to 66.67 %. However, for P23, the steering match has dropped to 0.00 % with the VAE-BLSTM-HA model, compared to 40.00 % with the VAR model. For P39, both models maintained a high steering match of 75.00 %.

For P11, the pedal match remained the same at 33.33 % with both models. The VAE-BLSTM-HA model improved the pedal match for P23 from 50.00 % to 60.00 %. These findings suggest that while the VAE-BLSTM-HA model enhances reaction time predictions across all participants, its effectiveness in predicting reaction types varies between individuals.

Participant-specific observations reveal that for P11, the VAE-BLSTM-HA model not only improves reaction time predictions but also shows a noticeable increase in steering action prediction accuracy, indicating the model's capacity to capture P11's steering behavior patterns more effectively. For P23, despite the significant improvement in reaction time predictions, the VAE-BLSTM-HA model experienced a decrease in steering match accuracy, dropping to 0.00 %, though there is a slight improvement in pedal action prediction. This discrepancy may be attributed to the model's difficulty in generalizing P23's steering responses, possibly due to imbalanced steering labels. For P39, reaction time predictions improved with the VAE-BLSTM-HA model, and the steering match remained consistently high at 75.00 % for both models. The pedal match improved from 0.00 % to 33.33 %, indicating better prediction of pedal actions with the proposed model.

The superior performance of the VAE-BLSTM-HA model in predicting reaction times across all participants underscores the effectiveness of this architecture in modeling tem-

poral dependencies and complex patterns in driver behavior. However, the mixed results in predicting reaction types highlight potential areas for model refinement. The decline in steering match accuracy for P23 suggests that the model may require additional tuning or incorporation of more discriminative features to capture the nuances of individual driving styles. It is important to note that the driver reaction data exhibit imbalanced labels, with actions steering left and braking being more prevalent. This imbalance is particularly evident in the dataset of Participant P23.
The enhanced reaction time predictions suggest that the VAE-BLSTM-HA model can serve as a reliable tool for anticipating driver responses in automated driving systems, which is critical for designing safer DVI. The ability to predict not only when but how a driver will react enables the development of adaptive TOR strategies for individual drivers.

8.6. Ablation study

In an effort to develop an efficient driver model for automated driving, a VAE-BLSTM-HA model is introduced, which integrates a VAE, BLSTM layers, a Hopfield network layer, and an attention mechanism. To discern the contribution of each component, an ablation study is conducted. This involves systematically removing one component at a time, retraining the model, and evaluating the performance impact. The ablation study encompasses six model variants, including

1. the full model with all components (VAE-BLSTM-HA),

2. variational autoencoder with BLSTM and Hopfield layers in encoder and decoder structure (VAE-BLSTM-H),

3. variational autoencoder with BLSTM in encoder and decoder structure and attention mechanism (VAE-BLSTM-A),

4. variational autoencoder with LSTM and Hopfield layers in encoder and decoder structure and attention mechanism (VAE-LSTM-HA),

5. autoencoder with BLSTM and Hopfield layers in encoder and decoder structure and attention mechanism (AE-BLSTM-HA), and

6. BLSTM and Hopfield layers with attention mechanism (BLSTM-HA).

Each model undergoes hyperparameter optimization using Bayesian optimization with the same configuration, except for the BLSTM-HA model. For this model, the latent size is removed from the hyperparameters since the model does not have an AE. The models are evaluated using metrics including RMSE and MAE for total predictions, reaction time RMSE and MAE, and reaction type match scores for steering and pedal actions. The performance metrics for each model variant are summarized in Table 8.5.

Table 8.5.: Performance metrics for ablation study models, averaged across all folds and the three participants

Model	Learning rate	Batch size	Hidden size	Latent size	Prediction horizon	Average RMSE	Average MAE	Reaction time RMSE	Reaction time MAE	Match score [%]	Steering match [%]	Pedal match [%]
VAE-BLSTM-HA	$2.25 \cdot 10^{-4}$	64	128	64	96	0.15	0.07	1.71	1.27	66.67	50.00	100.00
VAE-BLSTM-H	$1.05 \cdot 10^{-4}$	32	128	32	144	0.17	0.09	3.85	3.61	50.00	66.67	33.33
VAE-BLSTM-A	$7.41 \cdot 10^{-5}$	64	128	32	128	0.16	0.08	3.67	3.36	66.67	50.00	66.67
VAE-LSTM-HA	$4.56 \cdot 10^{-3}$	64	128	64	144	0.17	0.09	3.81	3.58	83.33	75.00	66.67
AE-BLSTM-HA	$3.15 \cdot 10^{-3}$	32	64	64	112	0.16	0.08	1.22	0.88	50.00	0.00	100.00
BLSTM-HA	$2.15 \cdot 10^{-3}$	32	32	-	80	0.14	0.07	1.26	0.98	50.00	0.00	100.00

Impact of the attention mechanism In this variation, shown in Figure 8.4a, the attention mechanism is removed to assess its impact on model performance. The absence of attention led to a decrease in the learning rate to $1.05 \cdot 10^{-4}$ and a reduction in batch size to 32. The latent size decreased to 32, and the prediction horizon extended to 144. The performance metrics indicated a slight increase in average RMSE and MAE to 0.17 and 0.09, respectively. More significantly, the reaction time RMSE and MAE increased to 3.85 s and 3.61 s, compared to 1.71 s and 1.27 s in the main model. The overall match score decreased from 66.67 % to 50.00 %, with the steering match improving to 66.67 % but the pedal match dropping to 33.33 %. These results suggest that the attention mechanism plays a crucial role in enhancing the model's ability to focus on relevant temporal features, thereby improving reaction time predictions and pedal action accuracy.

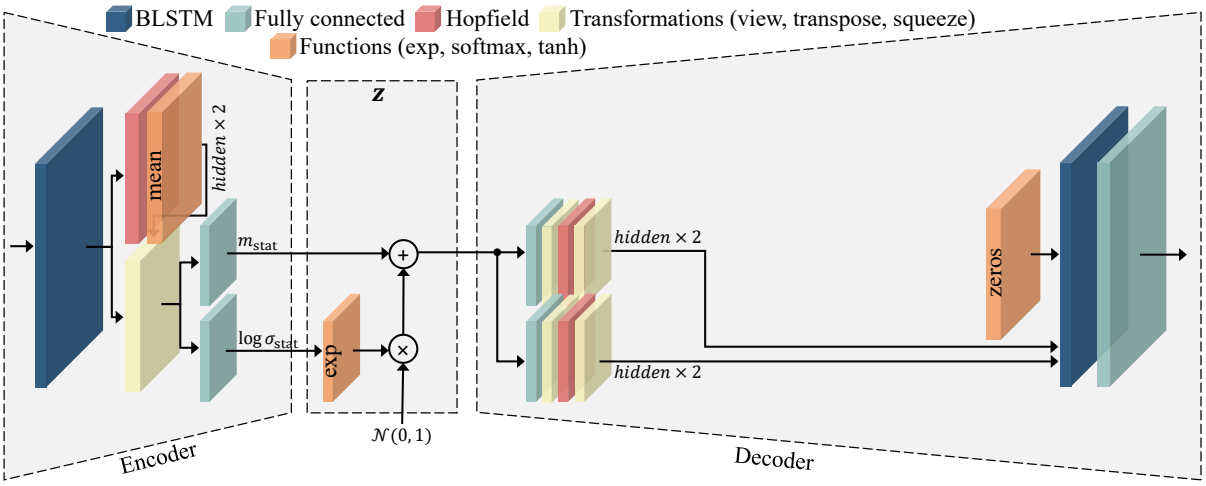
Role of the Hopfield network layer This model omits the Hopfield layer to evaluate its contribution (see Figure 8.4b). The learning rate further decreased to $7.41 \cdot 10^{-5}$, while the batch size remained at 64. The latent size is reduced to 32, and the prediction horizon is set at 128. The average RMSE and MAE are slightly elevated at 0.16 and 0.08, respectively. The reaction time RMSE and MAE increase to 3.67 s and 3.36 s, indicating diminished accuracy in predicting reaction times. The overall match score remained at 66.67 %, consistent with the main model, and the pedal match improved to 66.67 %. These findings imply that the Hopfield layer significantly contributes to capturing temporal dependencies crucial for accurate reaction time predictions, although its absence does not markedly affect the prediction of reaction types.

Effect of replacing BLSTM with LSTM Replacing the BLSTM with a unidirectional LSTM resulted in the VAE-LSTM-HA model. This change aims to assess the importance of bidirectional context in sequence modeling. The learning rate increased largely to $4.56 \cdot 10^{-3}$, and other hyperparameters remain similar to those of the main model. The average RMSE and MAE rise to 0.17 and 0.09, respectively, and the reaction time RMSE and MAE in-

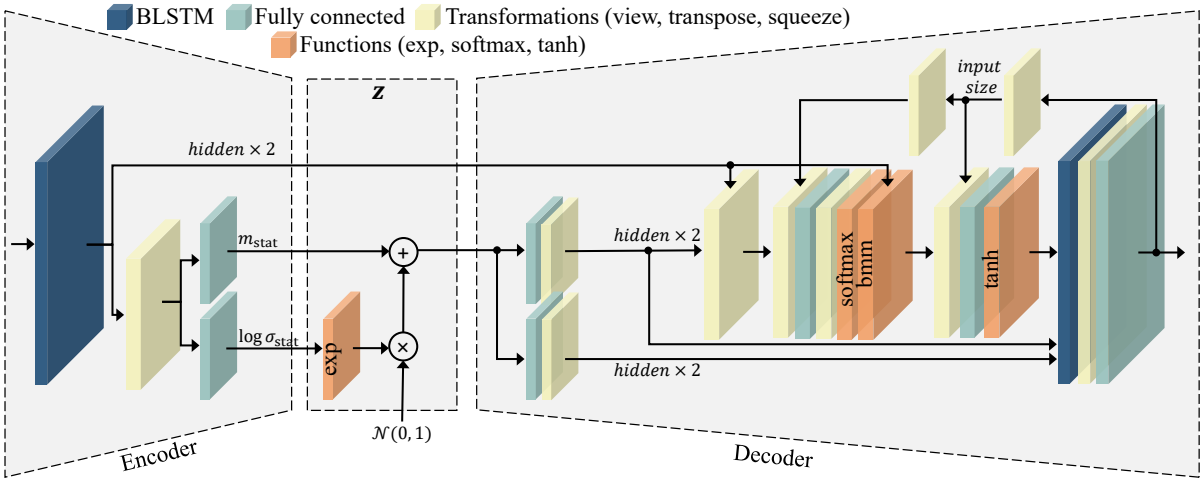
creased to 3.81 s and 3.58 s. Interestingly, the overall match score improved to 83.33 %, and the steering match increased to 75.00 %, while the pedal match decreased to 66.67 %. These outcomes suggest that while bidirectional processing enhances reaction time predictions by capturing information from both past and future contexts, the unidirectional LSTM may better capture certain aspects of reaction type prediction due to its focus on sequential dependencies.

Replacing VAE with AE This variant investigated the effect of the stochastic latent space on model performance. The learning rate increases to $3.15 \cdot 10^{-3}$, the batch size decreases to 32, and the hidden size is reduced to 64. The latent size remained at 64, and the prediction horizon is adjusted to 112 (7 s). The average RMSE and MAE are consistent with the main model at 0.16 and 0.08. Notably, the reaction time RMSE and MAE are improved to 1.22 s and 0.88 s, outperforming the main model. However, the overall match score has decreased to 50.00 %, and the steering match dropped to 0.00 %, while the pedal match remained high at 100.00 %. These results indicate that the deterministic latent space of the AE enhances reaction time prediction but may not capture the variability necessary for accurate steering action predictions.

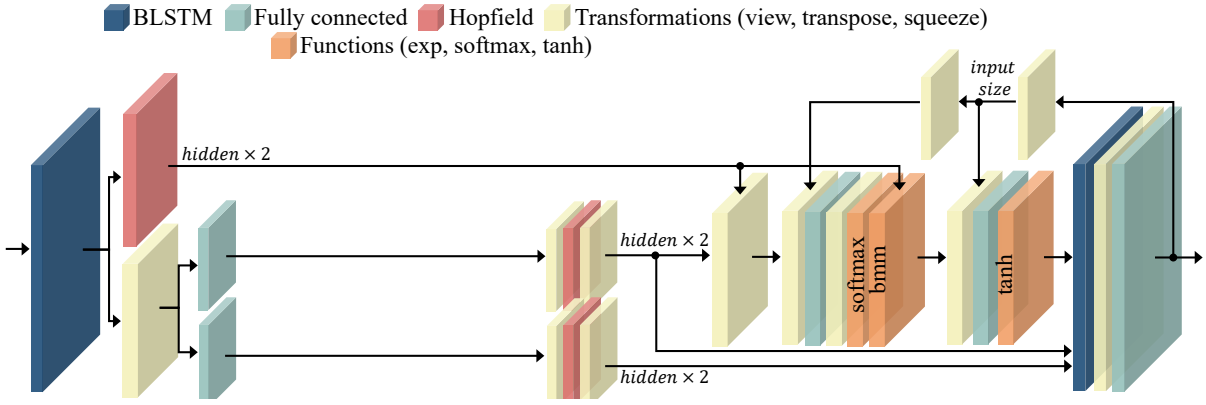
Removing the VAE component The final model removes the entire VAE structure as depicted in Figure 8.4c, focusing solely on the BLSTM with Hopfield layers and the attention mechanism. The optimized learning rate is set at $2.15 \cdot 10^{-3}$, with a batch size of 32 and a reduced hidden size of 32. The latent size is not applicable in this configuration, and the prediction horizon decreases to 80. The average RMSE and MAE slightly improve to 0.14 and 0.07, respectively. The reaction time RMSE and MAE are 1.26 s and 0.98 s, showing enhancements over the main model. Prediction of the reaction type shows a similar behavior to that of model AE-BLSTM-HA. The performance of pedal prediction remains high; however, the steering match comes down to 0.00 %. These findings suggest that the VAE component may not be critical for reaction time prediction in this context and that the BLSTM combined with Hopfield layers and the attention mechanism is effective in modeling driver reactions. Nonetheless, the VAE may still be valuable for capturing underlying data distributions, which could be significant for other predictive aspects such as reaction type.
Overall, the ablation study highlights the individual contributions of each component in the VAE-BLSTM-HA model. The attention mechanism and Hopfield layer are instrumental in enhancing reaction time predictions and action type accuracy. The bidirectional nature of the BLSTM is crucial for capturing temporal dependencies that improve reaction time predictions, while the stochastic latent space of the VAE is important for modeling the variability in driver behavior, particularly in steering actions. Removing these components adversely affects the model's performance, confirming their reinforcing role in accurately simulating driver responses to takeover requests.



(a) VAE-BLSTM-H model: removed attention mechanism



(b) VAE-BLSTM-A model: removed Hopfield layer



(c) BLSTM-HA model: removed VAE structure

Figure 8.4.: Architectures of different model variations trained for the ablation study

9

Conclusion and future directions

9.1. Summary and conclusion

This dissertation set out to develop a holistic driver model capable of predicting driver state and reactions during the critical transition from automated to manual driving. Recognizing the limitations of existing models, which often oversimplify complex driving scenarios and human decision-making processes, an approach is proposed that integrates both cognitive and emotional aspects of driver behavior. By grounding the model in the ACT-R cognitive architecture and the theory of constructed emotion, it is aimed to advance ACT-R architecture to capture the close-to-real-world driving experiences.
The integration of these two theories allows for the creation of a model that not only considers the cognitive processes involved in driving but also incorporates the dynamic affect and constructed nature of emotion. This dual-framework approach allows the model to account for the continuous variation in affect during decision-making, leading to more accurate predictions of human sensation development.
To validate the model, the comprehensive manD 1.0 dataset is collected through designed driving simulator studies. This dataset provides rich, synchronized information on driving situations, driver states, and vehicle dynamics. By training individualized models for each participant, focusing particularly on takeover performance, the personal differences in past experiences and physiological characteristics are accounted.
Comparative analyses demonstrated that the proposed driver model outperforms the baseline model in predicting driver states and reactions. The incorporation of psychological architectures improves the model's ability to generalize across diverse driving situations and increases its resilience to variability. These findings suggest that the presented approach can contribute to the development of advanced driver assistance systems, potentially enhancing safety during critical transitions in automated driving.

Highlights and contributions

This work aims to advance the field of driver modeling and automated driving through several contributions. A precisely designed subject study is conducted using a static driving simulator to collect synchronized data from individuals across a diverse array of driving contexts and driver states. This study is carefully crafted to encompass a broad

spectrum of emotional and cognitive changes, providing a controlled and reproducible environment where various scenarios could be systematically tested. The use of a driving simulator allows for the precise manipulation of driving conditions and the collection of detailed behavioral data, ensuring that the findings are robust and reproducible. This comprehensive approach addresses the limitations of previous studies that are confined to specific scenarios with limited situational criteria.

One of the central outcomes of this research is the development of the manD 1.0 dataset, an open-access resource that captures extensive data on driver behavior under varying conditions and states. This dataset includes synchronized information on driver actions, physiological signals, vehicle dynamics, and environmental factors collected during the subject study. By making the manD 1.0 dataset publicly available, this work provides a valuable resource for the research community, facilitating the replication of studies, validation of models, and fostering the development of new approaches in driver behavior analysis and automated driving systems. The dataset stands as a considerable contribution, enabling researchers to explore complex interactions between drivers and automated systems further.

The collected data undergo thorough analysis to identify and extract initial correlations between driver characteristics, driver states, driving contexts, and subsequent actions. This analysis provides initial insights into the patterns of driver behavior during automated driving and transitions from automated to manual driving. By uncovering how various emotional and cognitive factors influence driver reactions, the study offers a foundational understanding that informs the development of more accurate and robust driver models. These initial correlations serve as a basis for future research endeavors aiming to enhance predictive capabilities and understand the underlying mechanisms of driver behavior.

At the core of this dissertation is the development of a holistic driver model grounded in neuroscientific findings and psychological constructs—specifically, the ACT-R cognitive architecture developed by Anderson and the theory of constructed emotion proposed by Feldman Barrett. This innovative model simulates the human mind's capacity to integrate driver state and driving context, predicting subsequent changes in both. By incorporating both cognitive and emotional theories, the model moves beyond traditional approaches that often simplify driver behavior, providing a more nuanced and accurate prediction of driver reactions during critical takeover situations. This integration allows the model to handle a wide range of driver states, including those influenced by intense emotions.

Through individualized training on participant-specific data, the proposed driver model accounts for individual differences such as past experiences and physiological characteristics. Comparative analysis with baseline models demonstrated that incorporating psychological architectures enhances the accuracy of predicting driver states and reactions. The model showed improved robustness against variability and increased generalizability across diverse driving situations. This advancement holds potential for use in advanced driver assistance systems, particularly in critical situations where accurate prediction of driver state and reaction can prevent severe safety consequences.

9.2. Future directions

While this research has made several contributions to the field of driver modeling, several avenues for future work remain open. Although the model is validated on takeover situations, its design is inherently generalizable. Future research should apply the model to a broader range of driving scenarios, such as urban driving, highway merging, or adverse weather conditions, to test its versatility and adaptability.
Integrating the model into a feedback loop [Dar+20a] and testing it in real-time with interaction signals would provide valuable insights into its practical applicability. Additionally, real-world implementation as vehicle assistance systems can reveal further challenges and opportunities not evident in simulator-based research.
Incorporating additional data sources, such as biometric sensors, eye-tracking, or vocal analysis, could enrich the model's understanding of driver state. Multimodal data integration and investigating the interplay between different sensory inputs might enhance the prediction of complex behaviors and reactions.
Future research could also focus on refining and expanding the generative machine learning algorithms, which draw heavily on neuroscientific insights, by incorporating more advanced deep learning architectures or reinforcement learning methods to enhance their predictive power. A limitation of the current model is that it does not directly account for drivers' existing knowledge prior to the experiment; addressing this gap through transfer learning would allow the prefrontal cortex component to capture and leverage these pre-existing mental models. Ensuring that a hippocampal model supports robust contextual recall is also critical, as episodic memories are known to shape decision-making processes. Moreover, modeling the DLPFC could involve implementing a decision layer that weights inputs based on logical reasoning, working memory, and goal-directed objectives. Finally, incorporating a reinforcement learning agent to mimic the basal ganglia's mechanism for action selection (by learning from rewards and penalties) could further improve the overall fidelity and adaptability of the proposed cognitive architecture.
Conducting studies using the whole manD 1.0 dataset could provide insights into how driver behavior and reactions evolve over time, particularly with increased exposure to automated driving systems. Longitudinal data might reveal trends and patterns valuable for refining the model.
Testing the model in different cultural contexts could assess its universal applicability. Cultural factors can influence driving behavior, and understanding these nuances is important for creating globally relevant driver assistance systems.

Final thoughts

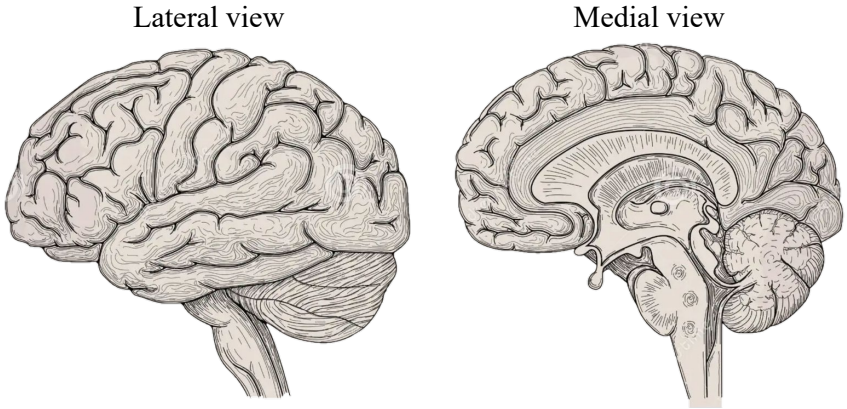
The development of a holistic driver model that effectively integrates cognitive and emotional aspects represents a significant step forward in understanding and predicting driver behavior during transitions from automated to manual driving. By addressing the complexities of real-world driving and acknowledging the dynamic nature of human affect and cognition, this research contributes to safer and more reliable driver assistance technologies. As automated driving systems become increasingly prevalent, the need for advanced models that can anticipate human actions becomes more critical. The findings of this dissertation

lay a foundation for future innovations in this field, with the ultimate goal of enhancing road safety and improving the interaction between humans and automated systems. Continued research and collaboration will be essential in realizing the full potential of these technologies and in navigating the challenges that lie ahead.

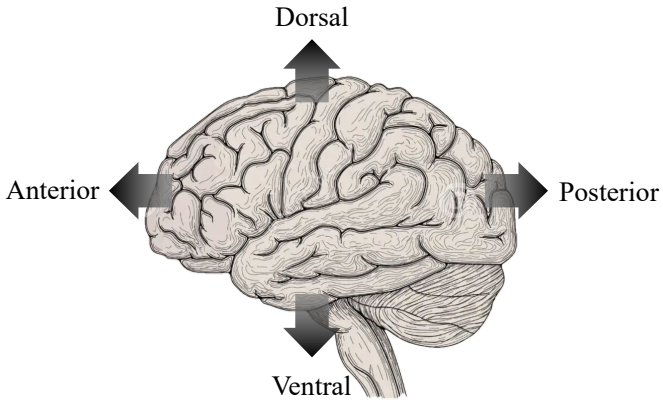
A

Human brain anatomical views

Figure A.1 shows the key anatomical views and directional terms related to the human brain. Figure A.1a presents both the lateral and medial views of the brain, the standard perspectives used in neuroanatomy. Figure A.1b illustrates the main anatomical directions, dorsal, ventral, anterior, and posterior, representing the spatial orientations within the brain.



(a) Lateral and medial views of the human brain, showing the external and internal structures visible from these perspectives



(b) The spatial orientations used within brain roadmap

Figure A.1.: Anatomical terms of location

B

Human sensations

The brain, being an organ enclosed within the skull, does not have direct contact with the external environment. It relies on three primary processes to receive and interpret information from both inside and outside the body. These processes (interoception, exteroception, and proprioception) are differentiated based on the types of sensations they convey to the brain. Exteroception pertains to the brain's perceptual inference about the state of the external environment, encompassing sensory inputs such as sight, sound, and touch [Pet+17]. Interoception refers to the brain's capacity to perceive, interpret, and integrate signals originating from within the body, such as changes in physiological states like hunger or heart rate [BK21]. Proprioception, on the other hand, involves the brain's sense of the relative positioning and movement of body parts, providing critical feedback for motor control and spatial awareness [TA18]. The following sections provide a brief explanation of each of these processes.

B.1. Exteroception

Extroception involves inputs from various sensory systems, including sight, hearing, smell, taste, and touch that provide information about the external world.

- Vision (sight) is facilitated by photoreceptors, specifically rods and cones, located in the retina of the eyes. These receptors detect light and enable the perception of various visual attributes such as color, shape, depth, and motion. Rods are more sensitive to light and help in low-light conditions, while cones are responsible for color vision and function best under brighter light.

- Audition (hearing) involves hair cells located in the inner ear, which serve as the primary receptors for sound waves. These hair cells detect variations in sound waves, translating them into neural signals that allow the brain to perceive different auditory characteristics such as pitch, volume, and tone. This process enables humans to interpret a wide range of sounds, from speech to music.

- Olfaction (smell) is mediated by olfactory receptors situated in the nasal cavity. These receptors detect airborne chemical molecules and convert them into neural signals, which are then processed by the brain to identify and differentiate between

various odors. Olfaction plays a crucial role in flavor perception, environmental awareness, and even memory formation.

- Gustation (taste) is based on the function of taste buds located on the tongue and other parts of the oral cavity. These receptors detect chemicals present in food and beverages, enabling the perception of tastes.

- Tactition (touch) is governed by mechanoreceptors embedded in the skin. These receptors respond to various forms of physical contact, such as pressure, vibration, and texture, which are translated into sensory signals. The brain interprets these signals to perceive touch, allowing humans to interact with their environment through tactile feedback and spatial awareness.

Exteroceptive inputs are critical for interacting with and navigating the environment, enabling organisms to respond to external changes and potential threats.

B.2. Interoception

Interoception is focused on physiological and biochemical states within the body. Interoceptive inputs are the sensory signals that inform the brain about the internal state of the body, such as temperature, heart rate, hunger, or respiratory rate. These inputs are critical for the regulation of internal conditions and are tightly coupled with homeostatic and allostatic processes.
The brain uses sensory inputs to make inferences about the internal states of the body, allowing it to regulate these states effectively. The distinction between interoception and exteroception, therefore, lies in the purpose of the sensory inputs rather than their origin: interoceptive inputs are used to control bodily states, while exteroceptive inputs are used to interact with and understand the environment.

B.3. Proprioception

Proprioception involves the brain's integration of sensory information primarily from muscles, tendons, and joints. This information is critical for maintaining posture, balance, and coordinated movements. Proprioceptive inputs are sensory signals that inform the brain about the relative position of body parts and their movement dynamics. These inputs provide feedback to the motor system, enabling adjustments and fine-tuning of motor actions.
Proprioception concerns the external spatial configuration of the body and is vital for motor control. It involves distinct sensor-effector loops, where the sensory inputs guide motor actions to adjust the body's position and movement in space. These loops do not directly contribute to homeostatic or allostatic regulation, which are associated with interoception. Instead, they support the continuous and dynamic coordination of the body's movements through motor actions that respond to proprioceptive feedback [Tou+24].

It is notable that sensory processes can be categorized in various ways depending on the focus and application of the research. For instance, nociception (pain perception), thermoception (temperature perception), equilibrioception (balance perception), mechanoreception (mechanical pressure perception), baroreception (blood pressure perception), and chemoreception (chemical detection) can each be considered distinct sensory modalities. These categorizations allow for a detailed understanding of how different sensory inputs are processed and integrated by the brain to maintain bodily functions and respond to environmental changes. However, for specific applications such as driver modeling, where the emphasis is on understanding the overall sensory-motor responses and decision-making processes rather than isolating each sensory input type, such detailed distinctions are not always necessary. In this context, a more generalized approach to categorizing sensory processes, focusing on broader categories like interoception, exteroception, and proprioception, can be more practical and effective for modeling driver behavior and responses.

C

Experimental apparatus

The study is carried out in a driving simulator, set up under tightly regulated laboratory conditions to maintain a consistent temperature of 22 °C. To replicate a real vehicle's interior environment, the driver's area is isolated from the external surroundings, effectively minimizing any potential distractions or disturbances. The ambient lighting in the laboratory is switched off; however, a controlled light source within the simulator is provided to the driver to emulate the typical lighting found in a vehicle. Details regarding the equipment utilized in the experiment are provided in the following sections.

C.1. Driving simulation system

The static driving simulator (see Figure C.1) used in this study is comprised of a driving setup that features three 55 inch screens positioned directly in front of the driver's seat. These screens are arranged at 120° angles relative to each other, creating a comprehensive field of view that closely mirrors a real driving experience. To enhance the driver's sense of immersion, the entire simulator setup is enclosed, effectively isolating it from the surrounding environment. The simulator is equipped to support automatic functions during manual driving, eliminating the need for drivers to use a gear stick or clutch. It provides a range of driving modes, from SAE L0 to SAE L3. For interaction with the automated system, an additional pedal (communication pedal) is installed to the right of the accelerator, which allows drivers to respond to binary prompts from the system. A tablet, positioned to the right of the steering wheel, serves as an interface for NDRT, such as gaming. Furthermore, an additional screen on the driver's right-hand side displays videos designed to elicit specific emotional responses from participants prior to their simulated drive. To account for comfort and realistic in-cabin interactions, items like a book, a bottle of water, and cookies are also placed within easy reach of the driver.

The simulation environment is powered by SCANeR Studio 2021, a real-time simulation software (developed by AVSimulation in Boulogne-Billancourt, France). This platform facilitates high-fidelity simulation experiences and is integrated with external sensors through a custom application programming interface (API) built using Python. This setup ensures seamless communication between the simulation software and the external devices, enabling a robust and dynamic simulation environment.

Figure C.1.: Static driving simulator employed to collect the manD 1.0 dataset

C.2. Driver monitoring sensors

Intel RealSense camera To monitor and analyze the facial expressions and behaviors of participants during the simulation, an Intel RealSense D435 camera (Intel Corporation, Santa Clara, California, U.S.; as in Figure C.2a) is strategically positioned above the central display, approximately 1.5 m away from the driver's seat. This placement ensures a clear and unobstructed view of the driver's face, allowing for precise data capture.
The camera captures high-resolution RGB images at 1,920 × 1,080 px, with a frame rate of 30 Hz. This configuration provides a detailed and continuous stream of visual data that is crucial for accurately assessing and interpreting the driver's facial movements and emotional responses during the simulation.

SmartEye eye tracking system To monitor eye movements and gather gaze data during the driving simulation, a SmartEye Aurora eye tracking system (SmartEye AB, Gothenburg, Sweden; as in Figure C.2b) is installed on the driving mockup, positioned atop the dashboard directly in front of the driver. The eye tracker is placed at an optimal distance of approximately 0.7 m from the driver's eyes, ensuring precise and reliable tracking. The system employs a combination of dark pupil detection and corneal reflection to accurately monitor eye movements, operating at a sampling rate of 120 Hz to capture fine details of the driver's visual attention.
The eye tracking hardware is integrated with SmartEye Pro 9.3 software, which receives the eye movement data through a wired connection. This software facilitates seamless real-time data transfer to the SCANeR simulation platform, allowing for synchronized analysis of eye tracking data alongside other simulator metrics. This setup is crucial for understanding driver behavior, attention, and response under varying simulated driving conditions.

(a) Intel RealSense D435 camera

(b) SmarEye eye tracking system

(c) Empatica E4 wristband

(d) BIOPAC EEG-ECG sensor

Figure C.2.: Driver monitoring sensors used for data collection

Empatica E4 wristband Physiological data from drivers are gathered using an Empatica E4 wristband (Empatica Inc., Boston, Massachusetts, U.S.; see Figure C.2c), a wearable device designed for continuous monitoring. The wristband is equipped with a PPG sensor that captures the BVP, enabling the calculation of key cardiovascular metrics such as interbeat interval (IBI), HR, and HRV. Additionally, the wristband includes an EDA sensor that tracks variations in the skin's electrical conductance, which can be influenced by the type of electrodes used, the sensor's placement, and external factors like ambient temperature and humidity. The device also incorporates a 3-axis accelerometer to detect hand movements and an infrared thermopile sensor to measure peripheral skin temperature. Data collected by the Empatica E4 wristband are transmitted via Bluetooth v4.0 to the E4 Streaming Server software, which then relays the information in real-time to the SCANeR simulation software through a Python-based API. This setup ensures synchronized physiological data collection alongside other experimental metrics for comprehensive analysis.

BIOPAC electroencephalogram (EEG)-ECG sensor For the purpose of capturing brain activity during the experiment, participants are equipped with a B-ALERT X10 wireless EEG sensor system (BIOPAC, Goleta, California, U.S.; see Figure C.2d). This device features nine wet EEG channels positioned at specific scalp locations (Poz, Fz, Cz, C3, C4, F3, F4, P3, P4) to monitor brain signals, alongside an ECG channel for recording cardiac activity. Additionally, two mastoid electrodes are positioned behind the ears to enhance signal quality. The EEG and ECG data are transmitted wirelessly via Bluetooth to the AcqKnowledge 4 software, which then streams the data in real-time to the SCANeR simulation software, ensuring synchronized integration of neural and physiological measurements throughout the experiment.

BodiTrak seat-pressure-sensor mats To monitor the pressure distribution during the driving simulation, the acquisition system incorporates two BodiTrak2 Pro seat-pressure-

sensor mats (Vista Medical Ltd, Winnipeg, Manitoba, Canada; see Figure C.3), with one mat positioned on the driver's seat and the other on the backrest. Each mat covers an area of $0.45 \times 0.45\,\mathrm{m}^2$ and is equipped with 1,024 sensors arranged in a 32×32 grid to measure the pressure distribution accurately. For enhanced performance, the seat mat is calibrated to a pressure level of 200 mmHg, while the backrest mat is set to 100 mmHg. The calibration ensures an accuracy of 10 % at the midpoint of each range; for instance, the backrest mat's calibration maintains an accuracy of ± 10 mmHg at a pressure of 50 mmHg. The pressure data are collected via USB 2.0 at a sampling rate of up to 150 Hz and are subsequently transferred in real-time to the SCANeR software through a Python-based API, enabling synchronized data analysis within the driving simulation framework.

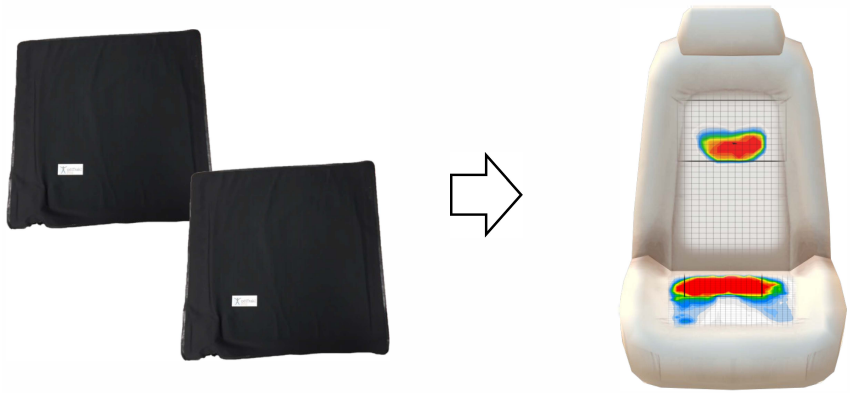


Figure C.3.: BodiTrak seat-pressure-sensor mats on the seat and backrest

D

Data preparation

This dissertation utilizes the manD 1.0 dataset to train the driver models. The following sections provide an overview of the data preparation steps for each type of sensor data taken from this dataset. All data within the manD 1.0 dataset are recorded at a sampling rate of 256 Hz. Synchronization of the data is achieved directly during the recording process, using SCANeR Studio's real-time simulation environment. The simulation software incorporates a communication package that supports custom-defined interactions with various sensors, enabling a synchronized data stream from the Empatica E4 wristband and BodiTrak seat-pressure-sensor mats.

D.1. Driving simulator output data

Data from the driving environment are gathered using SCANeR Studio real-time simulation software, which initially records these data at a frequency of 20 Hz, but to improve synchronization across all data sources, the data are resampled to 256 Hz in real-time and then recorded. Before employing the data in this research, they are again downsampled to 16 Hz.

A key aspect of data analysis centers around the area immediately surrounding the ego-vehicle. When an object is identified in the same lane ahead of the ego-vehicle, the distance to the object is computed as

$$\mathbf{d}_o = \sqrt{(x_{\text{ego}} - x_{\text{front}})^2 + (y_{\text{ego}} - y_{\text{front}})^2} \quad [\text{m}], \tag{D.1.1}$$

where $(x_{\text{ego}}, y_{\text{ego}})$ and $(x_{\text{front}}, y_{\text{front}})$ represent the positions of the ego-vehicle and the object in front, respectively. Additionally, the TTC with an approaching object or pedestrian is calculated as

$$\text{TTC} = \frac{\mathbf{d}_o}{v_{\text{ego}} - v_{\text{front}}} \quad [\text{s}], \tag{D.1.2}$$

where v_{ego} and v_{front} denote the velocities of the ego-vehicle and the vehicle in front, respectively. These metrics are crucial for understanding potential collision scenarios and driver behavior in response to nearby objects.

D.2. Physiological data from Empatica E4

The Empatica E4 wristband, worn on the dominant arm of each participant, is used to collect wrist acceleration data along with several physiological signals. The raw data obtained from the wristband undergoes preprocessing and is resampled to 256 Hz, after which it is stored in a .txt file format. The acceleration data from the Empatica E4 is recorded within a range of $\pm 2\, g_0$ and is originally represented in units of $\frac{1}{16}\, g_0$. However, in the manD 1.0 dataset, this data is already converted, and the unit of measurement is presented in meters per second squared ($\mathrm{m\,s^{-2}}$). Although the initial sampling rate for acceleration data is 32 Hz, it is upsampled before being recorded, then downsampled to 16 Hz in this study for model training purposes.
BVP data are recorded at a sampling rate of 64 Hz and expressed in nanoWatts (nW), with the sensor capable of capturing a range between ± 500 nW. EDA is sampled at a frequency of 4 Hz, with the data presented in microSiemens (µS). Additionally, the wristband records peripheral skin temperature at a sampling rate of 4 Hz, reported in degrees Celsius (°C).

D.3. Seat-pressure-sensor readings

The BodiTrak2 Pro seat-pressure-sensor mats capture pressure data using sensors that sample at a rate of 15 Hz. Although the data are collected from the sensor at this rate, the final dataset provides the data at a rate of 256 fps to ensure synchronization with other collected datasets. Before use in this study, the data are again downsampled to 16 Hz. After data collection, a preprocessing step is conducted to address any inaccuracies or errors. This process involves correcting frames that contain only zero values, which may occur due to network interruptions. Such frames are replaced with the data from the preceding frame to ensure continuity and accuracy in the dataset.
In this study, the center of mass of participants seated in the driver's seat is calculated using pressure distributions obtained from seat-pressure-sensor mats. The center of mass coordinates (x, y) are determined for both the seat and backrest by applying the command `center_of_mass` function from the scipy.ndimage package [Vir+20]. These coordinates are then provided to the models as part of the proprioceptive data.

E

Correlation matrices

The proposed driver model in this study is trained individually for three participants using data collected from these subjects in the manD 1.0 dataset. Figures E.1 to E.3 present the correlation matrix heatmaps of the features in the dataset for each driver.

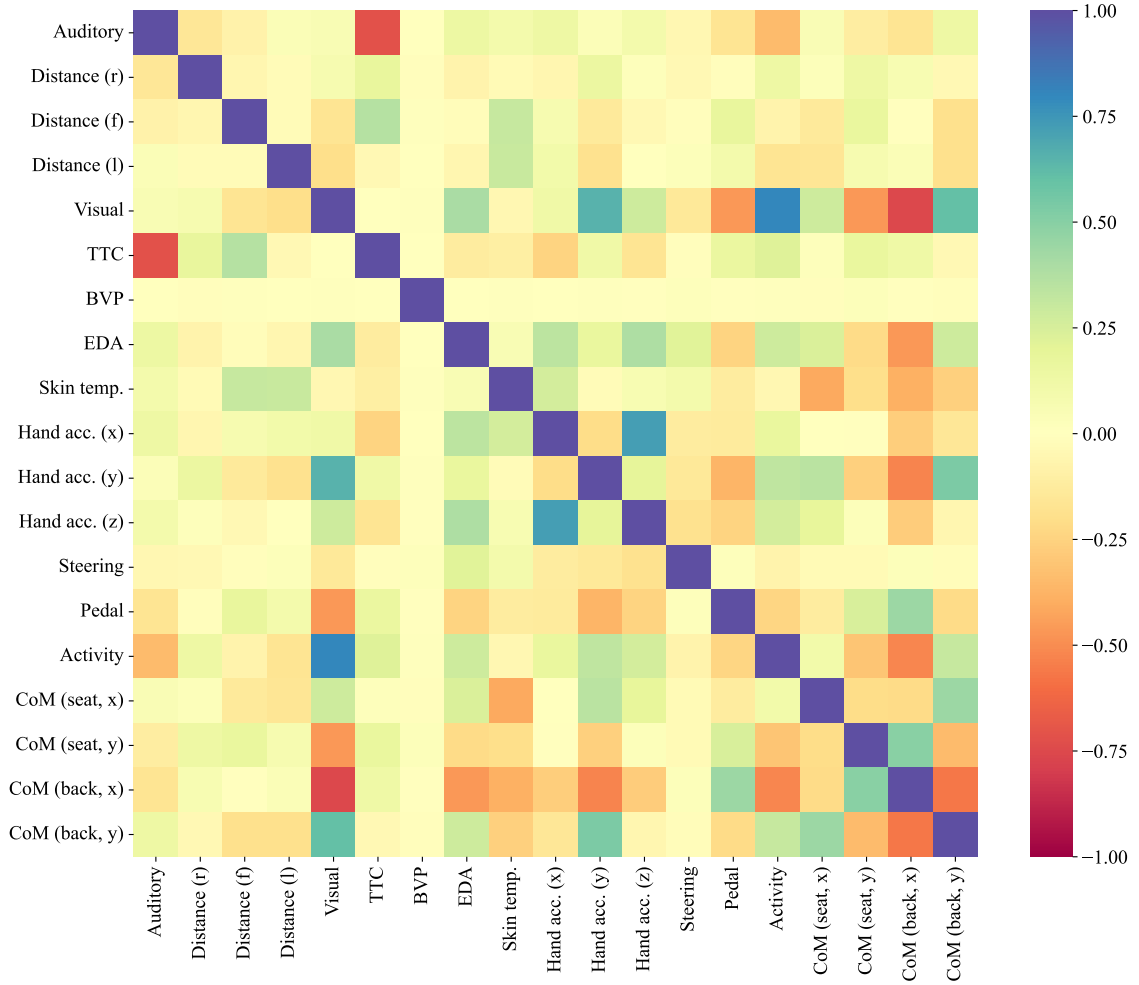


Figure E.1.: Correlation matrix heatmap of features collected from P11

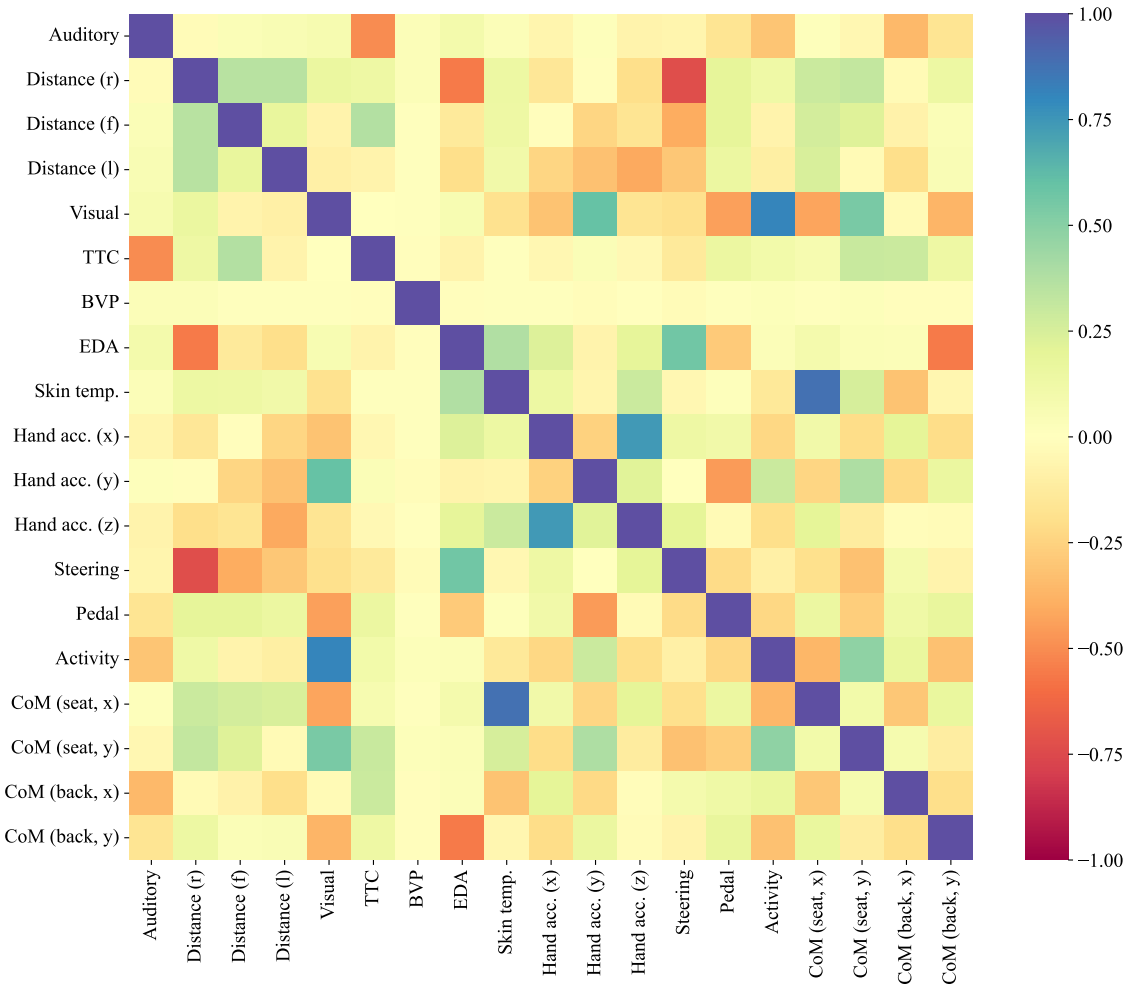


Figure E.2.: Correlation matrix heatmap of features collected from P23

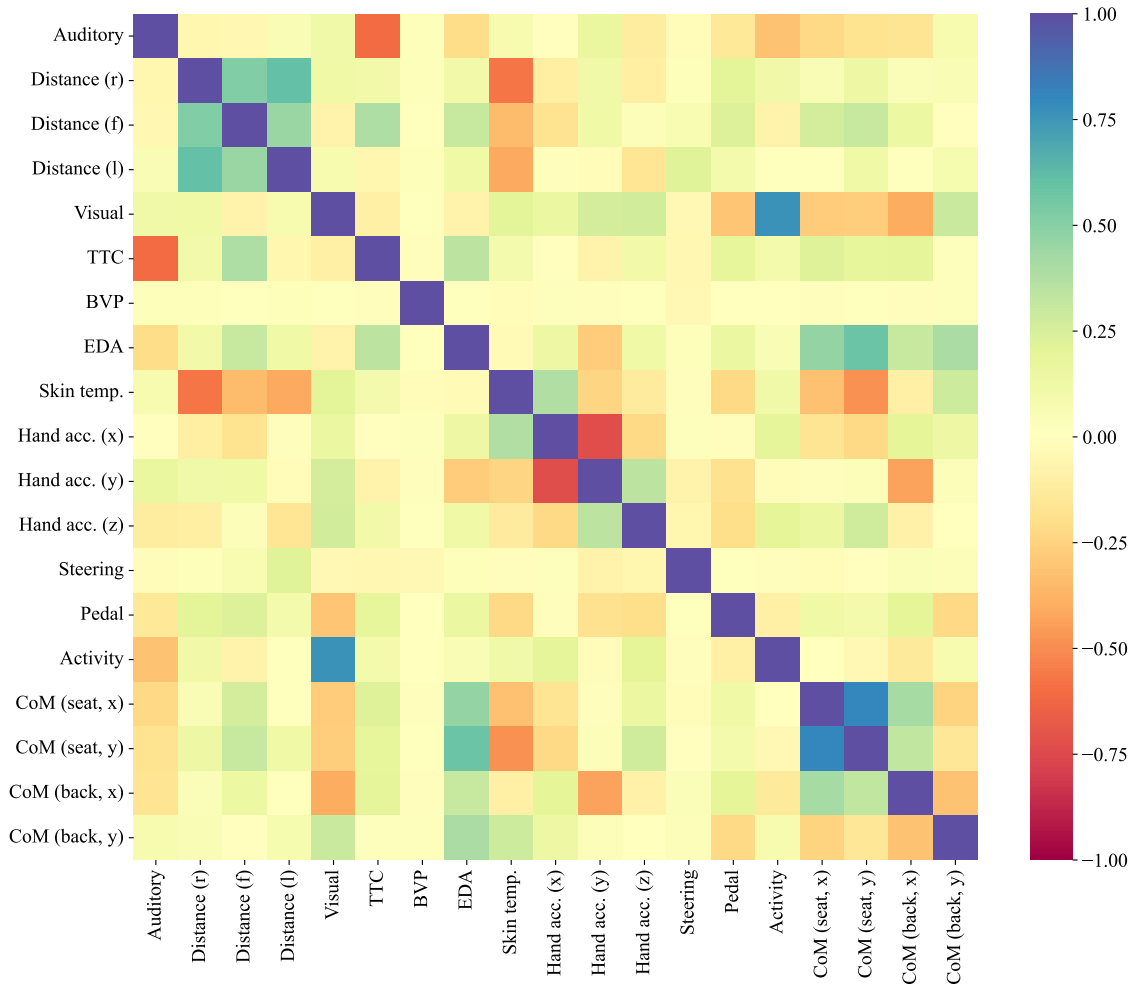


Figure E.3.: Correlation matrix heatmap of features collected from P39

F

Usage of generative AI - Affidavit

☐ not at all

☒ for correcting, optimizing, or restructuring the entire work (This eliminates the need for explicit marking of individual passages or sections, as this type of usage refers to the entire written work. Explicit marking in the text is not necessary, as this serves as the global indication.)

☐ Code optimization: Optimization or restructuring of software function

☐ Code generation: Creating entire software functions from a detailed functional description.

☐ Substance generation in code: Generating entire software source code

☐ Media optimization: Correction, optimization, or restructuring of entire passages

☒ Media generation: Creating entire passages from given content.

☐ Substance generation in media: Generating entire sections

☐

☒ More, namely:

to correct, optimize, and restructure all codes developed during this work for data analysis and model creation.

I assure that I have provided all usages completely. Missing or incorrect information may be considered an attempt to deceive.

place, date

Khazar Dargahi Nobari

Bibliography

[AB73] **J. R. Anderson and G. H. Bower:** *Human associative memory.* V. H. Winston, 1973.

[Aki+19] **T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama:** "Optuna: A next-generation hyperparameter optimization framework". In: *CoRR* abs/1907.10902 (2019). URL: http://arxiv.org/abs/1907.10902.

[And+04] **J. R. Anderson, D. Bothell, M. D. Byrne, S. Douglass, C. Lebiere, and Y. Qin:** "An integrated theory of the mind". In: *Psychological Review* 111.4 (2004), p. 1036.

[And+08] **J. R. Anderson, J. M. Fincham, Y. Qin, and A. Stocco:** "A central circuit of the mind". In: *Trends in Cognitive Sciences* 12.4 (2008), pp. 136–143. URL: http://dx.doi.org/10.1016/j.tics.2008.01.006.

[And07] **J. R. Anderson:** *How can the human mind occur in the physical universe?* Oxford University Press, 2007. URL: https://doi.org/10.1093/acprof:oso/9780195324259.001.0001.

[And13] **J. R. Anderson:** *The adaptive character of thought.* Psychology Press, 2013. URL: http://dx.doi.org/10.4324/9780203771730.

[And93] **J. Anderson:** *Rules of the mind.* Psychology Press, 1993.

[AR99] **J. R. Anderson and L. M. Reder:** "The fan effect: New results and new theories". In: *Journal of Experimental Psychology: General* 128.2 (1999), pp. 186–197. URL: https://doi.org/10.1037/0096-3445.128.2.186.

[Ash+01] **J. A. Ashton-Miller, E. M. Wojtys, L. J. Huston, and D. Fry-Welch:** "Can proprioception really be improved by exercises?" In: *Knee Surgery, Sports Traumatology, Arthroscopy* 9.3 (2001), pp. 128–136. URL: http://dx.doi.org/10.1007/s001670100208.

[Bah+15] **D. Bahdanau, K. Cho, and Y. Bengio:** "Neural machine translation by jointly learning to align and translate". In: *ICLR* (2015). URL: https://arxiv.org/abs/1409.0473.

[Bar06] **L. F. Barrett:** "Solving the emotion paradox: Categorization and the experience of emotion". In: *Personality and Social Psychology Review* 10.1 (2006), pp. 20–46. URL: http://dx.doi.org/10.1207/s15327957pspr1001_2.

[Bar17] **L. Barrett:** *How emotions are made: The secret life of the brain*. Harper-Collins, 2017. URL: https://books.google.de/books?id=hN8MBgAAQBAJ.

[BH93] **D. Barry and J. A. Hartigan:** "A bayesian analysis for change point problems". In: *Journal of the American Statistical Association* 88.421 (1993), pp. 309–319. URL: http://www.jstor.org/stable/2290726 (visited on 08/25/2024).

[BK21] **G. G. Berntson and S. S. Khalsa:** "Neural circuits of interoception". In: *Trends in Neurosciences* 44.1 (2021), pp. 17–28. URL: http://dx.doi.org/10.1016/j.tins.2020.09.011.

[BP66] **L. E. Baum and T. Petrie:** "Statistical inference for probabilistic functions of finite state markov chains". In: *The Annals of Mathematical Statistics* 37.6 (1966), pp. 1554–1563. URL: http://www.jstor.org/stable/2238772 (visited on 08/25/2024).

[BP70] **G. E. P. Box and D. A. Pierce:** "Distribution of Residual Autocorrelations in Autoregressive-Integrated Moving Average Time Series Models". In: *Journal of the American Statistical Association* 65.332 (1970), pp. 1509–1526. URL: https://www.tandfonline.com/doi/abs/10.1080/01621459.1970.10481180.

[BS13] **L. F. Barrett and A. B. Satpute:** "Large-scale brain networks in affective and social neuroscience: Towards an integrative functional architecture of the brain". In: *Current Opinion in Neurobiology* 23.3 (2013), pp. 361–372. URL: http://dx.doi.org/10.1016/j.conb.2012.12.012.

[Bud03] **A. Bud Craig:** "Interoception: The sense of the physiological condition of the body". In: *Current Opinion in Neurobiology* 13.4 (2003), pp. 500–505. URL: http://dx.doi.org/10.1016/s0959-4388(03)00090-4.

[BW07] **D. Badre and A. D. Wagner:** "Left ventrolateral prefrontal cortex and the cognitive control of memory". In: *Neuropsychologia* 45.13 (2007), pp. 2883–2901. URL: http://dx.doi.org/10.1016/j.neuropsychologia.2007.06.015.

[CD08] **N. Caporale and Y. Dan:** "Spike Timing–Dependent Plasticity: A Hebbian Learning Rule". In: *Annual Review of Neuroscience* 31.1 (2008), pp. 25–46. URL: http://dx.doi.org/10.1146/annurev.neuro.31.060407.125639.

[Cep+08] **N. J. Cepeda, E. Vul, D. Rohrer, J. T. Wixted, and H. Pashler:** "Spacing effects in learning: A temporal ridgeline of optimal retention". In: *Psychological Science* 19.11 (2008), pp. 1095–1102. URL: http://dx.doi.org/10.1111/j.1467-9280.2008.02209.x.

[CL95] **Y.-W. Cheung and K. S. Lai:** "Lag order and critical values of the augmented Dickey–Fuller test". In: *Journal of Business & Economic Statistics* 13.3 (1995), pp. 277–280.

[Cor02] **S. Corkin:** "What's new with the amnesic patient H.M.?" In: *Nature Reviews Neuroscience* 3.2 (2002), pp. 153–160. URL: http://dx.doi.org/10.1038/nrn726.

[Dar+20a] **K. Dargahi Nobari, F. Albers, K. Bartsch, and T. Bertram:** "Online feedback control for driver-vehicle interaction in automated driving". In: *Advances in Human Aspects of Transportation*. 2020, pp. 159–165. URL: https://doi.org/10.1007/978-3-030-50943-9_21.

[Dar+21a] **K. Dargahi Nobari, C. Velasquez, and T. Bertram:** "Emotion induction strategies in driving simulator for validated experiments". In: *Human Systems Engineering and Design (IHSED2021) Future Trends and Applications*. 2021. URL: http://dx.doi.org/10.54941/ahfe1001156.

[Dar+22a] **K. Dargahi Nobari, F. Albers, K. Bartsch, J. Braun, and T. Bertram:** "Modeling driver-vehicle interaction in automated driving". In: *Engineering Research* 86.1 (2022), pp. 65–79. URL: https://doi.org/10.1007/s10010-021-00576-6.

[Dar+22c] **K. Dargahi Nobari, A. Hugenroth, and T. Bertram:** "Position classification and in-vehicle activity detection using seat-pressure-sensor in automated driving". In: *AmE 2022 - Automotive Meets Electronics; 13. GMM-Symposium*. 2022, pp. 1–6. URL: https://ieeexplore.ieee.org/document/100259.

[DB19] **K. Dargahi Nobari and T. Bertram:** "Development of driver-vehicle interaction in automated driving: An optimization problem". 2019.

[DB20] **K. Dargahi Nobari and T. Bertram:** "Evolution of driver state through interaction in automated vehicles". 2020.

[DB21] **K. Dargahi Nobari and T. Bertram:** "Unobtrusive monitoring method for estimating the driver state using seat pressure sensors". 2021.

[DB22a] **K. Dargahi Nobari and T. Bertram:** "Driver modeling: Architecture of cognition and emotion". 2022.

[DB22b] **K. Dargahi Nobari and T. Bertram:** "Generalized model for driver activity recognition in automated vehicles using pressure sensor array". In: *Proceedings of the 8th International Conference on Human Interaction & Emerging Technologies (IHIET 2022): Artificial Intelligence & Future Applications*. 2022. URL: http://dx.doi.org/10.54941/ahfe1002733.

[DB23a] **K. Dargahi Nobari and T. Bertram:** "Benchmark dataset: Multimodal driver monitoring". 2023.

[DB23b] **K. Dargahi Nobari and T. Bertram:** *manD 1.0*. Version V2. 2023. URL: https://doi.org/10.7910/DVN/SG9TMD.

[DB23c] **K. Dargahi Nobari and T. Bertram:** "manD 1.0". 2023. URL: https://doi.org/10.7910/DVN/SG9TMD.

[DB24a] **K. Dargahi Nobari and T. Bertram:** "A multimodal driver monitoring benchmark dataset for driver modeling in assisted driving automation". In: *Scientific Data* 11.1 (2024), p. 327.

[DB24d] **K. Dargahi Nobari and T. Bertram:** "Understanding and predicting driver behavior: Insights into cognitive architecture and neurological patterns". 2024.

[Den+19] **C. Deng, S. Cao, C. Wu, and N. Lyu:** "Modeling driver take-over reaction time and emergency response time using an integrated cognitive architecture". In: *Transportation Research Record* 2673.12 (2019), pp. 380–390.

[DM96] **M. Daneman and P. M. Merikle:** "Working memory and language comprehension: A meta-analysis". In: *Psychonomic Bulletin & Review* 3.4 (1996), pp. 422–433. URL: http://dx.doi.org/10.3758/bf03214546.

[Dud12] **Y. Dudai:** "The restless engram: Consolidations never end". In: *Annual Review of Neuroscience* 35.1 (2012), pp. 227–247. URL: http://dx.doi.org/10.1146/annurev-neuro-062111-150500.

[EG01] **G. Edelman and J. Gally:** "Degeneracy and complexity in biological systems". In: *Proceedings of the National Academy of Sciences of the United States of America* 98.24 (2001), pp. 13763–13768. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-0035923604%5C&doi=10.1073%2fpnas.231499798%5C&partnerID=40%5C&md5=9b5d648e85a2cdbd9f8a36210f4af81d.

[Ekm+87] **P. Ekman, W. V. Friesen, M. O'Sullivan, A. Chan, I. Diacoyanni-Tarlatzis, K. Heider, R. Krause, W. A. LeCompte, T. Pitcairn, P. E. Ricci-Bitti, K. Scherer, M. Tomita, and A. Tzavaras:** "Universals and cultural differences in the judgments of facial expressions of emotion". In: *Journal of Personality and Social Psychology* 53.4 (1987), pp. 712–717. URL: http://dx.doi.org/10.1037/0022-3514.53.4.712.

[Ela+20] **Z. Elamrani Abou Elassad, H. Mousannif, H. Al Moatassime, and A. Karkouch:** "The application of machine learning techniques for driving behavior analysis: A conceptual framework and a systematic literature review". In: *Engineering Applications of Artificial Intelligence* 87 (2020), p. 103312. URL: https://www.sciencedirect.com/science/article/pii/S0952197619302672.

[Eva91] **L. Evans:** *Traffic safety and the driver*. Vol. 20. Van Nostrand Reinhold, 1991.

[Fin+02] **J. M. Fincham, C. S. Carter, V. van Veen, V. A. Stenger, and J. R. Anderson:** "Neural mechanisms of planning: A computational analysis using event-related fMRI". In: *Proceedings of the National Academy of Sciences* 99.5 (2002), pp. 3346–3351. URL: http://dx.doi.org/10.1073/pnas.052703399.

[Fle+11] **F. Flemisch, A. Schieben, N. Schoemig, M. Strauss, S. Lueke, and A. Heyden:** "Design of human computer interfaces for highly automated vehicles in the EU-project HAVEit". In: *Universal Access in Human-Computer Interaction*. 2011, pp. 270–279.

[FUL84] **R. FULLER:** "A conceptualization of driving behaviour as threat avoidance". In: *Ergonomics* 27.11 (1984), pp. 1139–1155. URL: https://doi.org/10.1080/00140138408963596.

[GC23] **A. Guskjolen and M. S. Cembrowski:** "Engram neurons: Encoding, consolidation, retrieval, and forgetting of memory". In: *Molecular Psychiatry* 28.8 (2023), pp. 3207–3219. URL: http://dx.doi.org/10.1038/s41380-023-02137-5.

[Gir92] **E. R. Girden:** *ANOVA: Repeated Measures*. 84. Sage Publications, Inc., 1992. URL: https://psycnet.apa.org/record/1992-97020-000.

[GK12] **S. Games and Kiloo:** *Subway Surfers*. Denmark. 2012. URL: https://sybogames.com/subway-surfers/.

[Gol16] **C. G. Gold:** "Modeling of take-over performance in highly automated vehicle guidance". en. PhD thesis. Technische Universität München, 2016, p. 187. URL: https://mediatum.ub.tum.de/1296132.

[Har+15] **S. Hareli, K. Kafetsios, and U. Hess:** "A cross-cultural study on emotion expression and the learning of social norms". In: *Frontiers in Psychology* 6 (2015). URL: http://dx.doi.org/10.3389/fpsyg.2015.01501.

[Har+20] **C. R. Harris, K. J. Millman, and et al.:** "Array programming with NumPy". In: *Nature* 585 (2020), pp. 357–362.

[Has+17] **M. S. Hassanpour, W. K. Simmons, J. S. Feinstein, Q. Luo, R. C. Lapidus, J. Bodurka, M. P. Paulus, and S. S. Khalsa:** "The insular cortex dynamically maps changes in cardiorespiratory interoception". In: *Neuropsychopharmacology* 43.2 (2017), pp. 426–434. URL: http://dx.doi.org/10.1038/npp.2017.154.

[Hay72] **J. C. Hayward:** "Near miss determination through use of a scale of danger". In: *Highway Res. Rec.* 384 (1972), pp. 24–34.

[HB19] **J. B. Hutchinson and L. F. Barrett:** "The power of predictions: An emerging paradigm for psychological research". In: *Current Directions in*

Psychological Science 28.3 (2019), pp. 280–291. URL: https://doi.org/10.1177/0963721419831992.

[HM15] **R. M. Hutchison and J. B. Morton:** "Tracking the brain's functional coupling dynamics over development". In: *The Journal of Neuroscience* 35.17 (2015), pp. 6849–6859. URL: http://dx.doi.org/10.1523/jneurosci.4638-14.2015.

[Hoc97] **S. Hochreiter:** "Long short-term memory". In: *Neural Computation MIT-Press* (1997).

[Hop82] **J. J. Hopfield:** "Neural networks and physical systems with emergent collective computational abilities." In: *Proceedings of the National Academy of Sciences* 79.8 (1982), pp. 2554–2558. eprint: https://www.pnas.org/doi/pdf/10.1073/pnas.79.8.2554. URL: https://www.pnas.org/doi/abs/10.1073/pnas.79.8.2554.

[HP12] **B. M. Herbert and O. Pollatos:** "The body in the mind: On the relationship between interoception and embodiment". In: *Topics in Cognitive Science* 4.4 (2012), pp. 692–704. URL: http://dx.doi.org/10.1111/j.1756-8765.2012.01189.x.

[HR11] **Y. Huang and R. P. N. Rao:** "Predictive coding". In: *WIREs Cognitive Science* 2.5 (2011), pp. 580–593. URL: http://dx.doi.org/10.1002/wcs.142.

[Hu+13] **T.-Y. Hu, X. Xie, and J. Li:** "Negative or positive? The effect of emotion and mood on risky driving". In: *Transportation Research Part F: Traffic Psychology and Behaviour* 16 (2013), pp. 29–40. URL: https://www.sciencedirect.com/science/article/pii/S1369847812000800.

[Iza+74] **C. Izard, F. Dougherty, B. Bloxom, and N. Kotsch:** "The differential emotions scale: A method of measuring the subjective experience of discrete emotions". In: *Nashville* (1974).

[Iza77] **C. E. Izard:** *Human emotions*. Springer US, 1977. URL: http://dx.doi.org/10.1007/978-1-4899-2209-0.

[Jeo+14] **M. Jeon, B. N. Walker, and J.-B. Yim:** "Effects of specific emotions on subjective judgment, driving performance, and perceived workload". In: *Transportation Research Part F: Traffic Psychology and Behaviour* 24 (2014), pp. 197–209. URL: https://www.sciencedirect.com/science/article/pii/S1369847814000412.

[Kal60] **R. E. Kalman:** "A new approach to linear filtering and prediction problems". In: *Journal of Basic Engineering* 82.1 (1960), pp. 35–45. URL: https://doi.org/10.1115/1.3662552.

[KB15] **D. Kingma and J. Ba:** "Adam: A method for stochastic optimization". In: *International Conference on Learning Representations (ICLR)*. 2015.

[Ker+22] **L. Kerruish, A. S. Cheng, K.-H. Ting, and K. P. Liu:** "Exploring the sustained and divided attention of novice versus experienced drivers". In: *Transportation Research Interdisciplinary Perspectives* 16 (2022), p. 100702. URL: https://www.sciencedirect.com/science/article/pii/S2590198222001622.

[Ker+99] **W. N. Kernan, C. M. Viscoli, R. W. Makuch, L. M. Brass, and R. I. Horwitz:** "Stratified randomization for clinical trials". In: *Journal of Clinical Epidemiology* 52.1 (1999), pp. 19–26.

[Kin13] **D. P. Kingma:** "Auto-encoding variational bayes". In: *arXiv preprint arXiv:1312.6114* (2013).

[Kir58] **W. K. Kirchner:** "Age differences in short-term retention of rapidly changing information". In: *Journal of Experimental Psychology* 55.4 (1958), pp. 352–358. URL: http://dx.doi.org/10.1037/h0043688.

[KL51] **S. Kullback and R. A. Leibler:** "On Information and Sufficiency". In: *The Annals of Mathematical Statistics* 22.1 (1951), pp. 79–86. URL: http://dx.doi.org/10.1214/aoms/1177729694.

[Krc17] **D. Krch:** "Tower of london". In: *Encyclopedia of Clinical Neuropsychology*. 2017, pp. 1–5. URL: http://dx.doi.org/10.1007/978-3-319-56782-2_1912-2.

[KW19] **D. P. Kingma and M. Welling:** *An Introduction to Variational Autoencoders*. Now Publishers, 2019. URL: http://dx.doi.org/10.1561/9781680836233.

[Lai+86] **J. Laird, P. Rosenbloom, and A. Newell:** *Universal subgoaling and chunking: The automatic generation and learning of goal hierarchies*. Springer New York, NY, 1986.

[Lap+16] **S. Lapoehn, M. Dziennus, A. Schieben, F. Utesch, T. Hesse, F. Köster, M. Dotzauer, and K. Johann:** "Interaction design for nomadic devices in highly automated vehicles". In: *Mensch und Computer 2016 Proceedings* (2016). URL: https://elib.dlr.de/103550/.

[Leb99] **C. Lebiere:** "The dynamics of cognition: An ACT-R model of cognitive arithmetic". In: *Kognitionswissenschaft* 8.1 (1999), pp. 5–19. URL: http://doi.org/10.1007/bf03354932.

[LeD23] **J. LeDoux:** "The amygdala". In: *Current Biology* 33.2 (2023), R48–R50. URL: http://doi.org/10.1016/j.cub.2022.12.010.

[Lee76] **D. N. Lee:** "A theory of visual control of braking based on information about time-to-collision". In: *Perception* 5.4 (1976), pp. 437–459. URL: https://doi.org/10.1068/p050437.

[Lin+12] **K. A. Lindquist, T. D. Wager, H. Kober, E. Bliss-Moreau, and L. F. Barrett:** "The brain basis of emotion: A meta-analytic review". In:

Behavioral and Brain Sciences 35.3 (2012), pp. 121–143. URL: http://dx.doi.org/10.1017/s0140525x11000446.

[Lio+14] **C.-Y. Liou, W.-C. Cheng, J.-W. Liou, and D.-R. Liou:** "Autoencoder for words". In: *Neurocomputing* 139 (2014), pp. 84–96.

[Liu+06] **Y. Liu, R. Feyen, and O. Tsimhoni:** "Queueing network-model human processor (QN-MHP): A computational architecture for multitask performance in human-machine systems". In: *ACM Trans. Comput.-Hum. Interact.* 13.1 (2006), pp. 37–70.

[Liu+14] **P. Liu, A. Kurt, and Ü. Özgüner:** "Trajectory prediction of a lane changing vehicle based on driver behavior estimation and classification". In: *17th International IEEE Conference on Intelligent Transportation Systems (ITSC)*. IEEE. 2014, pp. 942–947.

[LK00] **J. S. Lerner and D. Keltner:** "Beyond valence: Toward a model of emotion-specific influences on judgement and choice". In: *Cognition and Emotion* 14.4 (2000), pp. 473–493. URL: https://doi.org/10.1080/026999300402763.

[Mar+12] **G. Markkula, O. Benderius, K. Wolff, and M. Wahde:** "A review of near-collision driver behavior models". In: *Human Factors* 54.6 (2012), pp. 1117–1143.

[Mar+18] **C. Marberger, H. Mielenz, F. Naujoks, J. Radlmayr, K. Bengler, and B. Wandtner:** "Understanding and applying the concept of driver availability in automated driving". In: *Advances in Human Aspects of Transportation*. 2018, pp. 595–605.

[McK+10] **W. McKinney et al.:** "Data structures for statistical computing in python". In: *Proceedings of the 9th Python in Science Conference*. Vol. 445. 2010, pp. 51–56.

[Meh80] **A. Mehrabian:** *Basic dimensions for a general psychological theory: Implications for personality, social, environmental, and developmental studies*. Oelgeschlager, Gunn & Hain, 1980.

[Mic85] **J. A. Michon:** "A critical view of driver behavior models: What do we know, what should we do?" In: *Human Behavior and Traffic Safety*. 1985, pp. 485–524. URL: https://doi.org/10.1007/978-1-4613-2173-6_19.

[Mic89] **J. A. Michon:** "Explanatory pitfalls and rule-based driver models". In: *Accident Analysis & Prevention* 21.4 (1989), pp. 341–353.

[Mir19] **J. H. Mirman:** "A dynamical systems perspective on driver behavior". In: *Transportation Research Part F: Traffic Psychology and Behaviour* 63 (2019), pp. 193–203. URL: https://www.sciencedirect.com/science/article/pii/S1369847818307666.

[MK97] **D. E. Meyer and D. E. Kieras:** "A computational theory of executive cognitive processes and multiple-task performance: Part I. basic mechanisms". In: *Psychological Review* 104.1 (1997), p. 3.

[Mot+25] **T. H. Motal, K. Dargahi Nobari, and T. Bertram:** "Analysis and modeling of social interactions in automated public transport - Multisensor-based recording and psychological implications". 2025.

[NR81] **A. Newell and P. S. Rosenbloom:** "Mechanisms of skill acquisition and the law of practice". In: *Cognitive Skills and Their Acquisition*. 1981, pp. 1–55.

[Pal+18] **M. Pallebage-Gamarallage, S. Foxley, R. A. L. Menke, I. N. Huszar, M. Jenkinson, B. C. Tendler, C. Wang, S. Jbabdi, M. R. Turner, K. L. Miller, and O. Ansorge:** "Dissecting the pathobiology of altered MRI signal in amyotrophic lateral sclerosis: A post mortem whole brain sampling strategy for the integration of ultra-high-field MRI and quantitative neuropathology". In: *BMC Neuroscience* 19.1 (2018). URL: http://dx.doi.org/10.1186/s12868-018-0416-1.

[Par+00] **R. Parasuraman, T. Sheridan, and C. Wickens:** "A model for types and levels of human interaction with automation". In: *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans* 30.3 (2000), pp. 286–297.

[Pas+19] **A. Paszke, S. Gross, and et al.:** "PyTorch: An imperative style, high-performance deep learning library". In: *Advances in Neural Information Processing Systems 32*. 2019, pp. 8024–8035. URL: http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

[Pel+99] **M. Pelikan, D. E. Goldberg, and E. Cantú-Paz:** "BOA: The bayesian optimization algorithm". In: GECCO'99. 1999, pp. 525–532.

[Pet+17] **F. H. Petzschner, L. A. Weber, T. Gard, and K. E. Stephan:** "Computational psychosomatics and computational psychiatry: Toward a joint framework for differential diagnosis". In: *Biological Psychiatry* 82.6 (2017), pp. 421–430. URL: https://www.sciencedirect.com/science/article/pii/S0006322317315846.

[Pos+88] **M. I. Posner, S. E. Petersen, P. T. Fox, and M. E. Raichle:** "Localization of cognitive operations in the human brain". In: *Science* 240.4859 (1988), pp. 1627–1631. URL: https://www.science.org/doi/abs/10.1126/science.3289116.

[PR96] **J. K. Patel and C. B. Read:** *Handbook of the normal distribution*. Vol. 150. CRC Press, 1996.

[Qin+03] **Y. Qin, M.-H. Sohn, J. R. Anderson, V. A. Stenger, K. Fissell, A. Goode, and C. S. Carter:** "Predicting the practice effects on the

blood oxygenation level-dependent (BOLD) function of fMRI in a symbolic manipulation task". In: *Proceedings of the National Academy of Sciences* 100.8 (2003), pp. 4951–4956. URL: http://dx.doi.org/10.1073/pnas.0431053100.

[Qui+21] **K. S. Quigley, S. Kanoski, W. M. Grill, L. F. Barrett, and M. Tsakiris:** "Functions of interoception: From energy regulation to experience of the self". In: *Trends in Neurosciences* 44.1 (2021), pp. 29–38. URL: http://dx.doi.org/10.1016/j.tins.2020.09.008.

[Ram+20] **H. Ramsauer, B. Schäfl, J. Lehner, P. Seidl, M. Widrich, L. Gruber, M. Holzleitner, M. Pavlovic, G. K. Sandve, V. Greiff, D. P. Kreil, M. Kopp, G. Klambauer, J. Brandstetter, and S. Hochreiter:** "Hopfield Networks is All You Need". In: *CoRR* abs/2008.02217 (2020). arXiv: 2008.02217. URL: https://arxiv.org/abs/2008.02217.

[Ram+21] **E. Ramezanı-khansarı, F. Moghadas Nejad, and S. Moogeh:** "Comparing time to collision and time headway as safety criteria". In: *Pamukkale Üniversitesi Mühendislik Bilimleri Dergisi* 27.6 (2021), pp. 669–675.

[Ran94] **T. A. Ranney:** "Models of driving behavior: A review of their evolution". In: *Accident Analysis & Prevention* 26.6 (1994), pp. 733–750. URL: https://www.sciencedirect.com/science/article/pii/0001457594900515.

[Ras83] **J. Rasmussen:** "Skills, rules, and knowledge; signals, signs, and symbols, and other distinctions in human performance models". In: *IEEE Transactions on Systems, Man, and Cybernetics* 13.3 (1983), pp. 257–266.

[Rau+10] **N. Rauch, A. Kaussner, H. Krueger, S. Boverie, and F. Flemisch:** "Measures and countermeasures for impaired driver's state within highly automated driving". In: *Proceedings of the Transport Research Arena Europe* (2010).

[RB10] **M. Reimann and A. Bechara:** "The somatic marker framework as a neurological theory of decision-making: Review, conceptual comparisons, and future neuroeconomics research". In: *Journal of Economic Psychology* 31.5 (2010), pp. 767–776. URL: https://www.sciencedirect.com/science/article/pii/S0167487010000255.

[Rea90] **J. Reason:** *Human error*. Cambridge University Press, 1990.

[Rit+19] **F. E. Ritter, F. Tehranchi, and J. D. Oury:** "ACT-R: A cognitive architecture for modeling cognition". In: *WIREs Cognitive Science* 10.3 (2019), p. 1488. URL: https://wires.onlinelibrary.wiley.com/doi/abs/10.1002/wcs.1488.

[Rus80] **J. A. Russell:** "A circumplex model of affect". In: *Journal of Personality and Social Psychology* 39.6 (1980), pp. 1161–1178. URL: http://dx.doi.org/10.1037/h0077714.

[SAE18] **O. SAE International:** "Taxonomy and definitions for terms related to driving automation systems for on-road motor vehicles". In: *SAE Standard J3016_201806* (2018).

[SAL01] **D. D. SALVUCCI:** "Predicting the effects of in-car interface use on driver performance: an integrated model approach". In: *International Journal of Human-Computer Studies* 55.1 (2001), pp. 85–107. URL: https://www.sciencedirect.com/science/article/pii/S1071581901904720.

[Sal06] **D. D. Salvucci:** "Modeling driver behavior in a cognitive architecture". In: *Human Factors* 48.2 (2006), pp. 362–380.

[Sch+10] **A. Schaefer, F. Nils, X. Sanchez, and P. Philippot:** "Assessing the effectiveness of a large database of emotion-eliciting films: A new tool for emotion researchers". In: *Cognition & Emotion* 24.7 (2010), pp. 1153–1172. URL: http://dx.doi.org/10.1080/02699930903274322.

[Sch00] **N. Schwarz:** "Emotion, cognition, and decision making". In: *Cognition and Emotion* 14.4 (2000), pp. 433–440. URL: https://doi.org/10.1080/026999300402745.

[Sch10] **K. Schmidtke:** "Tower of hanoi problem". In: *The Corsini Encyclopedia of Psychology* (2010), pp. 1–2. URL: http://dx.doi.org/10.1002/9780470479216.corpsy1002.

[Sch86] **R. Schank:** *Explanation patterns: Understanding mechanically and creatively.* Psychology Press, 1986.

[See04] **M. Seeger:** "Gaussian processes for machine learning". In: *International Journal of Neural Systems* 14.02 (2004), pp. 69–106. URL: http://dx.doi.org/10.1142/s0129065704001899.

[She11] **C. S. Sherrington:** *The integrative action of the nervous system.* Yale University Press, 1911. URL: http://dx.doi.org/10.1037/13798-000.

[Šim+21] **G. Šimić, M. Tkalčić, V. Vukić, D. Mulc, E. Španić, M. Šagud, F. E. Olucha-Bordonau, M. Vukšić, and P. R. Hof:** "Understanding Emotions: Origins and Roles of the Amygdala". In: *Biomolecules* 11.6 (2021), p. 823. URL: http://doi.org/10.3390/biom11060823.

[Sme49] **R. J. Smeed:** "Some statistical aspects of road safety research". In: *Journal of the Royal Statistical Society* 112.1 (1949), pp. 1–34. URL: http://www.jstor.org/stable/2984177.

[SP98] **L. W. Swanson and G. D. Petrovich:** "What is the amygdala?" In: *Trends in Neurosciences* 21.8 (1998), pp. 323–331. URL: https://doi.org/10.1016/s0166-2236(98)01265-x.

[Sto+10] **A. Stocco, C. Lebiere, and J. R. Anderson:** "Conditional routing of information to the cortex: A model of the basal ganglia's role in cognitive

coordination." In: *Psychological Review* 117.2 (2010), pp. 541–574. URL: http://dx.doi.org/10.1037/a0019077.

[Sto+23] **A. Stocco, P. Rice, R. Thomson, B. Smith, D. Morrison, and C. Lebiere:** "An integrated computational framework for the neurobiology of memory based on the ACT-R declarative memory system". In: *Computational Brain & Behavior* 7.1 (2023), pp. 129–149. URL: http://dx.doi.org/10.1007/s42113-023-00189-y.

[Stu08] **Student:** "The Probable Error of a Mean". In: *Biometrika* 6.1 (1908), pp. 1–25. URL: http://www.jstor.org/stable/2331554 (visited on 11/06/2023).

[Suz08] **W. A. Suzuki:** "Chapter 19 Associative learning signals in the brain". In: *Essence of Memory*. 2008, pp. 305–320. URL: http://dx.doi.org/10.1016/s0079-6123(07)00019-2.

[TA18] **J. C. Tuthill and E. Azim:** "Proprioception". In: *Current Biology* 28.5 (2018), pp. 194–203.

[Tav+23] **A. Tavakoli, S. Boker, and A. Heydarian:** "Driver state modeling through latent variable state space framework in the wild". In: *IEEE Transactions on Intelligent Transportation Systems* 24.2 (2023), pp. 1879–1893. URL: https://doi.org/10.1109/TITS.2022.3221858.

[TH22] **A. Tavakoli and A. Heydarian:** "Multimodal driver state modeling through unsupervised learning". In: *Accident Analysis & Prevention* 170 (2022), p. 106640. URL: https://doi.org/10.1016/j.aap.2022.106640.

[Tom14] **S. S. Tomkins:** "Affect theory". In: *Approaches to Emotion*. 2014, pp. 163–195.

[Tom62] **S. Tomkins:** *Affect imagery consciousness: The positive affects*. Vol. 1. Springer Publishing Company, New York, N.Y., 1962.

[Tou+24] **B. Toussaint, J. Heinzle, and K. E. Stephan:** "A computationally informed distinction of interoception and exteroception". In: *Neuroscience & Biobehavioral Reviews* 159 (2024), p. 105608. URL: http://dx.doi.org/10.1016/j.neubiorev.2024.105608.

[Tra+15] **D. Tran, W. Sheng, L. Liu, and M. Liu:** "A hidden markov model based driver intention prediction system". In: *2015 IEEE International Conference on Cyber Technology in Automation, Control, and Intelligent Systems (CYBER)*. IEEE. 2015, pp. 115–120.

[Unn+15] **A. Unni, K. Ihme, H. Surm, L. Weber, A. Lüdtke, D. Nicklas, M. Jipp, and J. W. Rieger:** "Brain activity measured with fNIRS for the prediction of cognitive workload". In: *2015 6th IEEE International Conference on Cognitive Infocommunications (CogInfoCom)*. 2015, pp. 349–354. URL: http://doi.org/10.1109/CogInfoCom.2015.7390617.

[VD09] **G. Van Rossum and F. L. Drake:** *Python 3 Reference Manual*. CreateSpace, 2009.

[VH88] **C. Vlek and L. Hendrickx:** "Statistical risk versus personal control as conceptual bases for evaluating (traffic) safety". In: *International Conference on Road Safety*. 1988.

[Vin+09] **A. Vinciarelli, M. Pantic, and H. Bourlard:** "Social signal processing: Survey of an emerging domain". In: *Image and Vision Computing* 27.12 (2009), pp. 1743–1759. URL: http://dx.doi.org/10.1016/j.imavis.2008.11.007.

[Vir+20] **P. Virtanen, R. Gommers, and et al.:** "SciPy 1.0: Fundamental algorithms for scientific computing in python". In: *Nature Methods* 17 (2020), pp. 261–272.

[Web+17] **M. Weber, P. D. Maia, and J. N. Kutz:** "Estimating Memory Deterioration Rates Following Neurodegeneration and Traumatic Brain Injuries in a Hopfield Network Model". In: *Frontiers in Neuroscience* 11 (2017). URL: http://dx.doi.org/10.3389/fnins.2017.00623.

[Wes+23] **C. Westlin, J. E. Theriault, Y. Katsumi, A. Nieto-Castanon, A. Kucyi, S. F. Ruf, S. M. Brown, M. Pavel, D. Erdogmus, D. H. Brooks, et al.:** "Improving the study of brain-behavior relationships by revisiting basic assumptions". In: *Trends in Cognitive Sciences* 27.3 (2023), pp. 246–257. URL: https://doi.org/10.1016/j.tics.2022.12.015.

[WH00] **C. D. Wickens and W. S. Helton:** *Engineering psychology and human performance*. 3rd ed. Upper Saddle River, NJ : Prentice Hall, 2000.

[Wil82] **G. J. Wilde:** "The theory of risk homeostasis: Implications for safety and health". In: *Risk Analysis* 2.4 (1982), pp. 209–225.

[Wun97] **W. M. Wundt:** *Grundriss der Psychologie*. Leipzig : W. Engelmann, 1897.

[Zda14] **B. Zdaniuk:** "Ordinary Least-Squares (OLS) Model". In: *Encyclopedia of Quality of Life and Well-Being Research*. 2014, pp. 4515–4517.

[Zha+15] **S. Zhang, D. Zheng, X. Hu, and M. Yang:** "Bidirectional long short-term memory networks for relation classification". In: *Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation*. 2015, pp. 73–78.

[Zha+21] **M. Zhang, K. Ihme, U. Drewitz, and M. Jipp:** "Understanding the Multidimensional and Dynamic Nature of Facial Expressions Based on Indicators for Appraisal Components as Basis for Measuring Drivers' Fear". In: *Frontiers in Psychology* 12 (2021). URL: http://doi.org/10.3389/fpsyg.2021.622433.

[ZW03] **E. Zivot and J. Wang:** "Vector Autoregressive Models for Multivariate Time Series". In: *Modeling Financial Time Series with S-Plus®*. 2003, pp. 369–413. URL: https://doi.org/10.1007/978-0-387-21763-5_11.

Related peer-reviewed publications

K. Dargahi Nobari and T. Bertram: "Neuropsychologically inspired driver model informed by physiological dynamics within the context of automated driving". In: *Nature Human Behaviour* (in submission).

K. Dargahi Nobari and T. Bertram: "Analysis of Driver Workload on Risk Perception in Non-Safety-Critical Situations with Partial Automation". In: *2025 IEEE Conference on Cognitive and Computational Aspects of Situation Management (CogSIMA)*. 2025, pp. 56–63. URL: http://dx.doi.org/10.1109/cogsima64436.2025.11079506.

K. Dargahi Nobari and T. Bertram: "A multimodal driver monitoring benchmark dataset for driver modeling in assisted driving automation". In: *Scientific Data* 11.1 (2024), p. 327.

K. Dargahi Nobari and T. Bertram: "Influencing perceived risk and trust by changing driver workload in automated driving". In: *VDI Mechatronics*. 2024.

K. Dargahi Nobari, F. Albers, K. Bartsch, J. Braun, and T. Bertram: "Modeling driver-vehicle interaction in automated driving". In: *Engineering Research* 86.1 (2022), pp. 65–79. URL: https://doi.org/10.1007/s10010-021-00576-6.

K. Dargahi Nobari and T. Bertram: "Generalized model for driver activity recognition in automated vehicles using pressure sensor array". In: *Proceedings of the 8th International Conference on Human Interaction & Emerging Technologies (IHIET 2022): Artificial Intelligence & Future Applications*. 2022. URL: http://dx.doi.org/10.54941/ahfe1002733.

K. Dargahi Nobari, A. Hugenroth, and T. Bertram: "Position classification and in-vehicle activity detection using seat-pressure-sensor in automated driving". In: *AmE 2022 - Automotive Meets Electronics; 13. GMM-Symposium*. 2022, pp. 1–6. URL: https://ieeexplore.ieee.org/document/100259.

K. Dargahi Nobari, C. Velasquez, and T. Bertram: "Emotion induction strategies in driving simulator for validated experiments". In: *Human Systems Engineering and Design (IHSED2021) Future Trends and Applications*. 2021. URL: http://dx.doi.org/10.54941/ahfe1001156.

K. Dargahi Nobari, F. Albers, K. Bartsch, and T. Bertram: "Online feedback control for driver-vehicle interaction in automated driving". In: *Advances in Human Aspects of Transportation*. 2020, pp. 159–165. URL: https://doi.org/10.1007/978-3-030-50943-9_21.

K. Dargahi Nobari, K. Bartsch, F. Albers, and T. Bertram: "Driver state regulation via real-time neurofeedback in partially automated driving". In: *2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC)*. IEEE. 2020, pp. 1–6. URL: http://doi.org/10.1109/itsc45102.2020.9294349.

K. Dargahi Nobari, F. Albers, J. Braun, and T. Bertram: "Driver-vehicle-interaction in a control loop". In: *8th Interdisciplinary Workshop Cognitive Systems: Understanding, describing and Designing Cognitive (Technical) Systems*. 2019.

Related peer-reviewed publications with co-authorship

H. Renz, K. Dargahi Nobari, M. Faizan, and T. Bertram: "Does it Feel Safer? A Pilot Study on the Stress Levels of Humans for Varied Robot Control Strategies and Collaboration Scenarios". In: *Human Systems Engineering and Design (IHSED2024) Future Trends and Applications*. Vol. 158. IHSED 2024. 2024. URL: http://dx.doi.org/10.54941/ahfe1005525.

F. Albers, K. Dargahi Nobari, J. Braun, K. Bartsch, and T. Bertram: "Coordination of takeover maneuvers in highly automated driving considering driver availability". In: *Engineering Research* (2021).

Additional peer-reviewd publications with co-authorship

F. Tanshi, K. Dargahi Nobari, J. Wang, and D. Söffker: "Design of conditional driving automation variables to improve takeover performance". In: *IFAC-PapersOnLine* 52.8 (2019), pp. 170–175.

Open-access datasets

K. Dargahi Nobari and T. Bertram: "manD 1.0". 2023. URL: https://doi.org/10.7910/DVN/SG9TMD.

Posters

T. H. Motal, K. Dargahi Nobari, and T. Bertram: "Analysis and modeling of social interactions in automated public transport - Multisensor-based recording and psychological implications". 2025.

K. Dargahi Nobari and T. Bertram: "Understanding and predicting driver behavior: Insights into cognitive architecture and neurological patterns". 2024.

K. Dargahi Nobari and T. Bertram: "Benchmark dataset: Multimodal driver monitoring". 2023.

K. Dargahi Nobari and T. Bertram: "Driver modeling: Architecture of cognition and emotion". 2022.

K. Dargahi Nobari and T. Bertram: "Unobtrusive monitoring method for estimating the driver state using seat pressure sensors". 2021.

K. Dargahi Nobari and T. Bertram: "Evolution of driver state through interaction in automated vehicles". 2020.

K. Dargahi Nobari and T. Bertram: "Development of driver-vehicle interaction in automated driving: An optimization problem". 2019.

Supervised theses

E. Baka: "Derivation of objective metrics for human perception of high-definition matrix headlight distributions for automated driving". Master's thesis. 2025.

S. Mohamed Khan: "Early prediction of passenger exit intent to optimize door-opening and stop time". Master's thesis (on-going). 2025.

T. H. Motal: "Analysis and modeling of social interaction between passengers in an automated shuttle". Master's thesis (on-going). 2025.

M. Faizan: "Impact of varying collision avoidance strategies on human stress level in huamn-robot interaction". Master's thesis. 2024.

P. Surendra: "Robust driver heart rate monitoring system via unobtrusive sensors". Master's thesis. 2024.

A. Gaul: "Überwachung der physiologischen Daten von Fahrern über ein Drucksensor-Array". Bachelor's thesis. 2021.

G. Negoita: "Driver state detection based on seat-pressure-sensor". Master's thesis. 2021.

C. Velasquez: "Emotion detection through sensor fusion in automated driving". Master's thesis. 2021.

M. H. Slim: "Analysis of the cognitive workload of drivers during gaze-triggered takeover scenarios in a driving simulator". Master's thesis. 2020.

J. Schmitz: "Online-Klassifikation verschiedener Fahrertypen im Kontext des automatisierten Fahrens". Master's thesis. 2019.

Published media

K. Dargahi Nobari: *manD 1.0 dataset*. `https://rst.etit.tu-dortmund.de/forschung/datensaetze/mand-10/`. Accessed: 2025-08-19. 2024.

K. Dargahi Nobari: *Shuttle passenger cabin – Emulator*. `https://www.youtube.com/watch?v=gazv4YuuNyk`. Accessed: 2025-08-19. 2024.

K. Dargahi Nobari: *Driving simulator*. `https://www.youtube.com/watch?v=-B9_1fF4UYQ`. Accessed: 2025-08-19. 2023.