

01010
01010
01010

information

Special Issue Reprint

Test and Evaluation Methods for Human-Machine Interfaces of Automated Vehicles II

Edited by

Frederik Naujoks, Yannick Forster, Andreas Keinath, Nadja Schömig, Sebastian Hergeth and Katharina Wiedemann

mdpi.com/journal/information



Test and Evaluation Methods for Human-Machine Interfaces of Automated Vehicles II

Test and Evaluation Methods for Human-Machine Interfaces of Automated Vehicles II

Guest Editors

Frederik Naujoks

Yannick Forster

Andreas Keinath

Nadja Schömig

Sebastian Hergeth

Katharina Wiedemann



Basel • Beijing • Wuhan • Barcelona • Belgrade • Novi Sad • Cluj • Manchester

Guest Editors

Frederik Naujoks
BMW Group
Munich
Germany

Yannick Forster
BMW Group
Munich
Germany

Andreas Keinath
BMW Group
Munich
Germany

Nadja Schömig
Wuerzburg Institute for
Traffic Sciences
Veitshoechheim
Germany

Sebastian Hergeth
BMW Group
Munich
Germany

Katharina Wiedemann
Wuerzburg Institute for
Traffic Sciences
Veitshoechheim
Germany

Editorial Office

MDPI AG
Grosspeteranlage 5
4052 Basel, Switzerland

This is a reprint of the Special Issue, published open access by the journal *Information* (ISSN 2078-2489), freely accessible at: https://www.mdpi.com/journal/information/special_issues/Automated_Vehicles2.

For citation purposes, cite each article independently as indicated on the article page online and as indicated below:

Lastname, A.A.; Lastname, B.B. Article Title. <i>Journal Name</i> Year , Volume Number, Page Range.
--

ISBN 978-3-7258-4637-5 (Hbk)

ISBN 978-3-7258-4638-2 (PDF)

<https://doi.org/10.3390/books978-3-7258-4638-2>

© 2025 by the authors. Articles in this book are Open Access and distributed under the Creative Commons Attribution (CC BY) license. The book as a whole is distributed by MDPI under the terms and conditions of the Creative Commons Attribution-NonCommercial-NoDerivs (CC BY-NC-ND) license (<https://creativecommons.org/licenses/by-nc-nd/4.0/>).

Contents

About the Editors	vii
Denise Sogemeier, Yannick Forster, Frederik Naujoks, Josef F. Krems and Andreas Keinath Driving across Markets: An Analysis of a Human–Machine Interface in Different International Contexts Reprinted from: <i>Information</i> 2024 , 15, 349, https://doi.org/10.3390/info15060349	1
Marc Wilbrink, Stephan Cieler, Sebastian L. Weiß, Matthias Beggiato, Philip Joisten, Alexander Feierle, et al. Principles for External Human–Machine Interfaces Reprinted from: <i>Information</i> 2023 , 14, 463, https://doi.org/10.3390/info14080463	27
Naomi Y. Mbelekani and Klaus Bengler Learnability in Automated Driving (LiAD): Concepts for Applying Learnability Engineering (CALE) Based on Long-Term Learning Effects Reprinted from: <i>Information</i> 2023 , 14, 519, https://doi.org/10.3390/info14100519	44
Muhammad Zainul Abidin Kamaludin, Juffrizal Karjanto, Noryani Muhammad, Nidzamuddin Md Yusof, Muhammad Zahir Hassan, Mohamad Zairi Baharom, et al. The Correlation Between Self-Assessment and Observation in Driving Style Classification: An On-Road Case Study Reprinted from: <i>Information</i> 2025 , 14, 140, https://doi.org/10.3390/info16020140	72
Rafael Cirino Gonçalves, Tyron Louw, Yee Mun Lee, Ruth Madigan, Jonny Kuo, Mike Lenné, et al. Is Users’ Trust during Automated Driving Different When Using an Ambient Light HMI, Compared to an Auditory HMI? Reprinted from: <i>Information</i> 2025 , 14, 260, https://doi.org/10.3390/info14050260	86
Yuan-Cheng Liu, Nikol Figalová, Klaus Bengler Transparency Assessment on Level 2 Automated Vehicle HMIs Reprinted from: <i>Information</i> 2022 , 13, 489, https://doi.org/10.3390/info13100489	103
Marlene Susanne Lisa Scharfe-Scherf, Sebastian Wiese and Nele Russwinkel A Cognitive Model to Anticipate Variations of Situation Awareness and Attention for the Takeover in Highly Automated Driving Reprinted from: <i>Information</i> 2022 , 13, 418, https://doi.org/10.3390/info13090418	121
Yke Bauke Eisma, Dirk J. Eijssen and Joost C. F. de Winter What Attracts the Driver’s Eye? Attention as a Function of Task and Events Reprinted from: <i>Information</i> 2022 , 13, 333, https://doi.org/10.3390/info13070333	140
Nadja Schömig, Katharina Wiedemann, André Wiggerich and Alexandra Neukum S.A.D.E.—A Standardized, Scenario-Based Method for the Real-Time Assessment of Driver Interaction with Partially Automated Driving Systems Reprinted from: <i>Information</i> 2022 , 13, 538, https://doi.org/10.3390/info13110538	155

About the Editors

Frederik Naujoks

Frederik Naujoks earned his Diploma (2010) and PhD (2015) in Psychology from the University of Wuerzburg. He worked at the Center for Traffic Sciences (IZVW) and the Wuerzburg Institute for Traffic Sciences (WIVW) until 2017, before joining BMW. His research focuses on applied psychology in traffic, including driver distraction, usability, human-centered design, and the evaluation of assisted and automated driving systems.

Yannick Forster

Yannick Forster holds a Master's degree (2016) and PhD (2020) in Psychology. After working at WIVW, he pursued a PhD with BMW and TU Chemnitz (2017–2020). His research focuses on human interaction with automotive technologies, especially automation and driver monitoring.

Andreas Keinath

Andreas Keinath is Head of Concept Quality and Usability in the HMI department at BMW. He received his PhD in Psychology from Chemnitz University of Technology in 2003. His work focuses on cognitive and applied psychology in automotive system design.

Nadja Schömig

Nadja Schömig studied Psychology at the University of Wuerzburg (2003). She worked at IZVW (2003–2007) and became a Senior Researcher at WIVW in 2008. She received her PhD in 2009 on driver situation awareness. Her research centers on human factors in automated driving, including HMI design, evaluation methods, and driver-state assessment (e.g., fatigue, distraction).

Sebastian Hergeth

Sebastian Hergeth was born in Munich in 1986. He holds a BSc in Psychology (University of Salzburg, 2011), an MSc in Economic and Organizational Psychology (University of Munich, 2013), and a PhD from Chemnitz University of Technology (2016) on trust in automation. He has worked at BMW since 2016, focusing on human factors in automation, HMI, method development, and driver distraction.

Katharina Wiedemann

Katharina Wiedemann completed her Psychology degree at the University of Wuerzburg in 2014 and earned her PhD in 2020 on automated vehicle HMI design. She worked at WIVW and joined BMW in 2022. Her research includes HMI design/evaluation, driver distraction, and test method development for automated driving.

Article

Driving across Markets: An Analysis of a Human–Machine Interface in Different International Contexts

Denise Sogemeier ^{1,2,*}, Yannick Forster ¹, Frederik Naujoks ¹, Josef F. Krems ² and Andreas Keinath ¹

¹ BMW Group, 80937 Munich, Germany; yannick.forster@bmw.de (Y.F.); frederik.naujoks@bmw.de (F.N.); andreas.keinath@bmw.de (A.K.)

² Department of Psychology, Chemnitz University of Technology, 09111 Chemnitz, Germany; josef.krems@psychologie.tu-chemnitz.de

* Correspondence: denise.sogemeier@bmw.de; Tel.: +49-151-601-25986

Abstract: The design of automotive human–machine interfaces (HMIs) for global consumers’ needs to cater to a broad spectrum of drivers. This paper comprises benchmark studies and explores how users from international markets—Germany, China, and the United States—engage with the same automotive HMI. In real driving scenarios, $N = 301$ participants (premium vehicle owners) completed several tasks using different interaction modalities. The multi-method approach included both self-report measures to assess preference and satisfaction through well-established questionnaires and observational measures, namely experimenter ratings, to capture interaction performance. We observed a trend towards lower preference ratings in the Chinese sample. Further, interaction performance differed across the user groups, with self-reported preference not consistently aligning with observed performance. This dissociation accentuates the importance of integrating both measures in user studies. By employing benchmark data, we provide insights into varied market-based perspectives on automotive HMIs. The findings highlight the necessity for a nuanced approach to HMI design that considers diverse user preferences and interaction patterns.

Keywords: benchmarking; human–machine interaction; human–machine interface; usability; user experience

1. Introduction

Watson [1] emphasized quality, technology, and cost as crucial for surpassing global competition. To this day, the urgency to outperform competitors has further intensified. Benchmarking is a common method that allows companies to evaluate their products by comparing them to those of their direct competitors. While conventional usability studies aim at improving a product and eliminating errors in the development cycle [2], benchmarking is defined as a process of identifying, understanding, and adapting outstanding practices from organizations anywhere in the world. Benchmark studies pursue the overachieving goal of improving performance [3] and ensuring customer satisfaction [4]. The findings provide important impulses for product requirements in a product’s development cycle. Hence, benchmarking can be recognized “as a catalyst for improvement and innovation” [5] (p. 258). Furthermore, by benchmarking, researchers can identify best practices and industry standards that contribute to optimal human–system interaction. This knowledge helps in advancing the field of human factors by providing insights into effective design principles, ergonomic considerations, and user-centered approaches.

1.1. Evaluation and Development of Human–Machine Interfaces

Benchmarking can be conducted in a number of contexts, including the automotive industry. In the automotive context, it is of particular importance to evaluate human–machine interfaces (HMIs) since their design is critical in several aspects of safety, efficiency, and branding [6]. Essential components of a vehicle’s HMI are the in-vehicle information

systems (IVISs), which include functions such as navigation, media, radio, and communication, among others. These systems should be used safely while driving since interaction with them is only a secondary or tertiary task [7]. Sufficient mental resources should remain for the primary driving task [8]. Apart from safety considerations, IVISs should be easy to use and understand. Additionally, as automated driving systems (ADSs) become increasingly accessible to a large consumer population, it becomes imperative to assess the appropriateness of HMIs and to evaluate them accordingly [9]. Those systems are evaluated during the development cycle [10] and can also be compared to those of competitors in in-market evaluations. The latter refers to benchmarking. When comparing systems, different constructs can be evaluated to assess the user's attitude towards the system. Commonly investigated constructs are acceptance, trust, usability, or user experience (UX) [11,12]. Usability is defined as “[...] the extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use” [10]. UX expands the concept of usability by adding hedonic qualities such as joy and, thereby, going beyond mere pragmatic aspects [13]. Both concepts can be evaluated using standardized instruments such as rating scales and self-report questionnaires and can also be applied to evaluate ADSs [14]. Furthermore, observational measures should be included in user studies to derive a holistic picture of the in-vehicle HMI [15].

1.2. Benchmarking in Different Markets

As the automotive industry operates on a global level, benchmarking in an international context is essential. With respect to automotive HMIs, including ADSs, the design and development focus predominantly on the needs and preferences of drivers from Western markets [16]. Markets that differ from Western societies in, e.g., culture, language, driving environment, and behavior, may react differently to HMI solutions regarding their comprehension of and attitude towards the system. Designing a universal user interface (UI) that achieves comparable levels of user satisfaction is an immense challenge for original equipment manufacturers (OEMs) [17]. Regulation and legislation vary among countries, requiring market-specific solutions. The fact that market requirements differ also calls for specific solutions rather than a “one size fits all” HMI [6]. Differences in UI evaluation between countries have been acknowledged in previous research [17,18], showing that behavioral tendencies are associated with culture-influenced usability and design preferences [19]. Wang and colleagues [20], for example, found differences between Swedish and Chinese drivers regarding information requirements for advanced driving assistance systems. In complex traffic situations, the two cultural groups required different types of information. Khan and Williams [21] showed differences in expectations for HMI systems between Indian and British drivers: while usability was seen as a universal requirement to provide a satisfying product, Indian users held stronger opinions on functionality requirements such as user help. Similar results provided by Young and colleagues [22] showed differences between Australian and Chinese drivers in the comprehension of IVIS functions. Braun and colleagues [23] suggested considering cultural differences when designing effective automotive UIs, as these interfaces are perceived differently by German and Chinese users.

1.3. Research Question

Previous studies identified differences between different international user groups with prototypes in driving simulators using questionnaires. However, there is a dearth of research on how users from different markets perceive the same automotive UI integrated into a series production vehicle in a real driving context. As data derived from benchmark studies hold immense value, this paper attempts to process these datasets in a way that makes them not only accessible but also useful for research purposes. Over a span of three years, we collected diverse samples from three international markets, resulting in a substantial sample size. Subsequently, we meticulously processed and analyzed these datasets to extract greater utility and insights from the information they encapsulate. Does a global

user exist or are market-specific UI solutions in the automotive context promising? In order to gain a deeper understanding of this question and to identify potential avenues for further inquiry, we adopted an exploratory approach to target human–machine interaction with a focus on usability and UX. Besides self-report measures, the multi-method approach also included an observational measure as an indicator of interaction performance. The aim of this paper is to utilize benchmark data in order to provide preliminary insights into how user groups with distinct international backgrounds differ across a range of constructs. Based on the literature presented in Section 1.2, we assume that there are differences between user groups from three markets regarding self-report measures, i.e., satisfaction, hedonic qualities, overall UI evaluation, and interaction performance. We further propose that these differences may be attributed to different cultural standards, habits, and expectations prevalent in the different markets. Thereby, this paper aims to spark discussion about the underlying factors contributing to differences between international markets, addressing broader questions related to the root cause of such differences. By prioritizing the understanding of user preferences and interaction patterns across international markets, automotive manufacturers, designers, and researchers can develop more user-centric, globally relevant, and successful products and services. This deep user understanding can inspire new ideas and innovations that may not have been apparent from a single-market perspective. Cross-cultural insights can spark novel solutions to meet the evolving needs of automotive customers worldwide.

2. Materials and Methods

To gain further insights into whether real driving scenarios engender differences in the HMI evaluation, we used a multi-study research approach. An overview of the multi-study approach can be found in Table 1. Over the course of four years, we conducted a series of studies in three distinct markets, which were subsequently consolidated into three comprehensive studies labeled study 1, study 2, and study 3. We pursued similar procedures in each study and implemented a pre-defined set of use cases (UCs) and comparable measures. In study 2 and study 3, the vehicles tested were identical, whereas the vehicle in study 3 was equipped with an updated interface version. Consequently, comparisons were not performed across the three studies but within each study between the user groups. The primary difference in the updated interface was the design and layout, while the input modalities for operating various functions like navigation, media, comfort, and assistance functions remained unchanged. In light of the entry restrictions imposed by the government of the United States (US) in response to the COVID-19 pandemic, we were unable to include the US in study 2.

Table 1. Overview of the multi-study research approach.

Label	Market	Year
Study 1	Germany	2018
	China	2019
	US	2019
Study 2	Germany	2020
	China	2020
Study 3	Germany	2021
	China	2022
	US	2022

In each study, participants from each market interacted with the same vehicle. As benchmark studies aim at comparing products to those of competitors, we included three other reference vehicles in each market that were considered main competitors in the explicit market. Different vehicles were used as references in the different markets, as each market produces its own domestic brands. Currently, Chinese domestic brands, for

example, are not available in the US and are just at the doorstep of the European consumer market. Therefore, a strict separation between market and domestic brands within the scope of a real-world driving study seems inappropriate. Nevertheless, within the scope of the present work, we were not interested in comparing multiple vehicles but rather in comparing different user groups.

2.1. Participants

In study 1, $n = 102$ participants from Germany, China, and the US participated in the study. Study 2 included $n = 73$ participants from Germany and China. Study 3 included $n = 126$ participants from Germany, China, and the US. National subsamples were compared along several demographic variables (Table 2). The age difference was intentional since participants were recruited explicitly with respect to certain age criteria, e.g., the comparatively young average age of a Chinese new car buyer [24]. All participants were members of the public who were recruited through local field agencies and compensated for their participation. They did not receive any prior training. All participants were customers of premium vehicles whose current vehicle was not older than three years and not newer than six months. Thus, they had experience with novel digital user interfaces in general. In order to differentiate between premium-level and non-premium-level vehicles, we used brand reputation as a classification criterion. Furthermore, premium vehicles can be distinguished from non-premium vehicles by their higher price point [25].

Table 2. Demographics for study 1, study 2, and study 3.

Study	Market	<i>n</i>	Sex		Age	
			Female	Male	Mean (<i>M</i>)	Standard Deviation (<i>SD</i>)
1	Germany	30	3	27	54.0	11.2
	China	36	8	28	35.5	7.1
	US	36	12	24	39.5	8.3
	Total	102	23	89		
2	Germany	36	10	26	41.5	11.8
	China	37	9	28	35.2	7.9
	Total	73	19	54		
3	Germany	37	7	30	40.0	13.0
	China	39	9	30	35.0	7.2
	US	50	15	35	50.7	10.6
	Total	126	31	95		

2.2. Human–Machine Interface

The vehicles in the present work have typical solutions for modern interfaces. Thus, the insights generalize to a variety of modern vehicle interfaces. In each vehicle, interactions were available via tangible interfaces, i.e., remote controllers, hard keys, and steering wheel control switches, and via graphical user interfaces, i.e., an instrument cluster and a central information touchscreen. Figure 1 schematically illustrates the interior layout and highlights the relevant input modalities that will be analyzed in this work, including a touch display and remote-control element.

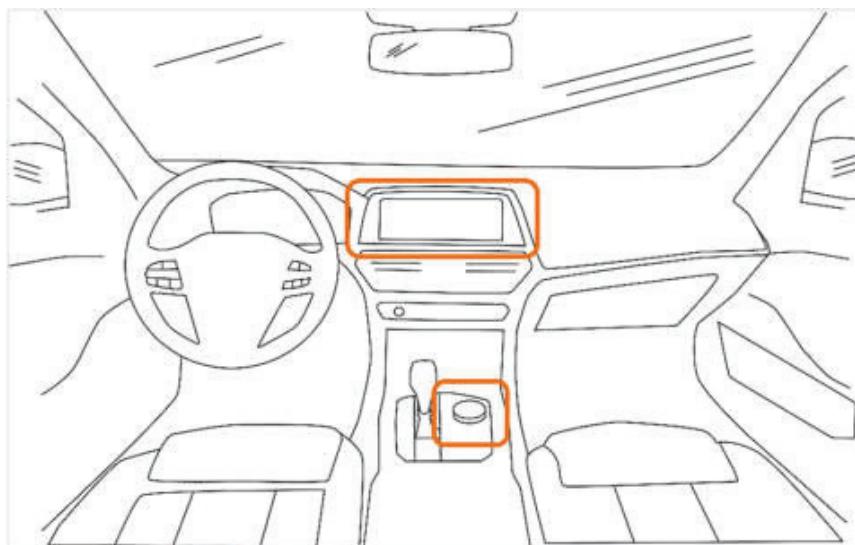


Figure 1. Layout of a modern HMI used in the studies including steering wheel control switches, hard keys, displays, and remote-control elements. Interaction modalities are highlighted in orange.

2.3. Material

2.3.1. Measurement of Satisfaction

Satisfaction is one of the three components of usability evaluation [10] and is considered an important evaluation criterion for HMIs [26]. Each participant was asked to rate the usability of the system using the system usability scale (SUS), which is considered the most widely used self-report measure of usability [27]. The SUS is a technology-agnostic tool and can therefore be used to assess the usability of a wide range of products, such as phones, IVISs [28], and automated driving [29]. The questionnaire consists of ten items, each to be answered on a five-point Likert scale from “strongly agree” to “strongly disagree”. Numerous studies have confirmed its excellent psychometric properties [30,31]. Moreover, prior research has shown the psychometric reliability and validity of the SUS in specific languages. Both the German and the Chinese translations seem to be capable of measuring the subjective usability of an infotainment system [32].

2.3.2. Measurement of Hedonic Qualities

The user experience questionnaire (UEQ) [33] was added to assess positive emotions and attitudes toward the human–machine interaction. The questionnaire covers usability aspects (efficiency, perspicuity, dependability) and UX aspects (originality, stimulation, novelty). The UEQ contains six scales with 26 items in total. The items have the form of a semantic differential, i.e., each item anchors with bipolar adjectives at each end, such as attractive versus unattractive. Prior research indicated acceptable to good reliability for all six subscales [34].

2.3.3. Measurement of Overall Evaluation

A widely used metric to rate the likelihood one would recommend a product, system, or service is the net promoter score (NPS). It is an aggregate-level measure that is derived from scores on a single survey item [35]. We included the NPS to provide a comprehensive global assessment, which takes into account multiple factors contributing to the user’s experience.

2.3.4. Measurement of Interaction Performance

There are two types of data (i.e., self-report versus behavior); therefore, we included both types in the studies. Besides reaction time and error rates as measures of speed and accuracy, observations by the experimenter can be a helpful tool to assess interaction

performance [36]. Experimenter ratings serve as one solution to the difficulty and extensive effort of collecting behavioral data [37,38]. The experimenter rated participants' behavior on a five-point rating scale for user interaction success (Table 3) right after each UC during the experiment [39].

Table 3. Experimenter rating with label and description.

Category	Value	Description
No problem	1	No problem
Hesitation	2	- Independent solution without errors - But: hesitation, very conscious operating, and full concentration
Minor errors	3	- Independent solution without or with minor errors, which were corrected confidently - But: longer pauses for reflection - Evaluation of potential operating steps
Massive errors	4	- One or multiple errors - Clearly impaired operation flow - Excessive correction of errors - No help from experimenter necessary
Help of experimenter	5	- Multiple errors - Massive errors require to restart task - Help of experimenter necessary

Prior to the experimental procedure, each category had been adjusted specifically to each UC so that the generic categories were linked to specific descriptions of behavior for each UC. During the pilot tests, the experimenters discussed reasons for potential deviations in the ratings. The training aimed to support consistent coding during the study procedure.

Considering the number of UCs (see Table 4) and available modalities, analyses of the observational data were limited to two main modalities: the touchscreen and remote-control elements. Three UCs were selected from the main categories of navigation, media, and communication (e.g., UC 1 for navigation, UC 6 for media, UC 8 for communication). In their entirety, these three UCs provide a comprehensive overview of the interaction logic and design of a vehicle's infotainment system.

Table 4. Overview of use cases for studies 1, 2, and 3.

Use Case Number	Task	Mode
1	Start navigation	P
2	Cancel navigation	P
3	View call list	P
4	Change volume/mute	D
5	Restaurant list	D
6	Skip radio station	D
7	Adjust temperature	D
8	Call contact	D
9	Play song	P
10 *	Send voice message	P

Note: all use cases were performed in each available modality; * only performed in study 3. The use cases of interest that have been evaluated in this work are presented in bold.

2.4. Study Design and Procedure

To compare the self-report and experimenter ratings between the user groups within each study, we conducted analyses with the market as a between-subjects factor. For the experimenter ratings, we further included the UCs (navigation-UC, media-UC, communication-UC) and modality (touch, remote) as within-subject factors.

The experiment began with participants signing a consent form and completing demographic questionnaires. The questionnaires used for the three cultural user groups were presented in the respective languages of the countries. After introducing the vehicle to the participants, they were asked to rate their first impression of the displays and operating modalities. Afterwards, the experimenter explained the tasks that should be performed. In each market, a corresponding native-speaking experimenter conducted the study. The studies followed the same UC set, which consisted of nine UCs (Table 4) in studies 1 and 2 and ten UCs in study 3. The experimenter sat in the passenger seat throughout the whole study and read the tasks out loud. The UCs were carried out in the given order: UCs 1, 2, 3, 9, and 10 were conducted in a parked situation, whereas the other UCs were conducted while driving. Each UC was performed with every available modality, but it was the participants' choice of which modality to use first. If participants felt unsafe conducting a UC using the touchscreen or the remote-control element while driving, they skipped the respective modality due to reasons of safety. The experimenter completed experimenter ratings after each UC during the experiment. After conducting a UC in each modality, the subsequent UC was carried out. The tasks were always started from the home screen of the IVIS. Participants completed the use cases while driving on a pre-defined route with a speed limit of 30 km/h. Driving safety had the highest priority; therefore, drivers were to refuse to complete UCs if they seemed too distracting. After conducting the UC set, the overall rating was executed in the vehicle by completing the SUS and the UEQ. Also, an in-vehicle evaluation of overall satisfaction and the voice assistant's visual, emotional, and functional value was collected. A final interview, including capturing the intention to recommend the system by the NPS, took place outside the vehicle. This resulted in a total length of about 60 min.

2.5. Statistical Procedure

Data from the questionnaires were analyzed using IBM SPSS Statistics Version 26, including, where appropriate [40], *t*-tests, univariate analysis of variance (ANOVA), univariate analysis of covariance (ANCOVA), multivariate analysis of variance (MANOVA), and multivariate analysis of covariance (MANCOVA). For ANCOVA and MANCOVA, we will report the adjusted means, which are calculated by taking into account the values of the covariates. When conducting multiple statistical tests, the risk of a type I error increases. Therefore, we employed appropriate correction techniques as a post hoc procedure to correct the family-wise error rate following the ANOVA [40,41]. For the analyses of the overall evaluation using the NPSs, the assumptions were violated. Thus, non-parametric tests, e.g., the Mann–Whitney U test and Kruskal–Wallis test, were calculated. As mentioned in Section 2.1, the age differences between the samples were intentional. However, age could still have a confounding impact on the effects. Therefore, the first step prior to the main analyses was to determine the correlation between the respective dependent variable (DV) and age. Where we found significant correlations, we chose a more complex approach and included age as a covariate in the analyses. Table 5 outlines pertinent constructs with their corresponding scales and subscales, the studies in which they have been included, and the analyses employed.

Table 5. Overview of relevant constructs, their respective data types, methods, and analyses.

Data Type	Construct	Method	Subscales	Source	Studies	Analyses Applied
Self-report measures	Satisfaction	SUS	Attractiveness Perspicuity Efficiency Dependability Stimulation Novelty	Brooke [27]	1, 2, 3	<i>t</i> -test, ANOVA
	Pragmatic and hedonic qualities	UEQ		Laugwitz et al. [33]	2, 3	MANCOVA
	Overall evaluation	NPS		Reichheld [42]	1, 2, 3	Mann–Whitney U test, Kruskal–Wallis test
Observational measures	Interaction performance	Experimenter rating	UC (Navi, Media, Communication) Modality (Touch, Remote)	Naujoks et al. [36]	1, 2, 3	Mixed within-between ANCOVA

3. Results

3.1. Results: Study 1

3.1.1. SUS

There was no significant correlation between age and the DV, $r = 0.08$, $p = 0.441$; therefore, there was no indication of a confounding influence of age, and we did not include age as a covariate. A univariate ANOVA was conducted to assess the effect of the between-subjects factor market (Germany, China, US) on usability ratings. The Levene test indicated equal variances ($p = 0.890$). The ANOVA showed no significant effect of the market on the SUS scores, $F(2,99) = 1.07$, $p = 0.346$, $\eta_p^2 = 0.02$, indicating that participants' satisfaction ratings did not differ significantly between the markets (Table 6).

Table 6. Descriptive statistics (i.e., *M* and *SD*) for the SUS scores in study 1.

Market	<i>M</i>	<i>SD</i>
Germany	72.08	14.26
China	73.89	19.50
US	78.89	21.22

3.1.2. NPS

There was no significant correlation between age and the DV, $r = 0.08$, $p = 0.489$. The NPS was only conducted in China and the US; therefore, only these two markets are included in the analysis. Since the assumptions of normal distribution and equality of variances were violated, a non-parametric approach was chosen. A Mann–Whitney U test revealed that the Chinese and American participants differed significantly in their intention to recommend the system ($z = -3.28$, $p = 0.001$, $r = 0.39$). Chinese participants ($Mdn = 8.00$) were less likely to recommend the system than American participants ($Mdn = 9.50$).

3.1.3. Experimenter Ratings

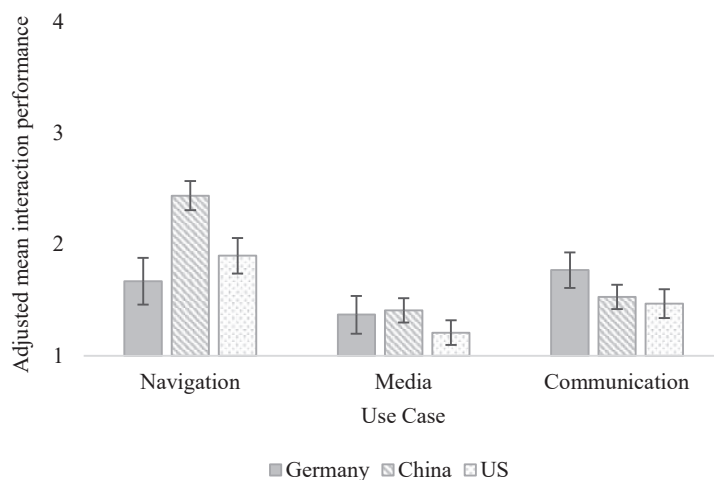
Since we found significant correlations between age and the DV (highest $r = 0.44$, $p < 0.001$), we calculated a mixed between-within ANCOVA with *market* (Germany, China, US) as the between-subjects factor and the *UC* (navigation-UC, media-UC, communication-UC) and *modality* (touch, remote control) as the within-subject factors. We further included *age* as a covariate. The assumptions of sphericity were not violated (all $\chi^2(2) < 1.85$, $p > 0.380$). The results of the $2 \times 3 \times 3$ ANCOVA can be found in Table 7.

Table 7. Inferential statistics (i.e., $df1$, $df2$, F , p -, and η_p^2 -values) for the mixed between-within ANCOVA for experimenter ratings in study 1.

Effect	$df1$	$df2$	F	p	η_p^2
Market	2	94	2.45	0.092	0.50
Age	1	94	12.30	<0.001	0.12
UC	2	188	1.27	0.283	0.01
Modality	1	94	0.92	0.340	0.01
UC $\times$ Market	4	188	2.53	0.042	0.05
UC $\times$ Age	2	188	0.71	0.495	0.007
Modality $\times$ Market	2	94	2.07	0.132	0.04
Modality $\times$ Age	1	94	0.02	0.904	0.00
UC $\times$ Modality	2	188	1.42	0.245	0.02
UC $\times$ Modality $\times$ Market	4	188	1.88	0.115	0.04
UC $\times$ Modality $\times$ Age	2	188	0.17	0.841	0.002

Note: Significant effects are in bold.

Interaction UC and Market. Pairwise comparisons using Bonferroni correction revealed a significant difference between German and Chinese participants for the navigation UC, $t(2) = 2.47$, $p = 0.044$, $d = 0.67$, indicating that the performance was significantly worse in the Chinese ($M = 2.44$, $SE = 0.17$) compared to the German sample ($M = 1.70$, $SE = 0.21$). According to Cohen (1988) [43], this is a medium-sized effect. While German participants showed the most difficulties initiating the phone call, Chinese and American participants had the most problems starting the navigation. To examine the interaction in more detail, Figure 2 presents the data in a graphical format. The examination revealed a disordinal interaction, meaning that the differences between the two markets can only be interpreted in combination with the levels of the UC factor.

**Figure 2.** Adjusted mean interaction performance scores for the three UCs split by market. Note: higher values indicate worse performance, i.e., having more trouble completing the tasks. The error bars represent the standard errors (SEs).

3.2. Results: Study 2

3.2.1. SUS

First, we checked for correlations between age and the DV. There was no significant correlation: $r = 0.21$, $p = 0.078$. We calculated a t -test for independent samples to examine if German and Chinese participants differ in satisfaction ratings. Since the Levene test was significant ($F(1,71) = 13.98$, $p < 0.001$), we could not assume equality of variance. Results of the t -test revealed a significant difference between the two markets: $t(71) = 4.08$, $p < 0.001$. German participants rated the interface significantly higher ($M = 78.68$, $SD = 11.32$) than Chinese participants ($M = 61.96$, $SD = 21.91$).

3.2.2. UEQ

A MANCOVA was calculated with *market* as the between-subjects factor. The six subscales of the UEQ, representing the construct of UX, formed the DVs. *Age* was included as a covariate because prior analyses revealed significant correlations between age and UEQ subscale *perspicuity* ($r = 0.24$, $p = 0.044$), *efficiency* ($r = 0.23$, $p = 0.048$), and *stimulation* ($r = 0.23$, $p = 0.046$). The MANCOVA revealed a significant effect of *market* on the UEQ subscales: $F(6,65) = 5.48$, $p < 0.001$, $\eta_p^2 = 0.34$, Wilk's $\Lambda = 0.66$. There was no significant effect of age: $F(6,65) = 0.48$, $p = 0.820$, $\eta_p^2 = 0.04$, Wilk's $\Lambda = 0.96$. Subsequent ANCOVAs showed significant differences between the German and Chinese samples in each subscale except for the subscale *novelty* (Table 8).

Table 8. Descriptive statistics (i.e., adjusted M , SE) for the UEQ subscales by market for study 2, and inferential statistics (i.e., $df1$, $df2$, F -, p -, and η_p^2 -values) for the univariate ANOVAs.

UEQ Scales	Market	Adjusted M	SE	$df1$	$df2$	F	p	η_p^2
Attractiveness	Germany	2.00	0.19	1	70	15.67	<0.001	0.18
	China	0.94	0.18					
Perspicuity	Germany	1.69	0.18	1	70	12.34	<0.001	0.15
	China	0.76	0.18					
Efficiency	Germany	1.62	0.19	1	70	6.68	0.012	0.09
	China	0.90	0.19					
Dependability	Germany	1.97	0.18	1	70	15.05	<0.001	0.18
	China	0.97	0.18					
Stimulation	Germany	1.67	0.19	1	70	17.68	<0.001	0.20
	China	0.54	0.18					
Novelty	Germany	0.81	0.23	1	70	2.52	0.117	0.04
	China	0.30	0.22					

Note: Significant effects are in bold.

According to Laugwitz and colleagues [33], certain subscales of the UEQ can be grouped into pragmatic qualities, i.e., *perspicuity*, *efficiency*, and *dependability*, and hedonic qualities, i.e., *stimulation* and *originality*. While pragmatic qualities describe task-related quality aspects, hedonic qualities represent non-task-related quality aspects. Furthermore, there are items that directly measure the perceived attractiveness. Figure 3 displays the distribution of pragmatic and hedonic quality aspects and perceived attractiveness in the German and Chinese samples.

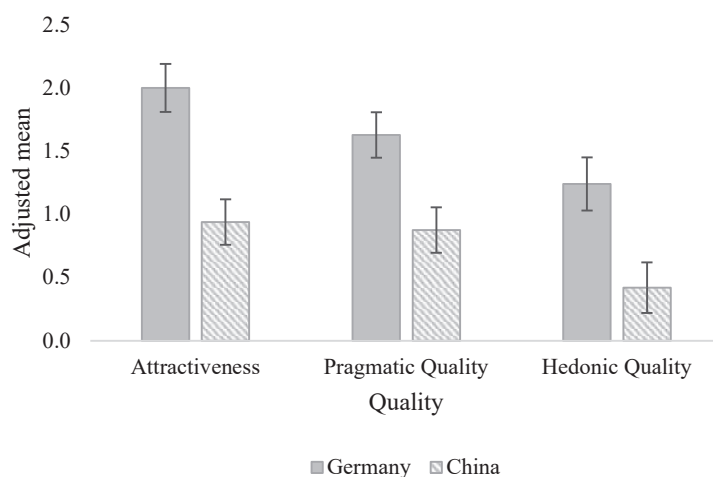


Figure 3. Distribution of pragmatic and hedonic qualities of participants from the German and Chinese markets for study 2. Note: the error bars represent the standard errors (SEs).

3.2.3. NPS

There was no significant correlation between age and the DV: $r = 0.16$, $p = 0.177$. Since the assumptions of normal distribution and equality of variances were violated, a non-parametric approach was chosen. A Mann–Whitney U test revealed that the German and Chinese participants differed significantly in their intention to recommend the system to others ($z = -5.22$, $p < 0.001$, $r = 0.61$). Chinese participants ($Mdn = 6.00$) were less likely to recommend the system than German participants ($Mdn = 8.50$).

3.2.4. Experimenter Ratings

A mixed between-within ANCOVA was calculated with *market* as a between-subjects factor and *UC* and *modality* as within-subject factors. We further included *age* as a covariate because we found a significant correlation between age and performance in communication UC using remote control ($r = 0.28$, $p = 0.017$). Therefore, the reported effects are adjusted for the influence of age. The assumption of sphericity was violated for the interaction of the factors UC and modality: $\chi^2(2) = 17.61$, $p < 0.001$; therefore, the Greenhouse–Geisser corrected values are reported for the interaction. The results of the $2 \times 2 \times 3$ ANCOVA can be found in Table 9.

Table 9. Inferential statistics (i.e., $df1$, $df2$, F -, p -, and η_p^2 -values) for the mixed between-within ANOVA for experimenter ratings in study 2.

Effect	$df1$	$df2$	F	p	η_p^2
Market	1	67	12.79	<0.001	0.16
Age	1	67	2.30	0.134	0.03
UC	2	134	4.60	0.012	0.06
Modality	1	67	1.47	0.230	0.02
UC $\times$ Market	2	134	9.87	<0.001	0.13
UC $\times$ Age	2	134	2.62	0.077	0.04
Modality $\times$ Market	1	67	6.33	0.014	0.09
Modality $\times$ Age	1	67	0.01	0.916	<0.01
UC $\times$ Modality	1.62	108.57	2.47	0.100	0.04
UC $\times$ Modality $\times$ Market	2	134	4.58	0.012	0.06
UC $\times$ Modality $\times$ Age	2	134	0.64	0.529	0.01

Note: Significant effects are in bold.

Market. Experimenter ratings were significantly higher in the Chinese sample ($M = 1.80$, $SE = 0.08$) than in the German sample ($M = 1.40$, $SE = 0.08$), indicating that performance was worse for Chinese participants.

Use Case. Pairwise comparisons using Bonferroni–Holm correction showed significant differences between the navigation UC and the media UC ($t(298) = 7.58$, $p < 0.001$, $d = 0.89$) and between the navigation UC and the communication UC ($t(299) = 6.75$, $p < 0.001$, $d = 0.79$). Experimenter ratings were significantly higher for the navigation UC ($M = 2.18$, $SE = 0.10$) than for the media UC ($M = 1.26$, $SE = 0.07$) and communication UC ($M = 1.36$, $SE = 0.08$). This result indicates that performance was the worst in the navigation UC across the two countries.

Interaction: UC and Market. Results revealed a significant interaction between the UC and market, which indicates that the effect of the UC is not equal for the two samples. Across both markets, experimenter ratings were highest for the navigation UC. However, in the Chinese sample, the difference in performance between the navigation UC and the media and communication UC was greater than the difference in the German sample. Detailed descriptive data on experimenter ratings and a figure displaying the interaction are provided in Appendix A (Table A1, Figure A1). Both main effects can be interpreted globally, as it was an ordinal interaction.

Interaction: Modality and Market. Results showed a significant ordinal interaction between modality and market. Across all three UCs, participants' performance was worse when using remote control to operate the system than it was when using touch. However,

we could not observe a main effect of modality since there was no significant difference between the two markets in the touch modality: $t(433) = 1.27, p = 0.200$. Only when using remote control to operate the system did the German and Chinese participants differ significantly in their performance: $t(433) = 3.61, p < 0.001$. Detailed descriptive data and a figure are provided in Appendix A (Table A1, Figure A1).

Interaction: UC, Modality, and Market. Inferential analyses revealed a significant three-way interaction. To further examine the interaction, we conducted pairwise comparisons using Bonferroni correction. Figure 4 displays the ratings split by modality to promote a better understanding of the higher-order interaction. The findings indicate that the $UC \times market$ interaction differs for one level of the factor *modality* to another level of the factor: Chinese participants had more difficulty starting the navigation compared to German participants, but solely when using remote control ($t(71) = 4.42, p < 0.001$) and not using touch control ($t(71) = 1.55, p = 0.130$). There were no significant differences between the German and Chinese participants for the other two UCs, i.e., media and communication, neither when using touch ($t(71) = 0.31, p = 0.733$) nor when using remote control ($t(71) = 0.30, p = 0.760$). For that reason, there was no main effect of *modality*. The modality exclusively impacted performance depending on the UC, resulting in a three-way interaction. Further, the $modality \times market$ interaction differed for the levels of factor *UC*: there were significant differences between the German and Chinese samples for operating the system using remote control compared to touch, but only for the navigation UC.

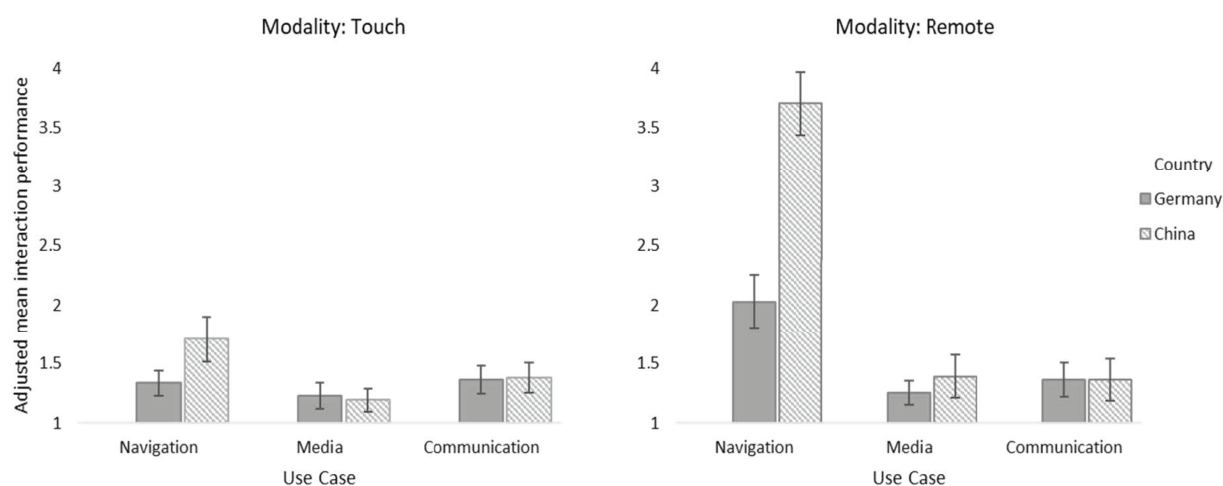


Figure 4. Adjusted mean interaction performance scores of German and Chinese participants for the three UCs split by modality in study 2. Note: higher values indicate worse performance, i.e., having more trouble completing the tasks. The error bars represent the standard errors (SEs).

3.3. Results: Study 3

3.3.1. SUS

There was no significant correlation between age and the DV: $r = 0.04, p = 0.663$. A univariate ANOVA was conducted to assess the effect of the between-subjects factor *market* (Germany, China, US) on usability ratings. The Levene test indicated unequal variances ($p = 0.005$). The ANOVA revealed a significant effect of the market on the SUS scores: $F(2,123) = 4.25, p = 0.016, \eta_p^2 = 0.07$. Post hoc t -tests using Bonferroni–Holm correction showed a significant difference between the German and Chinese participants ($t(62.44) = 2.97, p = 0.004, d = 0.65$), with participants from the German sample giving significantly higher scores than those from the Chinese sample. According to Cohen (1988) [43], this is a medium-sized effect. After correction, no additional statistically significant effects were observed. Table 10 shows the descriptive data.

Table 10. Descriptive statistics (i.e., *M* and *SD*) for the SUS in study 3.

Market	<i>M</i>	<i>SD</i>
Germany	79.73	12.77
China	67.76	21.48
US	71.65	18.94

3.3.2. UEQ

A MANCOVA was calculated with *market* as the between-subjects factor. The six subscales of the UEQ, representing the construct of UX, formed the DVs. *Age* was included as a covariate because prior analyses revealed significant correlations between age and the UEQ subscales *attractiveness* ($r = 0.22$, $p = 0.015$), *efficiency* ($r = 0.22$, $p = 0.015$), *dependability* ($r = 0.22$, $p = 0.015$), *stimulation* ($r = 0.26$, $p = 0.002$), and *novelty* ($r = 0.28$, $p = 0.002$). One participant was excluded because they gave the same answer for each item, resulting in a total of $N = 125$ completed questionnaires. There was no significant effect of *age* ($F(6,111) = 1.67$, $p = 0.135$, $\eta_p^2 = 0.08$, Wilk's $\Lambda = 0.92$), and no significant effect of *market* on the UEQ subscales ($F(12,222) = 1.72$, $p = 0.064$, $\eta_p^2 = 0.09$, Wilk's $\Lambda = 0.84$). The descriptive data, including adjusted means and *SE*, can be found in Table 11.

Table 11. Descriptive data (i.e., adjusted *M* and *SE*) for the six UEQ subscales grouped by market.

UEQ Subscale	Market	Adjusted <i>M</i>	<i>SE</i>
Attractiveness	Germany	2.02	0.16
	China	1.34	0.17
	US	1.74	0.17
Perspicuity	Germany	1.71	0.18
	China	1.34	0.19
	US	1.09	0.19
Efficiency	Germany	1.70	0.18
	China	1.22	0.18
	US	1.45	0.18
Dependability	Germany	1.82	0.17
	China	1.31	0.17
	US	1.40	0.17
Stimulation	Germany	1.67	0.17
	China	1.20	0.17
	US	1.56	0.17
Novelty	Germany	1.52	0.18
	China	1.04	0.18
	US	1.19	0.18

Figure 5 displays the distribution of pragmatic and hedonic quality aspects and perceived attractiveness in the German, Chinese, and American samples.

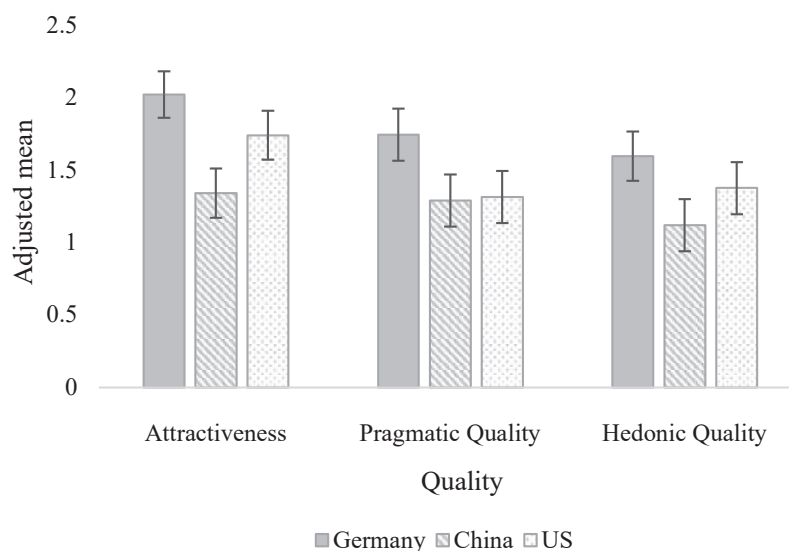


Figure 5. Distribution of pragmatic and hedonic qualities of participants from Germany, China, and the US for study 3. Note: the error bars represent the standard errors (SEs).

3.3.3. NPS

There was no significant correlation between age and the DV: $r = 0.09$, $p = 0.318$. Both the assumption of normal distribution and equality of variances were violated; hence, a non-parametric procedure was chosen. A Kruskal–Wallis test revealed a significant difference in the intention to recommend the system between the three markets ($H(2) = 12.51$, $p = 0.002$). A post hoc test, the Dunn–Bonferroni test, showed that the probability of Chinese participants ($Mdn = 6.00$) recommending the system was significantly lower compared to German participants ($(Mdn = 8.00)$, $z = 2.87$, $p = 0.012$, $d = 0.66$) and American participants ($(Mdn = 8.00)$, $z = -3.26$, $p = 0.003$, $d = 0.96$).

3.3.4. Experimenter Ratings

Touch. To compare experimenter ratings within the touch modality, a mixed between-within ANCOVA was calculated with *market* (Germany, China, US) as a between-subjects factor and *UC* as a within-subject factor. We included *age* as a covariate because prior analyses revealed significant correlations between age and the DV for the navigation UC ($r = 0.33$, $p < 0.001$) and the communication UC ($r = 0.35$, $p < 0.001$). The assumption of sphericity was violated for *UC* ($\chi^2(2) = 27.65$, $p < 0.001$); therefore, Greenhouse–Geisser corrected values are reported for this effect. The results of the 3×3 ANCOVA are displayed in Table 12.

Table 12. Inferential statistics (i.e., $df1$, $df2$, F -, p -, and η_p^2 -values) for the mixed between-within ANCOVA for experimenter ratings in study 3 (touch modality).

Effect	$df1$	$df2$	F	p	η_p^2
Market	2	115	11.10	<0.001	0.16
Age	1	115	22.47	<0.001	0.16
UC	1.65	189.25	5.48	0.005	0.05
UC $\times$ Market	4	230	7.08	<0.001	0.11
UC $\times$ Age	2	230	10.76	<0.001	0.09

Note: Significant effects are in bold.

Market. Pairwise comparisons using Bonferroni correction revealed a significant difference in interaction performance between German and Chinese participants ($t(73) = 3.9$, $p < 0.001$, $d = 0.90$) and between German and American participants ($t(79) = 3.7$, $p < 0.001$, $d = 0.83$). The findings indicate that German participants had more difficulty operating

the system using touch across all three UCs ($M = 1.60$, $SE = 0.07$) compared to Chinese ($M = 1.21$, $SE = 0.07$) and American participants ($M = 1.23$, $SE = 0.07$).

Use Case. As pairwise comparisons using Bonferroni correction showed, there was a significant difference in performance between the navigation UC and media UC ($t(236) = 5.44$, $p < 0.001$, $d = 0.71$) and the navigation UC and communication UC ($t(236) = 3.18$, $p = 0.006$, $d = 0.41$). The participants had significantly higher scores in the navigation UC ($M = 1.62$, $SE = 0.09$) than in the communication ($M = 1.14$, $SE = 0.04$) and media UCs ($M = 1.27$, $SE = 0.06$), indicating the worst performance was in the navigation UC.

Interaction: UC $\times$ Market. While German ($M = 2.27$, $SE = 0.16$) and Chinese participants ($M = 1.40$, $SE = 0.16$) struggled the most with starting the navigation, participants from the US had the most trouble with calling a contact ($M = 1.36$, $SE = 0.11$). A further examination of this interaction revealed a disordinal interaction, indicating that neither of these main effects can be interpreted globally. Differences between the three countries can only be interpreted in combination with the three UCs and vice versa. Detailed descriptive data are provided in Appendix B (Table A2, Figure A2).

Remote control. Because there were no experimenter ratings for the remote-control modality for the German sample in study 3, only the ratings from the Chinese and American samples were compared in the analysis. We found significant correlations between age and the DV (all $r > 0.22$, all $p < 0.042$) and included age as a covariate. The assumption of sphericity for the 2×3 ANCOVA with market as a between-subjects factor, UC as a within-subject factor, and age as a covariate was violated for UC: $\chi^2(2) = 46.01$, $p < 0.001$. Therefore, the Greenhouse–Geisser corrected values are reported. The ANCOVA revealed that only the covariate had a significant effect on interaction performance: $F(1,79) = 5.82$, $p = 0.018$, $\eta_p^2 = 0.07$. The remaining inferential statistics can be found in Appendix C (Table A3).

3.4. Overview of Satisfaction and Interaction Performance for German and Chinese Users across the Three Studies

In the following section, we will visualize the data from the three studies. No statistical analysis or inclusion of temporal factors (study 1, 2, 3) is conducted, as this is not a true between-design and different participants were involved in each year, with a different interface tested in study 3. The visualization serves the purpose of comparing self-report and observational measures. We illustrate satisfaction ratings and interaction performance over the course of the three studies for Germany and China (Figures 6 and 7). The figures do not include data from the US sample since study 2 was conducted in Germany and China only.

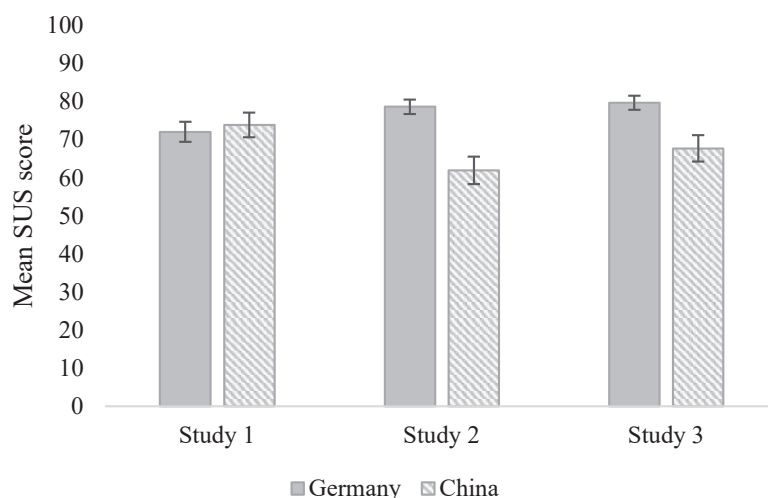


Figure 6. Satisfaction ratings over the course of time for German and Chinese users. Note: higher ratings indicate higher satisfaction ratings. The error bars represent the standard errors (SEs).

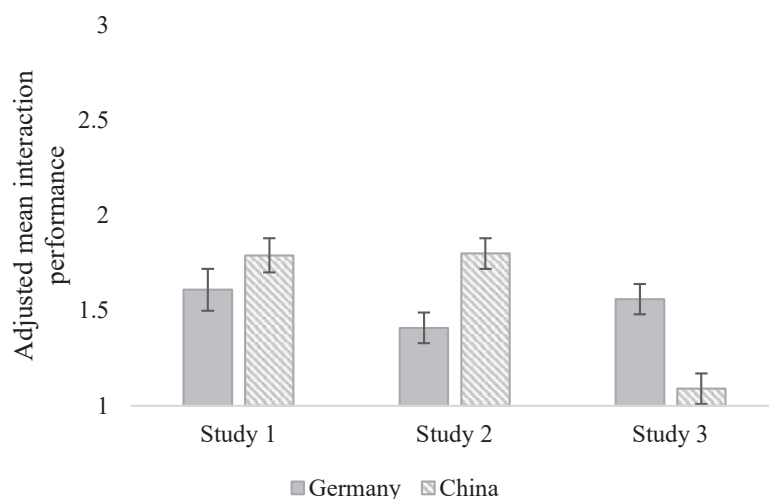


Figure 7. Interaction performance over the course of time for German and Chinese users. Note: higher values indicate worse performance, i.e., having more trouble completing the tasks. The error bars represent the standard errors (SEs).

4. General Discussion

This work outlined the findings of a series of international studies that investigated how users from different markets interact with the same in-vehicle infotainment system. In real driving scenarios, $N = 301$ participants from Germany, China, and the US completed several tasks using different interaction modalities and rated the system after the interaction. We identified differences in interaction performance, satisfaction, hedonic qualities, and overall ratings between the user groups. The following paragraphs discuss the present findings, considering study limitations, and point towards future research opportunities for benchmark studies in the automotive context.

4.1. Differences in Satisfaction

The results of the work at hand suggest the existence of differences between different user groups in satisfaction ratings. German participants reported the highest satisfaction with the interface, while the ratings were lowest in China (Figure 6). German users favored the overall usability and satisfaction of the interface compared to Chinese users. In terms of descriptive categorization, the scale provides a framework for identifying areas in need of action. The scores obtained from the questionnaires can be classified into different categories, ranging from the “worst imaginable” to the “best imaginable”. In our study, the scores ranged from “okay” to “good”. It is noteworthy that only the ratings of Chinese participants fell into the category of “okay”, which, according to the scale, implies a need for action and improvement. A German OEM developed the vehicle that has been evaluated in the studies. As Lindgren and colleagues [16] pointed out, HMI design and development often focus primarily on the needs and preferences of drivers from the Western market. Needs and expectations, in turn, can impact usability evaluations [44]. Mehler and colleagues [45] designed two interfaces: one adapted to the preferences of a Chinese sample and the other one adapted to the preferences of a German sample. The authors found that Chinese participants preferred the Chinese HMI over the German HMI, and vice versa. The interface used in this study was rather adapted to the needs of the Western market since it was developed by a German OEM and therefore received higher satisfaction from participants living and driving in the Western world. In a study by Khan and Williams [19], the authors compared two user groups from India and the United Kingdom and found some discrimination between the groups regarding usability factors, including satisfaction. They found that each group had a preferred approach to using information for problem-solving, which was determined by their respective cultural backgrounds. This determination subsequently impacts the expected visual presentation of information in a

vehicle HMI system. When there is a mismatch between the user's preferred approach and the visual information display, users tend to express dissatisfaction. Assuming there was a mismatch between the expectations of Chinese users and the displayed information in the studies, this mismatch could account for the lower satisfaction ratings. Through this, the necessity for an adapted market-specific HMI solution for the Eastern market becomes striking.

The findings of our multi-study approach revealed differences in satisfaction between German and Chinese users. However, when comparing with US users, the observed differences remained primarily descriptive in nature. A potential explanation could be found in models that address culture and cultural differences. Cultural frameworks aim to describe and define cross-cultural differences. Hofstede, for example, defines culture as the “collective programming of the mind that distinguishes the members of one group or category of people from others” [46] (p. 3). He describes culture along six dimensions and identifies differences between Germany, China, and the US [47]. Another cultural model by Trompenaars and Hampden-Turner [48] classifies culture into seven dimensions. The authors assume that differences between Germany, China, and the US occur in several dimensions, such as the perception of time. Nonetheless, our data did not reveal significant results for comparing satisfaction ratings with US users. This might be due to the methodology of this study. It is possible that the effects remained of a descriptive nature due to the given sample size. The effects are small in terms of the conventions introduced by Cohen [43]. To detect those small effects, a larger sample size would have been necessary. Conversely, this also underscores the practical significance of the effects that reached statistical significance in the between-subjects design with approximately 36 observations per cell (see Table 2). While the effect sizes were small, these effects emerged even with our relatively modest sample size. We selected medium-sized samples per market with the aim of detecting medium-sized effects. The fact that the apparently small effects were already significant given the present sample size can be considered meaningful.

There were no significant differences in satisfaction ratings in study 1 (Table 6). Each user group experienced a satisfactory interaction with the system. In the following study, one year later, differences occurred between German and Chinese users. Chang and colleagues [49] demonstrated that product preferences can be influenced by preconceived judgments due to the user's evaluative frame of reference. Those sequential effects might occur in the context of automotive benchmarking as well. Fundamental changes with switches to digital UI are developing rapidly. While the same UI was evaluated in study 1 and study 2, different reference vehicles were incorporated in these two studies. Overall, Chinese consumers showed that they appreciated digital innovations more than German consumers. They may have anticipated novel features, but the current UI failed to meet their expectations, resulting in diminished satisfaction ratings.

4.2. Differences in Hedonic Qualities

By means of a literature review, Souza and Bernardes [50] showed that cultural differences affect UX in several domains, such as product development, systems, or games. Our results of the UEQ may extend this insight to the automotive context: the user groups from Germany and China differed in their ratings of the UI's pragmatic and hedonic quality aspects in study 2 (Table 8). For each of the six UEQ subscales, the scores of the Chinese participants were below average, while the scores of the German participants were considered good or even excellent [51], with one exception: both user groups gave scores below average in the subscale *novelty*. This might indicate that both groups considered the interface to be rather conservative and conventional rather than innovative and inventive. In study 3 (Table 11), the differences were not significant. However, the descriptive data suggest a pattern of lower scores from Chinese users compared to the other two user groups. It is striking that, overall, the participants from China gave lower ratings than the participants from Germany and the US (Figures 3 and 5), which might lead to the assumption of different response tendencies in the Chinese sample. Chinese participants seemed to

be less euphoric when interacting with the system. One explanation might be again found in a cultural model: the sixth dimension, *indulgence vs. restraint*, of Hofstede's model [46] describes how members of a cultural group enjoy or suppress the gratification of natural human desires related to enjoying life and having fun. Sogemeier and colleagues [52] linked these cultural dimensions to usability criteria to gain a deeper comprehension of why the same products are rated differently in different markets. Considering cultural differences in this dimension to explain our results seems reasonable, as indulgence tends to prevail in Western Europe while restraint prevails in Asia. Less enthusiastic ratings in the Chinese sample might be rooted in more pessimistic attitudes that characterize restrained societies [46]. Indulgent cultures obtain more optimistic attitudes and express feelings of happiness and enjoyment more freely. This might account for the generally low ratings in the Chinese sample compared to the more positive ratings in the German and US samples.

Another dimension established by Hofstede [53], *uncertainty avoidance (UA)*, describes the extent to which members of a cultural group feel threatened by ambiguous situations. China has a low score on UA, while Germany rates this dimension highly. In weak UA cultures, people strive for novelty and are excited to try new things [54]. Chinese car manufacturers implement new features in their HMI, such as bigger screens, the integration of WeChat, or more exciting voice assistants. If new features are not being perceived as exciting or novel, hedonic qualities of the interaction might be missing.

4.3. Differences in Overall Ratings

The probability of recommending the vehicle to another person yielded significant differences between the user groups in each study. The results show that Chinese users were less likely to recommend the system to others than German and US users. These results corroborate the findings outlined above for UX measures in the sense that Chinese customers are somewhat more skeptical and less likely to be enthusiastic about vehicle technology.

4.4. Differences in Interaction Performance

Experimenter ratings were included as a measure of behavioral data to quantify interaction performance. The participants' age differences might have accounted for differences in interaction performance since age affects, among others, reaction time [55], computer-based task performance [56], and the usage of in-car information systems while driving [57]. Lerner and colleagues [58] further added that older drivers are less likely to engage in non-driving related tasks while driving compared to their younger or middle-aged counterparts. We therefore considered age's confounding influence on interaction performance in our analyses. The findings show that age should not be neglected as it can also explain observed differences to some extent (see Table 7, Table 12, and Table A3). Yet, these age effects may not be due to age per se, i.e., chronological age, but rather to age-related processes, such as a lack of experience with digital systems [59]. Hence, forthcoming research should inquire directly about the participants' familiarity and experience with digital systems. But age is not the only influencing factor. When eliminating the effect of age, participants from different countries still differed in their performance. Thus, the observed differences in performance go beyond age differences.

The results indicate that, across the three user groups, the navigation UC was the most difficult one to perform (in study 1, see Figure 2; in study 2, see Figure 4; in study 3, see Section 3.3.4). Experimenter ratings were highest for the navigation UC in each study, followed by the communication UC and the media UC. This finding supports considerations regarding the study design. The intent of the study was to implement a wide spectrum of UCs, with some being easier to perform than others. Although participants consistently experienced the greatest difficulty in the navigation UC, the interaction effects, as demonstrated in study 1 (refer to Table 7), study 2 (refer to Table 9), and study 3 (refer to Table 12), indicate that different user groups perceive distinct UCs to possess various

degrees of ease or difficulty. For instance, compared to German users, Chinese users perceived navigation as significantly more challenging than the other UCs (Figure 4).

With the results at hand, we cannot definitively conclude whether one operating method led to better or worse performance. In study 2, as illustrated in Figure 4, the interaction between market and modality showed that Chinese participants encountered more difficulty in completing tasks using the remote-control element compared to the touchscreen. The remaining results from study 1 and study 3 did not exhibit a distinct pattern.

The differences in experimenter ratings between the three user groups were ambiguous. As shown in Figures 6 and 7, satisfaction ratings assessed through the SUS did not consistently reflect measures of interaction performance. Only in study 2 did we observe a fit between self-report and observational measures, i.e., German participants showed better performance and rated the system higher. In study 3, however, even though Chinese participants were better at performing the UCs compared to German participants, they rated the interaction as significantly less satisfactory. Thus, the data at hand did not point towards a *reliable* predictive value of preference for performance. Roberts and colleagues [60] acknowledged a dissociation between preference and performance in a study that investigated the usability of metro maps: reaction times and error rates did not match self-reported satisfaction. Knapp [61] found that the performance of a Chinese user group was negatively affected if the system was based on a different user group's mental model, but even though the performance was worse in one experimental condition, participants did not perceive this condition as less attractive. Due to this gap, the author concludes that it is necessary to include a wide range of measures. In a literature review, Pettersson and colleagues [15] also concluded that the use of triangulation, i.e., including several methods in UX evaluation, is beneficial. Law and colleagues [62] stated that including quantitative measures while excluding qualitative measures and vice versa may lead to wrong implications. In our case, preference measures did not predict performance, and vice versa. Thus, collecting both self-report and observational data is essential to drawing a holistic picture of human-machine interactions. In addition, instead of relying on a single observational measure, it might be beneficial for researchers to include a broader array of methods, such as error rates, time on task, and reaction times. Likewise, Forster and colleagues [63] recommend a multi-method approach to provide superior insights into the quality of an HMI. Furthermore, performance indicators were aggregated over several UCs to generate a single measure for interaction performance. The data suggest that cutting down the number of variables for the sake of parsimony seems inappropriate due to an essential loss of information. An analysis of interaction on a more granular level, i.e., analyzing each UC individually, may provide more detailed information about the human-machine interaction to reveal further potential for the development of HMI. The results emphasize the need to include both self-reported preference measures and observational performance measures.

4.5. Limitations and Future Research

The present work comes with a few limitations. The studies were conducted over a period of approximately four years, and different interviewers guided the studies, which could have influenced objectivity. However, the studies followed a strict manual, and the interviewers received comprehensive training in advance to enable comparability. The questionnaires were given in the participants' respective languages, which may have impacted the results. Another limitation is that participants interacted with different reference vehicles in each study. The interaction with a previous product can foster sequence effects [49], which means that the evaluation of an object relates to what a participant has seen and experienced before. Those interaction effects between stimuli that are presented in successive order might occur in the context of automotive benchmarking, too. However, to avoid sequence effects in the present studies, the sequence of vehicles was permuted per participant [64]. Moreover, different reference vehicles in different markets were intended

for the sake of external validity. Today's domestic Chinese vehicle brands are not available in the US and are just at the doorstep of entering the European consumer market. We chose a real-world driving study and ensured an externally valid setting by respecting the fact that the availability of Eastern domestic brands in the Western market will still require time. In each market, we compared the newest HMI by a German OEM to the newest HMI by local competitors. This way, we provided a comparable framework by considering the technological progress and degree of novelty within each market. The majority of participants in our sample were male, which limits the generalizability of our findings. However, the participants were recruited with specific criteria regarding sex and age to represent a targeted customer group. Future research might include a more diverse sample to test the generalizability of the findings to a broader population.

The results of the work at hand highlight the potential of benchmark studies that enable the comparison of products such as IVISs in different international contexts. Future benchmark studies should also include ADSs to identify differences between user groups and to derive implications for future research and automotive HMI design.

In conclusion, from the results, a lot of valuable information can be summarized. There are differences between the three user groups, both in self-reporting and in behavioral measures. The findings allow initial insights into the perception of automotive UI from different international perspectives. The next consequential step is to provide more finely grained information on individual aspects, such as differential successes and failures during the interaction, to determine the underlying root causes of these differences. One potential reason could be grounded in cultural differences between user groups in international markets. Since culture impacts perception [65–67] and perception in turn affects evaluations, culture should be considered an influential factor in the perception of an automotive interface [6]. Thus, culture seems to be a feasible factor to explain differences in human–machine interactions. Future research should be conducted in diverse cultural contexts that differ in most of Hofstede's six cultural dimensions to ensure that they represent different sets of values and, hence, that they represent different cultures. Ensuring that the user groups are indeed representative of different cultures and thus attributing the observed differences to cultural variances exceeds the scope of the current findings. However, this is also not the aim of the benchmark data presented.

Besides choosing countries for future studies, researchers must define which components of the HMI will be further analyzed. IVISs include navigation, radio, or communication, among others. Mehler and colleagues [45], for example, focused on differences between German and Chinese users in the design of the navigation map. Lindgren and colleagues [16] identified the needs and requirements of the design of advanced driving assistance systems for Swedish and Chinese drivers, and Young and colleagues [22] identified different preferences for IVIS control types and labels (i.e., fan speed control) for Australian and Chinese drivers. Another HMI element that should be addressed in cross-cultural studies is the voice user interface (VUI). Gong and colleagues [54] already stated that Chinese customers would be willing to pay more for certain functions such as a physical voice assistant. It is yet unclear and needs to be investigated whether customers from Western markets desire those VUI features, too. Furthermore, instead of conducting another field study, future studies should instead comprise driving simulators. To better understand the differences between countries in UI evaluation, a more standardized setting should be chosen to control certain variables. Different factors can be included in a between-within study design to analyze specific impacts on the dependent measures, i.e., *market* as a between-subjects factor and, in the case of VUI, *visualization* as a within-subject factor. As outlined above (see Section 4.4), the results suggest that a triangulation of methods is necessary and that it is important to include both self-report as well as observational measures for a comprehensive understanding of human–machine interactions.

5. Conclusions

In conclusion, the results indicate that differences in the evaluation of an automotive UI exist between different user groups since the ratings differ both in self-report and behavioral measures. The preferences of specific user groups shall therefore be taken into consideration when designing interfaces for different international markets. For global automotive industries, the results highlight the challenge of respecting user expectations and needs in diversified markets while still benefiting from global processes. The results suggest that market-specific adaptations of HMI could be promising to enhance preference and interaction performance.

Author Contributions: Conceptualization, D.S., Y.F., F.N. and J.F.K.; methodology, D.S., Y.F., F.N. and J.F.K.; formal analysis, D.S.; data curation, D.S.; writing—original draft preparation, D.S.; writing—review and editing, D.S., Y.F., F.N. and J.F.K.; visualization, D.S.; supervision, J.F.K. and A.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: This study did not need ethical approval.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The data presented in this study are not available due to privacy reasons.

Conflicts of Interest: Authors Denise Sogemeier, Yannick Forster, Frederik Naujoks and Andreas Keinath were employed by the company BMW Group. All authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Appendix A

Table A1. Descriptive statistics (i.e., adjusted M , SE) for the interaction performance scores in study 2 for (a) the three UCs and (b) for the two modalities split by market.

(a)				(b)			
Market	Use Case	Adjusted M	SE	Market	Modality	Adjusted M	SE
Germany	Navigation	1.67	0.14	Germany	Touch	1.28	0.08
	Media	1.26	0.10		Remote	1.53	0.13
	Communication	1.28	0.12	China	Touch	1.42	0.08
China	Navigation	2.68	0.14		Remote	2.18	0.13
	Media	1.27	0.10				
	Communication	1.45	0.12				

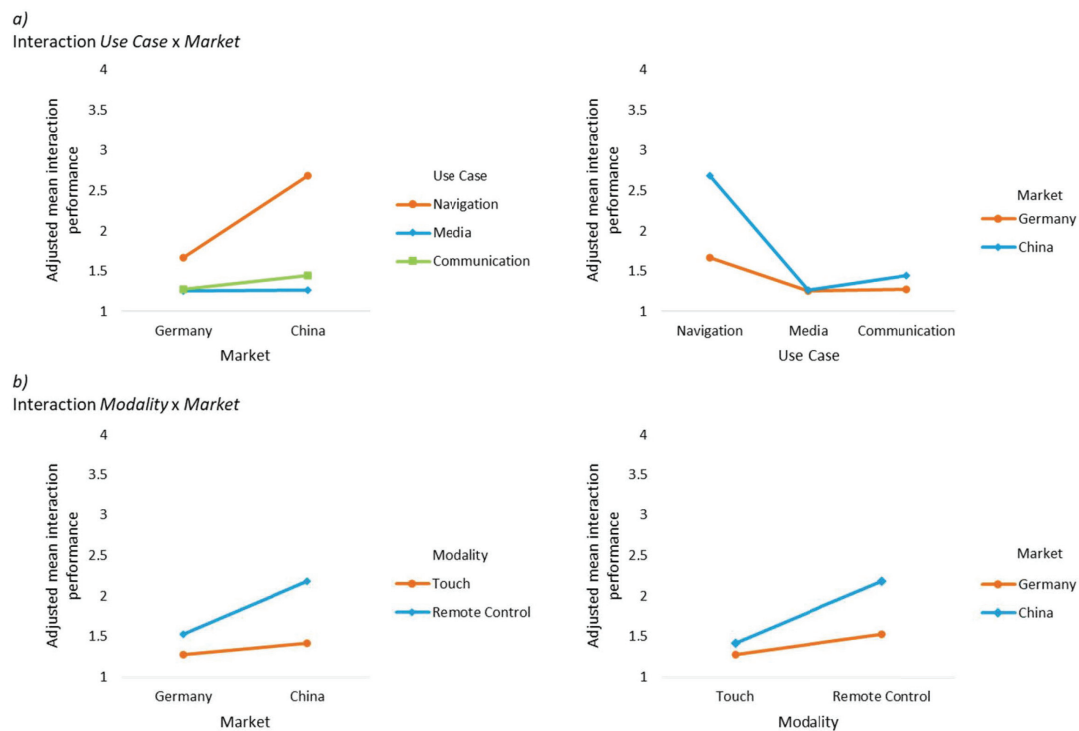


Figure A1. Adjusted mean interaction performance scores to visualize the interaction effect between (a) UC and market and (b) modality and market for study 2.

Appendix B

Table A2. Descriptive statistics (i.e., adjusted M , SE) for the interaction performance scores in study 3 for the three UCs split by market.

Market	Use Case	Adjusted M	SE
Germany	Navigation	2.27	0.16
	Media	1.23	0.07
	Communication	1.28	0.11
China	Navigation	1.40	0.16
	Media	1.05	0.08
	Communication	1.18	0.12
US	Navigation	1.20	0.16
	Media	1.13	0.07
	Communication	1.36	0.11

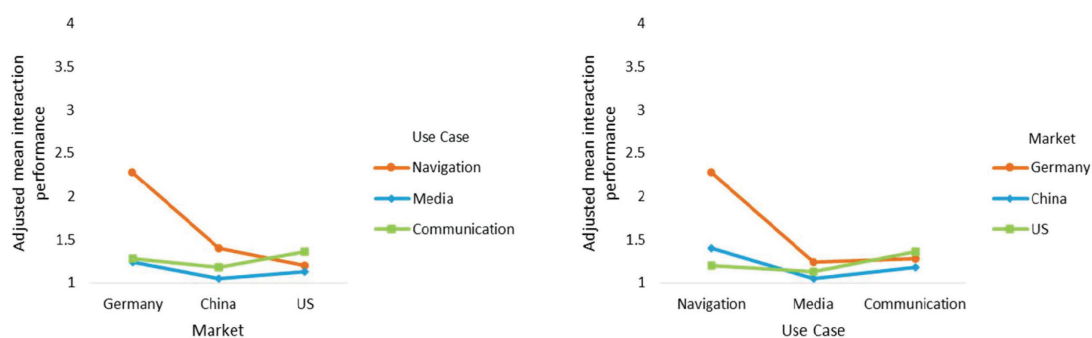


Figure A2. Adjusted mean interaction performance scores to visualize the interaction effect between UC and market for study 3.

Appendix C

Table A3. Inferential statistics (i.e., $df1$, $df2$, F -, p -, and η_p^2 -values) for the mixed between-within ANCOVA for experimenter ratings in study 3 (modality: remote control).

Effect	$df1$	$df2$	F	p	η_p^2
Market	1	79	0.04	0.842	0.001
Age	1	79	5.82	0.018	0.07
UC	1.38	109.30	1.49	0.230	0.02
UC $\times$ Market	2	158	0.02	0.982	0.00
UC $\times$ Age	2	158	0.54	0.585	0.01

Note: significant effects are in bold.

References

1. Watson, G.H. *Strategic Benchmarking: How to Rate Your Company's Performance Against the World's Best*; John Wiley & Sons Inc.: Hoboken, NJ, USA, 1993.
2. Sweeney, M.; Maguire, M.; Shackel, B. Evaluating user-computer interaction: A framework. *Int. J. Man-Mach. Stud.* **1993**, *38*, 689–711. [CrossRef]
3. Kumar, A.; Antony, J.; Dhakar, T.S. Integrating quality function deployment and benchmarking to achieve greater profitability. *BIJ* **2006**, *13*, 290–310. [CrossRef]
4. Erdil, A.; Erbiyik, H. The Importance of Benchmarking for the Management of the Firm: Evaluating the Relation between Total Quality Management and Benchmarking. *Procedia Comput. Sci.* **2019**, *158*, 705–714. [CrossRef]
5. Anand, G.; Kodali, R. Benchmarking the benchmarking models. *Benchmarking Int. J.* **2008**, *15*, 257–291. [CrossRef]
6. Rössger, P. Intercultural HMIs in Automotive: Do We Need Them?—An Analysis. In *HCI International 2021—Late Breaking Papers: Design and User Experience*; Stephanidis, C., Soares, M.M., Rosenzweig, E., Marcus, A., Yamamoto, S., Mori, H., Rau, P.-L.P., Meiselwitz, G., Fang, X., Moallem, A., Eds.; Springer International Publishing: Cham, Switzerland, 2021; pp. 584–596, ISBN 978-3-030-90237-7.
7. Kern, D.; Schmidt, A. Design Space for Driver-based Automotive User Interfaces. In Proceedings of the AutomotiveUI'09: 1st International Conference on Automotive User Interfaces and Interactive Vehicular Applications, Essen, Germany, 21–22 September 2009.
8. Young, K.; Regan, M. Driver Distraction: A Review of the Literature. In *Distracted Driving*; Australasian College of Road Safety: Sydney, NSW, Australia, 2007.
9. Forster, Y.; Hergeth, S.; Naujoks, F.; Krems, J.; Keinath, A. User Education in Automated Driving: Owner's Manual and Interactive Tutorial Support Mental Model Formation and Human-Automation Interaction. *Information* **2019**, *10*, 143. [CrossRef]
10. ISO 9241-210:2010; Prozess zur Gestaltung Gebrauchstauglicher Interaktiver Systeme. Deutsches Institut für Normung: Berlin, Germany, 2011.
11. Forster, Y.; Kraus, J.M.; Feinauer, S.; Baumann, M. Calibration of Trust Expectancies in Conditionally Automated Driving by Brand, Reliability Information and Introductory Videos. In Proceedings of the AutomotiveUI'18: 10th International Conference on Automotive User Interfaces and Interactive Vehicular Applications, Toronto, ON, Canada, 23–25 September 2018; Association for Computing Machinery: New York, NY, USA, 2018; pp. 118–128, ISBN 9781450359467.
12. Norman, D.A. Human-Centered Design Considered Harmful. *Interactions* **2005**, *12*, 14–19. [CrossRef]
13. Hassenzahl, M. The Effect of Perceived Hedonic Quality on Product Appealingness. *Int. J. Hum. Comput. Interact.* **2001**, *13*, 481–499. [CrossRef]
14. Forster, Y.; Hergeth, S.; Naujoks, F.; Krems, J.F.; Keinath, A. Self-report measures for the assessment of human-machine interfaces in automated driving. *Cogn. Technol. Work* **2020**, *22*, 703–720. [CrossRef]
15. Pettersson, I.; Lachner, F.; Frison, A.-K.; Riener, A.; Butz, A. A Bermuda Triangle?—A Review of Method Application and Triangulation in User Experience Evaluation. In Proceedings of the CHI'18: CHI Conference on Human Factors in Computing Systems, Montreal, QC, Canada, 21–26 April 2018; Mandryk, R., Hancock, M., Perry, M., Cox, A., Eds.; Association for Computing Machinery: New York, NY, USA, 2018; pp. 1–16, ISBN 9781450356206.
16. Lindgren, A.; Chen, F.; Jordan, P.W.; Zhang, H. Requirements for the Design of Advanced Driver Assistance Systems—The Difference between Swedish and Chinese Drivers. *Int. J. Des.* **2008**, *2*, 41–54.
17. Heimgaertner, R. Towards Cultural Adaptability in Driver Information and -Assistance Systems. In *Usability and Internationalization Part II*; Springer: Berlin/Heidelberg, Germany, 2007; pp. 372–381.
18. Lesch, M.F.; Rau, P.-L.P.; Zhao, Z.; Liu, C. A cross-cultural comparison of perceived hazard in response to warning components and configurations: US vs. China. *Appl. Ergon.* **2009**, *40*, 953–961. [CrossRef]
19. Khan, T.; Williams, M. A Study of Cultural Influence in Automotive HMI: Measuring Correlation between Culture and HMI Usability. *SAE Int. J. Passeng. Cars-Electron. Electr. Syst.* **2014**, *7*, 430–439. [CrossRef]

20. Wang, M.; Lyckvi, S.L.; Chen, F. Same, Same but Different: How Design Requirements for an Auditory Advisory Traffic Information System Differ Between Sweden and China. In Proceedings of the AutomotiveUI'16: 8th International Conference on Automotive User Interfaces and Interactive Vehicular Applications, Ann Arbor, MI, USA, 24–26 October 2016; Green, P., Boll, S., Gabbard, J., Osswald, S., Burnett, G., Borojeni, S.S., Löcken, A., Pradhan, A., Eds.; Association for Computing Machinery: New York, NY, USA, 2016; pp. 75–82, ISBN ISBN 9781450345330.
21. Khan, T.; Williams, M. Cross-Cultural Differences in Automotive HMI Design: A Comparative Study Between UK and Indian Users' Design Preferences. *J. Usability Stud.* **2016**, *11*, 45–65.
22. Young, K.L.; Rudin-Brown, C.M.; Lenné, M.G.; Williamson, A.R. The implications of cross-regional differences for the design of In-vehicle Information Systems: A comparison of Australian and Chinese drivers. *Appl. Ergon.* **2012**, *43*, 564–573. [CrossRef] [PubMed]
23. Braun, M.; Li, J.; Weber, F.; Pfleging, B.; Butz, A.; Alt, F. What If Your Car Would Care? Exploring Use Cases For Affective Automotive User Interfaces. In Proceedings of the MobileHCI'20: 22nd International Conference on Human-Computer Interaction with Mobile Devices and Services, Oldenburg, Germany, 5–8 October 2020; Association for Computing Machinery: New York, NY, USA, 2020. ISBN ISBN 9781450375160.
24. Zhu, D.; Wang, D.; Huang, R.; Jing, Y.; Qiao, L.; Liu, W. User Interface (UI) Design and User Experience Questionnaire (UEQ) Evaluation of a To-Do List Mobile Application to Support Day-To-Day Life of Older Adults. *Healthcare* **2022**, *10*, 2068. [CrossRef] [PubMed]
25. Niklas, U.; von Behren, S.; Eisenmann, C.; Chlond, B.; Vortisch, P. Premium factor—Analyzing usage of premium cars compared to conventional cars. *Res. Transp. Bus. Manag.* **2019**, *33*, 100456. [CrossRef]
26. François, M.; Osiurak, F.; Fort, A.; Crave, P.; Navarro, J. Automotive HMI design and participatory user involvement: Review and perspectives. *Ergonomics* **2016**, *60*, 541–552. [CrossRef] [PubMed]
27. Brooke, J. SUS—A quick and dirty usability scale. *Usability Eval. Ind.* **1996**, *189*, 4–7.
28. Li, R.; Chen, Y.V.; Sha, C.; Lu, Z. Effects of interface layout on the usability of In-Vehicle Information Systems and driving safety. *Displays* **2017**, *49*, 124–132. [CrossRef]
29. Forster, Y.; Hergeth, S.; Naujoks, F.; Beggiato, M.; Krems, J.F.; Keinath, A. Learning to use automation: Behavioral changes in interaction with automated driving systems. *Transp. Res. Part F Traff. Psychol. Behav.* **2019**, *62*, 599–614. [CrossRef]
30. Gao, M.; Kortum, P.; Oswald, F.L. Multi-Language Toolkit for the System Usability Scale. *Int. J. Hum. Comput.* **2020**, *36*, 1883–1901. [CrossRef]
31. Lewis, J.R. The System Usability Scale: Past, Present, and Future. *Int. J. Hum. Comput.* **2018**, *34*, 577–590. [CrossRef]
32. Loew, A.; Sogemeier, D.; Kulesa, S.; Forster, Y.; Naujoks, F.; Keinath, A. A Global Questionnaire? An International Comparison of the System Usability Scale in the Context of an Infotainment System. In Proceedings of the International Conference on Applied Human Factors and Ergonomics. International Conference on Applied Human Factors and Ergonomics, New York, NY, USA, 24–28 July 2022; pp. 224–232.
33. Laugwitz, B.; Held, T.; Schrepp, M. Construction and Evaluation of a User Experience Questionnaire. In Proceedings of the HCI and Usability for Education and Work: 4th Symposium of the Workgroup Human-Computer Interaction and Usability Engineering of the Austrian Computer Society, Graz, Austria, 20–21 November 2008; Springer: Berlin/Heidelberg, Germany, 2008; pp. 63–76.
34. Schankin, A.; Budde, M.; Riedel, T.; Beigl, M. Psychometric Properties of the User Experience Questionnaire (UEQ). In Proceedings of the CHI'22: CHI Conference on Human Factors in Computing Systems, New Orleans, LA, USA, 29 April–5 May 2022; Barbosa, S., Lampe, C., Appert, C., Shamma, D.A., Drucker, S., Williamson, J., Yatani, K., Eds.; Association for Computing Machinery: New York, NY, USA, 2022; pp. 1–11, ISBN ISBN 9781450391573.
35. Greve, G.; Benning-Rohnke, E. (Eds.) *Kundenorientierte Unternehmensführung: Konzept und Anwendung des Net Promoter® Score in der Praxis*, 1st ed.; Gabler: Wiesbaden, Germany, 2010; ISBN ISBN 9783834923196.
36. Naujoks, F.; Wiedemann, K.; Schömig, N.; Jarosch, O.; Gold, C. Expert-based controllability assessment of control transitions from automated to manual driving. *MethodsX* **2018**, *5*, 579–592. [CrossRef] [PubMed]
37. Kenntner-Mabiala, R.; Kaussner, Y.; Hoffmann, S.; Volk, M. Driving performance of elderly drivers in comparison to middle-aged drivers during a representative, standardized driving test in real traffic. *Z. Für Verkehrssicherheit* **2016**, *3*, 73.
38. Jarosch, O.; Bengler, K. Rating of Take-Over Performance in Conditionally Automated Driving Using an Expert-Rating System. In *Advances in Human Aspects of Transportation*; Stanton, N., Ed.; Springer International Publishing: Cham, Switzerland, 2019; pp. 283–294, ISBN 978-3-319-93884-4.
39. Forster, Y. Preference versus Performance in Automated Driving: A Challenge for Method Development. Doctoral Thesis, University of Technology Chemnitz, Chemnitz, Germany, 2019.
40. Field, A.P. *Discovering Statistics Using SPSS*, 3rd ed.; SAGE Publications: Los Angeles, CA, USA, 2009; ISBN 9781847879066.
41. Armstrong, R.A. When to use the Bonferroni correction. *Ophthalmic Physiol. Opt.* **2014**, *34*, 502–508. [CrossRef] [PubMed]
42. Reichheld, F.F. The one number you need to grow. *Harv. Bus. Rev.* **2003**, *81*, 46–55. [PubMed]
43. Cohen, J. *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed.; L. Erlbaum Associates: Hillsdale, NJ, USA, 1988.

44. Heimgärtner, R. ISO 9241-210 and Culture?—The Impact of Culture on the Standard Usability Engineering Process. In *Design, User Experience, and Usability. User Experience Design Practice*; Hutchison, D., Kanade, T., Kittler, J., Kleinberg, J.M., Kobsa, A., Mattern, F., Mitchell, J.C., Naor, M., Nierstrasz, O., Rangan, C.P., et al., Eds.; Springer International Publishing: Cham, Switzerland, 2014; pp. 39–48, ISBN 978-3-319-07637-9.
45. Mehler, J.; Guo, Z.; Zhang, A.; Rau, P.-L.P. Quick Buttons on Map-Based Human Machine Interface in Vehicles is Better or Not: A Cross-Cultural Comparative Study Between Chinese and Germans. In *Culture and Computing. Design Thinking and Cultural Computing*; Rauterberg, M., Ed.; Springer International Publishing: Cham, Switzerland, 2021; pp. 432–449, ISBN 978-3-030-77430-1.
46. Hofstede, G. Dimensionalizing Cultures: The Hofstede Model in Context. *Psychol. Cult.* **2011**, *2*, 8. [CrossRef]
47. Hofstede, G.J. Sixth Dimension Synthetic Culture Profiles. 2010. Available online: <https://geerthofstede.com/wp-content/uploads/2016/08/sixth-dimension-synthetic-culture-profiles.doc> (accessed on 11 May 2022).
48. Trompenaars, F.; Hampden-Turner, C. *Riding the Waves of Cultures: Understanding Cultural Diversity in Business*; Nicholas Brealey: London, UK, 2011.
49. Chang, S.; Kim, C.-Y.; Cho, Y.S. Sequential effects in preference decision: Prior preference assimilates current preference. *PLoS ONE* **2017**, *12*, e0182442. [CrossRef] [PubMed]
50. De Souza, T.R.C.B.; Bernardes, J.L. The Influences of Culture on User Experience. In *Cross-Cultural Design*; Rau, P.-L.P., Ed.; Springer International Publishing: Cham, Switzerland, 2016; pp. 43–52, ISBN 978-3-319-40092-1.
51. Schrepp, M.; Hinderks, A.; Thomaschewski, J. Applying the User Experience Questionnaire (UEQ) in Different Evaluation Scenarios. In *Design, User Experience, and Usability. Theories, Methods, and Tools for Designing the User Experience*; Hutchison, D., Kanade, T., Kittler, J., Kleinberg, J.M., Kobsa, A., Mattern, F., Mitchell, J.C., Naor, M., Nierstrasz, O., Rangan, C.P., et al., Eds.; Springer International Publishing: Cham, Switzerland, 2014; pp. 383–392, ISBN 978-3-319-07667-6.
52. Sogemeier, D.; Forster, Y.; Naujoks, F.; Krems, J.F.; Keinath, A. How to Map Cultural Dimensions to Usability Criteria: Implications for the Design of an Automotive Human-Machine Interface. In *Proceedings of the AutomotiveUI'22: 14th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, Seoul, Republic of Korea, 17–20 September 2022; Association for Computing Machinery: New York, NY, USA, 2022; pp. 123–126, ISBN 9781450394284.
53. Hofstede, G. Culture and Organizations. *Int. Stud. Manag. Organ.* **1980**, *10*, 15–41. [CrossRef]
54. Gong, Z.; Ma, J.; Zhang, Q.; Ding, Y.; Liu, L. Automotive HMI Guidelines For China Based On Culture Dimensions Interpretation. In *HCI International 2020—Late Breaking Papers: Digital Human Modeling and Ergonomics, Mobility and Intelligent Environments Proceedings of the 22nd HCI International Conference, HCII 2020, Copenhagen, Denmark, 19–24 July 2020*; LNCS; Springer: Cham, Switzerland, 2020; pp. 96–110.
55. Hultsch, D.F.; MacDonald, S.W.; Dixon, R.A. Variability in Reaction Time Performance of Younger and Older Adults. *J. Gerontol. B Psychol. Sci. Soc. Sci.* **2002**, *57B*, 101–115. [CrossRef]
56. Sharit, J.; Czaja, S.J. Ageing, computer-based task performance, and stress: Issues and challenges. *Ergonomics* **1994**, *37*, 559–577. [CrossRef]
57. Dingus, T.A.; Hulse, M.C.; Mollenhauer, M.A.; Fleischman, R.N.; Mcgehee, D.V.; Manakkal, N. Effects of Age, System Experience, and Navigation Technique on Driving with an Advanced Traveler Information System. *Hum. Factors* **1997**, *39*, 177–199. [CrossRef] [PubMed]
58. Lerner, N.; Singer, J.; Huey, R. *Driver Strategies for Engaging in Distracting Tasks Using In-Vehicle Technologies HS DOT 810 919*; U.S. Department of Transportation: Washington, DC, USA, 2008.
59. Totzke, I. Einfluss des Lernprozesses auf den Umgang mit Menügesteuerten Fahrerinformationssystemen. Doctoral Thesis, Julius-Maximilians-Universität Würzburg, Würzburg, Germany, 2013.
60. Roberts, M.J.; Gray, H.; Lesnik, J. Preference versus performance: Investigating the dissociation between objective measures and subjective ratings of usability for schematic metro maps and intuitive theories of design. *Int. J. Hum. Comput.* **2017**, *98*, 109–128. [CrossRef]
61. Knapp, B. Mental Models of Chinese and German Users and Their Implications for MMI: Experiences from the Case Study Navigation System. In *Human-Computer Interaction*; Springer: Berlin/Heidelberg, Germany, 2007; pp. 882–890.
62. Law, E.L.-C.; van Schaik, P.; Roto, V. Attitudes towards User Experience (UX) Measurement. *Int. J. Hum. Comput.* **2014**, *72*, 526–541. [CrossRef]
63. Forster, Y.; Naujoks, F.; Neukum, A. Increasing anthropomorphism and trust in automated driving functions by adding speech output. In *Proceedings of the 2017 IEEE Intelligent Vehicles Symposium (IV)*, Los Angeles, CA, USA, 11–14 June 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 365–372, ISBN ISBN 978-1-5090-4804-5.
64. Bortz, J.; Döring, N. *Forschungsmethoden und Evaluation*, 4th ed.; Springer: Berlin/Heidelberg, Germany, 2006; ISBN 978-3-540-33305-0.
65. Boduroglu, A.; Shah, P.; Nisbett, R.E. Cultural Differences in Allocation of Attention in Visual Information Processing. *J. Cross-Cult. Psychol.* **2009**, *40*, 349–360. [CrossRef] [PubMed]

- 66. Hadders-Algra, M. Human face and gaze perception is highly context specific and involves bottom-up and top-down neural processing. *Neurosci. Biobehav. Rev.* **2022**, *132*, 304–323. [CrossRef]
- 67. Nisbett, R.E.; Miyamoto, Y. The influence of culture: Holistic versus analytic perception. *Trends Cogn. Sci.* **2005**, *9*, 467–473. [CrossRef]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

Article

Principles for External Human–Machine Interfaces

Marc Wilbrink ^{1,*}, Stephan Cieler ², Sebastian L. Weiß ², Matthias Beggiato ³, Philip Joisten ⁴, Alexander Feierle ⁵ and Michael Oehl ¹

¹ German Aerospace Center (DLR), Institute of Transportation Systems, 38108 Braunschweig, Germany; michael.oehl@dlr.de

² Continental Automotive Technologies GmbH, Software and Central Technologies, 64832 Babenhausen, Germany; stephan.cielier@continental-corporation.com (S.C.); sebastian.2.weiss@continental-corporation.com (S.L.W.)

³ Department of Psychology, Cognitive and Engineering Psychology, Chemnitz University of Technology, 09120 Chemnitz, Germany; matthias.beggiato@psychologie.tu-chemnitz.de

⁴ Department of Mechanical Engineering, Institute of Ergonomics and Human Factors, Technical University of Darmstadt, 64287 Darmstadt, Germany; philip.joisten@tu-darmstadt.de

⁵ TUM School of Engineering and Design, Technical University of Munich, 85747 Garching, Germany; alexander.feierle@tum.de

* Correspondence: marc.wilbrink@dlr.de

Abstract: Automated vehicles will soon be integrated into our current traffic system. This development will lead to a novel mixed-traffic environment where connected and automated vehicles (CAVs) will have to interact with other road users (ORU). To enable this interaction, external human–machine interfaces (eHMIs) have been shown to have major benefits regarding the trust and acceptance of CAVs in multiple studies. However, a harmonization of eHMI signals seems to be necessary since the developed signals are extremely varied and sometimes even contradict each other. Therefore, the present paper proposes guidelines for designing eHMI signals, taking into account important factors such as how and in which situations a CAV needs to communicate with ORU. The authors propose 17 heuristics, the so-called eHMI-principles, as requirements for the safe and efficient use of eHMIs in a systematic and application-oriented manner.

Keywords: interaction with automated vehicles; external human–machine interface; design of communication signals for CAV; interaction of automated vehicles and other road users; guidelines for eHMI

1. Introduction

The introduction of connected automated vehicles (CAVs) into our traffic system will lead to a novel mixed-traffic environment where CAVs will coexist with other road users (ORU) [1]. This mixed-traffic environment creates new challenges for CAVs since they need the ability to interact with ORU in a safe and efficient way. Therefore, all traffic participants need to build up a common understanding of the ability and intentions of others [2]. Considering how humans interact in traffic environments, implicit and explicit communication are two promising methods for designing interactions between CAVs and ORU. Implicit communication needs to be interpreted by the receiver and often contains no obvious information (e.g., vehicle movements). Furthermore, implicit information is broadcast into the environment and is not directly addressed to one receiver. On the other hand, explicit communication is directed to a particular traffic participant and can involve different modalities (e.g., auditory, visual) to transfer further information (e.g., hand gestures and the use of turn indicators).

Research results in the fields of human–machine interaction and traffic psychology reveal the great potential of external human–machine interfaces (eHMIs). These eHMIs can play a significant role in enabling safe and efficient interactions between CAVs and

ORU, as they will enhance the current forms of communication and compensate for deficits in implicit communication, e.g., [3–5]. In the literature, a huge variety of different eHMI designs have been proposed and tested in various traffic scenarios [6]. These concepts differ in the number and modality (mostly visually) of signals, and in the directionality (signal for only one receiver vs. broadcast for the environment) of the information, as well as the taken perspective in which the messages are formulated (vehicle's point of view: "I am stopping"; pedestrian perspective: "You can cross the street"). Most of the empirical studies found benefits of eHMI concepts in terms of comprehensibility, safety, efficiency, and acceptability [7]. However, the variability in the eHMI concepts makes it hard to compare and transfer results between the studies. When considering the interaction between CAVs and ORU from a holistic point of view, a pure focus on explicit communication via eHMI seems insufficient. Therefore, the importance of the interplay between eHMIs and the vehicle's movement patterns, the so-called dynamic HMI (dHMI, e.g., acceleration and deceleration), should be highlighted.

2. Approval Aspects and Existing Guidelines

Currently, recommendations are organized as checklists or guidelines addressing product developers in the automotive and supplier industry. However, a surprisingly low number of design recommendations and legal requirements exist for eHMI signals.

In the USA, the Federal Motor Vehicle Safety Standard 108 regulates the installation and properties of "automotive lighting systems, signaling, and reflective devices" [8]. Among other things, the regulation prescribes the colors of emitting lights for different zones of the vehicle (e.g., white or amber is allowed in the front area of a vehicle). In addition, FMVSS 108 requires new lighting systems to be installed away from existing devices so that any confusion regarding light signals can be avoided. The UNECE document R48, authoritative for many countries in Europe, Asia, and beyond, regulates the integration of new external visual communication systems more restrictively [9]. In the current version, additional visual communication systems are not permitted. However, the question of whether and in what form CAVs should communicate with ORU has been discussed for several years in different working groups (e.g., the "Taskforce on Autonomous Vehicle Signaling Requirements", set up by the UNECE [10]). While the discussions in this panel proceed controversially, standardization institutes such as SAE and ISO have described concrete aspects for future communication between CAVs and ORU.

With the "recommended practice" J3134, SAE International has proposed an "Automated Driving System Marker Lamp" that lights up in cyan [10]. The lamp should be installed in the front area of the vehicle and light up permanently when automated driving is activated.

The International Organization for Standardization is also publishing documents that are to be understood as building blocks for future standardization. The Technical Report ISO/TR 23049 ("Road Vehicles—Ergonomic aspects of external visual communication from automated vehicles to other road users") analyzes communication patterns in non-automated traffic and compiles design dimensions that need to be considered when developing future eHMIs (e.g., communication of the driving mode, vehicle state, and vehicle's intention) [11]. Currently in preparation is the standard ISO/AWI PAS 23735 ("Road vehicles—Ergonomic design guidance for external visual communication from automated vehicles to other road users"), which deals with so-called design parameters for safe and well-accepted communication between CAVs and ORU. Among other things, a distinction is made as to which information should be transmitted on which communication media and in which concrete design. It is planned for the document to also contain an extensive list of traffic scenarios and use cases.

A harmonization of eHMI signals in a traffic system can play a major role regarding the acceptance of CAVs and traffic safety while interacting with them. Only if eHMI signals are consistent between vehicle manufacturers and traffic situations can ORU build up a matching mental model of CAVs and understand their present and future behavior.

3. Method for Deriving eHMI-Principles

Multiple approaches were followed to create the present list of eHMI-principles. As a first approach, the authors evaluated the existing research papers and literature reviews in the field of HMI and human factors for automated vehicles. Based on the research results, eHMI-principles were derived iteratively. As another approach, focus groups with drivers were conducted to discuss the need for, functionality, and design of eHMIs from a user's perspective. Finally, project partners identified open research questions and conducted over 40 user studies and expert workshops gaining new insights on the topic of eHMI as a main result of the German project @CITY-AF [12]. Over a period of 2 years, the authors revised the eHMI-principles iteratively based on these two approaches in several expert workshops.

4. Goal and Structure of the eHMI-Principles

Since guidelines for designing eHMI signals are still missing, the authors propose 17 heuristics, the so-called eHMI-principles, as requirements for the safe and efficient use of eHMIs in a systematic and application-oriented manner.

However, the eHMI-principles neither claim to be normative in the sense of homologation guidelines nor are they intended as instructions for detailed product designs and concrete design implementations. Rather, the eHMI-principles represent an application-oriented catalogue of recommendations that can be used to support the development and evaluation of eHMIs in a targeted manner.

The eHMI-principles are structured in two different sections covering the questions "How should a CAV communicate?" (category A) and "In which situations should a CAV communicate?" (category B). Accordingly, the eHMI-principles contain ten principles for "category A" and seven principles for "category B", each scientifically justified with an explanation and a rationale, and illustrated with examples of good or bad implementations. Finally, each eHMI-principle indicates open questions and areas for future research. Some of the eHMI-principles focus only on the interaction of external communication through eHMI, while others also take the more implicit communication through the vehicle's movement patterns (dHMI, e.g., acceleration and deceleration) into account; see [11]. The numeration and order of the eHMI-principles have no hierarchal character and are not based on each other. The principles are independent of each other and stand for themselves.

5. eHMI-Principles

The following section contains the eHMI-principles (see Table 1) starting with category A (how should a CAV communicate?) followed by eHMI-principles in category B (in which situation should an eHMI communicate).

Table 1. Overview of the eHMI-principles.

Principle ID	Name of Principle
A1	Clarity and unambiguity of used communication signals
A2	Side effects of communication
A3	Prosocial communication
A4	Consistency across different vehicle types
A5	Consistency of dHMI and eHMI
A6	Expectation conformity of dHMI communication signals
A7	Consideration of physical–psychological states
A8	Adaptivity to (traffic) environment
A9	Communication with passengers about the status of current external communication
A10	Changed intentions

Table 1. Cont.

Principle ID	Name of Principle
B1	Identification of the automation level
B2	Communication of intent
B3	Mixed traffic
B4	Informal traffic rules
B5	Restricted communication via the dHMI
B6	Minimal-risk condition
B7	External communication with other automated systems

5.1. Category A—How Should a CAV Communicate?

5.1.1. Principle A1: Clarity and Unambiguity of Used Communication Signals

«All signals sent through dHMI and eHMI should be as clear, distinct, and unambiguous as possible.»

Explanations: The signals sent by the CAV through dHMI and eHMI to ORU should be as clear, distinct, and unambiguous as possible. Ideally, the signals are standardized worldwide and uniform across manufacturers.

Rationale: The signals from the CAV must be clear and consistent. No ambiguous or contradicting messages should be sent to avoid accidents, conflicts, and uncertainties regarding the CAV's behavior.

Example of effective implementation: A CAV signals that it is yielding to allow pedestrians to cross before the CAV by slowing down and simultaneously showing a pulsing light signal on a standardized light band in a fixed monotone color and frequency.

Example of ineffective implementation: An already stationary CAV uses a non-uniform, manufacturer-specific light color and frequency on a light band to signalize that it will yield for ORU.

Open questions and areas for research: There is a particular need for research on the temporal coordination of dHMI and eHMI signals in different scenarios that contribute to the efficient design of interactions. Various subject groups (e.g., children, older adults) and their specific needs should be considered (e.g., different perspectives due to small body size) as well as possible cultural differences in various countries.

References: [13–24].

5.1.2. Principle A2: Side Effects of Communication

«When communicating via an eHMI, possible negative effects on ORU should be considered.»

Explanations: In order to exclude undesired or negative side effects, the explicit signals communicated by the CAV should be evaluated with regards to their effect on ORU.

Rationale: Phenomena related to the human–machine interaction such as “over-trust” and “controllability of automation failures” must be considered during the design of eHMI signals. Likewise, a falsely created feeling of being addressed by the eHMI must be prevented.

Example of effective implementation: A CAV communicates through a spatially directed eHMI which of the surrounding road users are detected and are being addressed by the eHMI signal.

Example of ineffective implementation: Non-addressed road users feel addressed by the eHMI signals. This spurious feeling of being addressed can lead to misunderstandings, uncertainty, and, ultimately, to accidents.

Open questions and areas for research: There is a particular need for research in complex traffic scenarios with multiple road users being present. Furthermore, the influence of learning aspects and long-term use on the behavior of road users in terms of “over-trust” and “controllability of automation failures” should be investigated.

References: [20,25–29].

5.1.3. Principle A3: Prosocial Communication

«Prosocial messages are to be sent to ORU via eHMI.»

Explanations: When determining the signals sent to ORU via eHMI, priority should be given to prosocial messages (e.g., “I give priority to you”, “I am considerate of you”, “I let you go first”, “Thank you kindly”).

Rationale: According to §1 (1) of the German Road Traffic Act (StVO), participation in road traffic requires constant caution and the mutual consideration of road users. Ideally, traffic participants behave in a way that helps each other share the traffic space, communicate, and act prosocially—this should also apply to communication via eHMIs. The contagious effect of role models in road traffic has been known for a long time in traffic psychology; see, e.g., [30]. Prosocial communication via an eHMI could serve as a role model for ORU and thus encourage considerate behavior on the road.

Example of effective implementation: In an ambiguous situation on a parking lot, the CAV communicates via eHMI that it will yield (e.g., when another vehicle is driving in reverse on a parking lot), allowing another vehicle to merge into its lane. The communication of the vehicle’s intention via eHMI is visible for ORU and adds a benefit beyond the pure dHMI (e.g., deceleration/yielding) by providing a prosocial reason for its behavior.

Example of ineffective implementation: The CAV does not communicate its prosocial intention to let another vehicle merge into its lane on the eHMI. The yielding behavior of the CAV creates an ambiguity because the ORU cannot grasp the vehicle’s prosocial behavior.

Open questions and areas for research: The principle is also applicable for “reverse” situations where CAVs benefit from the cooperative behaviors of ORU. In this case, the CAV could express “thank you” as a dialog display, which is common in manual traffic, e.g., via waving or smiling. However, the extent to which eHMIs can express gestures of saying thank you and which communicative problems may be associated with this still requires systematic analysis. Based on the current knowledge, it is advised to communicate the intention of giving priority to others.

References: [30–32].

5.1.4. Principle A4: Consistency across Different Vehicle Types

«The basic characteristic (e.g., modalities, designs) of an eHMI should be consistent across different vehicle types (e.g., bus, car, shuttles).»

Explanations: The eHMI signals of CAVs should be consistent across different vehicle types.

Rationale: CAVs may differ in size and vehicle type, but to avoid potential accidents, conflicts, and uncertainties, they should not differ in the basic characteristics of the eHMI to enable a predictable environment and the development of consistent mental models by ORU.

Example of effective implementation: CAVs of different types identically signal future behaviors and intentions. Vehicle type-specific signals (e.g., for busses or shuttles) fit consistently and without contradiction into other vehicle types’ existing communication patterns.

Example of ineffective implementation: A conventional-sized CAV communicates its intention to yield for ORU differently than a bus or shuttle.

Open questions and areas for research: There is a particular need for research in complex traffic scenarios with multiple, simultaneously present CAVs of different types. Additionally, the influence of possible vehicle type-specific signals (e.g., bus: departure at a stop) should be investigated.

References: [22,33–35].

5.1.5. Principle A5: Consistency of dHMI and eHMI

«The signals communicated through an eHMI should be in line with dHMI signals.»

Explanations: Vehicle signals communicated by dHMI (e.g., braking/accelerating) shall not contradict eHMI signals.

Rationale: The communication of the CAV must be consistent and shall not send any contradicting messages to avoid a feeling of uncertainty. In most scenarios, vehicle behavior (dHMI) affects the behavior of ORU. Therefore, the eHMI output should be adapted to the dHMI.

Example of effective implementation: A CAV signals through an eHMI its intention to stop in order to give priority. At the same time, the vehicle gradually slows down in a clearly perceptible manner.

Example of ineffective implementation: A CAV signals via the eHMI that it intends to yield and give priority. However, the intention of the CAV is not indicated by its approaching behavior (e.g., by showing an early braking behavior or lateral deviation from the road center).

Open questions and areas for research: There is a particular need for research on the synchronization of dHMI and eHMI signals in different scenarios. For example, an eHMI signal, such as the intention to brake, could be displayed as an announcement before the actual braking happens. The braking must occur in a narrow time span for vehicle behavior to be considered as decision guidance. In the event of discrepancies, trust and perceived safety will decrease.

References: [6,17,18,22,36–40].

5.1.6. Principle A6: Expectation Conformity of dHMI Communication Signals

«The dHMI signals of CAVs should be based on existing dHMI signals for manual driving.»

Explanations: The signals communicated by CAVs through dHMI (e.g., braking/accelerating, gap selection) should be oriented as closely as possible to the already established and familiar dHMI signals (assuming established dHMI is in line with the respective law) in order to avoid confusion/irritation and disruptions of established routines.

Rationale: According to §1 of the German Road Traffic Act (StVO), all road users should behave in such a way that no one is endangered, harmed, hindered, or disturbed. Therefore, CAVs must show predictable behavior, as disruptions of the usual interaction can otherwise occur. Variations can be offered within an accepted range of expected behavior; for example, through selectable automated driving style variants such as dynamic/defensive.

Example of effective implementation: When turning left with opposing traffic, a CAV selects a gap acceptable to both the vehicle occupants and ORU. It neither waits too long (thereby hindering traffic) nor chooses gaps that are too narrow, which would be perceived as aggressive or risky.

Example of ineffective implementation: A CAV turns left sharply and at high speed in front of approaching opposing traffic. From a technical viewpoint, the maneuver may not be critical because the gaps for the calculated execution are large enough. However, the fast turning is perceived as uncomfortable by the passengers of the vehicle, and the opposing traffic is forced to brake as a precaution.

Open questions and areas for research: Using subject tests, accepted parameters for specific (frequent) maneuvers can be determined. However, this is not possible for every scenario in real traffic. The parameters could be validated, extended, and adjusted, for example, by means of machine learning methods and real driving data.

References: [19,23,25,41–47].

5.1.7. Principle A7: Consideration of Physical–Psychological States

«The signals transmitted via eHMI should be designed in terms of sensory modality and representation format to ensure safe and efficient communication between the system and ORU.»

Explanations: In the event that an eHMI can output different sensory modalities (e.g., visual and auditory) and/or different representation formats (e.g., linguistic vs. non-linguistic), the information format that allows for the most reliable transmission of information shall be selected depending on the current physical–psychological state of the communication partner. To ensure maximum consistency in the behavior of the eHMI, switches between different modalities and representation formats should follow simple and clear rules.

Rationale: The success of communication between a sender (in this case: the eHMI) and a receiver (here: ORU) depends on the information receiver’s current psychological and physiological state. If the person’s attention is not focused on the CAV but on other objects, or if the person is in a state of intoxication or fatigue, different modalities may prove to be differentially successful means of communication. Furthermore, it is known from cognitive psychology that simultaneously occurring tasks that access the same processing mechanisms (e.g., codes for processing linguistic information) result in poorer performance than tasks that access different processing codes (e.g., for processing linguistic information on the one hand and pictorial information on the other). Based on these insights, multimodal eHMI concepts are to be developed in a way that ensures reliable and safe information acquisition and processing.

Example of effective implementation: A connected and automated bus wants to inform ORU about the departure at a bus stop. The system recognizes that the waiting people at the bus stop are distracted by other activities. In addition to the visual signal, an acoustic signal is emitted as well.

Example of ineffective implementation: The CAV’s detection system detects that a guide dog and a visually impaired person are at the pedestrian crossing. By default, the vehicle can only communicate visually, resulting in an impossible interaction with the person.

Open questions and areas for research: It should be examined in detail under which conditions an eHMI, which reacts differently depending on the situation, is perceived as understandable and reliable.

References: [28,48–52].

5.1.8. Principle A8: Adaptivity to (Traffic) Environment

«The eHMI signal should be adapted to the prevailing environmental conditions.»

Explanations: The eHMI signal should be adapted to the environmental conditions that are primarily determined by weather influences, time of day (e.g., brightness), infrastructure (e.g., terrain), and traffic context (e.g., occlusions and noise level) in order to ensure that the signal stands out from them and is always easily perceptible (ensuring the salience of the signal).

Rationale: Regarding explicit communication, it is essential that signals are perceivable, so ORU can be informed efficiently and safely. Adapting the eHMI signal to the respective environmental conditions ensures that ORU are able to perceive the intended signal. This can be supported by having redundant elements in the eHMI that guarantees that the signal is perceptible from various angles.

Example of effective implementation: An eHMI light strip that is mounted 360° around the vehicle, adapting its luminosity to the prevailing environmental conditions and is always visible regardless of the environment. The light strip adjusts its brightness during nighttime to avoid dazzling ORU and increases its intensity during intense sunlight.

Example of ineffective implementation: The luminosity of the eHMI light strip does not adapt to the ambient conditions making it hard for ORU to perceive information coming from the eHMI light strip. Furthermore, the eHMI is only mounted on the front of the vehicle, making it difficult to detect communication signals when a person stands at the side of the vehicle.

Open questions and areas for research: The interaction between the eHMI and other (visual) signals attached to the vehicle (such as turn signals and headlights) must be validated through testing in lighting laboratories and in real-world traffic.

References: [46,53–55].

5.1.9. Principle A9: Communication with Passengers about the Status of Current External Communication

«Signals shared with ORU through eHMI should be visible and understandable to passengers.»

Explanations: eHMI signals from CAVs should also be displayed for passengers inside the vehicle.

Rationale: By providing feedback from the CAV regarding the signals sent through eHMI, improved transparency regarding the behavior of the CAV can be achieved. This improves the situational awareness of the passengers, reduces subjective uncertainty regarding the vehicle's behavior, and enables a predictable environment with uniform mental models among road users.

Example of effective implementation: The CAV signals via eHMI that it will yield for an ORU. At the same time, this information is presented to the passengers inside the CAV. Passengers can then understand the reason for the deceleration.

Example of ineffective implementation: The CAV signals via eHMI that it will yield for an ORU. However, this information is not communicated to the passengers. Since the reason for the deceleration is not immediately clear, passengers will feel uncertain and may assume a possible malfunction of the CAV.

Open questions and areas for research: Novel, unobtrusive display and interaction concepts for system feedback (modality, timing, and positioning) remain to be investigated.

References: [56–59].

5.1.10. Principle A10: Changed Intentions

«When the CAV communicates right of way/priority to ORU, it should not change its intention or external message during the interaction.»

Explanations: After the CAV communicates the right of way to ORU, it should not change its eHMI signal and should proceed with the communicated maneuver until the ORU or the CAV have left the interaction zone.

Rationale: ORU perceive the signals of the CAV, rely on their accuracy, and perform actions accordingly. A change in intention is either no longer perceived or it takes some time after the awareness to process the changed intention and to adjust one's reaction. This can lead to reduced road safety and can negatively affect the trust in and acceptance of CAVs.

Example of effective implementation: The CAV approaches a pedestrian, communicates the intention to yield, and maintains this intention. The pedestrian then crosses the road. After the road is clear, the CAV continues its journey.

Example of ineffective implementation: The CAV approaches a pedestrian and communicates its yielding intention. The pedestrian starts to cross the road. Meanwhile, the CAV changes its intention due to a misperception and insists on having the right of way. As a result, a critical situation arises.

Open questions and areas for research: Research is needed on the “controllability of automation failures” and the behavior of human interaction participants in the case of a change in intention of the CAV.

References: [17,18,59,60].

5.2. Category B—In Which Situation Should an eHMI Communicate?

5.2.1. Principle B1: Identification of the Automation Level

«If a CAV is driving at SAE level L3 (conditional driving automation) or higher, this driving mode should be permanently communicated to the outside through an eHMI.»

Explanations: When a vehicle is moving at automation level L3 (SAE) or higher, the person at the steering wheel is no longer responsible for the driving task, does not have a constant monitoring function, and is therefore not available as an immediate communication person. This condition should be communicated to all ORU in the immediate vicinity.

Rational: The labeling of CAVs is desired by the majority of users (see references), allows, if necessary, an adaptation of the behavior of ORU, as is possible with, e.g., driving school vehicles, enables learning experiences with CAVs in mixed traffic, and avoids possible unclear or conflict-prone situations if people at the driver's seat, for example, turn to driving-unrelated activities and this is perceived from the outside.

Example of effective implementation: An LED light strip is installed 360° around the vehicle and is constantly lit when the vehicle is driving at L3 or higher (e.g., Autobahn chauffeur). ORU can infer from this that this vehicle leaves larger gaps and refrains from making sudden lane changes without prior announcement by flashing.

Example of ineffective implementation: The person behind the steering wheel of a CAV (SAEL3 and above) is engaged with a smartphone and not monitoring the road scene. Without eHMI, ORU do not know if this driver is distracted or if the vehicle is currently in charge of the driving task.

Open questions and areas for research: There are various options for placing an eHMI, e.g., as a signal visible from all sides. In addition to a 360° LED strip, a signal on the roof would also be conceivable (e.g., as with taxis, driving schools, and emergency vehicles) and/or coupled with acoustic signals in special situations such as parking. The design of the signal should ideally not create any confusion with other signals, such as those from emergency vehicles.

References: [21,34,35,46,61–70].

5.2.2. Principle B2: Communication of Intent

«The CAV should inform the environment via an eHMI about its intention and the behavior it will exhibit in certain situations.»

Explanations: The vehicle should inform the environment in situations where it will slow down, stop, grant priority, or start driving again via eHMI regarding its future intention. This allows ORU in the traffic environment of the CAV to perceive, understand, and adapt their behavior accordingly—without specific recommendations or instructions being given. The communication of the CAV's future intentions should be binding in this regard.

Rationale: The early communication of the intentions of the CAV creates a high sense of security for pedestrians and allows them to recognize the future behavior of the vehicle. As they can adapt their behavior accordingly, the overall efficiency of traffic flow is increased. Additionally, this prevents the vehicle from giving specific instructions to ORU. Messages instructing ORU to a specific behavior should be avoided as they raise legal and safety-related issues.

Example of effective implementation: A light signal in the front of the CAV informs the environment of the intention to stop. ORU perceive this signal and can adapt their behavior to the future trajectory of the CAV. They can rely on the information provided by the eHMI signal and do not have to wait until the vehicle shows the announced driving behavior.

Example of ineffective implementation: The CAV approaches ORU without informing them about its future behavior. The ORU has to wait until the vehicle actually slows down—e.g., communicates specifically through the dHMI.

Open questions and areas for research: Safety concerns need to be evaluated related to the resumption of the driving task when approaching the crosswalk after communicating with pedestrians via an eHMI. In addition, individual studies in the laboratory show that eHMIs that instruct a specific behavior and thus communicate from an egocentric perspective are understood more quickly and easily than those that provide information

about the future behavior of the CAV (allocentric perspective). These findings should be verified in real-world traffic.

References: [4,6,17,18,20,22,34,40,69–76].

5.2.3. Principle B3: Mixed Traffic

«As soon as a CAV is driving in mixed traffic, it should communicate with ORU via eHMI (mixed traffic = automated and non-automated road users, e.g., drivers, cyclists).»

Explanations: In mixed traffic, interaction between automated and non-automated road users is essential, and CAVs need to be able to communicate with ORU.

Rationale: The use of explicit communication by CAVs towards ORU becomes relevant when the vehicle moves in or enters mixed traffic. In segregated CAV networks, the use of explicit communication between CAVs would lead to efficiency losses because communication between the automated agents in these areas can be taken over more effectively and efficiently by other technologies (e.g., V2V communication interfaces). Furthermore, the use of explicit communication by CAVs in fully segregated traffic areas could negatively affect traffic safety, e.g., by distracting non-automated road users who are in the vicinity of this area.

Example of effective implementation: By leaving a delimited lane for automated road users and entering mixed traffic, the CAV will use the eHMI to communicate with non-automated road users.

Example of ineffective implementation: The CAV does not communicate with ORU through an eHMI in mixed traffic. Deadlocks can occur (e.g., in construction areas).

Open questions and areas for research: Open questions exist in the design of the HMIs of CAVs in the transitions between mixed traffic and fully separated areas.

References: [17,18,40,46,72,75,77].

5.2.4. Principle B4: Informal Traffic Rules

«The CAV should communicate via the eHMI in traffic situations where informal rules of interaction are predominant.»

Explanations: Regarding communication between road users, both formal and informal rules are applied. While formal rules stem from laws (e.g., the German StVO), informal rules arise from the interaction of road users who want to optimize a certain system.

Rationale: Informal rules of interaction between road users are situation-specific and are interpreted according to the characteristics of the situation. Due to the need for the interpretation of communication signals in these situations, explicit communication can promote mutual understanding, increase traffic efficiency, and ensure the satisfaction or acceptance of road users.

Example of effective implementation: The CAV interacts with ORU at an equal, unregulated bottleneck, using the eHMI in addition to the dHMI (e.g., braking of the vehicle). The ORU recognize that the CAV is giving priority.

Example of ineffective implementation: The CAV interacts with ORU at an equal, unstructured bottleneck by braking and stopping without using the eHMI. There is no clarity about who should go first.

Open questions and areas for research: Open questions may relate to the influence of other situational characteristics and other influencing factors, such as cultural variables.

References: [17,18,72,78–80].

5.2.5. Principle B5: Restricted Communication via the dHMI

«The CAV should communicate via the eHMI in scenarios in which communication via the dHMI is not clearly possible.»

Explanations: Communication through the dHMI (in particular, driving dynamics through acceleration and braking performed by the vehicle) can be limited or even impossible due to the driving environment conditions, especially at lower speeds. In these situations, a CAV should communicate explicitly through the eHMI.

Rationale: Explicit communication positively influences the willingness to cooperate and contributes to mutual understanding between road users in cooperative situations. Cooperative behavior in road traffic can minimize individual and global costs (e.g., optimize traffic flow) and contributes to fairness among road users. Especially at low speeds, it can be difficult for ORU to identify the intention of the CAV when only communicated by its dHMI.

Example of effective implementation: The CAV drives in a low speed area such as a shared space. Since communication by the dHMI is very limited here, the CAV communicates its intention via dHMI and eHMI.

Example of ineffective implementation: In a “deadlock” situation, where all road users have come to a standstill, the CAV does not communicate through the eHMI to resolve the deadlock.

Open questions and areas for research: Open questions concern the offensive signals emitted by the CAV and signals that increase the willingness to cooperate among all road users.

References: [74,76,81–85].

5.2.6. Principle B6: Minimal-Risk Condition

«The CAV should communicate through the eHMI when performing a minimal-risk maneuver.»

Explanations: A minimal-risk condition is defined as a stable, safe, or risk-minimizing state of a CAV, which it enters when a certain journey cannot or should not be continued. Through a minimal-risk maneuver, a CAV brings itself into a minimal-risk condition.

Rationale: In a minimal-risk condition, ORU should be informed and warned before the CAV performs the maneuver. This helps to maintain road safety, especially when the CAV performs unexpected maneuvers to achieve the minimal-risk condition. The means of explicit communication can also contribute to calibrating the trust of ORU in the CAV. Furthermore, explicit communication can support the identification of the intervention of a technical supervisor.

Example of effective implementation: A CAV bringing itself into a minimal-risk condition by exiting the lane and stopping on the emergency lane. During this maneuver, the vehicle informs ORU through the eHMI. When stopping on the emergency lane, the vehicle communicates its current state constantly to ORU.

Example of ineffective implementation: A CAV executes a minimal-risk maneuver without informing ORU through eHMI.

Open questions and areas for research: Open questions refer to the design of explicit signals during the implementation of a minimal-risk condition and the effects on road safety as well as on the attitudes (especially trust and acceptance) of ORU.

Reference: [86,87].

5.2.7. Principle B7: External Communication with Other Automated Systems

«When a CAV communicates with other CAVs or the infrastructure, the eHMI should serve as an additional source of information for the environment.»

Explanations: Communication between CAVs and/or infrastructure should be transparent and visible through an eHMI. This is particularly important in situations where ORU actively participate in traffic and are dependent on what was communicated between the vehicles/infrastructure.

Rationale: For ORU to build adequate situational awareness, they should have access to all possible information for predicting future system behavior. If the communication between CAVs remains a “black box”, it can lead to safety-critical situations.

Example of effective implementation: Two CAVs are facing each other at a two-way lane restriction. They communicate with each other to regulate priority. Pedestrians who want to cross the street at the same point are informed through the eHMIs of the CAVs as

to which vehicle has priority. This allows ORU to predict the traffic situation and adjust their behavior accordingly.

Example of ineffective implementation: Two CAVs are facing each other at a two-way lane restriction. Although they communicate with each other to clarify who has priority, the surrounding traffic is not informed. For pedestrians who want to cross the street at the same point, this communication remains invisible, and they must wait until the traffic situation has been resolved.

Open questions and areas for research: The situation in which an eHMI communicates in the sense of a broadcast with people in the environment who are not direct recipients must be further examined.

References: [6,16,60].

6. Discussion

The eHMI-principles should support the development of eHMI communication strategies by providing high-level design recommendations and a heuristic checklist for the evaluation of eHMI communication strategies. However, a possible limitation of the method used for creating the eHMI-principles is that they are based on German cultural practices and norms. As a result, there is a high chance that the eHMI-principles are picked selectively and lack consideration for cultural diversity. Therefore, the proposed eHMI-principles may need some design adaptations when implemented in other countries. Additionally, the proposed eHMI-principles might not enclose a comprehensive range of accessibility considerations for users with impairments, disabilities, or special needs. Individuals with visual/hearing impairments, cognitive disabilities, or motor disabilities may require specific design adaptations. Moreover, all eHMI-principles must comply with the road traffic laws of the respective country. There may be legal or safety reasons where deviation from the original eHMI-principles is required. Furthermore, the eHMI-principles do not claim to cover the topic of eHMI to the full extent nor are all principles transferable to all kinds of traffic situations or countries. Quite the contrary, the eHMI-principles should be updated and reviewed with growing research and evaluation in the field of eHMI. Additionally, it may not always be useful to comply with all eHMI-principles at once, as long as the developer reflects on the principles and considers them when designing an eHMI.

Future research on the development of eHMI should use these principles as a basis for designing safe and efficient interaction strategies while contributing new results and principles to the existing collection. Furthermore, an evaluation of existing and upcoming eHMI-principles regarding ethical issues should be conducted.

7. Conclusions

The proposed 17 eHMI-principles primarily aim to help practitioners design safe and efficient eHMIs in an application-oriented manner. Thus, the principles are intended to be used as a rather high-level checklist for a heuristic evaluation of eHMIs. The checklist can help in both (1) evaluating existing eHMI solutions with regard to, e.g., non-intended side effects, and (2) avoiding potential issues at early stages in the development process of new eHMIs. The rather high degree of abstraction as “principles” was chosen to cover a great variety, including potential future eHMI solutions even beyond passenger cars. This decision, in turn, does not allow the specification of very detailed product design aspects such as, e.g., colors or patterns of eHMI signals. The principles neither claim to be normative in the sense of homologation guidelines nor are they intended as instructions for detailed product designs and concrete design implementations. However, in addition to their use as a heuristic checklist, the principles should stimulate scientific as well as application-related discussions about eHMI guidelines and are therefore related to parallel activities in, e.g., standardization institutions such as ISO and SAE.

Author Contributions: Conceptualization, M.W., S.C., S.L.W., M.B., P.J., A.F. and M.O.; methodology, M.W., S.C., S.L.W., M.B., P.J., A.F. and M.O.; writing—original draft preparation, M.W. and M.O.; writing—review and editing, M.W., S.C., S.L.W., M.B., P.J., A.F. and M.O. All authors have read and agreed to the published version of the manuscript.

Funding: This work is a result of the research project @CITY-AF—Automated Cars and Intelligent Traffic in the City—Automatisierte Fahrfunktionen. The project is supported by the Federal Ministry for Economic Affairs and Climate Action (BMWK), based on a decision taken by the German Bundestag. The authors are solely responsible for the content of this publication.

Data Availability Statement: No new data were created or analyzed in this study. Data sharing is not applicable to this article.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study, in the collection, analyses, or interpretation of data, in the writing of the manuscript, or in the decision to publish the results.

References

1. Kauffmann, N.; Winkler, F.; Naujoks, F.; Vollrath, M. “What Makes a Cooperative Driver?” Identifying parameters of implicit and explicit forms of communication in a lane change scenario. *Transp. Res. Part F Traffic Psychol. Behav.* **2018**, *58*, 1031–1042. [CrossRef]
2. Färber, B. Communication and Communication Problems Between Autonomous Vehicles and Human Drivers. In *Autonomous Driving: Technical, Legal and Social Aspects*; Maurer, M., Gerdes, J.C., Lenz, B., Winner, H., Eds.; Springer Open: Berlin/Heidelberg, Germany, 2016; pp. 125–144, ISBN 978-3-662-48845-4.
3. Winter, J.; de Dodou, D. External human–machine interfaces: Gimmick or necessity? *Transp. Res. Interdiscip. Perspect.* **2022**, *15*, 100643. [CrossRef]
4. Tabone, W.; Winter, J.; de Ackermann, C.; Bärghman, J.; Baumann, M.; Deb, S.; Emmenegger, C.; Habibovic, A.; Hagenzieker, M.; Hancock, P.A.; et al. Vulnerable road users and the coming wave of automated vehicles: Expert perspectives. *Transp. Res. Interdiscip. Perspect.* **2021**, *9*, 100293. [CrossRef]
5. Schaarschmidt, E.; Yen, R.; Bosch, R.; Zwicker, L.; Schade, J.; Petzoldt, T. *Grundlagen zur Kommunikation zwischen automatisierten Kraftfahrzeugen und Verkehrsteilnehmern: Berichte der Bundesanstalt für Straßenwesen*; F138; Fachverlag NW: Bremen, Germany, 2021; ISBN 978-3-95606-570-5.
6. Dey, D.; Habibovic, A.; Löcken, A.; Wintersberger, P.; Pfleging, B.; Rienner, A.; Martens, M.; Terken, J. Taming the eHMI jungle: A classification taxonomy to guide, compare, and assess the design principles of automated vehicles’ external human-machine interfaces. *Transp. Res. Interdiscip. Perspect.* **2020**, *7*, 100174. [CrossRef]
7. Rouchitsas, A.; Alm, H. Corrigendum: External Human-Machine Interfaces for Autonomous Vehicle-to-Pedestrian Communication: A Review of Empirical Work. *Front. Psychol.* **2020**, *11*, 575151. [CrossRef] [PubMed]
8. National Highway Traffic Safety Administration, Department of Transportation. *Federal Motor Vehicle Safety Standards No. 108; Lamps, Reflective Devices, and Associated Equipment, Adaptive Driving Beam Headlamps*; 49 CFR Part 571; National Highway Traffic Safety Administration, Department of Transportation: Washington, DC, USA, 2022.
9. Economic Commission for Europe of the United Nations (UNECE). Regulation No 48—Uniform provisions concerning the approval of vehicles with regard to the installation of lighting and light-signalling devices—Uniform provisions concerning the approval of vehicles with regard to the installation of lighting and light-signalling devices [2016/1723]. *Off. J.* **2016**, 125–242. Available online: <http://data.europa.eu/eli/reg/2016/1723/oj> (accessed on 28 June 2023).
10. GRE AVSR. Report of the GRE Taskforce Autonomous Vehicle Signalling Requirements (AVSR): 8th Meeting, 17 October 2022. Available online: <https://wiki.unece.org/display/trans/AVSR++8th+session%2C+2022-10-17> (accessed on 30 March 2023).
11. ISO/TC 22/SC 39; Ergonomics; Road Vehicles—Ergonomic Aspects of External Visual Communication from Automated Vehicles to Other Road Users. International Standards Organization (ISO): Geneva, Switzerland, 2018.
12. WES-Office. Automated Driving in the City. Available online: <https://www.atcity-online.de/?language=en> (accessed on 27 April 2023).
13. Gauß, E.-M. Stehen oder Gehen?—Wenn kein Mensch mehr kommuniziert. Kommunikation Eines automatisierten Fahrzeugs am Zebrastreifen in der Interaktion mit Kindern und Erwachsenen. Bachelor’s Thesis, FernUniversität Hagen, Hagen, Germany, 2019.
14. Ackermann, C.; Beggiato, M.; Schubert, S.; Krems, J.F. An experimental study to investigate design and assessment criteria: What is important for communication between pedestrians and automated vehicles? *Appl. Ergon.* **2019**, *75*, 272–282. [CrossRef] [PubMed]
15. Deb, S.; Strawderman, L.J.; Carruth, D.W. Investigating pedestrian suggestions for external features on fully autonomous vehicles: A virtual reality experiment. *Transp. Res. Part F Traffic Psychol. Behav.* **2018**, *59*, 135–149. [CrossRef]
16. Habibovic, A.; Lundgren, V.M.; Andersson, J.; Klingegård, M.; Lagström, T.; Sirkka, A.; Fagerlönn, J.; Edgren, C.; Fredriksson, R.; Krupenia, S.; et al. Communicating Intent of Automated Vehicles to Pedestrians. *Front. Psychol.* **2018**, *9*, 1336. [CrossRef]

17. Rettenmaier, M.; Bengler, K. The Matter of How and When: Comparing Explicit and Implicit Communication Strategies of Automated Vehicles in Bottleneck Scenarios. *IEEE Open J. Intell. Transp. Syst.* **2021**, *2*, 282–293. [CrossRef]
18. Rettenmaier, M. Interaction between Automated Vehicles and Oncoming Human Drivers: Efficient and Safe Urban Driving in Bottleneck Scenarios. Ph.D. Thesis, Technical University of Munich, Munich, Germany, 2023.
19. Miller, L.; Leitner, J.; Kraus, J.; Baumann, M. Implicit intention communication as a design opportunity for automated vehicles: Understanding drivers' interpretation of vehicle trajectory at narrow passages. *Accid. Anal. Prev.* **2022**, *173*, 106691. [CrossRef]
20. Dey, D.; van Vastenhoven, A.; Cuijpers, R.H.; Martens, M.; Pfleging, B. Towards Scalable eHMIs: Designing for AV-VRU Communication Beyond One Pedestrian. In Proceedings of the 13th International Conference on Automotive User Interfaces and Interactive Vehicular Applications, Leeds, UK, 9–14 September 2021; ACM: New York, NY, USA, 2021; pp. 274–286.
21. Joisten, P.; Alexandi, E.; Drews, R.; Klassen, L.; Petersohn, P.; Pick, A.; Schwindt, S.; Abendroth, B. Displaying Vehicle Driving Mode—Effects on Pedestrian Behavior and Perceived Safety. In *Human Systems Engineering and Design II*; Ahram, T., Karwowski, W., Pickl, S., Taiar, R., Eds.; Springer International Publishing: Cham, Switzerland, 2020; pp. 250–256, ISBN 978-3-030-27927-1.
22. Lau, M.; Jipp, M.; Oehl, M. Toward a Holistic Communication Approach to an Automated Vehicle's Communication with Pedestrians: Combining Vehicle Kinematics with External Human-Machine Interfaces for Differently Sized Automated Vehicles. *Front. Psychol.* **2022**, *13*, 882394. [CrossRef] [PubMed]
23. Fuest, T.; Michalowski, L.; Schmidt, E.; Bengler, K. Reproducibility of Driving Profiles—Application of the Wizard of Oz Method for Vehicle Pedestrian Interaction. In Proceedings of the 2019 IEEE Intelligent Transportation Systems Conference (ITSC), Auckland, New Zealand, 27–30 October 2019; IEEE: New York, NY, USA; pp. 3954–3958.
24. Lee, Y.M.; Madigan, R.; Garcia, J.; Tomlinson, A.; Solernou, A.; Romano, R.; Markkula, G.; Merat, N.; Uttley, J. Understanding the Messages Conveyed by Automated Vehicles. In Proceedings of the 11th International Conference on Automotive User Interfaces and Interactive Vehicular Applications, Utrecht, The Netherlands, 21–25 September 2019; ACM: New York, NY, USA, 2019; pp. 134–143.
25. Lee, J.; Daimon, T.; Kitazaki, S. Negative Effect of External Human-Machine Interfaces in Automated Vehicles on Pedestrian Crossing Behaviour: A Virtual Reality Experiment. In *Proceedings of the 21st Congress of the International Ergonomics Association (IEA 2021)*; Black, N.L., Neumann, W.P., Noy, I., Eds.; Springer International Publishing: Cham, Switzerland, 2021; pp. 718–725, ISBN 978-3-030-74607-0.
26. Wilbrink, M.; Lau, M.; Illgner, J.; Schieben, A.; Oehl, M. Impact of External Human-Machine Interface Communication Strategies of Automated Vehicles on Pedestrians' Crossing Decisions and Behaviors in an Urban Environment. *Sustainability* **2021**, *13*, 8396. [CrossRef]
27. Hübner, M.; Feierle, A.; Rettenmaier, M.; Bengler, K. External communication of automated vehicles in mixed traffic: Addressing the right human interaction partner in multi-agent simulation. *Transp. Res. Part F Traffic Psychol. Behav.* **2022**, *87*, 365–378. [CrossRef]
28. Lanzer, M.; Koniakowsky, I.; Colley, M.; Baumann, M. Interaction Effects of Pedestrian Behavior, Smartphone Distraction and External Communication of Automated Vehicles on Crossing and Gaze Behavior. In Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, Hamburg, Germany, 23–28 April 2023; Schmidt, A., Väänänen, K., Goyal, T., Kristensson, P.O., Peters, A., Mueller, S., Williamson, J.R., Wilson, M.L., Eds.; ACM: New York, NY, USA, 2023; pp. 1–18.
29. Dietrich, A.; Willrodt, J.-H.; Wagner, K.; Bengler, K. Projection-Based External Human Machine Interfaces—Enabling Interaction between Automated Vehicles and Pedestrians. In Proceedings of the Driving Simulation Conference 2018 Europe VR, Driving Simulation Association, Antibes, France, 5–7 September 2018.
30. Echterhoff, W. Lernen im Straßenverkehr durch gezielt angebotenes Modell-Verhalten: Ein Feldexperiment. *Z. Für Verk.* **1989**, *35*, 156–159.
31. Stern, J. Evaluation eines dynamischen Rückmeldesystems an Fußgängerüberwegen und an einem Bahnübergang. In *Fortschritte der Verkehrspsychologie*; Schade, J., Engeln, A., Eds.; VS Verlag für Sozialwissenschaften: Wiesbaden, Germany, 2008; pp. 103–122, ISBN 978-3-531-15956-0.
32. Schlag, B.; Stern, J.; Butterwegge, B.; Degener, S. *“Lob und Tadel”: Wirkungen des Dialog-Display*; Forschungsbericht VV01; Gesamtverband der Deutschen Versicherungswirtschaft e. V.: Berlin, Germany, 2009.
33. Lau, M.; Jipp, M.; Oehl, M. Investigating the Interplay between eHMI and dHMI for Automated Buses: How Do Contradictory Signals Influence a Pedestrian's Willingness to Cross? In Proceedings of the 13th International Conference on Automotive User Interfaces and Interactive Vehicular Applications, Leeds, UK, 9–14 September 2021; ACM: New York, NY, USA, 2021; pp. 152–155.
34. Lau, M.; Jipp, M.; Oehl, M. One Solution Fits All? Evaluating Different Communication Strategies of a Light-based External Human-Machine Interface for Differently Sized Automated Vehicles from a Pedestrian's Perspective. *Accid. Anal. Prev.* **2022**, *171*, 106641. [CrossRef]
35. Oehl, M.; Lau, M.; Gehreke, L.; Wilbrink, M. Towards a Universal Explicit Communication Design of External Human-Machine Interfaces (eHMI) for Differently Sized Highly Automated Vehicles Evaluated by Different Pedestrian Age Groups. In *HCI International 2022—Late Breaking Posters*; Stephanidis, C., Antona, M., Ntoa, S., Salvendy, G., Eds.; Springer: Cham, Switzerland, 2022; pp. 391–398, ISBN 978-3-031-19681-2.
36. Hensch, A.-C.; Kreißig, I.; Beggiato, M.; Krems, J.F. The Effect of eHMI Malfunctions on Younger and Elderly Pedestrians' Trust and Acceptance of Automated Vehicle Communication Signals. *Front. Psychol.* **2022**, *13*, 866475. [CrossRef]

37. Holländer, K.; Wintersberger, P.; Butz, A. Overtrust in External Cues of Automated Vehicles. In Proceedings of the 11th International Conference on Automotive User Interfaces and Interactive Vehicular Applications, Utrecht, The Netherlands, 21–25 September 2019; ACM: New York, NY, USA, 2019; pp. 211–221.
38. Kaleefathullah, A.A.; Merat, N.; Lee, Y.M.; Eisma, Y.B.; Madigan, R.; Garcia, J.; de Winter, J. External Human-Machine Interfaces Can Be Misleading: An Examination of Trust Development and Misuse in a CAVE-Based Pedestrian Simulation Environment. *Hum. Factors* **2022**, *64*, 1070–1085. [CrossRef]
39. Bengler, K.; Rettenmaier, M.; Fritz, N.; Feierle, A. From HMI to HMIs: Towards an HMI Framework for Automated Driving. *Information* **2020**, *11*, 61. [CrossRef]
40. Dey, D.; Matvienko, A.; Berger, M.; Pfleging, B.; Martens, M.; Terken, J. Communicating the intention of an automated vehicle to pedestrians: The contributions of eHMI and vehicle behavior. *It—Inf. Technol.* **2021**, *63*, 123–141. [CrossRef]
41. Hensch, A.-C.; Beggiato, M.; Krems, J.F. Drive safely and comfortably—Gap Acceptance as a Basis for a user-centred Design of Driving Styles in Automated Vehicles. In Proceedings of the 7th HUMANIST Conference, Rhodes Island, Greece, 26–27 October 2021.
42. Hensch, A.-C.; Beggiato, M.; Krems, J.F. Drivers’ gap acceptance during parking maneuvers as a basis for initiating driving actions in automated vehicles. *Transp. Res. Part F Traffic Psychol. Behav.* **2023**, *92*, 133–142. [CrossRef]
43. Domeyer, J.; Dinparastdjadid, A.; Lee, J.D.; Douglas, G.; Alsaied, A.; Price, M. Proxemics and Kinesics in Automated Vehicle–Pedestrian Communication: Representing Ethnographic Observations. *Transp. Res. Rec.* **2019**, *2673*, 70–81. [CrossRef]
44. Domeyer, J.E.; Lee, J.D.; Toyoda, H. Vehicle Automation–Other Road User Communication and Coordination: Theory and Mechanisms. *IEEE Access* **2020**, *8*, 19860–19872. [CrossRef]
45. Rettenmaier, M.; Dinkel, S.; Bengler, K. Communication via motion—Suitability of automated vehicle movements to negotiate the right of way in road bottleneck scenarios. *Appl. Ergon.* **2021**, *95*, 103438. [CrossRef] [PubMed]
46. Schieben, A.; Wilbrink, M.; Kettwich, C.; Madigan, R.; Louw, T.; Merat, N. Designing the interaction of automated vehicles with other traffic participants: Design considerations based on human needs and expectations. *Cogn. Tech. Work* **2019**, *21*, 69–85. [CrossRef]
47. Fuest, T.; Michalowski, L.; Traris, L.; Bellem, H.; Bengler, K. Using the Driving Behavior of an Automated Vehicle to Communicate Intentions—A Wizard of Oz Study. In Proceedings of the 2018 21st International Conference on Intelligent Transportation Systems (ITSC), Maui, HI, USA, 4–7 November 2018; IEEE: New York, NY, USA, 2018; pp. 3596–3601.
48. Dou, J.; Chen, S.; Tang, Z.; Xu, C.; Xue, C. Evaluation of Multimodal External Human–Machine Interface for Driverless Vehicles in Virtual Reality. *Symmetry* **2021**, *13*, 687. [CrossRef]
49. Joisten, P.; Theobald, N.; Schwindt, S.; Walter, J.; Abendroth, B. Designing the Communication with Automated Vehicles: The Case of Elderly Pedestrians. In Proceedings of the Workshop on Inclusive Communication between Automated Vehicles and Vulnerable Road Users. In conjunction with MobileHCI 2020, Darmstadt, Germany, 5–9 October 2022.
50. Colley, M.; Walch, M.; Gugenheimer, J.; Rukzio, E. Including people with impairments from the start. In Proceedings of the 11th International Conference on Automotive User Interfaces and Interactive Vehicular Applications, Utrecht, The Netherlands, 21–25 September 2019; Janssen, C.P., Donker, S.F., Chuang, L.L., Ju, W., Eds.; ACM: New York, NY, USA, 2019; pp. 307–314.
51. Colley, M.; Walch, M.; Gugenheimer, J.; Askari, A.; Rukzio, E. Towards Inclusive External Communication of Autonomous Vehicles for Pedestrians with Vision Impairments. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, Honolulu, HI, USA, 25–30 April 2020; Bernhaupt, R., Mueller, F., Verweij, D., Andres, J., McGrenere, J., Cockburn, A., Avellino, I., Gogwey, A., Björn, P., Zhao, S., et al., Eds.; ACM: New York, NY, USA, 2020; pp. 1–14.
52. Löcken, A.; Golling, C.; Riener, A. How Should Automated Vehicles Interact with Pedestrians? In Proceedings of the 11th International Conference on Automotive User Interfaces and Interactive Vehicular Applications, Utrecht, The Netherlands, 21–25 September 2019; ACM: New York, NY, USA, 2019; pp. 262–274.
53. Eisele, D.; Petzoldt, T. Effects of traffic context on eHMI icon comprehension. *Transp. Res. Part F Traffic Psychol. Behav.* **2022**, *85*, 1–12. [CrossRef]
54. Faas, S.M.; Baumann, M. Light-Based External Human Machine Interface: Color Evaluation for Self-Driving Vehicle and Pedestrian Interaction. *Proc. Hum. Factors Ergon. Soc. Annu. Meet.* **2019**, *63*, 1232–1236. [CrossRef]
55. Rettenmaier, M.; Schulze, J.; Bengler, K. How Much Space Is Required? Effect of Distance, Content, and Color on External Human–Machine Interface Size. *Information* **2020**, *11*, 346. [CrossRef]
56. Hümmer, M. Automatisiertes Fahren in der Stadt: Auswirkungen eines iHMIs auf Vertrauen, Akzeptanz und Sicherheitsgefühl. Master’s Thesis, Hochschule Fresenius, Frankfurt, Germany, 2022.
57. Weber, F.; Sorokin, L.; Schmidt, E.; Schieben, A.; Wilbrink, M.; Kettwich, C.; Dodiya, J.; Oehl, M.; Kaup, M.; Willrodt, J.; et al. Final Human-Vehicle Interaction Strategies for the interACT Automated Vehicles: interACT D4.2, Project Deliverable. Available online: https://www.interact-roadautomation.eu/wp-content/uploads/interACT_WP4_D4.2_Final_Human_Vehicle_Interaction_Strategies_v1.1_uploadWebsiteApproved.pdf (accessed on 28 June 2023).
58. Wilbrink, M.; Schieben, A.; Oehl, M. Reflecting the automated vehicle’s perception and intention. In Proceedings of the 25th International Conference on Intelligent User Interfaces Companion, Cagliari, Italy, 17–20 March 2020; ACM: New York, NY, USA, 2020; pp. 105–107.
59. Feierle, A.; Rettenmaier, M.; Zeitlmeir, F.; Bengler, K. Multi-Vehicle Simulation in Urban Automated Driving: Technical Implementation and Added Benefit. *Information* **2020**, *11*, 272. [CrossRef]

60. Rettenmaier, M.; Albers, D.; Bengler, K. After you?!—Use of external human-machine interfaces in road bottleneck scenarios. *Transp. Res. Part F Traffic Psychol. Behav.* **2020**, *70*, 175–190. [CrossRef]
61. Faas, S.M.; Stange, V.; Baumann, M. Self-Driving Vehicles and Pedestrian Interaction: Does an External Human-Machine Interface Mitigate the Threat of a Tinted Windshield or a Distracted Driver? *Int. J. Hum.-Comput. Interact.* **2021**, *37*, 1364–1374. [CrossRef]
62. Lagström, T.; Lundgren, V.M. AVIP—Autonomous Vehicles' Interaction with Pedestrians—An Investigation of Pedestrian-Driver Communication and Development of a Vehicle External Interface. Master's Thesis, Chalmers University of Technology, Gothenburg, Sweden, 2016.
63. Hudson, C.R.; Deb, S.; Carruth, D.W.; McGinley, J.; Frey, D. Pedestrian Perception of Autonomous Vehicles with External Interacting Features. In *Advances in Human Factors and Systems Interaction*; Nunes, I.L., Ed.; Springer International Publishing: Cham, Switzerland, 2019; pp. 33–39, ISBN 978-3-319-94333-6.
64. Hensch, A.-C.; Neumann, I.; Beggiato, M.; Halama, J.; Krems, J.F. Effects of a light-based communication approach as an external HMI for Automated Vehicles—A Wizard-of-Oz Study. *ToTS* **2020**, *10*, 18–32. [CrossRef]
65. Faas, S.M.; Mattes, S.; Kao, A.C.; Baumann, M. Efficient Paradigm to Measure Street-Crossing Onset Time of Pedestrians in Video-Based Interactions with Vehicles. *Information* **2020**, *11*, 360. [CrossRef]
66. Faas, S.M.; Mathis, L.-A.; Baumann, M. External HMI for self-driving vehicles: Which information shall be displayed? *Transp. Res. Part F Traffic Psychol. Behav.* **2020**, *68*, 171–186. [CrossRef]
67. de Clercq, K.; Dietrich, A.; Núñez Velasco, J.P.; de Winter, J.; Happee, R. External Human-Machine Interfaces on Automated Vehicles: Effects on Pedestrian Crossing Decisions. *Hum. Factors* **2019**, *61*, 1353–1370. [CrossRef]
68. Fuest, T.; Feierle, A.; Schmidt, E.; Bengler, K. Effects of Marking Automated Vehicles on Human Drivers on Highways. *Information* **2020**, *11*, 286. [CrossRef]
69. Kaß, C.; Schoch, S.; Naujoks, F.; Hergeth, S.; Keinath, A.; Neukum, A.A. Methodological Approach to Determine the Benefits of External HMI During Interactions Between Cyclists and Automated Vehicles: A Bicycle Simulator Study. In *HCI in Mobility, Transport, and Automotive Systems. Driving Behavior, Urban and Smart Mobility*; Krömker, H., Ed.; Springer International Publishing: Cham, Switzerland, 2020; pp. 211–227, ISBN 978-3-030-50536-3.
70. Joisten, P.; Schwindt, S.S.; Theobald, N.; Abendroth, B. Pedestrians' Mental Model Development after Initial Encounters with Automated Driving Systems. In Proceedings of the 33rd European Conference on Cognitive Ergonomics, Kaiserslautern, Germany, 4–7 October 2022; Ebert, A., Lachmann, T., Dreßler, K., Lindblom, J., Reinhard, R., Eds.; ACM: New York, NY, USA, 2022; pp. 1–4.
71. Weiß, S.L.; Eisele, D.; Petzoldt, T. External Human-Machine-Interfaces on Automated Vehicles: Which message and perspective do pedestrians in crossing situations understand best? In *Intelligent Human Systems Integration (IHSI 2022): Integrating People and Intelligent Systems*; Ahram, T., Karwowski, W., Di Bucchianico, P., Taiar, R., Casarotto, L., Costa, P., Eds.; AHFE International: Venice, Italy, 2022. [CrossRef]
72. Rettenmaier, M.; Pietsch, M.; Schmidler, J.; Bengler, K. Passing through the Bottleneck—The Potential of External Human-Machine Interfaces. In Proceedings of the 2019 IEEE Intelligent Vehicles Symposium (IV), Paris, France, 9–12 June 2019; IEEE: New York, NY, USA, 2019; pp. 1687–1692.
73. Wilbrink, M.; Nuttelmann, M.; Oehl, M. Scaling up Automated Vehicles' eHMI Communication Designs to Interactions with Multiple Pedestrians—Putting eHMIs to the Test. In Proceedings of the 13th International Conference on Automotive User Interfaces and Interactive Vehicular Applications, Leeds, UK, 9–14 September 2021; ACM: New York, NY, USA, 2021; pp. 119–122.
74. Joisten, P.; Theobald, N.; Abendroth, B. Pedestrians' Crossing Decisions While Interacting with Automated Vehicles—Insights from a Longitudinal Study. In *Human Factors in Transportation, Proceedings of the 13th International Conference on Applied Human Factors and Ergonomics (AHFE 2022)*, New York, NY, USA, 24–28 July 2022; AHFE International: New York, NY, USA; p. 2022.
75. Colley, M.; Bajrovic, E.; Rukzio, E. Effects of Pedestrian Behavior, Time Pressure, and Repeated Exposure on Crossing Decisions in Front of Automated Vehicles Equipped with External Communication. In Proceedings of the CHI Conference on Human Factors in Computing Systems, New Orleans, LA, USA, 29 April–5 May 2022; Barbosa, S., Lampe, C., Appert, C., Shamma, D.A., Drucker, S., Williamson, J., Yatani, K., Eds.; ACM: New York, NY, USA, 2022; pp. 1–11.
76. Schieben, A.M.; Wilbrink, M.; Kettwich, C.; Dodiya, J.; Oehl, M.; Weber, F.; Sorokin, L.; Lee, Y.M.; Madigan, R.; Markula, G.; et al. *Testing External HMI Designs for Automated Vehicles—An Overview on User Study Results from the EU Project interACT*; 19. Tagung Automatisiertes Fahren: Munich, Germany, 2019. Available online: <https://mediatum.ub.tum.de/doc/1535145/1535145.pdf> (accessed on 28 June 2023).
77. Parkin, J.; Clark, B.; Clayton, W.; Ricci, M.; Parkhurst, G. Autonomous vehicle interactions in the urban street environment: A research agenda. *Proc. Inst. Civ. Eng.—Munic. Eng.* **2018**, *171*, 15–25. [CrossRef]
78. Björklund, G.M.; Åberg, L. Driver behaviour in intersections: Formal and informal traffic rules. *Transp. Res. Part F Traffic Psychol. Behav.* **2005**, *8*, 239–253. [CrossRef]
79. Imbsweiler, J.; Ruesch, M.; Weinreuter, H.; Puente León, F.; Deml, B. Cooperation behaviour of road users in t-intersections during deadlock situations. *Transp. Res. Part F Traffic Psychol. Behav.* **2018**, *58*, 665–677. [CrossRef]
80. Risser, R. Behavior in traffic conflict situations. *Accid. Anal. Prev.* **1985**, *17*, 179–197. [CrossRef]
81. Xing, Y.; Lv, C.; Cao, D.; Hang, P. Toward human-vehicle collaboration: Review and perspectives on human-centered collaborative automated driving. *Transp. Res. Part C Emerg. Technol.* **2021**, *128*, 103199. [CrossRef]
82. Stoll, T.; Lanzer, M.; Baumann, M. Situational influencing factors on understanding cooperative actions in automated driving. *Transp. Res. Part F Traffic Psychol. Behav.* **2020**, *70*, 223–234. [CrossRef]

83. Li, Y.; Cheng, H.; Zeng, Z.; Liu, H.; Sester, M. Autonomous Vehicles Drive into Shared Spaces: eHMI Design Concept Focusing on Vulnerable Road Users. In Proceedings of the IEEE International Intelligent Transportation Systems Conference (ITSC), Indianapolis, IN, USA, 19–22 September 2021. [CrossRef]
84. Mertens, J.C.; Knies, C.; Diermeyer, F.; Escherle, S.; Kraus, S. The Need for Cooperative Automated Driving. *Electronics* **2020**, *9*, 754. [CrossRef]
85. Matsunaga, N.; Daimon, T.; Yokota, N.; Kitazaki, S. Effect of External Human Machine Interface (eHMI) of Automated Vehicle on Pedestrian's Recognition. *Proc. Int. Disp. Work.* **2019**, *26*. [CrossRef]
86. SAE. *Ground Vehicle Standard: Taxonomy and Definitions for Terms Related to Driving Automation Systems for on-Road Motor Vehicles*; SAE International: Warrendale, PA, USA, 2021; (J3016_202104). Available online: https://saemobilus.sae.org/content/j3016_202104 (accessed on 30 March 2023).
87. Schindler, J.; Herbig, D.L.; Lau, M.; Oehl, M. Communicating Issues in Automated Driving to Surrounding Traffic—How should an Automated Vehicle Communicate a Minimum Risk Maneuver via eHMI and/or dHMI? In *HCI International 2020—Late Breaking Posters*; Stephanidis, C., Antona, M., Ntoa, S., Eds.; Springer International Publishing: Cham, Switzerland, 2020; pp. 619–626, ISBN 978-3-030-60702-9.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

Article

Learnability in Automated Driving (LiAD): Concepts for Applying Learnability Engineering (CALE) Based on Long-Term Learning Effects

Naomi Y. Mbelekani * and Klaus Bengler

School of Engineering and Design, Technical University of Munich, 85748 Munich, Germany; bengler@tum.de

* Correspondence: ny.mbelekani@tum.de

Abstract: Learnability in Automated Driving (LiAD) is a neglected research topic, especially when considering the unpredictable and intricate ways humans learn to interact and use automated driving systems (ADS) over the sequence of time. Moreover, there is a scarcity of publications dedicated to LiAD (specifically extended learnability methods) to guide the scientific paradigm. As a result, this generates scientific discord and, thus, leaves many facets of long-term learning effects associated with automated driving in dire need of significant research courtesy. This, we believe, is a constraint to knowledge discovery on quality interaction design phenomena. In a sense, it is imperative to abstract knowledge on how long-term effects and learning effects may affect (negatively and positively) users' learning and mental models. As well as induce changeable behavioural configurations and performances. In view of that, it may be imperative to examine operational concepts that may help researchers envision future scenarios with automation by assessing users' learning ability, how they learn and what they learn over the sequence of time. As well as constructing a theory of effects (from micro, meso and macro perspectives), which may help profile ergonomic quality design aspects that stand the test of time. As a result, we reviewed the literature on learnability, which we mined for LiAD knowledge discovery from the experience perspective of long-term learning effects. Therefore, the paper offers the reader the resulting discussion points formulated under the Learnability Engineering Life Cycle. For instance, firstly, contextualisation of LiAD with emphasis on extended LiAD. Secondly, conceptualisation and operationalisation of the operational mechanics of LiAD as a concept in ergonomic quality engineering (with an introduction of *Concepts for Applying Learnability Engineering* (CALE) research based on LiAD knowledge discovery). Thirdly, the systemisation of implementable long-term research strategies towards comprehending behaviour modification associated with extended LiAD. As the vehicle industry revolutionises at a rapid pace towards automation and artificially intelligent (AI) systems, this knowledge is useful for illuminating and instructing quality interaction strategies and Quality Automated Driving (QAD).

Keywords: automated vehicles; learnability in automated driving (LiAD); learning effects; concepts for applying learnability engineering (CALE); Quality Automated Driving; framework

1. Introduction

Vehicle automation is evolving at a higher level, with advanced systems set to hit the market in the near future. For example, the level 3 automated driving system (ADS) (L3, according to the SAE J3016_202104 [1]) will be introduced on public roads soon. Thus, in order to comprehend human-automation interactions (HAI) and how humans learn to interact with or use automation over time, there are several aspects that need to be considered. Arguably, automation systems should be designed in such a way that encourages efficiency, satisfaction, and effectiveness, as well as safety-inducing trust, behaviour-based safety, agility, and comfortability over time. User experiences (UX) are important to consider [2], such as the power law of learning, trust, reliance, and acceptance. In the automotive field,

trust is described in terms of automation and reliance in terms of the ADS [3]. According to [4], trust is considered a key element in continuous use and acceptance. Therefore, how well trust matches the true capabilities of automation has an effect on inappropriate reliance, which is associated with misuse and disuse [4,5]. This begs us to question different user characteristics, demographic factors, perceived aims of learning and the context of use [6]. Consequently, assessing different changing aspects of automated driving may result in advance knowledge discovery on users' behaviour adaptability and/or changeability (BAC) towards automation. For instance, assessing whether adaptive cruise control (ACC) induces behavioural adaptation in drivers (see [7]). When considering change in behaviour, Ref. [8] hypothesised that "change will be for the better and that performance will be more effective and decisions will be improved". Moreover, it can be argued that states of 'negative hedonicity' ("which is frustration due to user-hostile aspects of the interface and surprise when the automation does things the person does not understand") can be induced by intervention [8]. For instance, see Table 1 for other aspects to consider.

Table 1. Quality aspects.

Quality Aspects	Characteristics
Automated system aspects	Hardware and software design, transparency, limitations and capabilities, positive pragmaticity (easy-to-learn, easy-to-understand, easy-to-use, efficient, organised, straight-forward, comprehensive, understandable, useful, convenient, etc.), negative pragmaticity (hard-to-learn, difficult-to-understand, hard-to-use, ineffective, time-consuming, incomprehensible, irrelevant, poor quality, etc.), positive hedonicity (innovative, engaging, creative, inviting, fun, attractive, clean, etc.), negative hedonicity (overwhelming, frustrating, unattractive, complex, intimidating, etc.).
Subjective aspects	User types, age, gender, attitudes, emotional and cognitive states, personality, skills, user experience (quality of domain knowledge, level of experience with computers, level of experience with interface, and experience with similar software), intentions, trust, etc.
Internal aspects	Health, fatigue, drowsiness, mental fog, anxiety, stress, etc.
Learning aspects	Users' knowledge-seeking process, the learning process and the result of the learning process, the speed and quality of the skill development, type and sequence of effects as a result of their learning abilities.
External aspects	Distractions, the presence of external stress, weather conditions, road type, environmental factors, the difference between structured and unstructured ubiquitous settings, etc.
Social aspects	Individual and social learning, context of use, content-driven and process-driven approaches, isolated and integrated use, intra-active and interactive, and fixed tools and unfixed devices, etc.

As users are continuously exposed to automation, they learn new ways to interact and use it. Thus, their behaviours, mental models/user models, expectations and preferences evolve with time. This is grounded on a process of reengineering or redesigning their own preferred interaction configurations with automation—keeping in mind their personalised goals. For instance, this has the potential to result in 'resilient misbehaviours' or escalated misuses based on 'destructive learning effects' and misdirected/misapplied trust as a result of an undesirable change in behaviour towards automation. When discussing users' misbehaviours with ADS, one expert stated that "they [users] learn the speed, they learn at speed—how fast do I have to act, how often does it flash, how fast do I have to look up, and they actually get a little faster as they learn, and then that makes them feel more comfortable" [5]. Thus, as long as the automated system affords the user the opportunity to misuse it, this may be stimulus enough for them to continue the misbehaviours. Some undesirable use behaviours are due to resounding subjective expectations, peculiar philosophies on the systems, the system not being transparent and understandable enough, easily modified system features, and functions being erroneous, among others. Therefore, users' subjective expectations may change based on system mechanisms unaccounted for. In addition, this may influence overall performance, with 'unsatisfied' being the leading cause

of procedural inaccuracies, in part by users. Part of the issue is that users are presented with the power law of ‘learning and relearning’ new ways to interact and use automation systems. This then influences them to ‘redesign’ their own personalised interaction patterns with the system. For example, as the user matures in experience with the system, new branches of ideas are formed and compartmentalised in their mental model. Usually, these personalised interaction/use patterns are outside the system’s design scope, as we have seen with various YouTube videos on Tesla autopilot, for example.

As a result, there is a dire need for long-term research that assesses extended learnability. Essentially, considers the easiness with which users learn to interact/use automation, as well as how they behave and misbehave over time. This then requires quality engineering researchers to assess the quality of the designed interaction strategy. Moreover, how users ‘easily learn’ to redesign their own interaction patterns with automation. As well as how the automated system affords users specific learning configurations and behaviours, which is highly neglected in research. Principally, the automation affordances are fundamental to understanding users’ BAC, apathy, and inclinations. Therefore, it is important to assess and comprehend how the user gradually learns new abilities and the influence of passive or active learning. As such, pioneering data on long-term learning effects and behaviour modification towards automation over time may advance knowledge discovery in designing Quality Automated Driving (QAD) experiences to extraordinary heights. In view of that, it is imperative that researchers derive research strategies that link theory (e.g., on learning) and practise (e.g., behaviour modification). This was also argued by [9] as important in achieving optimal usability. Additionally, Ref. [10] reasoned that it is important to choose the right set of metrics and questions when it comes to measuring UX towards a system design. For example, using the following metrics: learnability, time/effort required to achieve maximum efficiency, measuring previous efficiency metrics over time, as well as behavioural (verbal and nonverbal behaviours) and physiological metrics [10].

Due to the anticipated higher automation design prototypes that are set to appear on the market in the imminent future, this raises initiatives for useful long-term research that assists in developing safely usable systems and human–machine interfaces (HMIs). Thus, designs consistently limit misbehaviours, destructive learning effects, irresponsible usages, unexpected/undesirable behaviours, intended/unintended misuses, and incorrect mental models. Accordingly, inspires interoperability across different original equipment manufacturer (OEM) products.

1.1. Motivation

This paper aims to fill the knowledge gap explicitly on extended learnability and automated driving. Additionally, it should not be overshadowed by general literature on technological systems and learning dynamics. Consequently, we argue that not all technological systems will be experienced or learned the same, and different technologies or automated systems will induce different learning effects and user behaviour outcomes. Thus, in the collection of reviews, the aim was to consider the extended learnability of automated driving as an operative concept. Moreover, comprehends how it is operationalised and defined rather than addressing learnability in the general sense of technology. Learning, as a concept, can be seen as merely the user learning process towards a system; however, with learnability, there is an underlying layer to that understanding. For example, the affordances the system design provides to users for “easy” learning (Figure 1). The most visible attempts to theorise ADS draw on the idea of affordance. Affordance was described by Gibson (1979) in the field of ecological psychology as “the *affordances* of the environment are what it *offers* the animal, what it *provides* or *furnishes*, either for good or ill” [11]. Thus, the *affordances* of an automated system are what it *offers* the user and what it *provides* or *furnishes*, either for good (e.g., easy-to-learn, safe to use) or bad (e.g., hard-to-learn, risky to use).

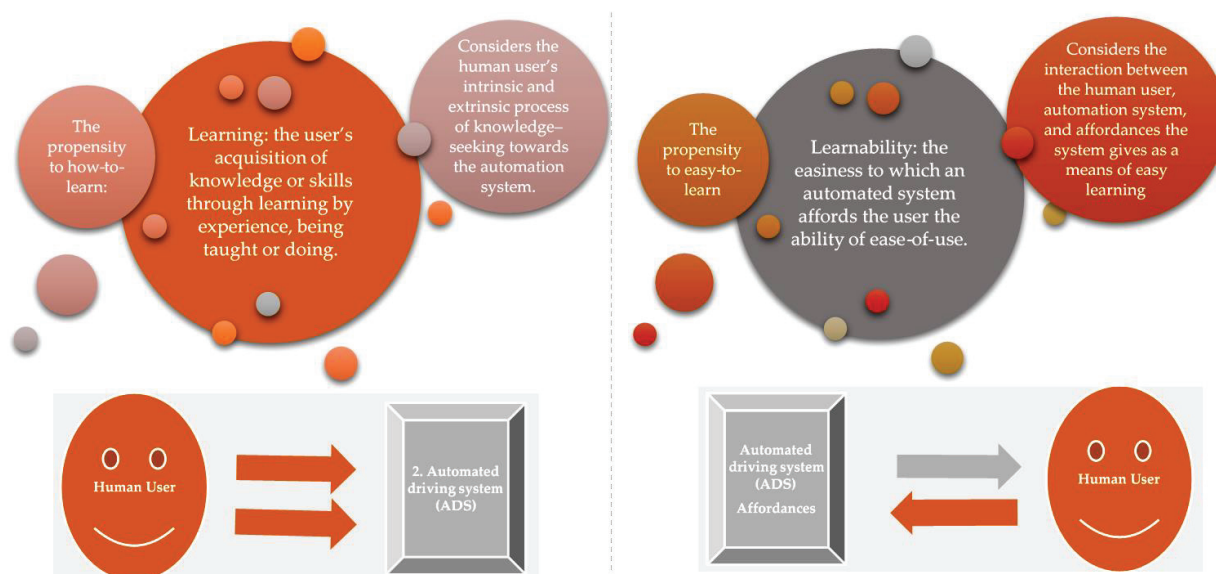


Figure 1. Theorising learning and learnability in terms of affordances.

We could not find sufficient publications suitable for this review. The idea was to examine the papers using thematic analysis, a core method for qualitative research. The trustworthiness of this type of analysis is greatly determined by precision, consistency and exhaustiveness in the data analysis. The papers would be screened individually, and the most relevant references related to the research question would be extracted into a universally applied form with several items that serve as preliminary codes, which include category, type of study, type of ADS used/discussed, and guiding the concept extended learnability. The papers are coded according to the data collected in each of the items of the form. The initial codes are based on a review of the literature and adjusted throughout the coding process, where new information emerges. The codes pertaining to extended learnability are further coded according to the purpose and context of use. The content of this data extraction is then the subject of a content analysis, with the presentation of the results being textual, tabular or diagrammatic.

1.2. Contribution

This paper addresses the questions: Why and how should we measure the learnability of ADS? We understand that high-quality learnability induces satisfaction, as users may feel confident in their abilities [12]. Thus, it is important that a system design's task amenability is efficiently simple and clear. In a sense, learnability studies are time and budget-consuming [12]; however, they are useful to lower costs (e.g., for training users) and promote satisfaction. Conversely, there are different theories, models and definitions, and most are contested. Clear and unambiguous definitions are important for both theoretical and practical considerations of any phenomenon, and learnability is no exception. In this paper, we seek to shed light on the concept within the context of automated driving, specifically addressing the following issues.

- Identify core operational concepts in order to consider the contributions, implications and boundaries for researching 'learnability in the context of automated driving',
- Propose a working definition for learnability that underpins long-term research as a way to advocate for the consequence of long-term effects on user behaviour.

To satisfy these objectives, the paper is divided into three sections:

1. Section 1 (Introduction): An introductory discussion of the topic of interest.
2. Section 2 ('What Is Already Known'): A literature overview on learnability is discussed.
3. Section 3 ('What This Paper Adds'): Engineering knowledge that underpins the operationalisation of the concept. Furthermore, we elaborate on how automated

driving engineers and researchers may need to employ extended assessment tools when designing for lifelong behaviour-based safety.

The concept of learnability is widely researched in the human–computer interaction (HCI) field; however, the opposite rings true in the automated driving field. As a result, we hope this paper brings this concept to the forefront of academic and industry research interests. In effect, it encourages more studies, especially pertaining to extended learnability and fundamental accounts that may underpin changes in user behaviour. We thus position this paper as a foundational condition for quality interaction design strategies that consider learnability as a prerequisite for safety. Moreover, highlights a set of challenges for metrics and suggests a class of possible solutionism as a multi-level phenomenon.

2. What Is Already Known?

This section aims to theorise learning abilities and profusely comprehend the translation of learnability theory into practise. Thus, considering the correlation between learning effects and behaviour modification. Therefore, it is significant that we theorise the power law of learning from an intellectual (behavioural and cognitive) point of view. The idea that learning has an impact on mental models is also crucial to consider and should be researched over the sequence of time. Furthermore, it is important to consider the different user types, mental models, HMI and automation system design, environment, as well as learning effects that contribute to users learning abilities and skill acquisitions over time. The observation that learning as the product of interactions between the cognitive activities of users as learners, the social behaviours they see modelled, the system stimuli and the wider environment are imperative.

2.1. Learnability

While there exists an abundance of published work on learnability, they are scattered across HCI and software engineering [13], informatics [14], technical communication [15], and Programmable User Models (PUMS) [16], etc. In a sense, HCI researchers preserve the concept of learnability and have relentlessly popularised it. However, there is neglect in the automated driving field. Learnability is primarily used as a transformative and regulatory concept that guides or regulates the easiness of learning a system. From understanding individual human reactions to automation systems or HMIs to system design and understanding of reaction and recovery to the understanding of learnability in system functioning, to AI/PUMS and understanding task-action mappings, among others. Most authors consider learnability in terms of errors, task time, error recovery time, and time to mastery. Author [9] described learnability as covering the time to mastery and error avoidance or recovery. Ref. [17] evaluated learnability by counting errors and measuring the time that is needed to perform a task as well as the time that is needed for error recovery. Author [18] wrote that learnability could be operationalised as the time that a new user needs to reach a predefined level of proficiency. Moreover, scholars argue that learnability concerns the system features that allow users to understand how to use it and then how to attain a maximal level of performance. Essentially, learnability literature has long focused on comprehensibility. In short, if a user cannot understand it, then the user cannot learn it and furthermore cannot use it (or rather use it correctly). Thus, comprehensibility influences learnability, as well as learnability yields, augmented comprehensibility of a system design. With reference to [13], they explained learnability as “allowing users to reach a reasonable level of usage proficiency within a short time” by assessing novice users and measuring the time it takes them to learn the basic tasks of the system. For instance, the effort required for a novice user to perform basic tasks. Therefore, the time and effort required to be able to perform specified functionalities. Author [19] saw learnability as allowing users to quickly begin to work with the system. Most scholars state that an automated system should be easy to learn by the type of users for whom it is intended. However, a major concern is the lack of longitudinal evaluations of learnability that further evaluate the user types’ learning ability over time. Ref. [13] focused on different user types

as a way to comprehend how different users may experience an automated system. For example, unexperienced (novice) users as well as experienced users.

Generic learnability is understood as “easy-to-learn” [20] and described as a quality attribute of usability, while generic usability is understood as “easy-to-use” [21]. Thus, there exists “a need to determine if learnability and usability are mutually exclusive” [22]. In a sense, to comprehend learnability, we must first comprehend usability (see Figure 2). The International Organization for Standardization (ISO) has developed a variety of models to identify and measure software usability [1]. The term usability refers to “a set of multiple concepts, such as execution time, performance, user satisfaction and ease of learning (“learnability”), taken together”. However, the researchers or by the standardization bodies have yet to define it homogeneously [23]. Usability is defined (ISO 9241-11 cited in [23]) as “the extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency, and satisfaction in a specified context of use”. Usability is dependent on the context of use, users, tasks, equipment, and environment (ISO 9241-11, 1991 cited in [23]), and it is regarded as a quality attribute that assesses how easy systems are to be used and methods for improving ease-of-use [24].

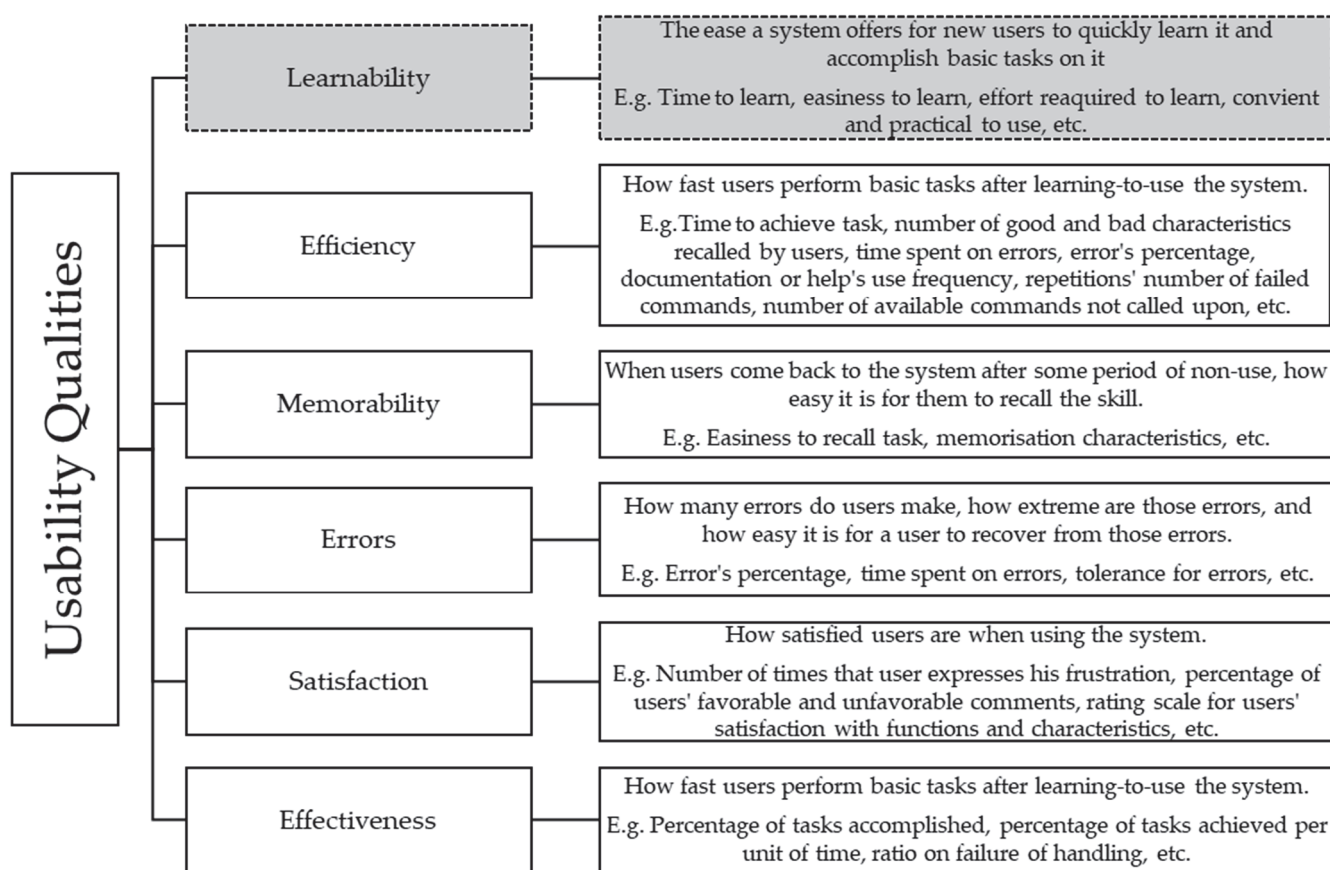


Figure 2. A baseline model of usability qualities.

Learnability as an aspect of usability can be traced back to 1976 [2]. However, the first substantial attempt to describe it was by [22], denoting that “a system should be easy-to-learn by the class of users for whom it is intended”. This description is rather vague, as the idea that a system should be “easy-to-learn” leaves room for many interpretations. It emphasises the need to evaluate learnability while considering the level of user experience with a system [25]. Undeniably, learnability is a significant component of usability and use satisfaction, with some classifying it as the most fundamental attribute [23]. Learnability is a quality attribute that evaluates ‘ease-of-learning’ and assesses user efficiency and effectiveness in carrying out specific tasks. ISO described learnability as the extent to which

a system can be easily learned by specific users to achieve task-specific goals efficiently, effectively, and with a high level of satisfaction. The quality-in-use (QinU) directly relates to the capability of a system to influence users. In addition, holds five attributes: effectiveness, productivity or efficiency, safety or free of risk, satisfaction, and context of use coverage (ISO 25010:2011 cited in [23]). Interaction quality is achieved when users meet operational tasks in a safe and comprehensive manner. In that, learnability is characterised by the capability of a system to enable the user to learn how to use it within a reasonable amount of time [26]. Ref. [26] “proposes an action-based technique to improve users’ learnability and understandability by studying users’ interactions while they are interacting with the user interface of a Web application”. This is because the quality of interactions directly relates and reflects the user’s knowledge and understanding of the system [26]. As a result, “the system should be easy to learn” [22]. System quality is described as the degree to which the system satisfies the intended needs of users in a valuable manner (ISO/IEC 25010:2011). Ref. [13] argues that there is generally no accepted operational definition of learnability; however, it describes it as the quality of being easily learnable and as a quality attribute that assesses how easy a system is to learn, as well as deduces methods for improving ease-of-learning. Even so, some common features that are inherently related to it can be known:

- Learnability is a function of the user’s experience, i.e., it depends on the type of user for which the learning occurs;
- Learnability can be evaluated on a single usage period (initial learnability) or after several usages (extended learnability);
- Learnability measures the performance of user interaction in terms of completion times, error and success rates or percentage of functionality understood.

The majority of learnability definitions are framed as initial or extended experience (Table 2). For example, it is categorised as first-time user performance, first-time performance after instruction, easy-to-use, easy-to-learn, the ability to master a system, a change in performance over time and the ability to remember skills over time [13].

Table 2. Learnability descriptions.

Scope	Description	Reference
<i>Initial learnability</i>	The initial learning experiences or initial performance of a user interacting with a system.	
	As a novice user’s initial performance with a system after instruction.	[13]
	“...allows users to reach a reasonable level of usage proficiency within a short time”	[18]
	“Allowing users to rapidly begin to work with the system”	[19]
	“The effort required for a typical user to be able to perform a set of tasks using an interactive system with a predefined level of proficiency”	[27]
<i>Extended learnability</i>	“The time it takes members of the user community to learn how to use the commands relevant to a set of tasks”	[25]
	The changes of performances over the sequence of time, and after several interactions with the system.	
	A user’s performance after they have become familiar with the system.	[13]
	“The ease at which new users can begin effective interaction and achieve maximal performance”,	[25]
	“Initial user performance based on self-instruction [allowing] experienced users to select an alternate model that involved fewer screens or keystrokes”	[9]
	“Minimally useful with no formal training, and should be possible to master the software”,	[28]
	The “quality of use for users over time”	[21]

Essentially, there is no accepted or agreed model of learnability. Moreover, “despite the consensus that learnability is an important aspect of usability, there is little consensus among scholars as to how learnability should be defined, evaluated, and improved” [13].

Primarily, there are several groups of scholars with different views concerning learning experiences; for example,

- (1) Group A consider ease-of-learning to be a sub-concept of ease-of-use.
- (2) Group B sees ease-of-learning and ease-of-use as competing attributes that can seldom be fulfilled at the same time.
- (3) Group C sees ease-of-learning as an attribute that covers the whole usage process (e.g., a system can be easy-to-learn for beginners; however, for experienced users, it may constantly assist users in finding new and more ways to use).

Human factors/ergonomics scholars have argued that the notorious frustrations and failures triggered by software interventions [8,29–31] have led to significant concerns with evaluation processes [8]. Thus, usability methods are considered to be useful methodological options [18,19,25,32–36]. Moreover, additional consideration has been metrics for ‘key performance indicators’ [37,38] and assessing performance in terms of satisfaction, efficiency and effectiveness [8]. Furthermore, other performance methods are self-assessment questionnaires and, in real-time, using user profiling and experience sampling methods (ESM). Furthermore, think-out-loud (TOL) methods for eliciting verbalizations related to thinking and problem solving [39]. Essentially, TOL verbal protocol techniques have wide applicability in cognitive psychology, and “have proved to be useful techniques for probing possible cognitive mechanisms underlying user performance and problems” [39]. The following table (Table 3) summarises some research approaches that have been used to assess learnability. As much as these are influential to consider, they were, however, tailored towards HCI and not necessarily automated driving. Granting, this is not to assert that there are no studies in the automated driving domain (see [24,40–44]); however, most of these studies do not focus on learnability as a concept but rather learning, with the exception of [45]. In a sense, there is little, if not none, that specifically focuses on how to define learnability in the context of automated driving, and especially extended learnability.

Therefore, in this paper, the aim is to shed light on extended learnability and reason potential factors that impact the concept and its effects on user BAC. These topics of interest are usually not approached in a clear and conscious manner. The aim is to comprehend research approaches that can be implemented, as well as evaluation metrics within the context of automated driving. Extended learnability is yet to occupy a high point of reference in research, as measuring it does not have a standardised set of strategies.

Table 3. Learnability research approaches tailored in HCI.

Methods/Parameters	Code	Factors	References
Research Methods			
<i>Empirical approach: Strategies based on data collection</i>	E1	In a lab experiment and survey method: Perform tasks (with a system), and after answering the questionnaire, formative/summative evaluation, and record completion time.	[9,13,46]
	E2	In a lab experiment and think-aloud method, Novice users interact with the system and vocalise their thoughts. The tutor sits next to them for questions during system use.	[47–49]
	E3	Naturalistic study and diary method: Interact with the system in their natural habitat and fill diary entries to keep all records of all the learning activities.	
	E4	Interviews: Cover subjective data and also for missing or incomplete study entries.	
	E5	In a lab experiment and observations: Trained and during test tasks, they are observed on performance. Summative evaluation on performing a task and then evaluated.	[50]
	E7	Online survey: Online questionnaire evaluation with questions related to learnability characteristics, task match, and interface understandability.	[51]

Table 3. Cont.

Methods/Parameters	Code	Factors	References
Assessment Metrics (adapted from [21])			
<i>Task Metrics:</i> Metrics based on task performance	T1	Percentage of users who complete a task optimally.	[52]
	T2	Percentage of users who complete a task without any help.	[52]
	T3	Ability to complete tasks optimally after a certain timeframe.	[9]
	T4	Decrease in task errors made over certain time intervals.	[22]
	T5	Time until the user completes a certain task successfully.	[18]
	T6	Time until the user completes a set of tasks within a timeframe.	[18]
	T7	Quality of work performed during a task, as scored by judges.	[50]
<i>Command Metrics:</i> Metrics based on command usage	C1	Success rate of commands after being trained.	[20]
	C2	Increase in commands used over certain time intervals.	[18]
	C3	Increase in complexity of commands over a time interval.	[18]
	C4	Percent of commands known to the user.	[53]
	C5	Percent of commands used by user.	[53]
<i>Mental Metrics:</i> Metrics based on cognitive processes	M1	Decrease in average think times over certain time intervals.	[18]
	M2	Alpha vs. beta waves in EEG patterns during usage.	[14]
	M3	Change in chunk size over time.	[27]
	M4	Mental Model questionnaire pre-test and post-test results.	[54]
<i>Subjective Metrics:</i> Metrics based on user feedback	S1	Number of learnability-related user comments.	[18]
	S2	Learnability questionnaire responses.	[42,46]
	S3	Twenty-six Likert statements.	[6]
<i>Documentation Metrics:</i> Metrics based on documentation usage	D1	Decrease in help commands used over certain time intervals.	[18]
	D2	Time taken to review documentation until starting a task.	[18]
	D3	Time to complete a task after reviewing documentation.	[18]
<i>Usability Metrics:</i> Metrics based on change in usability	U1	Comparing “quality of use” over time.	[21]
	U2	Comparing “usability” for novice and expert users.	[21]
<i>Rule Metrics:</i> Metrics based on specific rules	R1	Number of rules required to describe the system.	[16,55]

2.2. Theories of Learning and Behaviour

The normative possibilities of learning lie in understanding the multidimensionality of issues and experimenting with new approaches to mould accurate behaviour (responsible usage) and mitigate misbehaviours (irresponsible usage). The science of successful learning (see [56]) based on human learning and memory (see [57]) is important to consider. Our efforts to expand knowledge towards learnability (e.g., in the context of automated driving) can be seen as a form of learning within and about the context of HAI, with the aim of galvanising holistic knowledge discovery. When studying the learning maps of humans concerning automation or AI systems, it is important to explore and expand knowledge on how a human thinks they behave or misbehave in response to the system in a particular situational context. Thus, coming up with theories and questions that examine this outcome is beneficial in understanding learning effects and BAC. Ref. [58] noted that technological evolution has influenced learning theories in that learning theories and technologies are:

- Situated in a somewhat vague conceptual field, as well as connected and intertwined by information processing and knowledge acquisition;
- Learning theories represent a fuzzy mixture of principles and applications.

Thus, bottlenecks of knowledge discovery in how humans learn and the broader information structures that govern knowledge composition and practises are significant to consider. Ref. [59] argues that “instead of regarding the learning theories as discordant, it is rather proposed that human cognition is complex and that there is a role for behaviourist, cognitivist, constructivist and social constructivist models of learning based on the objectives and context of learning”. Theories (Figure 3) are a good foundation for evaluating users’ learning routes.

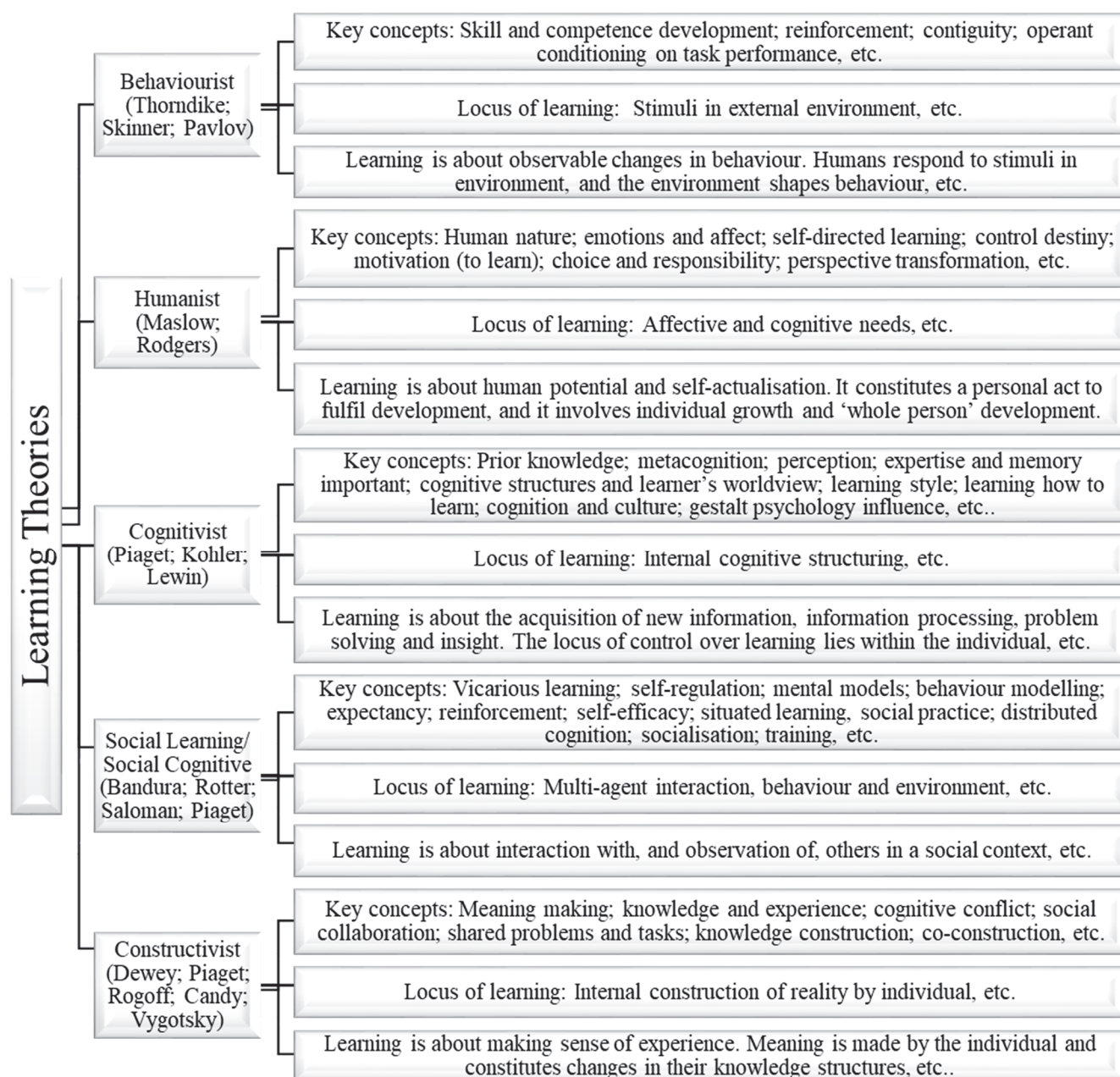


Figure 3. The network of learning theories.

Behaviourism views learning based on external factors and acknowledges that preferred learning can be controlled with reinforcements (rewards) and punishments. Humanistic learning comprises creative self-fulfilment and suggests that the learning process explores the self-reflection of the learner. Information processing views the human learning process as comparable to computer information processing functionality and stresses the path of information through the memory stages: first, the sensory memory, then working

memory for further elaboration and stored in long-term memory due to repetitive, active processing. Thus, design, selection, coordination, use, understanding, and information barriers are key to consider [60]. Cognitive psychology is believed to “have moved further away from its roots in information processing toward a stance that emphasises individual and group construction of knowledge” [17]. The view that “the mind as a computer has fallen into disfavour largely due to the mechanistic representation of a human endeavour and the emphasis on the mind–body separation” is one to also consider [17]. What Churchland cited in [17] has called “popular dualism”, that the “person” or mind is a “ghost in the machine”, is also important to consider. These theories help in comprehending how learnability in the context of automated driving evolves with time, as well as skills and knowledge acquisition. The theories add to cognitive theory by asserting that knowledge is never independent of the human that has the knowledge and the situation where the knowledge was acquired. The user as the learner is seen as an active information processor, and their previous knowledge is seen as the basis for their learning. As well as their ongoing learning influences new forms of knowledge and skills towards automation use and interaction. Thus, the user as a learner continuously adopts new automated driving knowledge and combines it with their prior knowledge to reform their mental model of the automation and driving experience. These learning theories hold an interesting outlook on different user types’ mental processes and mappings, as well as consider the influence of different external factors.

The goal here is knowledge that can be transferred to different situational contexts and user types. In addition, the ‘theory of mental models’ is important as it is useful in evaluating the learning maps from a user’s perspective. For example, two users can be exposed to an automation system, but those two users may go through different learning experiences. As a result, we reason that learnability and mental models are equivalent. Ref. [61] explored mental models and user models and thus argued that “mental models are usually considered to be the ways in which people model processes”. We agree with [61] that mental models may be incomplete and may even be internally inconsistent. Thus, in congruence, researchers need to derive long-term evaluation criteria for assessing user models, in that predicting behaviour modification, user’s performance as a result of learning effects. Ref. [61] further argued that focusing on mental/user models highlights the intentionality of the interaction between the user and the system. However, studying these models can be a challenge as “mental models are inside the user’s head, they are not accessible to direct inspection, and it is difficult to have confidence about how a mental model is constructed or how it can be modified” [61]. Contrary to popular belief, some scholars assert that “there are no mental models per se, but only generalisations from behaviourally conditioned expectations” [61]. Ref. [61] argues that user models can be studied and modified; however, it might be difficult to capture and represent information about the user and the task. The evolution of the mental model, trust and acceptance is essential to assess [62]. Ref. [59] noted that “a stronger orientation in the development and introduction of automated systems, especially considering existing mental models, will increase acceptance and system trust”. According to [63], “the dependability and usability of systems are not isolated constructs”, thus “a clearer picture of how these core attributes are dynamically related remains a vital area of future research”. Conversely, even though there is an enormous amount of research on learning and learning theories, its relationship to extended learnability in the context of automated driving is lacking.

2.3. “X about X”: Linking Learning Effects and BAC

To ensure a strong connection between theory and practise, different scholars have called for ‘linking science’ and employing an engineering analogy to aid in translating theory into practise. The “value of such a bridging function would be its ability to translate relevant aspects of the learning theories into optimal instructional actions” [64]. Thus highlighting potential learning theories and the persistent issues with practical learning, as well as the use of theoretical learning to facilitate solutions for practical learning. The term

learning is used in a variety of topics; thus, it occupies several definitions. For example, such as learning as a child how to speak (vocal-based learning) or to speak a language (language-based learning), learning how to interact with others (social-based learning), learning how to regulate emotions (emotive-based learning), learning how to ride a bicycle (somatic-based learning), and learning chemistry (mental-based learning). Thus, it is not surprising that the concept of learning is a prevalent field in educational psychology and engineering research. There already exist multiple theories of learning constructed by different schools of thought, and some of the theories somewhat overlap with each other. Learning theories grounded on learnability in the context of automated driving may describe the dynamics of learning-to-use an automation system mainly from a pedagogical, cognitive and behavioural point of view. Moreover, it can be argued that the sociotechnical context for learning is “dynamic and makes great demands on those trying to seize the opportunities presented by emerging technologies” [65].

In a sense, learning is described as a change in the user’s mental model or behaviour, resulting in behaviour modification towards an automation system based on continuous knowledge acquisition/information processing. Thus, there is a causal effect relationship between learning and BAC, with learning as a cause and behaviour modification as an effect (i.e., a change in behaviour due to internal and external influences). There is an assumed causal-effect relation between learning and behaviour built into the definition of learning. Learning is defined as “the observable change in behaviour of a specific organism as a consequence of regularities in the environment of that organism” [66]. It is seen as an effect—that is, as an observable change in behaviour that is attributed to an element in the environment. Adaptation is used as an example. To identify that learning has occurred, two conditions are believed to occur: (1) an observation change in behaviour must occur during the lifetime of the entity, and (2) the change in behaviour must be due to regularities in the environment [66]. Ref. [16] further argued that when assessing if learning has occurred, it is important to note that the observed change in behaviour can occur at any point during the lifetime of an organism. The impact of regularity on behaviour might be evident:

- Immediately, or
- Only after a short delay (e.g., one hour), or
- Even after a long delay (e.g., one year) [66].

Moreover, there are differences in how behaviour is defined, as well as what constitutes a change in behaviour. We, thus, adopt a broad definition that includes any observable reactions, regardless of whether that reaction is produced by the somatic nervous system (regulates voluntary physiologic processes, e.g., pressing a pedal), the autonomic nervous system (regulates involuntary physiologic processes, e.g., blood pressure or heart rates), or neural processes (the way the brain works, e.g., thoughts, memories, and feelings). In principle, the concept of behaviour also refers to reactions that are observable only by the individual themselves (e.g., a conscious mental image or thought) [66]. Learning is an occurrence that may go unnoticed in many cases [67]. Thus, in a way, it’s important to question how humans learn, what they learn, and what happens when skills develop [49]. In a sense, when looking closer at learning and knowledge or skills acquisition, it becomes clear that it’s important to assess learning effects based on long-term exposure to automation systems. Moreover, how these effects evolve over time. Thus, embracing future technologies requires a clearer understanding of the pedagogies that inform learning, as well as the learning strategies used to facilitate learning [68]. Additionally, people with different learning patterns (i.e., participatory, collaborative, and independent) approach learning differently [69]. Learning abilities (the ability to learn, characterised by the user’s learning success, i.e., taking into account the efficiency of this process) are different from knowledge acquisition (the quality of the knowledge the user learns, characterised by the speed and quality of the skill development, i.e., taking into account the outcome of the learning process). In acquiring comprehensive knowledge, we thus reflect on skills or skilfulness (as well as deskilling, upskilling, and reskilling) as factors to consider. The lack of studies covering concepts such as deskilling, upskilling, and reskilling as part of

learnability in the context of automated driving limits the opportunity for quality interactions (QAD and QinU), where the strengths and weaknesses of automation and human systems can be considered. Ref. [70] discussed mechanisms of skill acquisition and the law of practise, and considered theoretical and experimental approaches employed in assessing the mechanisms underlying these factors, for instance,

- On the experimental side, “it is argued that a single law, the power law of practise, adequately describes all of the practise data.
- On the theoretical side, “a model of practise rooted in modern cognitive psychology, the chunking theory of learning, is formulated”.

Skill acquisition is divided into three phases that support the idea of conceptual knowledge and procedural skills as being interrelated,

- (1) Cognitive stage: learner stores to memory (memorisation) a set of details relevant to the skill,
- (2) Associative stage: learner forms the details into a procedural model (contains step-by-step instructions for performing a certain action),
- (3) Autonomous phase: the procedure becomes more automated and rapid and, in the end, requires very little processing resources.

Essentially, conceptual knowledge develops into an efficient skill when it is practised. When considering acquiring conceptual information, the ability to learn means acquiring new information and combining it with previous knowledge to form a revised model of automated driving components to be learned. We also have to consider the aspect that learning requires active processing of information on automated driving/automation use to adopt it as a part of the users’ mental model. Concerning the ability to learn, repetition of an action is not enough, but the contextual background needs to be understood as well. Thus, combining conceptual knowledge and procedural skills is important for learning. For example, a user might have an inclination to take over the driving during automated driving mode but will need technical ability to carry it out without enduring risk. Using specific ADS generally requires memorising a lot of automated driving guff of knowledge, handover and takeover information, time delays, and automation HMI action sequences, among others [5]. In a sense, it is much easier for a user to recall details that are related to a thematic entity and connected to existing knowledge than to remember disconnected and non-existing details. Thus, users need to know how to interact with an automated system, how to interpret its actions/reactions, and, as a result, how to manipulate this information to their benefit. For example, interdisciplinary design strategies for HAI encompassing the KISS method (Keep It Simple and Simpler) [71]. This information should be understood on a conceptual level, whether as novice or experienced. Moreover, considering practising procedural abilities and skills acquisition, we deduce that skills must be practised for development in the mental model.

Another issue to consider is that the user may learn to solve only problems that are structured in a specific situation, and when exposed to new situations, they may not know the correct way to react. Thus resulting in explorative learning strategies with a relatively free-formed learning process. This may be the result of some automated driving crash causations in specific situations. The user may not have been fully aware of the high level of risk in that specific situation, for example. To avoid this, automation capabilities and limitations should be apparent to the users in all possible situations and contexts of use. Learning strategies are seen as either explorative learning with unsystematic directives or structured learning with systematic directives. Additionally, passive learning and active learning processes are important to consider. The user thus learns the automated system often in pursuit of knowledge and understanding. For instance, where a learning strategy is not efficient and does not compose thorough knowledge, some users may fill the gaps with their own beliefs. Thus, risk management techniques are crucial, as users may possibly face obstacles and partake in risk-taking behaviours during automated driving mode. Therefore, substantial knowledge may be helpful to recuperate from misjudgement.

3. What This Paper Adds? The Learnability Engineering Life Cycle

This paper introduces the ‘learnability engineering life cycle’, with a special interest in learnability in automated driving (LiAD). The objective of this discussion is to provide an exploration of LiAD and foundational strategies to consider. Based on the above sections (‘What Is Already Known’), our aim is to demonstrably cogitate the learnability of automated driving tasks on the basis of quantifying the quality of performance, learning curve and (negative and positive) learning effects (destructive and constructive), and, implicitly, the behaviour maturation/modification on the basis of experience sampling. The resulting sections contribute to the learnability engineering life cycle (see Figure 4).

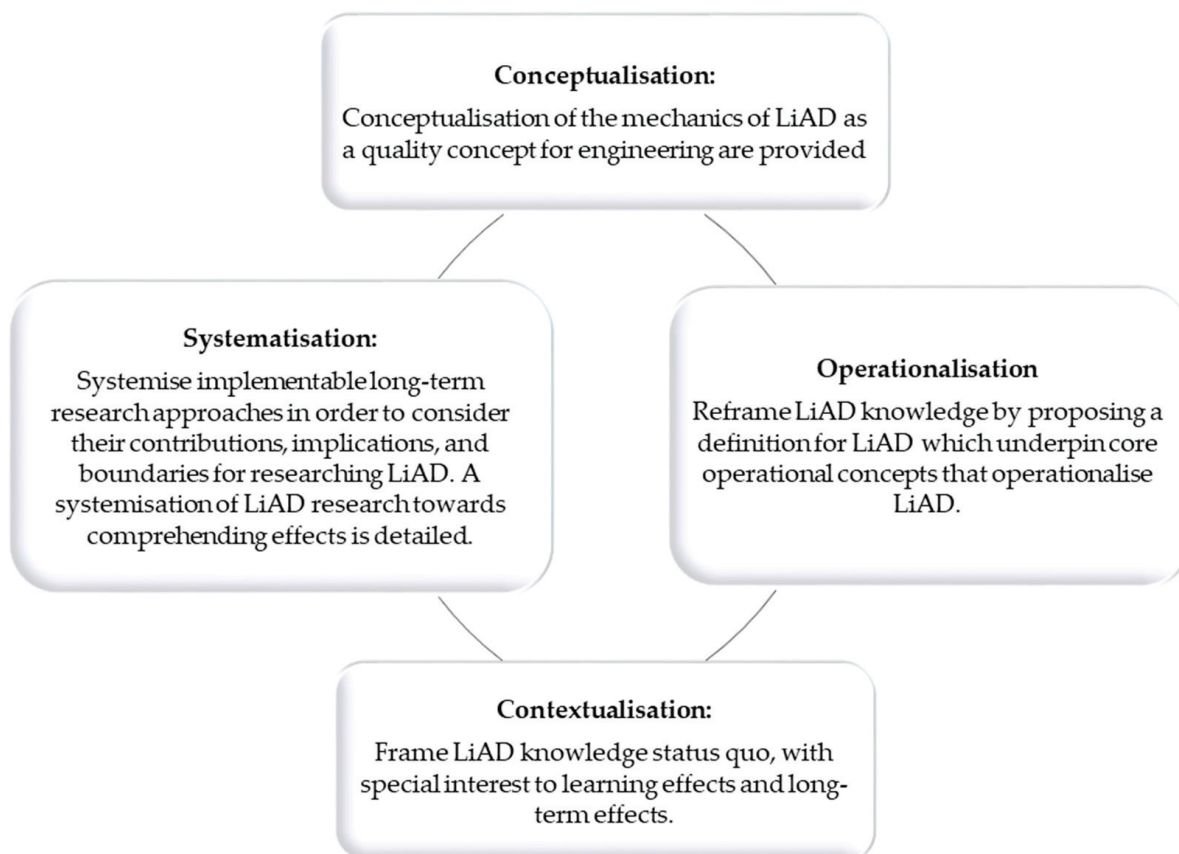


Figure 4. The Learnability Engineering Life Cycle.

These discussion points or mechanisms consider different layers of knowledge discovery in the context of LiAD research. Integrating a wide range of information aids a multidimensional protocol that infers LiAD knowledge configurations.

3.1. Conceptualisation

We consider different mechanisms that influence LiAD over time in order to classify users’ abilities in response to situational dynamics, system performance, contextual elements, environment, and user states. Resultantly highlighting behaviour modification, conditions to change, and new interaction norms, forms, and practises. Essentially, the emphasis is centrally on user/mental models, abilities to learn and change coupled with the transformation of usage patterns and learning effects. Thus, we believe that any integrative view of LiAD should be able to accommodate different factors at all levels of automated driving understanding. Since ADS are such a big part of current driving culture, users carry many deeply held assumptions on how they should be used. Moreover, the operation of automation systems involves nuanced and complex interactions between humans and automation in unpredictable situations, thus proving significant in assessing the quality of learning over time. We consider the concept LiAD, with emphasis on initial (short-term)

learning to extend (long-term) learning curves. We frame LiAD knowledge as a process of transforming quality engineering practices.

- Our conceptualisation:
 - LiAD infers a characteristic of the user's quality of learning, taking into account the easiness of learning-to-use automation systems within the spectrum of short-term or initial learnability (initial LiAD: [i]LiAD) to long-term or extended learnability (extended LiAD: [e]LiAD). It can be measured based on learning curves (considering its slope and its plateau) and behaviour modification that emphasises a scale of desirable (efficient, effective, satisfying, error-free and safety-taking behaviours with the system) and undesirable (inefficient/deficient, ineffective, unsatisfying, error-prone and risk-taking behaviours) behaviours as a result of learning effects/long-term effects.

This conceptualisation has been constructed by considering definitions given by different scholars [13], as well as considering human factors, long-term effects, mental models, etc., from the experience perspectives of different users and environmental factors. Thus, combining these factors into a coherent description is useful in identifying possible learning pitfalls that hinder the competent use of automation. Moreover, our description frames knowledge from the experience perspectives of safety-inducing or risk-inducing learning over time, bearing in mind trust (over-trust, distrust, mistrust, trust calibration) and use (over-use, misuse, disuse) patterns.

Learnability is a key concept in safety research, and learnability engineering is an accepted domain within safety science. It attracts considerable interest, as it advocates for new ways of understanding HAI processes in complex sociotechnical environments. If the aim is towards risk-free use, then there is a strong case for understanding the mechanics of user behaviour. Thus defining risk-taking behaviours, emphasising learned behaviours as well as learning effects. In a sense, "risk is construed as a probabilistic and penalising contingency on a specific class of responses (like driving), and risk-taking as making responses to that class" [72]. We may consider most user behaviours as risky, where only the amount of risk varies [72]. Risk, defined as "the probability of accident involvement, differs between individuals and causes them to have quite different perceptions" [72]. Moreover, risk can be categorised as either physical, cognitive, or emotional. In addition, classifying behaviour as either 'risky-misbehaved' (risk-based misbehaviour) or 'safely-behaved' (safety-based behaviour) during automated driving is important to consider in formulating an in-depth description. As a result, encouraging research towards LiAD and thus addressing the intellectual peculiarity of [e]LiAD is crucial. ISO 9241-11 described learnability using three attributes (effectiveness, efficiency, and satisfaction). However, emphasis should be directed towards the event of a negative validity and undesired effect. For instance, effectiveness vs. ineffectiveness, efficiency vs. inefficiency, and satisfaction vs. dissatisfaction. For each attribute, one or more measures should be presented. It is also important to recognise the significant equivalence (high or low probability) or weight associated with each attribute. Consequently, it is important to classify the characteristics that make for quality LiAD.

3.2. Operationalisation

Fundamentally, learning theories have helped in framing learnability, as well as informed and shaped LiAD as a concept. We describe [i]LiAD and [e]LiAD based on 'Concepts for Applying Learnability Engineering' (CALE) (see Figure 5).

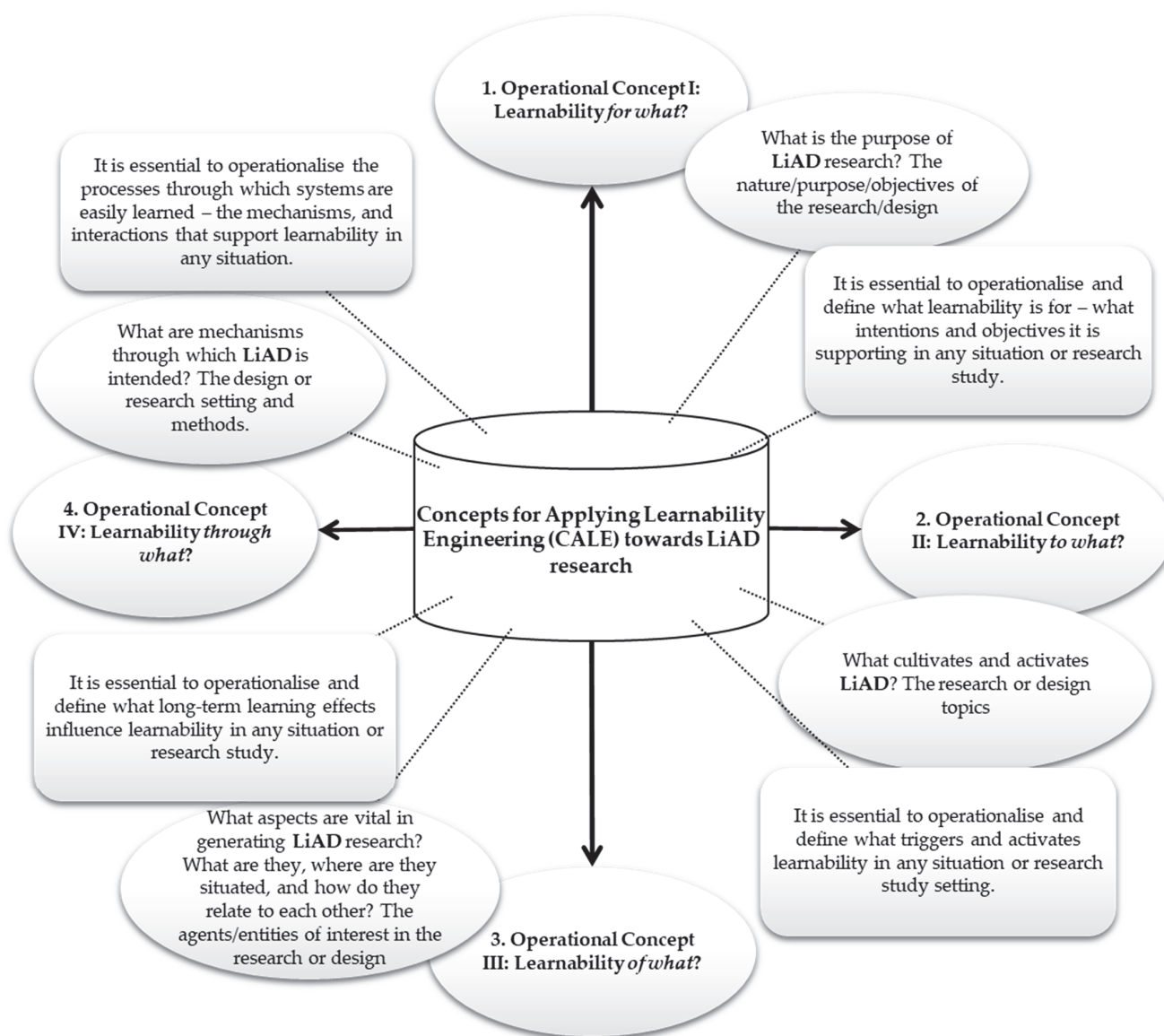


Figure 5. CALE model: Sub-concepts for guiding LiAD research.

- Our operationalisation:
 - CALE aligned to LiAD research considers learnability based on the following: the nature/purpose/objectives of the research or design (for what), the research or design topics (to what), the agents/entities of interest in the research or design (of what) and the design or research setting and methods (through what). These represent important fundamentals for operationalising LiAD research. Scientifically exploring these questions provides rich knowledge about the quality aspects of LiAD.

Thus, LiAD, as a neglected research field, seeks to understand and improve HAI in order to deliver high QAD and desirable BAC. We acknowledge that it is relatively challenging to systematically operationalise LiAD in an immaculate manner, as the concept is unregulated and not prolifically defined in the literature. We, therefore, advocate for using the CALE model to guide LiAD research. As well as to derive assessment scenarios that are equivalent to real-world situations. In the following sub-segments, we will discuss each operational concept, how it is operationalised and why it is important to consider. In this case, LiAD (‘for whom’) concerns an emergent ability of different user types. Achieving the primary objective likely depends on the system’s capacity to be easily learned by

different users. Additionally, channel that information in designing systems that induce efficiency.

3.2.1. Operational Concept I: Learnability for What?

This emphasises the nature/purpose/objectives of the research/design. Investigating LiAD comprises examining the purpose of learnability. An important question, therefore, refers to what research components are important in this process. This might include behavioural and cognitive resources (e.g., competence, epistemic and informational resources, knowledge) or emotional and social resources (e.g., trust, attitude, etc.), etc. These factors and more might be relevant for understanding, reinforcing and maintaining LiAD. Based on the CALE model, we describe the considerations needed to align ‘automated driving-as-imagined’ with ‘automated driving-as-done’, as well as ‘learnability-as-imagined’ with ‘learnability-as-done’. These highlight differences in ‘learnability as intended’ and described based on long-term research, compared to how learnability is performed in practice over the sequence of time. The key concepts to analyse are, therefore, the safety capacity and risk capacity for driving performance, and it is imperative to identify and characterise these capacities. Furthermore, it is grounded in a trilogy process of human factors, environmental factors, and system factors—which represent processes that reinforce or (re)structure learning, evolution, and transformation. Thus, considers research strategies that are open-ended enough to accommodate the application of diverse concepts from different literature lenses and support various methods and approaches over time.

3.2.2. Operational Concept II: Learnability to What?

This emphasises the research or design topics. It is key to understand what reinforces abilities. This advocates for research that integrates ‘what goes well’ and ‘what goes wrong’ as fundamental aspects to consider. This can be broadened to also consider different nuances of understanding, which is consistent with behaviour-based safety and safety science. This depends on the nature, goals and objectives of the learning. Thus, should consider LiAD as characteristically ‘good’ (safety-inducing learnability) or ‘bad’ (risk-inducing learnability) attributes. If safety is easily affected, then the automation system cannot reasonably claim to be able to provide quality-in-learning (QoL), quality-in-use (QoU) or QAD. As a sub-characteristic of usability, it constitutes internal and external qualities. Thus, learnability is the QinL, and this directly relates to the capability of automation to influence the overall user experience. Thus, the QinL holds two qualities: (1) poor learnability and (2) good learnability. These hallmarks (see Figure 6) comprise important blueprints in understanding learning mental maps.

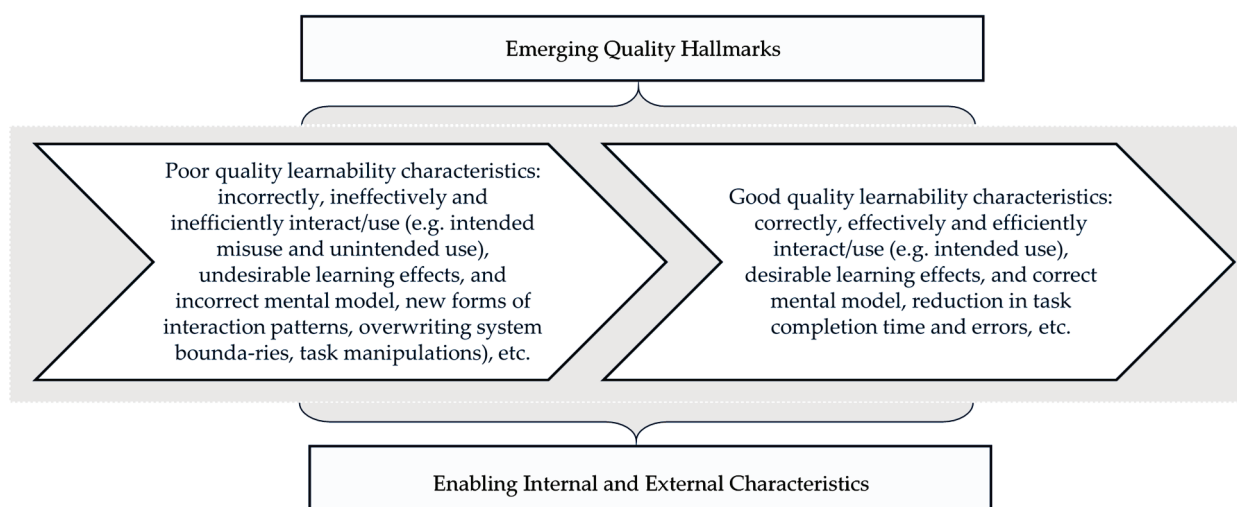


Figure 6. Emerging quality hallmarks and the enabling external and internal factors.

Essentially, good learnability results in efficient performance and plays a vital role in trust and acceptance. Thus, in order to evaluate and improve learnability, “understanding the factors affecting learnability is necessary” [51].

3.2.3. Operational Concept III: Learnability of What?

This emphasises the agents/entities of interest in the research or design. Considerable research investments have been vested in the development of new interaction concepts, system functionalities and features for new automation system designs. However, little research attention has been expended to explore the potential issues associated with user types and learning patterns over time. The aim is to identify and integrate problematic areas (from both human factors and system engineering) and to better understand aspects of design modalities that impede different users’ performance from these design systems with attributes that maximise accurate user behaviour, performance, and epistemology.

3.2.4. Operational Concept IV: Learnability through What?

This emphasises the design or research setting and methods. How LiAD unfolds can take a wide variety of methodological processes. Thus, understanding the nature and the evolution of these processes is important to understanding LiAD. The process should consider the speed of learning (fast vs. slow), the QoL, etc. Moreover, re-learning and un-learning is a process of continuous learning. Unravelling and deciphering these issues via long-term research approaches can help provide a clear understanding of LiAD. It is of importance to evaluate the different learning patterns (as novice and experienced), as we assume that learning takes different forms and that not all learning should be classified as a singular and linear phenomenon.

3.3. Contextualisation

It is apparent that learnability is a complex concept. Ref. [73] argues that “what most of the definitions have in common is that they address the initial usage experience and include a criterion such as effectiveness or efficiency that can be used to measure the learning results”. However, most of these definitions are neglectful of the long-term experience and how human factors, learning effects, and external factors may influence users’ learning curves and behaviour. For example, procedural scenarios and the context of use may trigger different reactions based on various factors (e.g., environmental situations, weather conditions, road types, system functionality, human factors, user states, etc.). In addition, how the system may continuously afford the users new ways to use is important to consider. Moreover, “some researchers have emphasised that the term learnability should also cover expert users’ ability to learn functions that are new to them” [73]. It is probable to deduce long-term knowledge frames from learning effects based on [i]LiAD (a novice user’s initial performance with a system after instruction) and based on [e]LiAD (a user’s performance after continuous exposure to the system). In a sense, we cannot only measure LiAD-based short-term usage by defining a level of ability/efficiency or inability/inefficiency on first-time use because reaching a prolific level of proficiency may take hours, days or years—depending on different internal and external factors. Accordingly, LiAD research encompasses three benchmarks (Figure 7). Thus, the initial learning curve (first-use/exposure) assesses LiAD by defining a level of learning proficiency based on the first encounter, as well as gathering knowledge on the type of learning effects encountered during this short-term period. The transitional learning curve (incline/decline slope of the learning curve) assesses LiAD based on repeated measures, as well as gathering knowledge on the type of learning effects encountered during this mid-term period). The long-term learning curve (the ultimate plateau) assesses LiAD based on extended/lifelong learning, as well as gathering knowledge on the type of learning effects encountered during this period. See Figure 8 for an example of an automated driving experience-time frame of use graph showing changes from moving at one uniform speed/slope to moving at a different uniform speed/slope. Line 1–3 shows the greater speed/slope or steeper

speed/slope, and line 3–4 shows the lower speed/slope. In measuring learnability, it is essential to define the metric, gather the data, and plot the learning effect curve in view of its slope and plateau. However, the number of repeated usages needed to reach saturation may vary from LOA-to-LOA, ADS-to-ADS, HMI-to-HMI, case-to-case, and even user types.

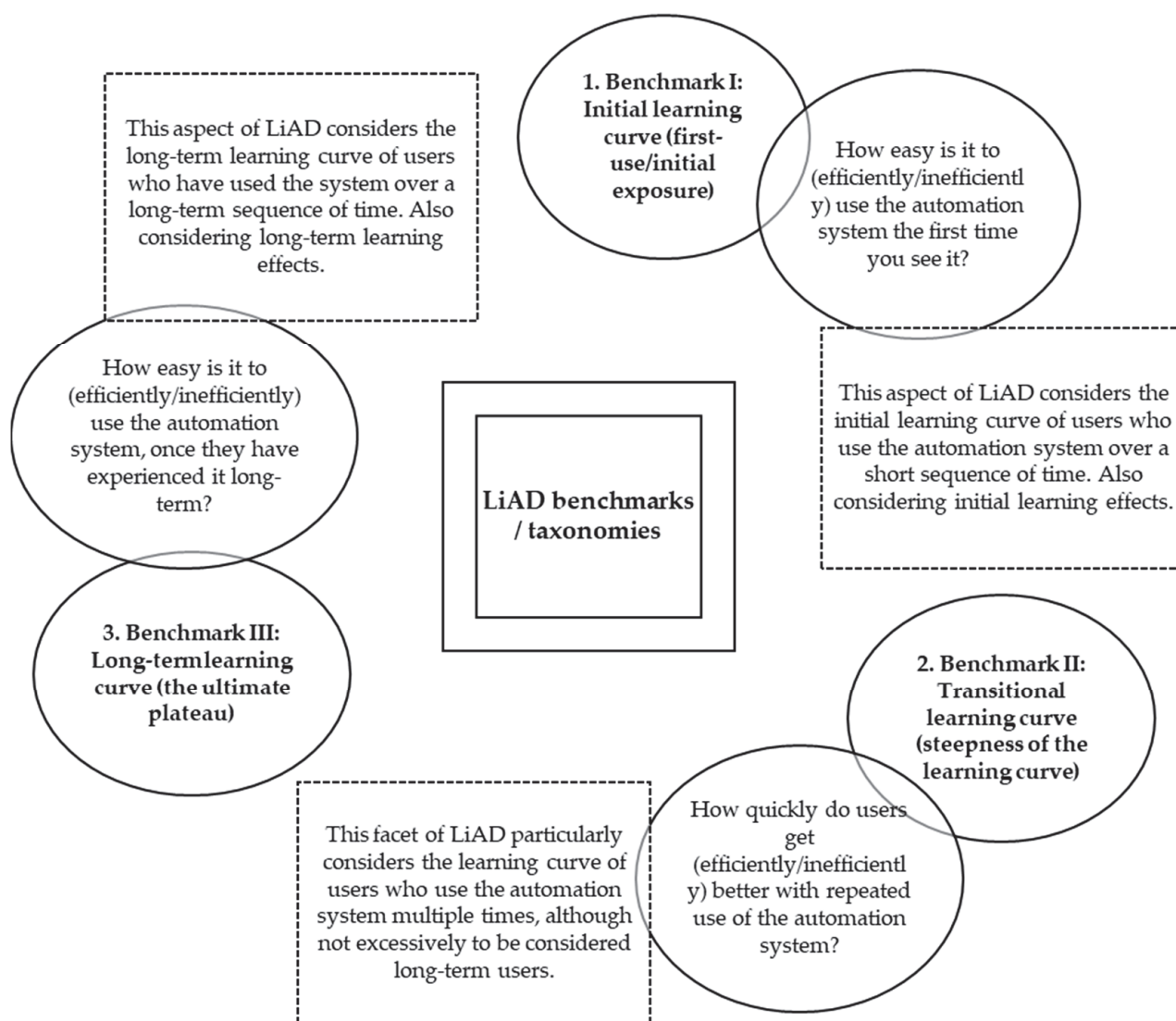


Figure 7. Methodology benchmarks: Quantifying LiAD research.

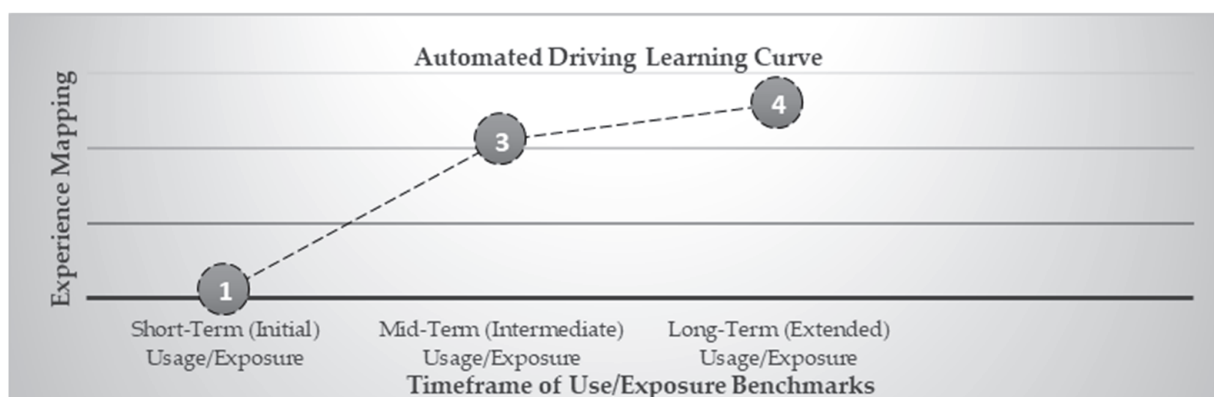


Figure 8. Learning curve: experience-timeframe graph.

Arguably, design decisions affecting the learning curve of a system depend on various factors, such as the target user type, goals of the automated system, existence of time-critical tasks, training, and other system design considerations [27]. It can be assumed that for systems having good learnability, the learning curve rises steeply from the beginning, which means that ability develops quickly over time. For systems having poor learnability, the learning curve may stay on a low level for a certain period of time, which means that ability develops very slowly. Efficiency is usually considered to refer to the right part of the learning curve [27].

$$\text{Slope/Speed} = \frac{\text{Change in learning experience (efficiently, effectively, satisfaction)}}{\text{Change in time benchmarks (initial to extended)}}$$

Assessing or measuring LiAD is important for designing automated systems, as gaining knowledge on how easily users can acclimate or adapt to the system is valuable for even objectively simple systems. Moreover, considering how easy it is for users to accomplish a task and how many repetitions it takes them to become efficient. The aim is to produce a learning curve, which reveals changes in UX and thus identifies how long is long enough for users to reach saturation—a plateau. For example, level 2 automated driving (based on SAE J3016_202104 [1]) with ACC. The aim would be to determine how users with different UX learn to drive with ACC over time or repeated usage. For instance, we recruited user types (with different usage timeframes) to participate in our current long-term study. The participants are repetitively asked to perform automated driving-related tasks, and as a result, we measure the learning curve and learning effects. Therefore, as much as it is important to measure users' task completion time, the learning effects are also important to consider. Thus, it's not only how fast they learn but also how efficiently they learn, as well as what and when they learn. Ref. [25] noted that it is important "to quantify degree of learnability over time and to identify potential learning issues".

Theorising Effects: A Theory of Effects

Notably, Ref. [74] argued, "research on the usability of automated driving HMIs has shown that strong learning effects are at play when novice users interact with the driving automation systems for the first time", as well as "appropriate understanding of driving automation requires repeated interactions". The relative types of effects depend on UX. Novice users may want to learn the system quickly and reach the point of optimal (plateau) performance speedily, but experienced users may have different expectations, for example, tweak the design in different directions. Thus, an ADS may not always be used efficiently using system shortcuts. As users become well versed in the system, the design may feel monotonous and result in negative effects, etc. It is essential to assess learning effects (e.g., destructive/constructive) as a general rule of thumb. These first-order effects have a residual component, where they induce other layers/forms of second-order effects.

In a situation where automation is learned by the user (correctly or incorrectly), this may result in the inducement of different effects of causality (Figure 9). Moreover, first-order effects have the potential to either transform, conform or deform the user to behave or misbehave in specific ways that might not have been anticipated. Thus, this may result in secondary effects (e.g., high insurance due to crash accidents, physical disabilities, etc.). For instance, transformational effects (right information), deformational effects (wrong information), and conformational effects (common information) are important to assess. Even so, we believe it is better for the user to be transformed than deformed, as when the user is transformed, they are able to see the system as it is and use it to its true potential. Long-term and learning effects are fundamental to consider when framing users' behaviour modification (BAC) and abilities toward automation systems over time. It is important to note that these effects come in different formulas or evaluative processes (Figure 10).

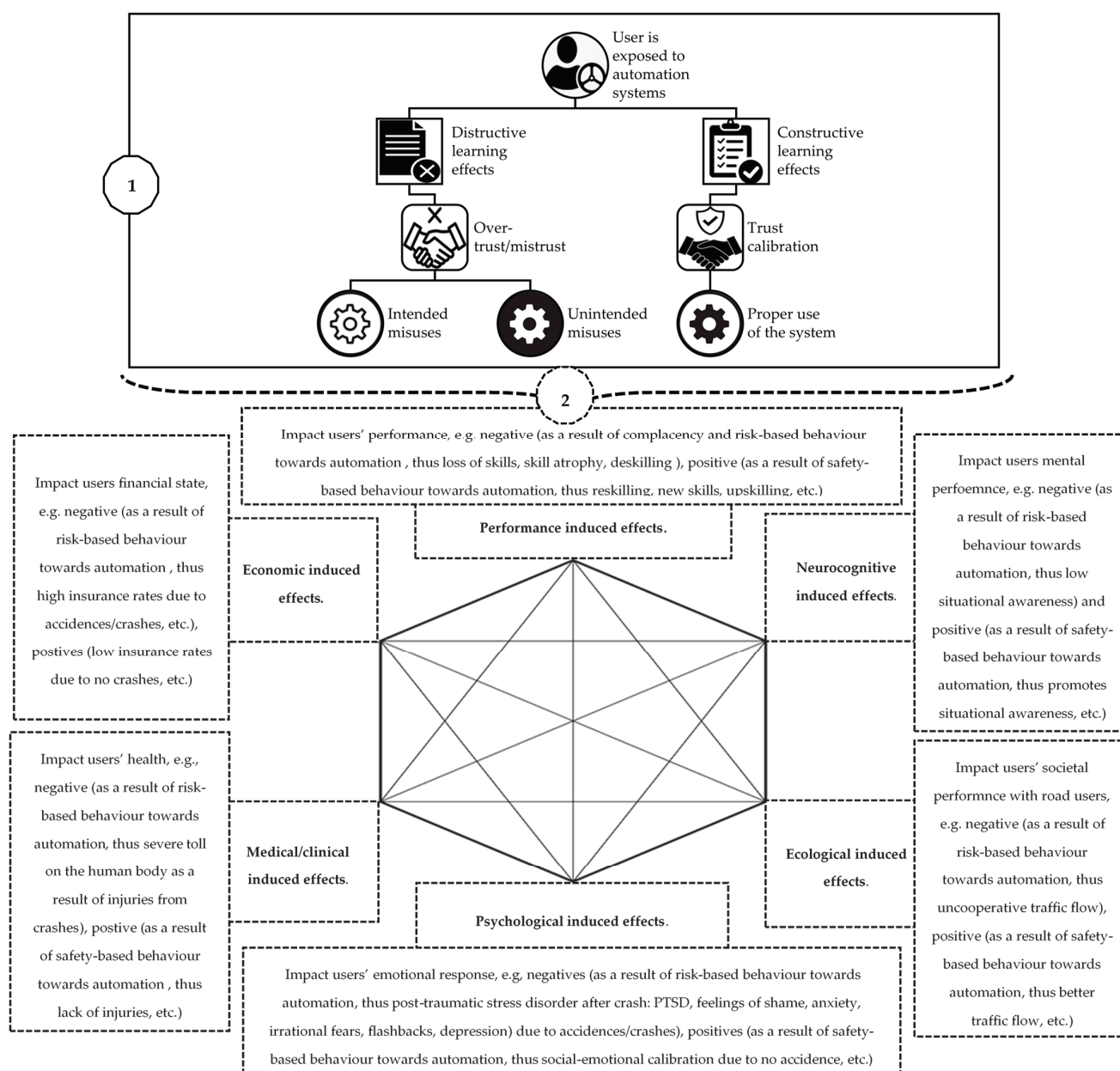


Figure 9. Locus of effects: first-order effects (circled 1), second-order effects (dotted line, circled 2).

In order to comprehend LiAD, we need to consider the procedural effects associated with BAC over time. There is no argument that the aim is to design ADS that users find 'easy-to-learn-to-use' in keeping with safety appearances. However, what is usually neglected is the 'easy-to-learn-to-misuse' (intendedly or unintendedly)—resulting in risk-taking behaviours. The neglected link between ease-of-learning and misuse, as well as its association with risk-taking behaviour, is vital. Therefore, it is important to investigate LiAD based on the etymological 'efficient learning' (i.e., easy-to-learn to use) and 'inefficient learning' (i.e., easy-to-learn to misuse) to fully understand learning effects encountered over time, particularly where the concept 'easy' is the context or standard. The goal is to fill the knowledge gap on learning effects and LiAD, based on the initial and long-term validity of effects, by considering the spectrum of positive (desirable) and negative (undesirable) as a way to manage learning configurations.

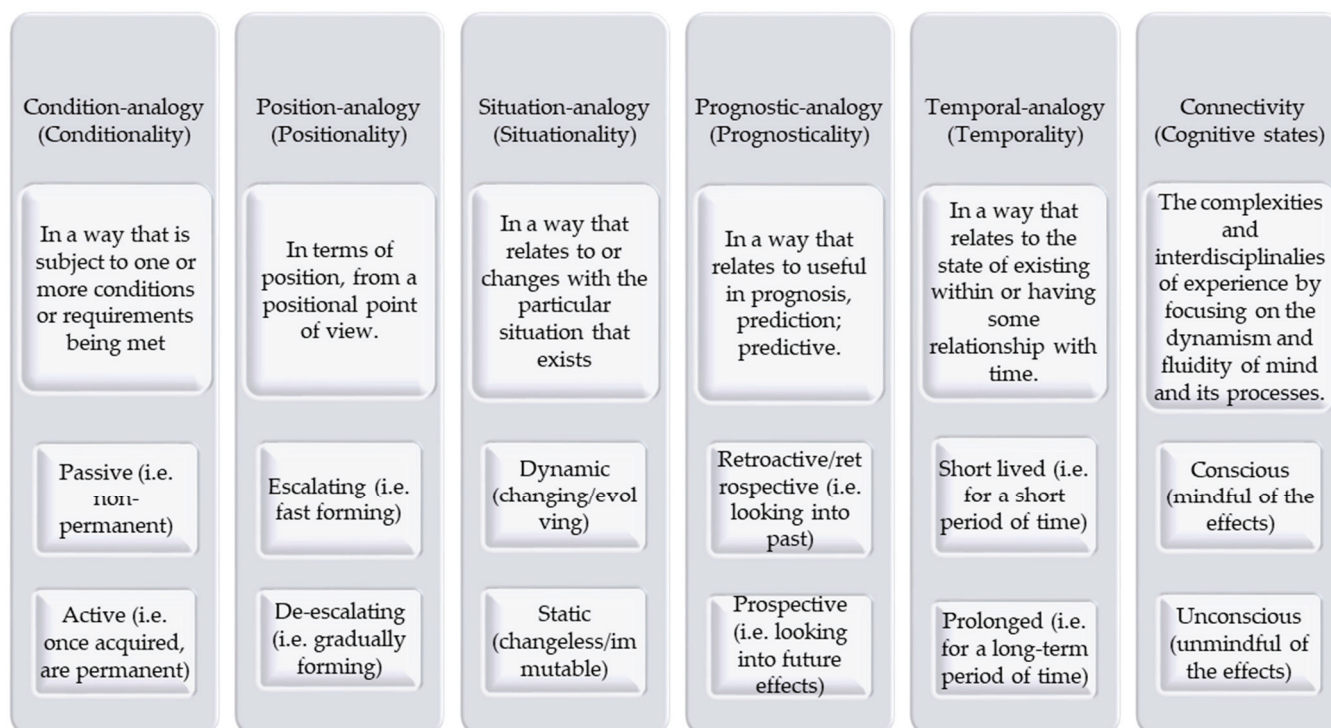


Figure 10. A sequence of regressive and progressive effects.

Thus, in conjunction with learnability (to learn), it's important to consider understandability (to understand), operability (to operate/control), and usability compliance (to adhere/use). In essence, we consider the 'ability to learn', 'the compatibility in learning', 'the meaning in learning', 'the consistency in learning', 'what they learn', 'how they learn', and 'when and where they learn' as interesting questions to consider in the pursuit of LiAD knowledge frontiers. Therefore, long-term research can be a good basis for generating requirements for improving LiAD and quality-based interaction design strategies.

Concerning LiAD, assessing the learning phases of an ACC would be ideal, as in the case of Ref. [42]. Thus, two phases are suggested:

- First phase: The user learns how to operate the system, identifies system limits, and internalises the system functionality, and this learning phase can be influenced by how the automation system is introduced to the user [75].
- Second phase: The "integration phase", the driver integrates the system into the management of the overall operating task via increasing experience in different situations [76].

Ref. [75] argued that the first encounter phase "depends greatly on how intuitive and self-explaining the HMI is". Thus, the learning phase "still depends highly on the HMI, especially in terms of required system input" [75]. Additionally, this corresponds with the compatibility phase. For instance, considering the Stimulus-Response Compatibility (S-RC), which refers to the phenomenon that "the subjects' responses were faster and more accurate for the pairings of stimulus sets and response sets that corresponded naturally than for those that did not" [77]. Essentially, compatibility effects depend on physical and conceptual correspondence, and these effects "occur whenever implicit or explicit spatial relationships exist among stimuli and among responses" [77]. Ref. [77] notes that "by attributing compatibility effects to the cognitive codes used to represent the stimulus and response sets, central cognitive processes must be responsible for the effects". Furthermore, the mediating processes (cognitive processes) generally reflect the user's mental model of the task. The implication of compatibility effects on LiAD is that user performance may be negatively influenced if the information displayed on the HMI and response required is

not compatible. Thus, making the HMI a central part of positively reinforced LiAD. Thus, it is important to consider quality design strategies for human-automation compatibility (HAC). Moreover, the *trust* phase is mainly “characterised by a shift in locus of control” from the user to the automation system, and in the adjustment and readjustment phases, users “adjust their adapted behaviour depending on their experience of (critical) situations and system limitations” [75].

3.4. Systematisation of LiAD and Multi-Level Long-Term Research

Many scholars argue that conducting long-term studies can be highly expensive and tedious, pertaining to budgets or cost, effort, and timelines [10,25,27]. Ref. [27] argued that observing and measuring users over an extended period consumes huge resources, costs and time consumption. Moreover, mentioned that “even if observation could be fully automated (through, for example, the use of event logging), a detailed analysis to study learnability would probably be very exorbitant” [27]. Ref. [25] noted that the major obstacle to evaluating extended learnability is that it traditionally requires expensive and time-consuming techniques for observing users over an extended period of time. Consequently, due to its prohibitive costs, researchers have gradually refrained from performing extended learnability evaluation [27]. Despite all the difficulties associated with long-term studies, we believe that the evaluation of [e]LiAD is valuable and will produce fruitful and insightful results, as the continuous trend of a learning curve is crucial. Thus, better and more useful long-term research strategies should be considered, as this type of research is important for sustaining safety and predicting behaviour over time.

In addition, “too many studies are still conducted in a laboratory situation using traditional methodological paradigms” [78]. Thus, “more learnability studies with new methodological approaches in the natural environments are needed where the human, learning and non-linear process of learnability are in focus” [78]. Ref. [75] argued that the “learning process is crucial for drivers to gain an appropriate understanding of the system’s functionality as well as system limits and helps to build an appropriate level of trust”. Thus, this learning process is presumed to “take some time and requires experience of the system in different situations and different environments” [75]. Further, noted that “field observations consist of usage data that are collected after the system is deployed and released”; user observation, monitoring of customer support calls, and user feedback are all methods that can be used to monitor usage over a long period of time, and estimate learning curves [27]. Another approach that should be considered is the retrospective reconstruction of personally meaningful experiences from memory. The recruitment of participants infers two main phases:

- (1) Explorative phase with screening, synthesis, and validation of participants’ UX,
- (2) Intervention phase with implementable quantitative and qualitative measures that support study design.

It is important to understand more deeply the process of learnability and investigate more in detail the key elements that enhance learnability and, on the other hand, cause learnability problems for users [78]. To this end, based on theoretical models of learnability, “we can develop tools for the commercial user interface world in order to measure and test the learnability process more precisely and better understand how skills are actually learnt and how to click learnability” [78]. We denote that evaluating [i]LiAD may be an easier task than [e]LiAD. Ref. [18] states, “one simply picks some users who have not used the system before and measures the time it takes them to reach a specified level of proficiency in using it”. However, with [e]LiAD, the evaluation is complex and a challenging task, as it can require repeated measures (e.g., on performance, etc.) over time, with a daunting question as to *how long is long enough?* To proficiently conduct long-term studies on [e]LiAD, researchers need to track users’ use or interaction with the system over a sequence of time, a potentially taxing process. Ref. [13] argues that “with a large, enough sample size, and an adequate range of user experience levels studied, this type of study could reveal important information on how user performance progresses over long periods of time”. Such a metric

could be used when comparing users at different learning phases to gain a sense of not only the progression of performance over time but also the occurrence of learned effects. This could be significant in understanding what triggers intended or unintended misuses (based on internal and external factors, etc.). As well as how, why, and when do users learn to misuse an automation system? It is believed that, whether good or bad, just about all human behaviours are learned. Thus, long-term studies are important for recognising LiAD and learning/long-term effects from multi-level spectrums: micro-perspective (short-term), meso-perspective (mid-term), and macro-perspective (long-term) (Figure 11). Tracking users' learning abilities and effects over a long period of time is crucial in determining users' mental maps. Thus, our aim is to interrogate and negotiate the process to fully discern automation effects based on LiAD.

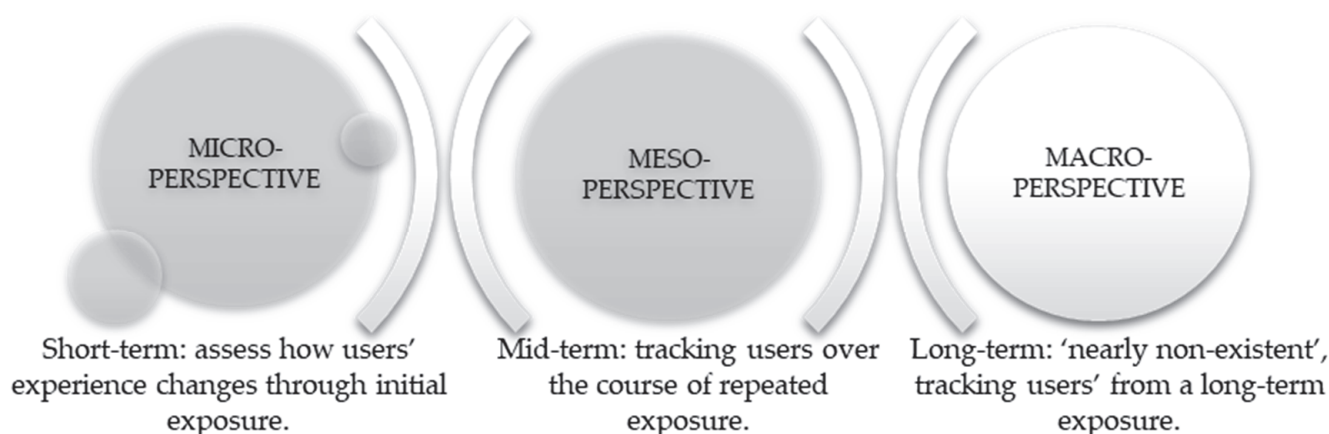


Figure 11. Multi-level perspective studies.

Ref. [27] highlights an extrapolation analysis of short-term performance to infer long-term performance. Moreover, how users approach 'easy' learning, 'perceived ease-of-use', perceived usefulness, 'perceived ease-of-learning', as well as 'perceived ease of misuse' are significant to address. Essentially, as learning is an ongoing process model, users remain learners throughout their entire use of the system; as a result, new thinking, new knowledge formation, attitudes and behaviours are formulated, which are important to consider. Furthermore, what user behaviour attributes and learning effects/curves are evoked (e.g., over-use, misuse, disuse, trust, etc.). A complete centralisation of UX in research developments should be highlighted; thus, a critical view of this phenomenon beyond its theorisation should be explored, including the nature of its effects. To support this framework's arguments, it is significant to normalise long-term research and [e]LiAD. Understanding that, without an appropriate view of long-term research, it is impossible to build a body of transformative and informative knowledge relating to LiAD. Thus, this paper is dedicated towards the naturalisation of [e]LiAD.

4. Conclusions

We framed the discussion based on different factors that influence LiAD; thus, a familiarisation of the concept is formulated. This paper may help with prognostication or extrapolation of future LiAD studies. The framework discussion provides an opportunity to add to knowledge discovery on learning/long-term effects and LiAD. To perform this effectively across the broad spectrum of automated driving research, it is important to define a set of core concepts that can inform and guide the research strategy. We have introduced the learnability engineering life cycle and CALE for a better understanding of LiAD research. Our intent for this framework is to encourage researchers, developers, and all relevant stakeholders in automated driving to engage in more LiAD research. In addition, we have proposed several fundamental strategies that can guide study approaches. Conducting rigorous empirical work on LiAD in all its forms, as well as

actively engaging with stakeholders across the diverse landscape of human factors and ergonomics engineering, will be fundamental to reforming knowledge. Developing an empirically grounded and theoretically informed approach to LiAD may help in making trade-off decisions on design more expedient to enhance safe HAI and QAD. In conclusion, we hope the reader attained discernment and conceptual understanding of LiAD. As well as encouraged scholars to open discussions on this type of research.

Author Contributions: Conceptualisation, N.Y.M.; methodology, N.Y.M.; software, N.Y.M.; validation, N.Y.M.; formal analysis, N.Y.M.; investigation, N.Y.M.; resources, N.Y.M. and K.B.; data curation, N.Y.M.; writing—original draft preparation, N.Y.M.; writing—review and editing, N.Y.M.; visualisation, N.Y.M.; supervision, K.B.; project administration, K.B.; funding acquisition, K.B. All authors have read and agreed to the published version of the manuscript.

Funding: This project has received funding from the European Union’s Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No. 860410.

Data Availability Statement: No new data were created or analysed in this study. Data sharing is not applicable to this article.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript in non-alphabetical order:

ACC	Adaptive Cruise Control
ADS	Automated Driving System
AI	Artificially Intelligent
BAC	Behaviour Adaptability and/or Changeability
CALE	Concepts for Applying Learnability Engineering
ESM	Experience Sampling Methods
HAI	Human-Automation Interaction
HAC	human-Automation Compatibility
HCI	Human-Computer Interaction
HMI	Human-Machine Interface
LiAD	Learnability in Automated Driving
OEM	Original Equipment Manufacturer
PUMS	Programmable User Models
QAD	Quality Automated Driving
QinU	Quality in Usability
QinL	Quality in Learnability
UX	User Experience
ISO	International Standard Organisation
[i]LiAD	Initial Learnability in Automated Driving (initial LiAD)
[e]LiAD	Extended Learnability in Automated Driving (extended LiAD)
S-RC	Stimulus-Response Compatibility

References

1. SAE International. *Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles (J3016_202104)*; SAE International: Warrendale, PA, USA, 2021.
2. Kyriakidis, M.; de Winter, J.C.F.; Stanton, N.; Bellet, T.; van Arem, B.; Brookhuis, K.; Martens, M.H.; Bengler, K.; Andersson, J.; Merat, N.; et al. A human factors perspective on automated driving. *Theor. Issues Ergon. Sci.* **2019**, *20*, 223–249. [CrossRef]
3. Körber, M.; Baseler, E.; Bengler, K. Introduction matters: Manipulating trust in automation and reliance in automated driving. *Appl. Ergon.* **2018**, *66*, 18–31. [CrossRef] [PubMed]
4. Lee, J.D.; See, K.A. Trust in automation: Designing for appropriate reliance. *Hum. Factors* **2004**, *46*, 50–80. [CrossRef] [PubMed]
5. Mbelekani, N.Y.; Bengler, K. Learning Design Strategies for Optimizing User Behaviour towards Automation: Architecting Quality Interactions from Concept to Prototype. In Proceedings of the International Conference on Human-Computer Interaction, Copenhagen, Denmark, 23–28 July 2023; Springer Nature: Cham, Switzerland, 2023; pp. 90–111.
6. Heimgärtner, R. Human Factors of ISO 9241-110 in the Intercultural Context. *Adv. Ergon. Des. Usability Spec. Popul. Part* **2014**, *3*, 18.

7. Rudin-Brown, C.M.; Parker, H.A.; Malisia, A.R. Behavioral adaptation to adaptive cruise control. In Proceedings of the Human Factors and Ergonomics Society 47th Annual Meeting, Denver, CO, USA, 13–17 October 2003; SAGE Publications: Los Angeles, CA, USA, 2003; Volume 47, pp. 1850–1854.
8. Hoffman, R.R.; Marx, M.; Amin, R.; McDermott, P.L. Measurement for evaluating the learnability and resilience of methods of cognitive work. *Theor. Issues Ergon. Sci.* **2010**, *11*, 561–575. [CrossRef]
9. Butler, K.A. Connecting theory and practice: A case study of achieving usability goals. *ACM SIGCHI Bull.* **1985**, *16*, 85–88. [CrossRef]
10. Albert, B.; Tullis, T. *Measuring the User Experience: Collecting, Analyzing, and Presenting UX Metrics*; Morgan Kaufmann: Burlington, MA, USA, 2022.
11. Oliver, M. Learning technology: Theorising the tools we study. *Br. J. Educ. Technol.* **2013**, *44*, 31–43. [CrossRef]
12. Joyce, A. *How to Measure Learnability of a User Interface*; Nielsen Norman Group: Dover, DE, USA, 2019.
13. Grossman, T.; Fitzmaurice, G.; Attar, R. A survey of software learnability: Metrics, methodologies and guidelines. In Proceedings of the Sigchi Conference on Human Factors in Computing Systems, Boston, MA, USA, 4–9 April 2009; pp. 649–658.
14. Stickel, C.; Fink, J.; Holzinger, A. Enhancing universal access—EEG based learnability assessment. In *Universal Access in Human-Computer Interaction. Applications and Services: 4th International Conference on Universal Access in Human-Computer Interaction, UAHCI 2007 Held as Part of HCI International 2007, Beijing, China, 22–27 July 2007*; Proceedings, Part III 4; Springer: Berlin/Heidelberg, Germany, 2007; pp. 813–822.
15. Haramundanis, K. Learnability in information design. In Proceedings of the 19th Annual International Conference on Computer Documentation, Sante Fe, NM, USA, 21–24 October 2001; pp. 7–11.
16. Howes, A.; Young, R.M. *Predicting the Learnability of Task-Action Mappings. The Soar Papers (Vol. II): Research on Integrated Intelligence*; MIT Press: Cambridge, MA, USA, 1993; pp. 1204–1209.
17. Burton, J.K.; Moore, D.M.M.; Magliaro, S.G. Behaviorism and instructional technology. In *Handbook of Research for Educational Communications and Technology*, 2nd ed.; Jonassen, D.H., Ed.; Lawrence Erlbaum Associates, Publishers: London, UK, 2004; pp. 46–73.
18. Nielsen, J. *Usability Engineering*; Academic Press: Boston, MA, USA, 1994.
19. Holzinger, A. Usability engineering methods for software developers. *Commun. ACM* **2005**, *48*, 71–74. [CrossRef]
20. Carroll, J.M.; Carrithers, C. Training wheels in a user interface. *Commun. ACM* **1984**, *27*, 800–806. [CrossRef]
21. Bevan, N.; Macleod, M. Usability measurement in context. *Behav. Inf. Technol.* **1994**, *13*, 132–145. [CrossRef]
22. Michelsen, C.D.; Dominick, W.D.; Urban, J.E. A methodology for the objective evaluation of the user/system interfaces of the MADAM system using software engineering principles. In Proceedings of the 18th Annual Southeast Regional Conference, Tallahassee, FL, USA, 24–26 March 1980; pp. 103–109.
23. Abran, A.; Khelifi, A.; Suryan, W.; Seffah, A. Usability meanings and interpretations in ISO standards. *Softw. Qual. J.* **2003**, *11*, 325–338. [CrossRef]
24. Nielsen, M.B.; Winkler, B.; Colonel, L. *Addressing Future Technology Challenges through Innovation and Investment*; Air University: Islamabad, Pakistan, 2012.
25. Marrella, A.; Catarci, T. Measuring the learnability of interactive systems using a Petri Net based approach. In Proceedings of the 2018 Designing Interactive Systems Conference, Hong Kong, China, 9–13 June 2018; pp. 1309–1319.
26. Shamsuddin, N.A.; Sulaiman, S.; Syed-Mohamad, S.M.; Zamli, K.Z. Improving learnability and understandability of a web application using an action-based technique. In Proceedings of the 2011 Malaysian Conference in Software Engineering, Johor Bahru, Malaysia, 13–14 December 2011; IEEE: Piscataway, NJ, USA, 2011; pp. 245–250.
27. Santos, P.J.; Albert, N.B. *Discount Learnability Evaluation*; Graphics, Visualization & Usability Center, Georgia Institute of Technology: Atlanta, GA, USA, 1995.
28. Rieman, J. A field study of exploratory learning strategies. *ACM Trans. Comput.-Hum. Interact. (TOCHI)* **1996**, *3*, 189–218. [CrossRef]
29. Goguen, J.A. Toward a Social, Ethical Theory of Information. In *Social Science, Technical Systems, and Co-Operative Work*; Bowker, G.C., Star, S.L., Turner, W., Eds.; Psychology Press: London, UK, 1997.
30. Hoffman, R.R.; Elm, W.C. HCC implications for the procurement process. *IEEE Intell. Syst.* **2006**, *21*, 74–81. [CrossRef]
31. Neville, K.; Hoffman, R.R.; Linde, C.; Elm, W.C.; Fowlkes, J. The procurement woes revisited. *IEEE Intell. Syst.* **2008**, *23*, 72–75. [CrossRef]
32. Bruun, A. Training software developers in usability engineering: A literature review. In Proceedings of the 6th Nordic Conference on Human-Computer Interaction: Extending Boundaries, Reykjavik, Iceland, 16–20 October 2010; pp. 82–91.
33. Grudin, J. Utility and usability: Research issues and development contexts. *Interact. Comput.* **1992**, *4*, 209–217. [CrossRef]
34. Mayhew, D.J. The usability engineering lifecycle. In Proceedings of the CHI’99 Extended Abstracts on Human Factors in Computing Systems, Pittsburgh, PA, USA, 15–20 May 1999; pp. 147–148.
35. Nielsen, J. The usability engineering life cycle. *Computer* **1992**, *25*, 12–22. [CrossRef]
36. Sohaib, O.; Khan, K. Integrating usability engineering and agile software development: A literature review. In Proceedings of the 2010 International Conference on Computer Design and Applications, Qinhuaungdao, China, 25–27 June 2010; IEEE: Piscataway, NJ, USA, 2010; Volume 2, pp. V2–V32.
37. Schaffer, E. *Institutionalization of Usability: A Step-by-Step Guide*; Addison-Wesley Professional: Boston, MA, USA, 2004.

38. Schaffer, E.; Lahiri, A. *Institutionalization of UX: A Step-by-Step Guide to a User Experience Practice*; Addison-Wesley: Boston, MA, USA, 2013.
39. Mack, R.; Robinson, J.B. When novices elicit knowledge: Question asking in designing, evaluating, and learning to use software. In *The Psychology of Expertise: Cognitive Research and Empirical AI*; Springer: New York, NY, USA, 1992; pp. 245–268.
40. Beggiato, M.; Pereira, M.; Petzoldt, T.; Krems, J. Learning and development of trust, acceptance and the mental model of ACC. A longitudinal on-road study. *Transp. Res. Part F Traffic Psychol. Behav.* **2015**, *35*, 75–84. [CrossRef]
41. Haritos, T. A Study of Human-Machine Interface (HMI) Learnability for Unmanned Aircraft Systems Command and Control. Ph.D. Thesis, Nova Southeastern University, Fort Lauderdale, FL, USA, 2017.
42. Ojeda, L.; Nathan, F. Studying learning phases of an ACC through verbal reports. In *Driver Support and Information Systems: Experiments on Learning, Appropriation and Effects of Adaptiveness*; Brouwer, R., Hoedemaeker, D.M., Eds.; 2006; pp. 47–73.
43. Sauro, J.; Lewis, J.R. Standardized usability questionnaires. In *Quantifying the User Experience: Practical Smartphone App for Anxious Youth 83 Statistics for User Research*; Elsevier: Waltham, MA, USA, 2012; pp. 198–212.
44. Weinberger, M.; Winner, H.; Bubbs, H. Adaptive cruise control field operational test—The learning phase. *JSAE Rev.* **2001**, *22*, 487–494. [CrossRef]
45. Simon, J.H. Learning to Drive with Advanced Driver Assistance Systems. Empirical Studies of an Online Tutor and a Personalised Warning Display on the Effects of Learnability and the Acquisition of Skill. Ph.D. Thesis, Philosophische Fakultät, Technische Universität Chemnitz, Chemnitz, Germany, 2005.
46. Elliott, G.J.; Jones, E.; Barker, P. A grounded theory approach to modelling learnability of hypermedia authoring tools. *Interact. Comput.* **2002**, *14*, 547–574. [CrossRef]
47. Kato, T. What “question-asking protocols” can say about the user interface. *Int. J. Man-Mach. Stud.* **1986**, *25*, 659–673. [CrossRef]
48. Mack, R.L.; Lewis, C.H.; Carroll, J.M. Learning to use word processors: Problems and prospects. *ACM Trans. Inf. Syst. (TOIS)* **1983**, *1*, 254–271. [CrossRef]
49. Unsöld, M. Measuring Learnability in Human-Computer Interaction. Ph.D. Thesis, Ulm University, Ulm, Germany, 2018.
50. Davis, S.; Wiedenbeck, S. The effect of interaction style and training method on end user learning of software packages. *Interact. Comput.* **1998**, *11*, 147–172. [CrossRef]
51. Rafique, I.; Weng, J.; Wang, Y.; Abbasi, M.Q.; Lew, P.; Wang, X. Evaluating software learnability: A learnability attributes model. In Proceedings of the 2012 International Conference on Systems and Informatics (ICSAI2012), Yantai, China, 19–20 May 2012; IEEE: Piscataway, NJ, USA, 2012; pp. 2443–2447.
52. Linja-aho, M. Evaluating and Improving the Learnability of a Building Modeling System. Master’s Thesis, Helsinki University of Technology, Helsinki, Finland, 2005.
53. Baecker, R.; Booth, K.; Jovicic, S.; McGrenere, J.; Moore, G. Reducing the gap between what users know and what they need to know. In Proceedings of the 2000 Conference on Universal Usability, Arlington, VA, USA, 16–17 November 2000; pp. 17–23.
54. Paymans, T.F.; Lindenberg, J.; Neerincx, M. Usability trade-offs for adaptive user interfaces: Ease of use and learnability. In Proceedings of the 9th International Conference on Intelligent User Interfaces, Funchal, Portugal, 13–16 January 2004; pp. 301–303.
55. Kieras, D.; Polson, P.G. An approach to the formal analysis of user complexity. *Int. J. Man-Mach. Stud.* **1985**, *22*, 365–394. [CrossRef]
56. Brown, P.C.; Roediger, H.L., III; McDaniel, M.A. *Make It Stick: The Science of Successful Learning*; Harvard University Press: Cambridge, MA, USA, 2014.
57. Lieberman, D.A. *Human Learning and Memory*; Cambridge University Press: Cambridge, UK, 2012.
58. Lowyck, J. Bridging learning theories and technology-enhanced environments: A critical appraisal of its history. In *Handbook of Research on Educational Communications and Technology*; Springer Science & Business Media: New York, NY, USA, 2014; pp. 3–20.
59. Homans, H.; Radlmayr, J.; Bengler, K. Levels of driving automation from a user’s perspective: How are the levels represented in the user’s mental model? In *Human Interaction and Emerging Technologies: Proceedings of the 1st International Conference on Human Interaction and Emerging Technologies (IHiet 2019), Nice, France, 22–24 August 2019*; Springer International Publishing: Cham, Switzerland, 2020; pp. 21–27.
60. Ko, A.J.; Myers, B.A.; Aung, H.H. Six learning barriers in end-user programming systems. In Proceedings of the 2004 IEEE Symposium on Visual Languages-Human Centric Computing, Rome, Italy, 26–29 September 2004; IEEE: Piscataway, NJ, USA, 2004; pp. 199–206.
61. Allen, R.B. Mental models and user models. In *Handbook of Human-Computer Interaction*, 2nd ed.; Helander, M., Landauer, T.K., Prabhu, P., Eds.; Elsevier Science B.V.: North-Holland, The Netherlands, 1997; pp. 49–63.
62. Beggiato, M.; Krems, J.F. The evolution of mental model, trust and acceptance of adaptive cruise control in relation to initial information. *Transp. Res. Part F Traffic Psychol. Behav.* **2013**, *18*, 47–57. [CrossRef]
63. Wozney, L.; McGrath, P.J.; Newton, A.; Huguet, A.; Franklin, M.; Perri, K.; Leuschen, K.; Toombs, E.; Lingley-Pottie, P. Usability, learnability and performance evaluation of Intelligent Research and Intervention Software: A delivery platform for eHealth interventions. *Health Inform. J.* **2016**, *22*, 730–743. [CrossRef]
64. Ertmer, P.A.; Newby, T.J. Behaviorism, cognitivism, constructivism: Comparing critical features from an instructional design perspective. *Perform. Improv. Q.* **2013**, *26*, 43–71. [CrossRef]
65. Bell, F. Connectivism: Its place in theory-informed research and innovation in technology-enabled learning. *Int. Rev. Res. Open Distrib. Learn.* **2011**, *12*, 98–118. [CrossRef]

66. De Houwer, J.; Hughes, S. *The Psychology of Learning: An Introduction from a Functional-Cognitive Perspective*; MIT Press: Cambridge, MA, USA, 2020.
67. Pritchard, A. *Ways of Learning: Learning Theories and Learning Styles in the Classroom*; Routledge: London, UK, 2009.
68. Lyons, J. Learning with technology: Theoretical foundations underpinning simulations in higher education. In *Proceedings of the Future Challenges, Sustainable Futures, Proceedings ASCILITE, Wellington, New Zealand, 25–28 November 2012*; pp. 582–586.
69. Balakrishnan, V.; Gan, C.L. Students' learning styles and their effects on the use of social media technology for learning. *Telemat. Inform.* **2016**, *33*, 808–821. [CrossRef]
70. Newell, A.; Rosenbloom, P.S. Mechanisms of skill acquisition and the law of practice. In *Cognitive Skills and Their Acquisition*; Psychology Press: London, UK, 2013; pp. 1–55.
71. Mbelekani, N.Y.; Bengler, K. Interdisciplinary Industrial Design Strategies for Human-Automation Interaction: Industry Experts' Perspectives. *Interdiscip. Pract. Ind. Des.* **2022**, *48*, 132–141.
72. Simonet, S.; Wilde, G.J.S. Risk: Perception, acceptance and homeostasis. *Appl. Psychol. Int. Rev.* **1997**, *46*, 235–252. [CrossRef]
73. Linja-aho, M. Creating a framework for improving the learnability of a complex system. *Hum. Technol.* **2006**, *2*, 202–224. [CrossRef]
74. Forster, Y.; Hergeth, S.; Naujoks, F.; Krems, J.; Keinath, A. User education in automated driving: Owner's manual and interactive tutorial support mental model formation and human-automation interaction. *Information* **2019**, *10*, 143. [CrossRef]
75. Metz, B.; Wörle, J.; Hanig, M.; Schmitt, M.; Lutz, A.; Neukum, A. Repeated usage of a motorway automated driving function: Automation level and behavioural adaption. *Transp. Res. Part F Traffic Psychol. Behav.* **2021**, *81*, 82–100. [CrossRef]
76. Saad, F. Some critical issues when studying behavioural adaptations to new driver support systems. *Cogn. Technol. Work* **2006**, *8*, 175–181. [CrossRef]
77. Lin, H.X.; Choong, Y.; Salvendy, G. A proposed index of usability: A method for comparing the relative usability of different software systems. *Behav. Inf. Technol.* **1997**, *16*, 267–277. [CrossRef]
78. Laakkonen, M. *Learnability Makes Things Click: A Grounded Theory Approach to the Software Product Evaluation*; University of Lapland: Rovaniemi, Finland, 2007.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

Article

The Correlation Between Self-Assessment and Observation in Driving Style Classification: An On-Road Case Study

Muhammad Zainul Abidin Kamaludin ¹, Juffrizal Karjanto ^{1,2,*}, Noryani Muhammad ^{1,2}, Nidzamuddin Md Yusof ^{1,2}, Muhammad Zahir Hassan ^{1,2}, Mohamad Zairi Baharom ³, Zulhaidi Mohd Jawi ⁴ and Matthias Rauterberg ⁵

¹ Fakulti Teknologi dan Kejuruteraan Mekanikal, Universiti Teknikal Malaysia Melaka, Durian Tunggal 76100, Melaka, Malaysia; muhammadzainulabidin97@gmail.com (M.Z.A.K.); noryani@utem.edu.my (N.M.); nidzamuddin@utem.edu.my (N.M.Y.); zahir@utem.edu.my (M.Z.H.)

² Centre for Advanced Research on Energy, Universiti Teknikal Malaysia Melaka, Durian Tunggal 76100, Melaka, Malaysia

³ Faculty of Mechanical & Automotive Engineering Technology, Universiti Malaysia Pahang Al-Sultan Abdullah, Pekan 26600, Pahang, Malaysia; mohamadzairi@ump.edu.my

⁴ Malaysian Institute of Road Safety Research (MIROS), Kajang 43000, Selangor, Malaysia; zulhaidi@miros.gov.my

⁵ Department of Industrial Design, Eindhoven University of Technology, P.O. Box 513, 5600 MB Eindhoven, The Netherlands; g.w.m.rauterberg@tue.nl

* Correspondence: juffrizal@utem.edu.my

Abstract: A typical classification of driving style from a human driver is conducted via self-assessment, which begs the question of the possibility of bias from the respondents. Although some research has been carried out validating the questionnaire, no controlled studies have yet to be reported to validate the Malaysian driving style. This study aimed to validate the Malaysian driver using the Multidimensional Driving Style Inventory (MDSI) with five-factor driving styles (careful, risky, angry, anxious, and dissociative) in on-road situations. Forty-one respondents completed the experiment on two designated routes recorded over 45 min of driving. A modest correlation existed between the MDSI and the score retrieved from the on-road observation assessment. The result showed a low-to-medium correlation collected from acceleration in longitudinal directions compared with correlation analysis utilizing the MDSI scale. Exploring such latent traits is essential for precisely classifying human driver styles without bias.

Keywords: driving style; self-assessment; on-road observation; acceleration; Malaysian driver

1. Introduction

Driving style can be defined as the way or method a driver chooses to drive, and it depends on the physical and emotional conditions of the driver while driving [1–3]. For example, the way a driver drives a vehicle when overtaking a vehicle on the road, approaching a junction, accelerating when a traffic light changes from red to green, or a driver's reaction when following a slower vehicle ahead depends on their driving style. All these examples can be seen in daily life. Some people drive vehicles hastily and riskily, such as following too closely behind another vehicle. But some people drive a vehicle slowly and carefully.

Some of the importance of identifying human driving style is improving the riding experience [4], enhancing trust in humans and machines [5], and improving road and vehicle safety [6,7].

Several existing studies in the literature have attempted to objectify the human driving style. Some approaches have a simple recognition of driving styles, such as studying the steering wheel angle [8]. One study characterized the driving style of Moroccan drivers using vehicle sensor data collected from the steering wheel angle. Another approach is implementing more complex and expensive equipment inside the designated instrumented vehicle. The approach was designed to detect the sensitive parts of the human body, such as biosensors, retina scans, and fingerprints, to investigate driving style [9]. Recent studies have proven that the best way to classify driving style is from objective driving data derived from acceleration [10–12].

In contrast, many previous studies have shown that self-assessment can also be used to identify driving styles because it is easier to access and economical for collecting a larger respondent group [10–13]. For example, several self-report measures of driver behavior and driving styles have been constructed in the last decade, such as the driving behavior inventory (DBI) [14], driving style questionnaire (DSQ) [15], attitudes to driving violations (ADVSs) [16] and the driver behavior questionnaire (DBQ) [13]. Hence, Taubman-Ben-Ari [10] created the multidimensional driving style inventory (MDSI) to incorporate all the driving styles into one questionnaire. Since then, MDSI has been widely translated in various countries and has found different driving styles based on these countries. For example, in Romania, the researcher found seven driving styles [17], while in China, Spain, and the Netherlands, they found six driving styles [11,18,19], and in Malaysia, they discovered five driving styles [20]. Although, understandably, different countries might have different driving styles, these dissimilar outcomes raise the question of whether a self-assessment (questionnaire) is a proper means to identify a person's driving style or whether there is bias from the respondents. The possibility of bias might arise from the respondents being unaware of the questionnaire objective [21–23]. It could be that the respondents completed the self-assessment without performing the actual task, with the possibility that the respondents just answered the questionnaire based on their point of view at that moment. Therefore, such predicted performance evaluations may have a hazy reference point problem since drivers lack objective feedback directly linked to their driving activities [24,25].

Several studies have suggested a method to validate the self-assessment of driving styles by comparing self-reported driving data with actual driving performance in a simulation. Interestingly, only a few studies have shown a more significant proportion of driving performance with a positive correlation [26]. Meanwhile, most studies have shown moderate and low correlations [24,27,28]. In addition, other studies have used a different approach to tackle the specificity of the driver, such as using a designated observer to validate self-assessment through the observation measure [28]. So far, most studies have stated that the simulator studies validated the application. However, when using a simulator, the disadvantages include limited physical, perceptual, and behavioral fidelity and a lack of research demonstrating the simulation's validity [29].

Recently, ref. [20] validated the fact that Malaysian drivers can be classified into five (5) categories of drivers (careful, risky, angry, anxious, and dissociative), with 27 items using the MDSI self-assessment. Motivated by the recent literature, this paper aims to validate the Malaysian driving style by comparing the results of self-assessment using the MDSI questionnaire from [20] and on-road driving styles observed by a designated observer. This study also investigates drivers' self-conception of driving skills based on objective driving data collected using the self-assessment [28] and the vehicle's acceleration [28,30–34]. According to recent literature, subjective judgments are reliable in supporting the formulation of a driver's driving style [8,24,29,35,36]. Two hypotheses were considered for this study: First, this study was expected to find correlations between self-assessments and

the designated observer on driving skills/behavior evaluations. The second hypothesis is about the correlations between the driving style from the MDSI scale and the acceleration data collected. Assuming that the participant's driving styles have relative validity, the expectation from the MDSI is that it will have the most robust connection with self-assessment after driving activities and acceleration choice [24]. This study attempted to validate the significance of the self-assessment of the Malaysian driving style [20] with the driver's behavior in on-road experiments.

2. Method

The general methodology can be divided into two (2) phases. The first phase began with the classification of the driving styles of participants. The classification of THE driving style was performed by instructing the participant to answer a Multidimensional Driving Style Inventory (MDSI) self-assessment questionnaire. Since all the potential participants were Malaysian, the MDSI for Malaysian (MDMI-M) drivers was answered [20] to determine whether the participants were careful, risky, angry, anxious, or dissociative drivers. The driving style is determined before the experiment (second phase). For the second phase, 43 participants (from various types of drivers, i.e., careful, risky, angry, anxious, or dissociative) were invited to perform the driving task on-road while being evaluated by the designated observer.

On the experiment day, a brief explanation of the rules, tasks, and ethical review with written informed consent was given following the approval of the Human Committee of Universiti Teknikal Malaysia Melaka (UTeM) on a study with human subjects (UTeM.23.01/500/25/4Jilid5(30)). The participant was then escorted to the instrumented vehicle, where another briefing was given about the task and the driving route for the experiment. An experimenter sat in the back and directed the participant (if needed) during the experiment phase (see Figure 1). At the same time, another experimenter (designated observer) evaluated the participant's driving performance during the driving session using a dedicated questionnaire (see Figure 1). The designated observer was an experienced driver with a vehicle safety background (>10 years) and was accustomed to the instrumented vehicle used in this study. The participant needed to perform two (2) sessions: the familiarization and experiment session. Only the acceleration from the experiment session was analyzed. In addition, during the briefing process, the participants were instructed to drive naturally on the defined route while obeying all the traffic rules. After the driving session, the participant needed to answer a questionnaire to self-evaluate their performance in the driving task.

This study collected data from a self-assessment questionnaire (from both the designated observer and participant) and the acceleration data from the accelerometer and programmable Data Acquisition System (DAQ). The gathered acceleration data were later processed and analyzed using LabVIEW Diadem (2018) with a 250 Hz sampling rate. SPSS software (version 26) was used to analyze the correlation between the data.

At the end of the experiment, a debriefing session was performed, and the participant was rewarded as a token of appreciation with Ringgit Malaysia 30 (RM 30~USD 6).

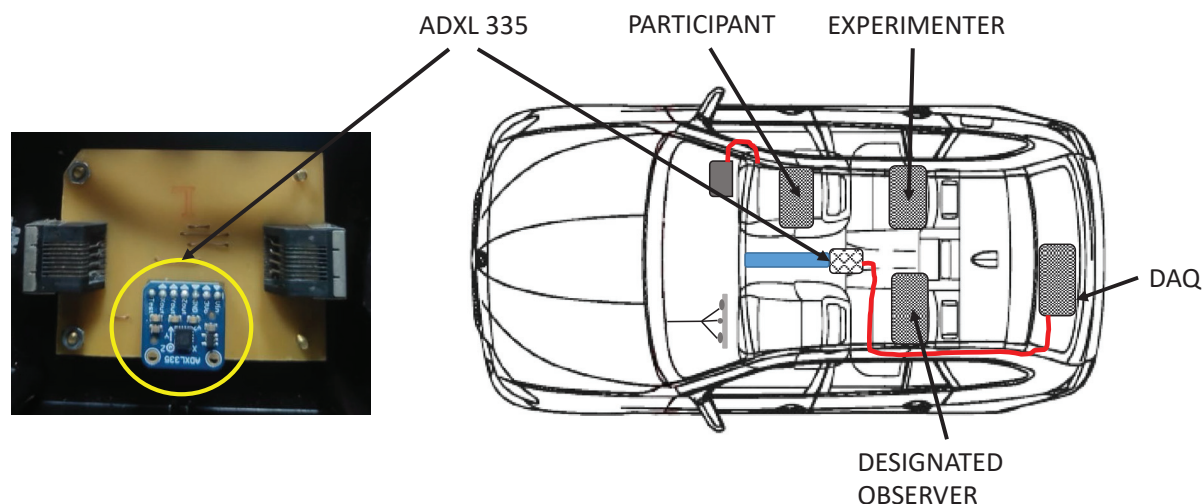


Figure 1. Location of the accelerometer (ADXL 335, Analog Devices, Inc., Wilmington, MA, USA), DAQ, participant, experimenter, and designated observer inside the test vehicle.

2.1. Participant

The participants consisted of 30 males and 11 females. The study focused on young drivers aged 17 to 29 (Mean age = 22.67, SD = 1.59). They had valid driving licenses from at least three months to nine years (Mean = 4.96, SD = 1.84). Based on the duration of the driving license, the participants were categorized into three types of drivers. First, those who had a driving license below three years were considered novice drivers (23.81%). Second, those with a driving license between four and six years were considered moderate drivers (59.52%). Third, those with over seven years of licenses were categorized as expert drivers (16.67%). The driving mileage per year was up to 100,000 km with a minimum of 5000 km (Mean age = 390.86, SD = 1702.78).

2.2. Instrumented Vehicle

A 2010 Perodua Myvi (B-Segment car made by Perodua in Rawang, Selangor, Malaysia) was installed with a data acquisition system (DAQ) as an instrumented vehicle to perform the driving activities and collect the participants' raw data. The DAQ system included an accelerometer to collect acceleration data. The accelerometer was placed inside the vehicle between the passenger and driver seats (see Figure 1).

2.3. Designated Route

The participants were required to perform the driving sessions under two (2) sessions, including the familiarization and the experiment session. The experiment started with the familiarization session (about 5 min), which aimed to explain the purpose of the experiment and the objectives to the participants (see Figure 2). In addition, the familiarization session was used to familiarize the participants with the roundabout, cornering, straight line, and the vehicle used for this study. During this session, the participant could ask questions and clarify doubts. Then, the participant proceeded to the experiment session, where the participant went through the course for an estimated 25 min (see Figure 3).

The route consisted of roundabouts, cornering, and straight lines to imitate typical suburban roads in Malaysia. During the experiment session, the participant was instructed to complete the designated route two (2) times. The objective of this was to ensure that the data collected were dependable and accurate. The total distance covered was approximately 15 km, including 3 km for the familiarization session and 12 km for the experiment session.



Figure 2. The familiarization route, started at A and ended at E (source: Google Maps).



Figure 3. The experiment route, started at B and ended at J (source: Google Maps).

2.4. Questionnaire

Four (4) questionnaires were used in this study (see Figure 4), namely the Multi-dimensional Driving Style Inventory-Malaysian (MDSI-M), On-Road Observation form (ORO), Driver Self-Evaluation Questionnaire (DSEQ), and Driver Evaluation Questionnaire (DEQ) [28]. In this study, participants answered the MDSI before undertaking the driving session. The designated observer evaluated the participants during the driving session using the DEQ and ORO (labeled as ORO-O, from here onwards). After the driving session, the participants completed a self-assessment of their driving by answering the DSEQ and ORO (labeled as ORO-P, from here onwards).

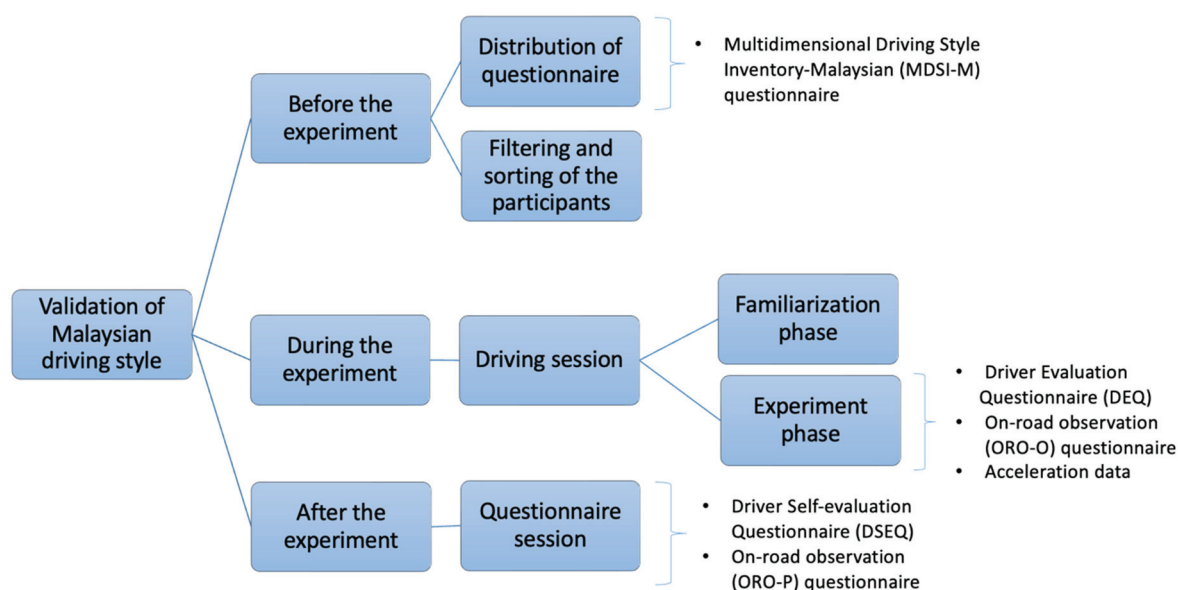


Figure 4. The methodological approach to validate Malaysian driving styles using self-assessment.

The MDSI-M summarizes the fact that the Malaysian driving style contains only 27 items with five (5) factors that should be rated on a Likert scale [20]. Based on the responses given from the seven (7)-point scale, these data indicate how frequently the driver engages in a list of driving habits using the given response scale (never, almost, sometimes, often, very often, and always). The result showed that the internal consistency questionnaire is reliable, indicating adequate to acceptable assessments for the following five (5) factors: careful driving (8 items, Cronbach's alpha 0.73), risky driving (5 items, Cronbach's alpha 0.69), angry driving (6 items, Cronbach's 0.72), dissociative driving (5 items, Cronbach's alpha 0.73), and anxious driving (3 items, Cronbach's alpha 0.60) [20].

An ORO evaluates the driving behavior to indicate the error/violation of participants while operating in varied road situations and environments. The ORO evaluates five (5) dimensions of violations or errors: speed, lane use and passing, road signs, intersections and their usage, clearance and checking, and brakes and gears. The questionnaire was initially constructed based on previous research, with some items modified or removed to fit the objective based on the situation of Malaysian drivers [28]. For example, the dimensions of a traffic light were changed to road signs because the experiment was conducted in the university area. The item "Driving at the red light" was changed to "Driving when there is a stop sign". The designated observer (see Figure 1) was required to rate the five (5) dimensions and 24 items of the ORO form during the driving session in terms of a five (5)-point Likert scale ("1" never, "5" very frequently) when the participant was driving the instrumented vehicle. The designated observer reported his observations on the form while moving between the different road segments at least two (2) times or more in each road segment.

The Driver Evaluation Questionnaire (DEQ) is a questionnaire form used by the designated experimenter to evaluate the participant's performance. The Driver Self Evaluation Questionnaire (DSEQ) is a questionnaire that assesses the participant. This form comprises 18 items regarding driving style, behavior, and abilities. The questionnaire was initially constructed based on previous research [28] with some modifications to fit with the objective and designated route based on the situation of Malaysian drivers. For example, item number two was changed from the initial meaning of "perceiving hazards in traffic" to "perceiving hazards in the intersection" in terms of the five (5)-point Likert scale (where "1" indicates low and "5" indicates high-skill performance). The designated observer

evaluated the DEQ during the driving session, while the participants evaluated themselves after the driving session using the DSEQ.

3. Results and Discussion

3.1. Pearson Correlation for Self-Assessment Questionnaire

Pearson's correlation analysis was used to compare the relationship between the MDSI-M self-assessment and the participant's driving activities ascertained from the observer's judgment (ORO-O). The results for the MDSI-M score with error and violation scores from ORO-O were correlated (see Table 1). Firstly, the careful driving style showed a significantly low positive correlation with road sign usage ($r = 0.37, p < 0.05$). This suggests that careful drivers use road signs more frequently while driving. Conversely, the dissociative ($r = -0.03, p < 0.05$) and anxious ($r = -0.03, p < 0.05$) driving styles displayed low negative correlations with road sign usage, indicating that drivers with these styles tend to have lower scores for utilizing road signs [24]. Secondly, the dissociative driving style exhibited significant low-to-medium positive correlations with speed ($r = 0.47, p < 0.05$), clearance ($r = 0.34, p < 0.05$), and checking ($r = 0.34, p < 0.05$). This suggests that drivers with dissociative tendencies are more likely to engage in behaviors such as checking intersections and maintaining clearance during driving activities. Similarly, the anxious driving style displayed significant low-to-medium positive correlations with speed ($r = 0.38, p < 0.05$), clearance ($r = 0.49, p < 0.05$), and checking ($r = 0.38, p < 0.05$). This indicates that anxious drivers exhibit higher speeds and consistently check intersections and clearances [24].

Table 1. Correlation between MDSI-M score with error and violation from ORO form.

		MDSI-M				
	Scale	Careful	Risky	Angry	Anxious	Dissociative
ORO-O	Speed	−0.15	0.23	0.06	0.47 *	0.38 *
	Lane use and passing	−0.07	0.07	−0.01	0.21	0.21
	Road sign	0.34 *	−0.29	−0.17	−0.49 *	−0.32 *
	Intersection and usage	−0.18	−0.00	0.05	0.17	0.16
	Clearance and checking	−0.05	0.11	0.06	0.34 *	0.48 *
	Brakes and gears	0.04	0.09	0.02	0.04	0.16
ORO-P	Speed	−0.21	0.11	−0.01	0.01	0.28
	Lane use and passing	−0.39 *	−0.01	−0.03	0.27	0.48 *
	Road sign	0.15	−0.24	−0.34 *	−0.32 *	−0.27
	Intersection usage	−0.28	0.02	−0.05	0.02	0.39 *
	Clearance and checking	−0.08	0.09	0.14	−0.01	0.10
	Brakes and gears	−0.22	0.08	−0.01	0.15	0.37 *

$p < 0.05$ * indicates a significant effect.

Furthermore, the anxious driving style demonstrated significant low-to-medium positive correlations with lane use ($r = 0.48, p < 0.05$), passing ($r = 0.37, p < 0.05$), brakes ($r = 0.48, p < 0.05$), and gears ($r = 0.37, p < 0.05$). This implies that drivers with anxious tendencies have a higher risk of making mistakes in lane changes, passing maneuvers, braking, and gear selection compared to other driving styles [X]. Conversely, the careful driving style displayed significantly low negative correlations with lane use ($r = -0.39, p < 0.05$) and passing ($r = -0.34, p < 0.05$), indicating that careful drivers tend to use fewer lanes and pass less frequently. Additionally, the angry and dissociative driving styles exhibited significantly low negative correlations with road sign usage ($r = -0.34, p < 0.05$ and $r = -0.32, p < 0.05$), respectively. This suggests that angry and dissociative drivers tend to utilize road signs less frequently during their driving activities. These findings contribute to understanding the relationship between driving styles and specific driving behaviors.

These findings also support previous research by [24,28] emphasizing the influence of personality traits on driving behavior. However, it is important to acknowledge this study's limitations, including the reliance on self-assessments and the context-specific nature of the research conducted in Malaysia.

Meanwhile, there were also correlations between the MDSI-M with DEQ and DSEQ (see Table 2). The results showed a significant medium negative correlation with the “dissociative driver” ($r = -0.39, p < 0.05$). Next, a significant medium positive correlation was found between the DSEQ and careful driving ($r = 0.47, p < 0.05$). This result indicates that self-assessment was consistent before and after the experiment in showing the participant's driving style, in general, for driving tasks.

Table 2. Correlation MDSI-M with DEQ and DSEQ.

	Careful	Risky	Angry	Anxious	Dissociative
DEQ	0.17	−0.30	−0.19	−0.34 *	−0.39 *
DSEQ	0.47 *	−0.10	0.14	−0.25	−0.50 *

$p < 0.05$ * indicates a significant effect.

3.1.1. Evaluation Between DEQ (Observer) and DSEQ (Participant)

ANOVA could be used to understand who is better at driving style evaluation, comparing different groups across two dependent variables: the DEQ and DSEQ. One-way ANOVA was used in this analysis because each independent variable (gender, driving experience, annual mileage) was being tested separately to see if it affected the dependent variables (DEQ and DSEQ). The mean item scores from DEQ and DSEQ were compared based on demographic data. Group means (see Table 3) show that gender driving experience and annual millage were covariates. The participant's self-evaluation was higher compared to the observer, whereby only gender was significantly related to the difference between the DEQ and DSEQ ($F(1, 39) = 1.63, p < 0.05$). These results indicate that male participants evaluate their driving and driving skill levels much better than the observers.

Table 3. ANOVA analysis for DEQ and DSEQ.

		Population Size (N)	DEQ			DSEQ		
			Mean	SD	ANOVA	Mean	SD	ANOVA
Gender	Male	30	3.89	0.56	F(1,39) = 0.699 <i>p</i> = 0.408	3.93	0.43	F(1,39) = 1.627 <i>p</i> = 0.002 *
	Female	11	3.73	0.38		3.48	0.21	
Driving experience	Novice driver	10	3.83	0.50	F(2,38) = 0.825 <i>p</i> = 0.448	3.95	0.40	F(2,38) = 2.135 <i>p</i> = 0.132
	Moderate driver	24	3.78	0.58		3.70	0.38	
	Experienced driver	7	4.07	0.27		4.00	0.56	
Annual mileage	Rarely drive	5	3.96	0.53	F(1,39) = 0.260 <i>p</i> = 0.613	3.68	0.56	F(1,39) = 0.510 <i>p</i> = 0.480
	Daily driver	36	3.83	0.53		3.83	0.41	

$p < 0.05$ * indicates a significant effect.

The next section of the questionnaire concerns each DEQ and DSEQ item. Each item was compared individually (see Table 4) using the Wilcoxon signed ranked test. It was apparent that only three items were positively significant compared to eighteen items in the driving performance of all drivers.

Table 4. Wilcoxon analysis for DEQ and DSEQ.

Items	DEQ		DSEQ		Z
	M	SD	M	SD	
Steering	4.41	0.59	4.10	0.66	−2.42
Anger toward other drivers	2.05	0.89	2.44	1.12	−2.12
Supportive driving	3.95	1.05	3.88	0.75	−0.22
Anxiety	2.49	1.12	2.39	1.22	−0.41
Vigilance	4.39	0.97	4.10	0.70	−1.96
Safety	3.93	0.99	4.32	0.61	−2.13
Fluent driving	4.63	0.70	4.24	0.70	−2.57
Perceiving hazards in an intersection	4.51	0.78	4.34	0.79	−0.98
Careful towards other road users	4.34	0.73	4.49	0.51	−0.93
Driving behind a car without being impatient	2.88	1.23	2.71	1.10	−0.89
Predicting traffic situations ahead	4.22	0.65	4.00	0.78	−0.28
Fluent lane changes	4.07	0.91	4.10	0.66	−1.60
Adjusting speed	4.07	0.93	4.12	0.68	−0.05
Signal	4.56	0.84	4.05	0.81	−3.25 *
Relinquishing legitimate rights when necessary	4.41	0.67	3.76	0.86	−3.41 *
Avoiding risk	4.22	1.01	4.29	0.72	−0.04
Conforming to speed limits	3.07	1.59	3.39	0.83	−1.33
Obedying traffic lights	2.98	1.39	3.85	0.73	−3.41 *

$p < 0.01$ * indicates a significant effect.

In summary, these results indicate that most of the driving evaluations conducted by participants overestimated their driving performance on signals, relinquishing legitimate rights when necessary, and traffic lights. Moreover, a difference score was calculated to determine how many participants overestimated their driving performance in this research. The mean difference of 18 matching questions was labeled as “difference scores” after subtracting the observer’s rating (DEQ) from the participant’s rating (DSEQ) for each item. Drivers who scored a difference greater than zero rated their driving performance better than the expert’s opinion. The response rate was $N = 29$ for the total number of drivers who evaluated their driving performance as better than the expert’s. In contrast, ($N = 12$) drivers evaluated their driving performance as lower or the same compared to the expert’s evaluation.

3.1.2. Correlation of MDSI-M with the Acceleration Data

Pearson’s correlation was used to investigate the relationships between driving style derived from the MDSI-M and acceleration data in the longitudinal direction (see Table 5). There was a significantly low negative correlation between the MDSI’s careful driver score ($r = -0.28$) and the acceleration of cautious driving. Meanwhile, there was a significantly low positive correlation ($r = 0.18$) when comparing the MDSI to the acceleration of a “risky driver”. Together, these results provide important insights into the “careful driver”, who scored higher in the self-assessment for careful driving and had lower acceleration. At the same time, there was a high acceleration for the “angry driver”. Next, the MDSI’s “risky driver” score showed a significantly low positive correlation ($r = 0.07$) with the acceleration risk. These results indicate that the “risky driver” has higher self-assessment scores and acceleration.

Table 5. Correlation between MDSI and acceleration data (longitudinal direction).

MDSI	Correlation (r)				
	Acceleration_Careful	Acceleration_Risky	Acceleration_Angry	Acceleration_Anxious	Acceleration_Dissociative
Careful	−0.28 *	0.18 *	0.10	0.28	−0.05
Risky	0.23	0.07 *	0.33	0.09	0.02
Angry	0.20	0.23	0.22	0.04	0.23
Anxious	−0.37	0.19	0.09	0.00	−0.11
Dissociative	0.13	−0.01	−0.03	0.23	−0.33 *

$p < 0.05$ * indicates a significant effect.

Interestingly, there was no significant correlation between the MDSI's "angry driver" and MDSI's "anxious driver" scores. Lastly, a significantly low negative correlation existed between the MDSI's "anxious driver" score ($r = -0.33$) and the "dissociative driver" score for acceleration. Together, these results indicate that the "anxious driver" has a higher score but lower acceleration than the "dissociative driver".

3.2. Discussion

The current study addresses the validation of Malaysian driving styles from the self-assessment questionnaire and acceleration data to assess the performance of driving style classification in on-road studies. Within the established Malaysian language and culture, it was discovered that the MDSI-M is reliable and valid in determining the types of Malaysian drivers [20]. However, this raises the question of whether a questionnaire is the best way to evaluate a person's driving style or if it is possible to precisely define it due to the possibility of bias while responding to that question. The driving style was evaluated using self-assessment and a vehicle data recorder, which included many items about particular driving characteristics to date.

Firstly, before constructing validity, the question that needed to be answered was whether the self-assessment questionnaire was reliable or otherwise in reflecting Malaysian driving style. Therefore, Cronbach's alpha was performed to measure the internal consistency (a measure of reliability) of each factor in the MDSI-M scale, and the result was good and acceptable (from $r = 0.52$ to $r = 0.77$). Each scale of the self-assessments was also analyzed to determine whether they were reliable, and the results also indicated a high level of internal consistency (from $r = 0.70$ to $r = 0.89$). The results showed that self-assessments were reliable for the study. The result is consistent with the recent findings in the literature [10,28].

Secondly, the self-assessment questionnaire was checked for any possible bias because the respondents might answer it based on their experience of driving activities. Therefore, in this study, a designated observer and participant evaluated their driving performance during and after the driving session. The results showed no significance between the observer's (ORO-O) and participant's evaluation (ORO-P). In addition, when driving skills and behaviors were evaluated using a questionnaire relating to general driving competence that was not particular to one driving session, the result showed no significance between the driver's (DEQ) and observer's ratings (DSEQ) for most items. However, when moving to the criterion validity for MDSI-M, the results showed a modest correlation between them (see Table 1). A low-to-medium correlation illustrated in this study proved that there was a correlation between the participant's answers before and after the experiment was completed. However, a slight correlation between the observer's evaluation and the participant's self-rating in the results was also reasonable and was to be expected.

The first reason for this was that the sampling size used in this study was minimal ($n = 41$). Using the GPower sample size calculator, if a medium effect size was selected ($f = 0.25$) with an alpha error probability of 0.05 and power of 95%, it reveals that a sample of around 210 participants was needed [37].

A previous study showed that male drivers were more likely to have a higher correlation for risk behavior than female drivers because of the considerable gap in the sample size [38]. Secondly, the findings showed that the observer reports were based on small conversation tasks in a brief period. A possible explanation for the results may be that the on-road evaluation focused only on 30 min without including the familiarization phase. The differences can be partly explained by the proximity of the study [28] when it was performed over a more extended period. The results showed a small-to-medium correlation between the observer and the participant evaluations. Most recent studies also showed

at least a correlation between the self-assessment of the observer and participant when increasing the experiment period [24,38,39]. Hence, some behaviors may change if the driving activity period is extended.

Next, the result showed an expected outcome when most drivers appraised their driving competence as overly positive. The response rate was $N = 29$, which is the total number of drivers who evaluated their driving performance as better than the expert's evaluation. At the same time, the value $N = 12$ was either lower or the same as the observer's evaluation number. These results match those of previous work [28], reiterating the earlier finding that drivers tend to overestimate their driving skills. Next, an analysis was performed to determine what demographic factors caused the participant to overestimate their driving skill. The results showed that one unexpected finding was found from the ANOVA analysis (see Table 3). The difference between genders was significantly related to the difference between DEQ and DSEQ. Even though there was a new contribution shown in these studies, contrary expectations supported the previous research [28], which found a significant difference between age and annual mileage. The lack of age difference between the participants may be a possible explanation for these results. For example, ref. [40] did not have a requirement to set an age limit for each participant, while this study focused only on novice drivers, most of whom were students (aged 17–29 years old). There are several possible explanations for this: if the variety of age ranges increased, the possibility of seeing a high correlation for each type of driver would also increase [41]; next, the objective focused on novice drivers because they are the future users of automated vehicles. So, it was relevant to use the age limitation to increase the possibility of validating the hypothesis of these studies, indicating a need to broaden the sample to include a more diverse and wide age range. Next, the Wilcoxon test was used to compare each item in the driving evaluation and driving self-evaluation questionnaire (see Table 4). It is interesting to note that only three items were significant, while the rest were not. Table 4 also shows that most items evaluated were the same for the observer and participant. This means there was no difference in the evaluation for the participant and the designated observer. In conclusion, there was no bias when the study was performed for driving style in the self-assessment questionnaire because the error tendency for the self-assessment was lower.

The findings from the current study are consistent with findings from the literature, indicating substantial relationships between self-reported driving style ratings and driving behavior in a driving simulator [10,24]. This finding supports the notion that the MDSI results are predictive of driving behavior in a simulator. The MDSI scale could distinguish each type of driver using Barlett's method [12]. The results in Table 5 show that there is a significant relationship between them. The results provided a significantly low negative correlation for the MDSI's "careful driver" score ($r = -0.28$) with cautious acceleration.

Meanwhile, there was a significantly low positive correlation ($r = 0.18$) comparing the acceleration from "risky driver" and so on (see Table 5). This result was in line with the previous literature by observing the correlation between acceleration data and scores from the self-assessment [10,31]. Together, these results provide important insights for careful drivers, with a higher score for cautious driving with lower acceleration, while there was high acceleration for "angry drivers". These findings further support the idea that a driver who accelerates quickly must apply the brakes more forcefully, illustrated by the higher acceleration for the "risky driver" [42]. Meanwhile, the "careful driver" slowly used the lower gears and brakes, resulting in a low acceleration [24].

4. Conclusions

In conclusion, the findings indicate that ongoing research is appropriate to investigate driving style when the self-assessment method is performed based on objective, sensitive,

and strongly correlated indicators. In addition, the results were generally consistent with prior studies on driving simulators; even the current research was performed using on-road situations, demonstrating that the MDSI is predictive of driving style. Next, the study suggested that MDSI can be utilized as a diagnostic instrument to detect the usual driving behavior of individuals by collecting acceleration data.

The results from this study indicate that for the first hypothesis, there are no concrete correlations between the observation and self-rating evaluation. The results for the second hypothesis show that the driving styles of “careful”, “risky”, and “dissociative” correlate with the respective acceleration observed, while the driving styles of “angry” and “anxious” do not have any correlation with the same comparison.

Meanwhile, a possible limitation of this study might be the ability to control all the experiment variables, such as road situation and environment, when the driving activities are performed on the road compared to a simulation. This limitation is due to the fact that in a simulation study, the setup is under control, reproducible, and standardized. Meanwhile, the situation on-road is not the same every time the study is conducted, unlike driving circumstances utilizing simulators. Hence, each participant may experience different conditions during the experiment using on-road conditions.

Author Contributions: Conceived and designed the experiments, M.Z.A.K. and J.K.; Performed the experiments, M.Z.A.K. and J.K.; Analyzed and interpreted the data, M.Z.A.K., J.K. and N.M.; Wrote the original paper, M.Z.A.K. and J.K.; Wrote the revised manuscript, M.Z.A.K., J.K., N.M., N.M.Y., M.Z.H., M.Z.B., Z.M.J. and M.R. All authors have read and agreed to the published version of the manuscript.

Funding: This research is fully supported by a national grant from Ministry of Higher Education (MOHE), Malaysia with funding number of FRGS/1/2020/TK02/UTEM/02/1.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki, and approved by the Institutional Review Board (or Ethics Committee) of Universiti Teknikal Malaysia Melaka on 1 April 2019 (Reference letter: UTeM.23.01/500-25/4 Jilid 5(30)).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: All data are available from the authors.

Acknowledgments: The authors thank the Ministry of Higher Education (MOHE) and Universiti Teknikal Malaysia Melaka (UTeM) for the approved funds and support to make this vital research viable and effective. This research is fully supported by a national grant, FRGS/1/2020/TK02/UTEM/02/1. The ethics committee has approved this study (UTeM.23.01/500-25/4 Jilid 5(30)).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Ishibashi, M.; Okuwa, M.; Doi, S.; Akamatsu, M. Indices for Characterizing Driving Style and Their Relevance to Car Following Behavior. In Proceedings of the SICE Annual Conference, Takamatsu, Japan, 17–20 September 2007; pp. 1132–1137. [CrossRef]
2. Yigitcanlar, T.; Kamruzzaman, M. Smart cities and mobility: Does the smartness of Australian cities lead to sustainable commuting patterns? *J. Urban Technol.* **2019**, *26*, 21–46. [CrossRef]
3. Qiao, Z.; Sun, X.; Loeb, H.; Mangharam, R. Drive Right: Shaping Public’s Trust, Understanding, and Preference Towards Autonomous Vehicles Using a Virtual Reality Driving Simulator. In Proceedings of the 2023 IEEE Intelligent Vehicles Symposium (IV), Anchorage, AK, USA, 4–7 June 2023; pp. 1–8.
4. Delmas, M.; Camps, V.; Lemerrier, C. Personalizing automated driving speed to enhance user experience and performance in intermediate-level automated driving. *Accid. Anal. Prev.* **2024**, *199*, 107512. [CrossRef] [PubMed]
5. Yang, H.; Zhou, Y.; Wu, J.; Liu, H.; Yang, L.; Lv, C. Human-Guided Continual Learning for Personalized Decision-Making of Autonomous Driving. *IEEE Trans. Intell. Transp. Syst. Early Access* **2025**. [CrossRef]
6. Sagberg, F.; Selpi; Piccinini, G.F.B.; Engström, J. A Review of Research on Driving Styles and Road Safety. *Hum. Factors* **2015**, *57*, 1248–1275. [CrossRef]

7. Cordero, J.; Aguilar, J.; Aguilar, K.; Chávez, D.; Puerto, E. Recognition of the Driving Style in Vehicle Drivers. *Sensors* **2020**, *20*, 2597. [CrossRef] [PubMed] [PubMed Central]
8. Yang, L.; Ma, R.; Zhang, H.M.; Guan, W.; Jiang, S. Driving behavior recognition using EEG data from a simulated car-following experiment. *Accid. Anal. Prev.* **2018**, *116*, 30–40. [CrossRef]
9. Basu, C.; Yang, Q.; Hungerman, D.; Singhal, M.; Dragan, A.D. Do You Want Your Autonomous Car to Drive Like You? In Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction, Vienna, Austria, 6–9 March 2017; pp. 417–425. [CrossRef]
10. Taubman-Ben-Ari, O.; Mikulincer, M.; Gillath, O. The multidimensional driving style inventory—Scale construct and validation. *Accid. Anal. Prev.* **2004**, *36*, 323–332. [CrossRef]
11. van Huysduynen, H.H.; Terken, J.; Martens, J.-B.; Eggen, B. Measuring Driving Styles. In Proceedings of the 7th International Conference on Automotive User Interfaces and Interactive Vehicular Applications, Nottingham, UK, 1–3 September 2015; pp. 257–264. [CrossRef]
12. Karjanto, J.; Yusof, N.M.; Terken, J.; Hassan, M.Z.; Delbressine, F.; van Huysduynen, H.H.; Rauterberg, M. The identification of Malaysian driving styles using the multidimensional driving style inventory. *MATEC Web Conf.* **2016**, *90*, 01004. [CrossRef]
13. Martinussen, L.M.; Hakamies-Blomqvist, L.; Møller, M.; Özkan, T.; Lajunen, T. Age, gender, mileage and the DBQ: The validity of the Driver Behavior Questionnaire in different driver groups. *Accid. Anal. Prev.* **2013**, *52*, 228–236. [CrossRef]
14. Glendon, A.I.; Dorn, L.; Matihews, G.; Gulian, E.; Davies, D.R.; Debney, L.M. Reliability of the driving behaviour inventory. *Ergonomics* **1993**, *36*, 719–726. [CrossRef]
15. French, D.J.; West, R.J.; Elander, J.; Wilding, J.M. Decision-making style, driving style, and self-reported involvement in road traffic accidents. *Ergonomics* **1993**, *36*, 627–644. [CrossRef] [PubMed]
16. West, R.; Hall, J. The role of personality and attitudes in traffic accident risk. *Appl. Psychol.* **1997**, *46*, 253–264. [CrossRef]
17. Holman, A.C.; Havârneanu, C.E. The Romanian version of the multidimensional driving style inventory: Psychometric properties and cultural specificities. *Transp. Res. Part F Traffic Psychol. Behav.* **2015**, *35*, 45–59. [CrossRef]
18. Long, S.; Ruosong, C. Reliability and validity of the Multidimensional Driving Style Inventory in Chinese drivers. *Traffic Inj. Prev.* **2019**, *20*, 152–157. [CrossRef]
19. Padilla, J.L.; Castro, C.; Doncel, P.; Taubman, O. Adaptation of the multidimensional driving styles inventory for Spanish drivers: Convergent and predictive validity evidence for detecting safe and unsafeAC driving styles. *Accid. Anal. Prev.* **2020**, *136*, 105413. [CrossRef]
20. Kamaludin, M.Z.A.; Bin Karjanto, J.; Muhammad, N.; Yusof, N.M.; Hassan, M.Z.; Zulkifli, A.F.; Ab Rashid, A.A. Exploration of Automated Vehicle’s Driving Style by Adapting Multidimensional Driving Styles Inventory for Malaysian Drivers. *Adv. Transp. Stud.* **2022**, *58*, 229–244. [CrossRef]
21. Sümer, N.; Özkan, T.; Lajunen, T. Asymmetric relationship between driving and safety skills. *Accid. Anal. Prev.* **2006**, *38*, 703–711. [CrossRef]
22. Shanker RJonas, A.; Devitt, S.; Huberty, K.; Flannery, S.; Greene, W.; Swinburne, B.; Locraft, G.; Wood, A.; Weiss, K.; Moore, J. Autonomous cars: Self-driving the new auto industry paradigm. *Morgan Stanley Blue Pap.* **2013**, 1–109.
23. Beiker, S.A. Legal aspects of autonomous driving. *Santa Clara L. Rev.* **2012**, *52*, 1145. [CrossRef]
24. van Huysduynen, H.H.; Terken, J.; Eggen, B. The relation between self-reported driving style and driving behaviour. A simulator study. *Transp. Res. Part F Traffic Psychol. Behav.* **2018**, *56*, 245–255. [CrossRef]
25. WÅhlberg, A.A.F. *Driver Behaviour and Accident Research Methodology: Unresolved Problems*; CRC Press: Boca Raton, FL, USA, 2017.
26. West, R.; French, D.; Kemp, R.; Elander, J. Direct observation of driving, self reports of driver behaviour, and accident involvement. *Ergonomics* **1993**, *36*, 557–567. [CrossRef] [PubMed]
27. Holland, C.; Geraghty, J.; Shah, K. Differential moderating effect of locus of control on effect of driving experience in young male and female drivers. *Personal. Individ. Differ.* **2010**, *48*, 821–826. [CrossRef]
28. Amado, S.; Arian, E.; Kaça, G.; Koyuncu, M.; Turkan, B.N. How accurately do drivers evaluate their own driving behavior? An on-road observational study. *Accid. Anal. Prev.* **2014**, *63*, 65–73. [CrossRef] [PubMed]
29. Möller, H.-J. Effectiveness studies: Advantages and disadvantages. *Dialogues Clin. Neurosci.* **2011**, *13*, 199–207. [CrossRef]
30. Vaitkus, V.; Lengvenis, P.; Zylius, G. Driving style classification using long-term accelerometer information. In *2014 19th International Conference on Methods and Models in Automation and Robotics (MMAR)*; IEEE: Piscataway, NJ, USA, 2014; pp. 641–644. [CrossRef]
31. Murphey, Y.L.; Milton, R.; Kiliaris, L. Driver’s Style Classification Using Jerk Analysis. In Proceedings of the 2009 IEEE Workshop on Computational Intelligence in Vehicles and Vehicular Systems, CIVVS 2009—Proceedings, Nashville, TN, USA, 30 March–2 April 2009; pp. 23–28. [CrossRef]
32. Khorram, B.; WÅhlberg, A.E.A.; Kashani, A.T. Longitudinal jerk and acceleration as measures of safety in bus rapid transit drivers in Tehran. *Theor. Issues Ergon. Sci.* **2020**, *21*, 577–594. [CrossRef]

33. Itkonen, T.H.; Pekkanen, J.; Lappi, O.; Kosonen, I.; Luttinen, T.; Summala, H. Trade-off between jerk and time headway as an indicator of driving style. *PLOS ONE* **2017**, *12*, e0185856. [CrossRef]
34. Feng, F.; Bao, S.; Sayer, J.R.; Flannagan, C.; Manser, M.; Wunderlich, R. Can vehicle longitudinal jerk be used to identify aggressive drivers? An examination using naturalistic driving data. *Accid. Anal. Prev.* **2017**, *104*, 125–136. [CrossRef]
35. Pande, A.; Chand, S.; Saxena, N.; Dixit, V.; Loy, J.; Wolshon, B.; Kent, J.D. A preliminary investigation of the relationships between historical crash and naturalistic driving. *Accid. Anal. Prev.* **2017**, *101*, 107–116. [CrossRef]
36. Yan, F.; Liu, M.; Ding, C.; Wang, Y.; Yan, L. Driving style recognition based on electroencephalography data from a simulated driving experiment. *Front. Psychol.* **2019**, *10*, 1254. [CrossRef]
37. Faul, F.; Erdfelder, E.; Lang, A.-G.; Buchner, A. G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behav. Res. Methods* **2007**, *39*, 175–191. [CrossRef]
38. Zardosht, M.; Beauchemin, S.S.; Bauer, M.A. Identifying Driver Behavior in Preturning Maneuvers Using In-Vehicle CANbus Signals. *J. Adv. Transp.* **2018**, *2018*, 5020648. [CrossRef]
39. Choi, S.; Kim, J.; Kwak, D.; Angkititrakul, P.; Hansen, J.H.L. Analysis and classification of driver behavior using in-vehicle can-bus information. *Bienn. Workshop DSP Veh. Mob. Syst.* **2015**, *2007*, 17–19.
40. Harris, P.B.; Houston, J.M.; Vazquez, J.A.; Smither, J.A.; Harms, A.; Dahlke, J.A.; Sachau, D.A. The Prosocial and Aggressive Driving Inventory (PADI): A self-report measure of safe and unsafe driving behaviors. *Accid. Anal. Prev.* **2014**, *72*, 1–8. [CrossRef]
41. Zimbardo, P.G.; Keough, K.A.; Boyd, J.N. Present time perspective as a predictor of risky driving. *Personal. Individ. Differ.* **1997**, *23*, 1007–1023. [CrossRef]
42. Götze, M.; Bißbort, F.; Petermann-Stock, I.; Bengler, K. ‘A careful driver is one who looks in both directions when he passes a red light’—increased demands in urban traffic. In *International Conference on Human Interface and the Management of Information*; Springer: Berlin/Heidelberg, Germany, 2014; pp. 229–240.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

Article

Is Users' Trust during Automated Driving Different When Using an Ambient Light HMI, Compared to an Auditory HMI?

Rafael Cirino Gonçalves ^{1,*}, Tyron Louw ¹, Yee Mun Lee ¹, Ruth Madigan ¹, Jonny Kuo ², Mike Lenné ² and Natasha Merat ¹

¹ Institute for Transport Studies, Faculty of Environment, University of Leeds, Leeds LS2 9JT, UK

² Seeing Machines, Melbourne, VIC 3066, Australia

* Correspondence: trarg@leeds.ac.uk; Tel.: +44-791-928-2338

Abstract: The aim of this study was to compare the success of two different Human Machine Interfaces (HMIs) in attracting drivers' attention when they were engaged in a Non-Driving-Related Task (NDRT) during SAE Level 3 driving. We also assessed the value of each on drivers' perceived safety and trust. A driving simulator experiment was used to investigate drivers' response to a non-safety-critical transition of control and five cut-in events (one hard; deceleration of 2.4 m/s², and 4 subtle; deceleration of ~1.16 m/s²) over the course of the automated drive. The experiment used two types of HMI to trigger a takeover request (TOR): one Light-band display that flashed whenever the drivers needed to takeover control; and one auditory warning. Results showed that drivers' levels of trust in automation were similar for both HMI conditions, in all scenarios, except during a hard cut-in event. Regarding the HMI's capabilities to support a takeover process, the study found no differences in drivers' takeover performance or overall gaze distribution. However, with the Light-band HMI, drivers were more likely to focus their attention to the road centre first after a takeover request. Although a high proportion of glances towards the dashboard of the vehicle was seen for both HMIs during the takeover process, the value of these ambient lighting signals for conveying automation status and takeover messages may be useful to help drivers direct their visual attention to the most suitable area after a takeover, such as the forward roadway.

Keywords: automated vehicles; transition of control; trust in automation; human machine interfaces; gaze behaviour; Light-band displays

1. Introduction

The latest automated driving technology currently available on the market assumes both lateral and longitudinal vehicle control, allowing the driver to physically disengage from the driving task (SAE L2, SAE, [1]). However, these automated driving systems still require the driver to monitor the driving task and surrounding environment and intervene when necessary or when requested to do so by the system [1]. The technology enabling higher levels of automated driving is constantly developing, with some flagship vehicle models now equipped with SAE Level 3 automated driving systems that are legal to be used in some road and driving conditions, e.g., the newly-released Mercedes-Benz Drive Pilot [2]. These systems (and current recommendations—e.g., SAE, 2021) allow the driver to physically and cognitively disengage from the driving task while automation is active, engaging in Non-Driving-Related Tasks (NDRTs), such as reading an email or text message. However, drivers are still responsible for resuming control if requested by the system, for example, due to an operational limitation (e.g., road works, inclement weather), the end of the trip, or a system malfunction.

However, one of the primary human factors challenges for Level 3 (L3) automated driving systems (as defined by [1]) is ensuring that drivers are capable of responding to requests to intervene (RtI) in a timely and safe fashion. The main touted benefit of L3

automated driving is that the driver can engage in other activities (NDRTs), which is likely to take their eyes (and mind) away from the forward roadway and the main driving task. Simulator and test-track studies with L2 driving have shown that, if drivers spend too much time looking away from the forward road, they are more at risk of colliding into an impending lead obstacle, especially if the driver is not expecting to resume control from the vehicle or not given sufficient time to do so [3]. Level 3 functionalities, on the other hand, are expected to provide drivers with sufficient time to take over control, bringing them back into the control loop, with a suitably long time budget for a response (currently recommended to be around 10 s [4]). In this case, drivers should feel comfortable to engage in NDRTs while the automation is engaged. This may pose a challenge for system designers, ensuring that the automated driving system engenders the appropriate level of trust that balances the driver's desire to engage in an NDRT against the system's limitations and its requirement that the driver responds to a takeover or request to intervene (RtI). One approach to tackle this issue is to somehow keep drivers in the "control loop" by informing them of the automated system's status, allowing them to engage in NDRTs while the automation is engaged, helping them to maintain enough situation awareness [5] to safely resume control, if needed. In this paper, we compare driver response to a takeover request (TOR) by a more traditionally used auditory alert with a novel peripheral light-based display, intended to inform drivers of automation status, enhancing their trust in the system during automated driving and facilitating a smooth transition between L3 and manual driving.

In today's vehicles, system status and functionality, including engagement of driver assistance systems, is primarily communicated through an instrument cluster or Human-Machine Interface (HMI), normally placed behind the steering wheel or on a central console. This information is within the drivers' line of sight and allows them to detect potential incoming hazards with their peripheral vision [6]. Using a manual driving simulator study, ref. [7] found that engaging in a self-paced, visual-manual tracking task with high visual eccentricity (60 degrees) still allowed drivers sufficient on-road glances, suggesting that drivers use peripheral vision to collect evidence for braking during off-road glances.

However, the design solution to convey information about an L3 automated system behind the steering wheel may present issues, given that drivers are unlikely to be facing the road, but rather would be facing the interior of the vehicle, engaging with an NDRT. When drivers are engaged in a visual NDRT, it is difficult for them to detect any changes in the visual information provided by the dash-based HMI. Therefore, they often need to shift their visual attention between the dash area and the NDRT, to confirm system status or an RtI. This is at odds with the intended user-related benefits of higher levels of automation (i.e., comfort and ability to engage in NDRTs). Moreover, if drivers are engaged in an NDRT, with their eyes away from the forward roadway and the dash area, they may need to be alerted with auditory cues to attend to more critical situations, for example, if the system reaches a technical limitation.

Relying on auditory cues alone is potentially problematic, since these could be masked by other auditory noise in the environment, or their meaning could be difficult to decipher amongst a range of similar auditory alerts. Therefore, to enhance comprehension, system notifications/alerts are typically multi-modal, e.g., audio-visual. However, as discussed above, as drivers in L3 automation are allowed to be engaged in NDRTs, they alter their posture and direct their visual and cognitive attention away from the main driving task and road environment. This also means that they may miss the visual element of an audio-visual alert, which is normally placed in an area that would be perceived by a forward-facing driver. This could be problematic for Automated Vehicles (AVs) if system limitations cannot detect sudden or unexpected obstacles which may have been noticed by a forward facing, alert driver. In a manual driving study, ref. [8] showed that mean time-to-collision to a decelerating lead vehicle decreased significantly with increasing eccentricity of an in-car foveal task when measuring brake response for each position of the in-car task. Therefore,

the placement and saliency of existing visual alerts may not be appropriate or perceptible in this new context, where drivers may be out of the traditional driving position in L3 AVs.

According to [9], two of the core design principles of human–automation interaction are to keep drivers in the loop where possible and to support smooth re-entry into the loop. If a system does not allow drivers to maintain adequate system updates or detect important system notifications, it would fail to adhere to these principles and may cause drivers to interrupt their NDRT to check system status, which may lead to lower acceptance and, therefore, use of the system. Studies from [10] have shown that inappropriate levels of trust may change the way humans interact with an automated system, enforcing them to perform more frequent system information checks and therefore increasing the levels of workload for the operator.

In recent years, a number of novel HMI concepts have been proposed to both keep the driver in the loop and support drivers' re-entry into the loop after automated driving. These include tactile [11], visual [12], auditory (including speech- and non-speech-based [13]), and even olfactory cues [14]. Given drivers' likely high engagement in visual NDRTs in the future, the use of ambient Light-Emitting Displays (LEDs) has been explored as a means to provide peripheral light-based cues to drivers about an AV's current state, along with requests for a driver's attention or action. These solutions come after evidence from allied domains suggest that the peripheral visual channel can be used to relay system-related information when the operator is not monitoring the system. In the aviation domain, for example, ref. [15] assessed the effectiveness of current foveal feedback and two implementations of peripheral visual feedback for keeping pilots informed about changes in the status of an automated cockpit system. They found that peripheral visual displays resulted in higher detection rates and faster response times without interfering with the performance of concurrent visual tasks any more than currently available automation feedback.

In the driving context, peripheral ambient light displays have been investigated as potential collision warning tools [16], as lane change decision aids [17], as a means to help modulate drivers' speed [18,19], to guide drivers' attention for identifying targets (road users/obstacles), and for indicating vehicle intention [20,21]. Peripheral ambient light displays have also been used to inform drivers of malfunctioning Advanced Driver Assistance Systems (ADAS) [22] and to facilitate collaborative driving tasks between the driver and the co-driver [18].

Recently, light-based displays have also been applied in the context of automated driving. For example, ref. [23] conveyed contextual information through ambient displays, to assist drivers during takeover requests, and found that this resulted in shorter reaction times and longer times to collision, without increasing driver workload. More commonly, light displays have been able to provide information/warnings to drivers about other road users or the AV's intentions [24]. Research from both manual and automated driving shows that, in general, ambient lights are rated highly by drivers and drivers are responsive to these peripheral cues [17]. Therefore, adding a peripheral light display indicating the current automation state may have potential benefits in terms of reducing the frequency of drivers' glances to the road during automation, increased engagement in NDRTs, and higher perceptions of trust in and comfort during automated driving. However, few studies have investigated the use of these displays to improve drivers' perceptions of trust and safety during automated driving and to facilitate transitions between L3 automation and manual driving.

Therefore, the current driving simulator study addressed this gap via the following research questions.

1. Can an ambient peripheral light display (Light-band HMI) be used to improve drivers' perceived safety and trust during L3 automated driving?
2. How effective is a Light-band HMI for facilitating an effective transition of control between L3 automated driving and manual driving, when compared to an auditory HMI alert?

3. What is the pattern of drivers' eye movements during the takeover process for each type of HMI?

2. Methods

2.1. Participants

Following approval from the University of Leeds Research Ethics Committee (Reference Number: LTTRAN-132), we recruited 41 drivers via an online social media platform. Participant demographic details are displayed in Table 1. Participants received GBP 30 for taking part in the experiment and were free to withdraw at any point.

Table 1. Participant demographics information.

	Males (N = 20; Mean (SD))	Females (N = 21; Mean (SD))
Age (years)	44 (13)	44 (13)
Years with license	25 (13)	24 (12)
Miles driven annually	10,300 (5332)	6642 (3350)

2.2. Equipment

The experiment was conducted in the fully motion-based University of Leeds Driving Simulator (UoLDS), which consists of a Jaguar S-type cab housed in a 4 m-diameter spherical projection dome with a 300°-field-of-view projection system. The simulator also incorporates an 8-degrees-of-freedom electrical motion system. This consists of a 500 mm-stroke-length hexapod motion platform carrying the 2.5T payload of the dome and vehicle cab combination, allowing movement in all six orthogonal degrees of freedom of the Cartesian inertial frame. Additionally, the platform is mounted on a railed gantry that allows a further 5 m of effective travel in surge and sway. Inside the simulator's vehicle cabin, a Lilliput 7" VGA touchscreen with 800 × 480 resolution was installed near the gear shift and used for a non-driving-related, secondary task, described below.

A Seeing Machines driver-monitoring system was used to record the participants' eye movements, at 60 Hz. Figure 1 shows the system's world model from the driver's point of view, with annotated areas of interest. Included in this world map is the On Road area, which encompassed the full width of the road in the drivers' current and projected future path.

2.3. Experimental Design

In this experiment, participants completed two experimental drives on a three-lane UK motorway with ambient traffic. Each experimental drive lasted ~17 min, with five ~2 min automation segments interspersed with ~1 min manual driving segments (Figure 2). There were five non-critical takeover requests per drive and 10 takeovers in total. The entire experiment lasted approximately 2 h. In both experimental drives, during automated driving, participants were instructed to engage in a visual non-driving-related "Arrows" task (NDRT [25]). The Arrows task requires participants to search for, and touch, the upward-facing Arrow, displayed in a 4 × 4 grid of Arrows, using a touch screen in the centre console. Each time the upward-facing arrow is correctly identified and selected, a new grid of arrows is generated. The screen displayed the current participant's cumulative score and a 'score to beat' to keep them engaged in the task. Participants were informed that they would receive an additional GBP 10 if they managed to beat this score, in addition to the agreed GBP 20 received for their participation. However, due to ethical reasons, all participants received the full payment of GBP 30 at the end of the experiment, regardless of their final score on the Arrows task. Analysis of performance in the Arrows task indicated drivers had constant attention to the NDRT throughout the automation period.

A



B

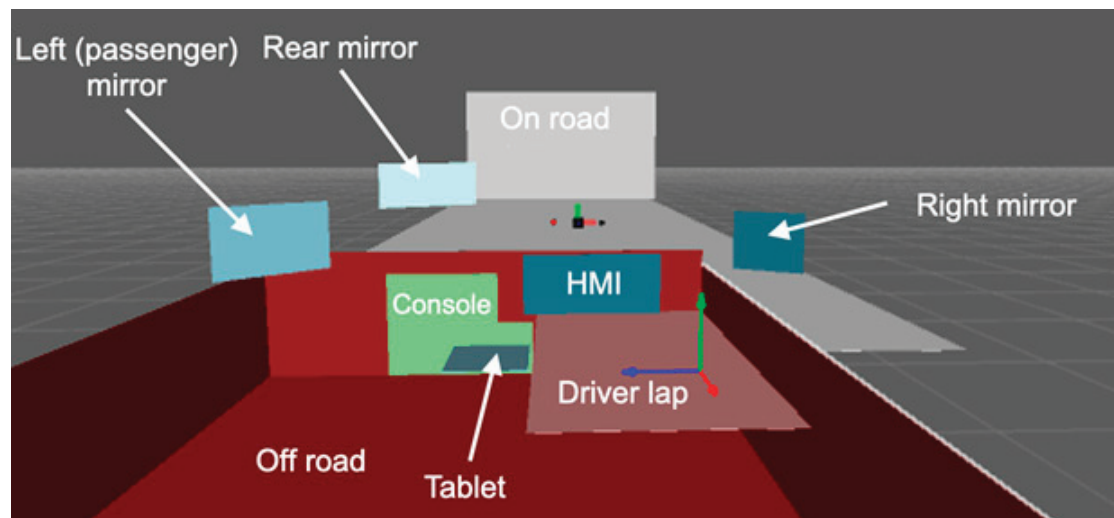


Figure 1. Visual representation of the interior of the simulator cab (A), and a screenshot of the Seeing Machines driver-monitoring system world model, with annotated areas of interest (B).

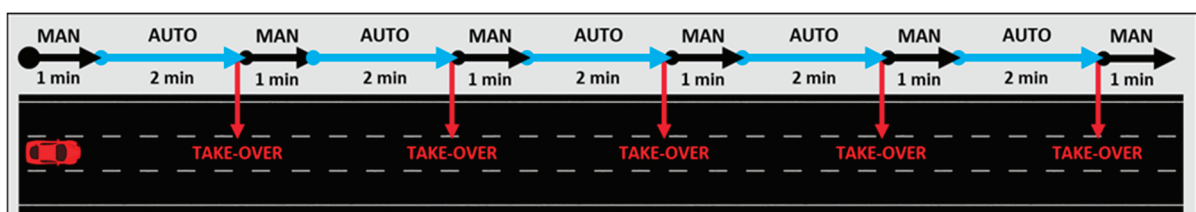


Figure 2. Schematic representation of each experimental drive, with five 2 min automation driving segments interspersed with six 1 min manual driving events, separated by five takeover events.

A 2×5 within-participant design was used for this study, with the factors HMI type (Light-band, Auditory) and takeover number (1–5), which specified the number of times drivers resumed control during the experimental drive, for each HMI condition (Figure 2).

HMI type was fully counterbalanced across participants and specifies the HMI that the drivers were presented with during automated driving and used for the takeover, i.e., Light-band HMI or Auditory HMI. Selection of these two HMIs was motivated by previous work [3] suggesting that drivers in L3 automation do not continue to monitor the environment and the vehicle's dashboard (where information about system status is normally located) during an automated drive. Therefore, efficient Rtl strategies must be able to draw drivers' attention away from other activities (e.g., NDRTs) using salient stimuli, reducing the need for drivers to constantly scan the dash-based interface. For this reason, once drivers' attention was captured by one of the TORs (Auditory or Light-band), we provided the same dash-based information for both experimental conditions after the Rtl in order to keep the additional information provided by the dash-based HMI constant (see Figure 3). Our aim was to compare the benefit of the LED and Auditory stimuli for grabbing drivers' attention, keeping any additional information from the dashboard after the TOR constant.

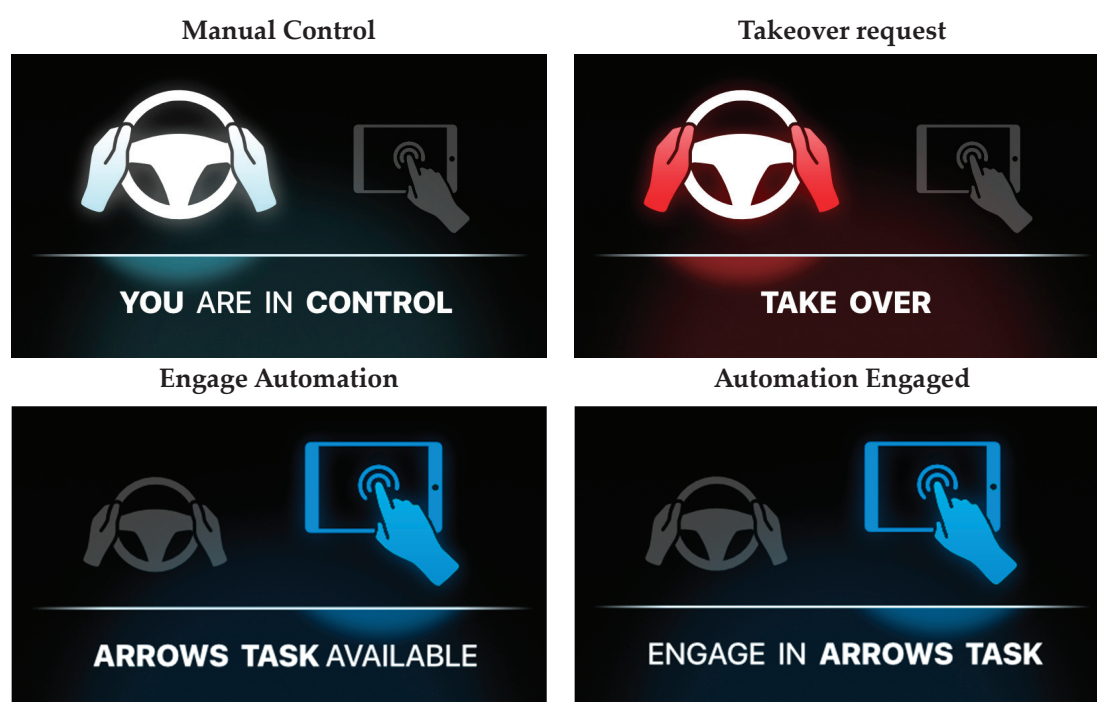


Figure 3. The dash-based HMIs used for the study.

In terms of the instructions provided to drivers for takeover, the same text and symbols (Figure 3) were displayed in the vehicle's dashboard display for both conditions. The dashboard symbols for "Takeover request" and "Engage automation" pulsed at a rate of 2 Hz (in parallel with either the Light-band HMI or Auditory HMI) until the driver resumed control or engaged automation, as required. When active, the display of the symbols for "Manual control" and "Automation engaged" remained constant.

For the Light-band condition, an LED-based Light-band notification system was displayed in the vehicle cabin during automated driving and for signalling takeovers. As shown in Figure 4, the Light-band was placed horizontally along the top of the whole dashboard. To allow a clear distinction between the different stages of a takeover, three different settings were used for the Light-band, which were briefly piloted before the study:

1. When automation was available to be engaged, the Light-band pulsed with a blue light at 2 Hz until the driver turned the automation on.
2. During automated driving, the Light-band displayed a solid blue light to indicate that the automation was operating normally.
3. During takeover requests, the Light-band pulsed with a red light at 2 Hz until the driver resumed manual control.



Figure 4. Placement of the automation status symbol and vehicle speed in the dashboard display (left), and the driver performing the Arrows task during automated driving in the Light-band HMI condition (right). The left picture also shows the Seeing Machines driver-monitoring system on the dashboard, above the steering wheel (white arrow).

The Light-band HMI was not accompanied by any auditory warnings. During manual driving, the Light-band was not active. For the Auditory HMI condition, participants only received an auditory alert (880 Hz, lasting 0.2 s) for engaging/disengaging the automated driving system. No other stimuli were provided during automation in this condition.

2.3.1. Vehicle Cut-In Scenarios

To assess if drivers' desire to engage in the NDRT would be disturbed by changes in the driving environment, and whether this engagement/disturbance would be different for the two HMI conditions, we implemented a number of vehicle cut-in scenarios when automation was engaged. Approximately 60 s into each 2 min automated drive, a vehicle entered the driver's lane from the left-hand lane. This lane change occurred when the time-to-collision between the subject car and the lead vehicle was either 10 s (events 1, 2, 4, and 5) or 4 s (event 3). For events 1, 2, 4, and 5, the desired headway of the subject car was 1.5 s, resulting in a deceleration of 1.14–1.19 m/s² during the cut-in scenario, which was a subtle, but sufficient, cue to indicate the presence of the lead vehicle. However, for event 3, the headway was 0.5 s, resulting in a deceleration of 2.4 m/s², which was a much more salient cue. We hypothesised that each cut-in would signify a change in the driving environment and act as a cue for drivers to pay more attention to the road ahead, perhaps indicating the need to resume control. We assumed that, if drivers trusted the automated system, they would look up to the scene less often during these cut-in scenarios, apart from the more salient event 3. We also hypothesized that, because Light-band provided information about automation status, it would encourage continued engagement in the NDRT, reducing the need for drivers to look up and ensure the automation was still active.

2.4. Driver Disengagement Algorithm

At higher levels of automation, drivers are allowed to engage in NDRTs, but they may still be required to resume control from the vehicle, e.g., at the end of an ODD. In these situations, it is important to ensure that the human driver is fully ready to reengage in the driving task and provides an appropriate level of attention and control, with their hands on the wheel and their eyes and attention to the forward roadway. Therefore, in this study, we developed an algorithm that determined when it was safe to hand control back to the driver. This also prevented any accidental/unwanted disengagements. The algorithm used information from the touch-sensitive steering wheel and eye-gaze data from the driver monitoring system to determine in real time whether the driver had their hands on the wheel and was looking at the road ahead. When the takeover request was issued, the system would only disengage if the participant:

1. Was pulling the stalk to turn automation off.
2. Had at least one hand on the steering wheel (determined via the capacitive steering wheel).

3. Was looking ahead (i.e., On Road) for a sufficient length of time (see below for thresholds).

All three criteria had to be met before disengagement was possible. We implemented two different thresholds for glances ahead (criterion 3 above) depending on whether the participant was already looking On Road.

Threshold 1: if, at the point criteria 1 and 2 were met, the participant was not looking On Road, then at least 60% of the glances in the previous 2 s would need to be On Road before the algorithm would consider that criterion 3 was met.

Threshold 2: if, at the point criteria 1 and 2 were met, the participant was already looking On Road, then at least 40% of the glances in the subsequent 2 s would need to be On Road before the algorithm would consider that criterion 3 was met.

2.5. Procedure

As part of the recruitment process, participants were emailed a screening and demographics questionnaire, which included questions about age, gender, driving experience, and their experience with different Advanced Driver Assistance Systems (ADAS). To be eligible for taking part in the experiment, participants had to hold a valid driving license, have at least one year's experience driving in the UK, and not have participated in a driving simulator study that included interaction with automated vehicles. Prior to arrival, participants were emailed a description of the study and information about COVID-19 procedures during the experiment and were asked to sign a consent form.

Upon arrival at the simulator, the experimenter asked the participant a series of questions to ensure COVID-19 compliance. They were then taken into the building, where the experiment was explained in more detail, and they were given the opportunity to ask questions. Participants were then taken into the simulator dome and the experimenter explained all the safety procedures, the driving controls of the vehicle, and various dashboard icons and gave a description of how to do the Arrows task, as well as engage and disengage the automated driving system. To enable automation, participants were asked to drive in the centre of the middle lane, maintaining the 70 mph speed limit and adhering to the standard rules of the road, ensuring safe operation of the vehicle, throughout the drive. When active, the automated driving system (ADS) assumed lateral and longitudinal vehicle control and maintained a maximum velocity of 70 mph.

Before each of the two experimental drives, participants performed a short practice drive. To avoid confusing participants by showing them both takeover HMIs at the start of the experiment, they were only shown the HMI system that they would experience in the subsequent experimental drive (i.e., Light-band or Auditory HMI).

They were then left in the simulator dome to perform the first practice drive, allowing them to become familiar with the simulator controls and motion system. The experimenter talked with the participant via an intercom system, when required.

The experiment began with the participant driving in manual mode for a couple of minutes, after which they received an instruction from the automated driving system to turn the automation on. This was achieved by pulling the left indicator stalk towards them. Once automation was engaged, participants began performing the Arrows task. After approximately 2 min from the start of each automated driving segment, participants received a notification to take over control. Participants had to meet the criteria of the disengagement algorithm to turn the automation off. To allow a non-critical takeover, there was no lead vehicle or obstacle during the takeovers.

After the practice drive, and after each experimental drive, participants took a short break while they were seated in the driving simulator. At this point, they were asked to complete a short set of questions, outlined in Table 2. The questionnaire assessed drivers' trust and acceptance of the driving automation system for each of the two takeover HMIs. This questionnaire was tailored specifically for this experiment as part of a series of studies for the L3Pilot project [26], and was validated by a group of experts from the project consortium.

Table 2. Experiment questionnaire.

Question	Response Format	Baseline (Pre-Experiment)	Post Drive 1	Post Drive 2
I trust that the vehicle will drive safely; while I do the Arrows task	5-point scale (Strongly disagree-Strongly agree)	x		
I trusted that the vehicle would drive safely while I did the Arrows task	5-point scale (Strongly disagree-Strongly agree)		x	x
If your level of trust in the automated driving system changed since the start of the experiment, please explain why.	Free text		x	x
I felt safe while doing the Arrows task during automated driving	5-point scale (Strongly disagree-Strongly agree)		x	x
In this drive, the Light-band/ Auditory signal was...	5-point scale for each Van der Laan Scale item		x	x
How engaged were you with the Arrows task while automation was on?	10-point (Not at all engaged-Highly engaged)			x
Apart from takeover requests, was there anything that interrupted your engagement in the Arrows task while automation was on? If so, please explain briefly.	Free text			x
Which warning system did you prefer?	Light-band/ Auditory			x

2.6. Statistical Analyses

We analysed data with SPSS V.24 (IBM, Armonk, NY, USA), and generated the visualizations in R and Microsoft Excel. An α -value of 0.05 was used as the criterion for statistical significance, and partial eta-squared was computed as an effect size statistic. Unless otherwise stated, variance of the data was homogenous, as assessed by Levene's test of equality of error variance.

3. Results

3.1. Visual Attention during Automated Driving

We were interested in understanding whether the Light-band HMI increased participants' understanding of automation status and encouraged their trust in its operation, resulting in a more sustained engagement in the Arrows task during the lead vehicle cut-in events, when compared to the Auditory HMI drives (note that the Auditory HMI was only used during takeover requests). To assess this, we calculated drivers' visual attention to the On Road area after each cut-in event and conducted a Chi-Squared analysis on the number of drivers who made at least one fixation to the On Road area of interest (see Section 2.3.1) in the 10 s after the lead vehicle braked, comparing between the two takeover HMIs, across the five lead vehicle cut-in events. Fixations were calculated using a 200 ms threshold, with a standard deviation below 1° for the gaze position [27].

Results showed no difference in the number of fixations to the On Road area [$X^2 = 2.35$, $p = 0.125$] between the two HMI conditions. For both conditions, the proportion of drivers who looked On Road during the cut-in events was relatively low (11% of drivers made at least 1 fixation). There was a significant difference in fixations between the five lead vehicle cut-in events [$X^2 = 37.226$, $p = 0.0001$], with many more fixations to the On Road area after the third event (Figure 5).

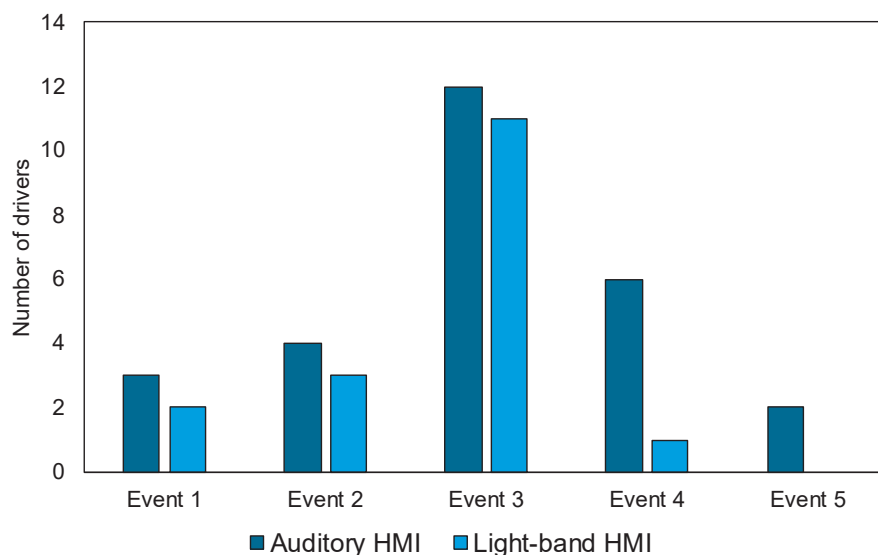


Figure 5. Number of drivers who fixated on the road at least once in the 10 s after a cut-in event.

3.2. Behaviour during the Takeover

To assess the ability of either takeover HMI to capture drivers' visual attention during the takeover, we compared the distribution of drivers' visual attention allocation to different areas of interest (AoIs) (see Figure 6), as well as how long it took drivers to make a fixation to the On Road area and Instrument Cluster, after the TOR. We also examined how long it took drivers to place their hands on the steering wheel (hands on wheel time) and disengage automation (automation disengagement time).

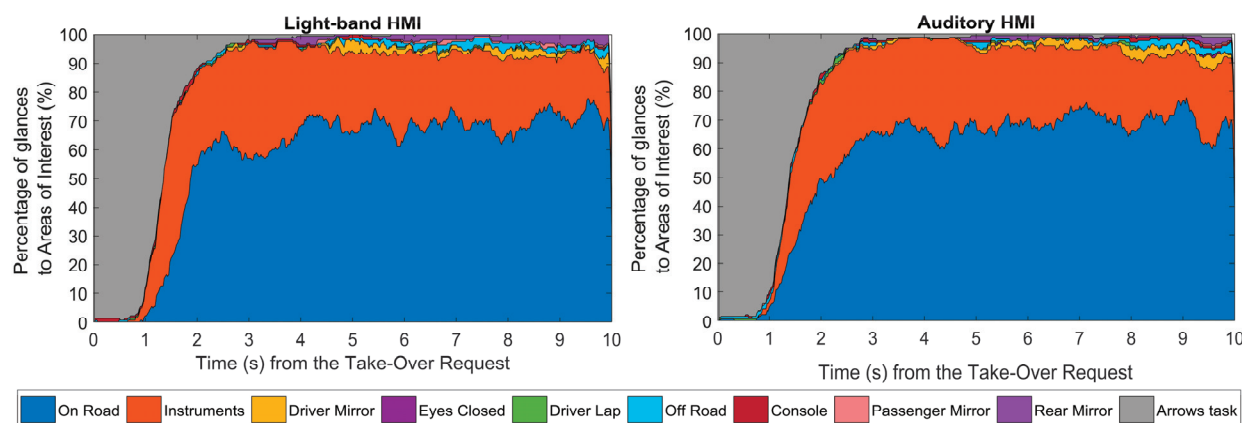


Figure 6. Drivers' visual attention allocation to different Areas of Interest in the 10 s after receiving a takeover request (TOR) for the Light-band HMI (Left) and Auditory HMI (Right) conditions.

An initial analysis of drivers' visual attention distribution to different AoIs showed that both groups took a similar time to shift their attention away from the Arrows task in response to the TORs, glancing primarily between the On Road area and Instrument Cluster. The only observable difference was a small increase in the proportion of glances to the Instrument Cluster (and an associated decrease in glances to the On Road area) approximately 1–2 s after the TOR for the Auditory HMI condition.

We also conducted a Markov-chain gaze transition analysis (see [28]) between areas of interest to investigate any differences between the two HMIs in terms of where drivers were most likely to look first after receiving a TOR. Results confirmed that, for both HMIs, the majority of drivers' first fixations after the Arrows task were directed towards the Instrument Cluster or On Road areas (Table 3). However, drivers were significantly ($p < 0.0001$) more likely to glance towards the Instrument Cluster first after receiving the Auditory TOR (65%), compared to the Light-band TOR (50%). Conversely, drivers were significantly ($p = 0.004$) more likely to glance On Road first after receiving the Light-band TOR, (35%), compared to the Auditory TOR (25%).

To measure the physical response time to the TORs, we assessed hands on wheel time and automation disengagement time. Hands on wheel time was taken from the onset of the TOR to the point at which both hands were detected on the steering wheel. Automation disengagement time was taken from the onset of the TOR to the point at which the driver disengaged the automation using the stalk, and the driving mode turned to manual. We conducted two repeated-measures ANOVAs (2 (HMI Type) $\times$ 5 (Takeover Number)), to investigate the effect of HMIs and takeover number on drivers' hands on wheel time and automation disengagement time, respectively.

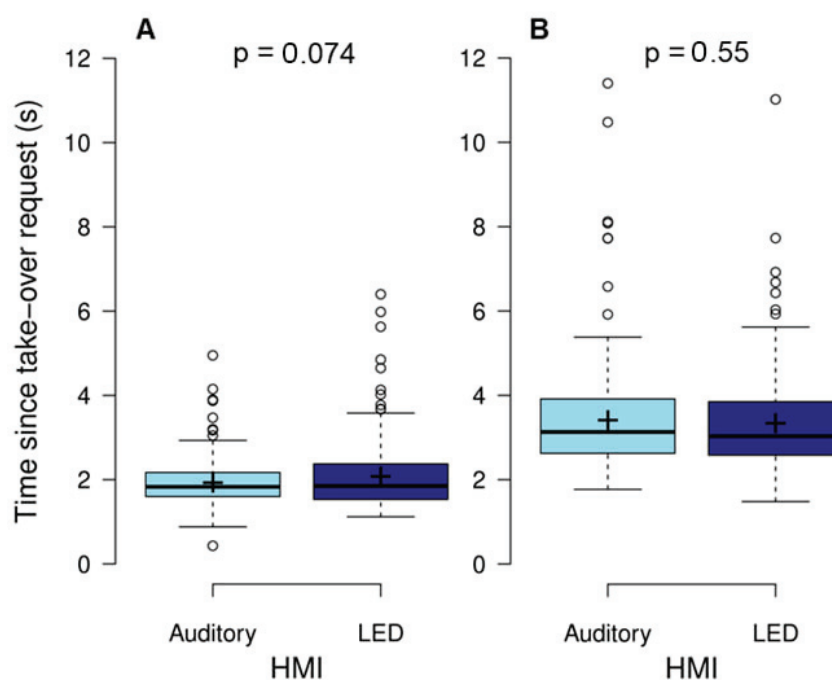
As shown in Table 4 and Figure 7, there was no effect of HMI type on hands on wheel time, although the mean RT was higher for the Light-band HMI condition ($M = 2.16$ s, $SD = 0.14$), compared to the Auditory HMI condition ($M = 1.93$ s, $SD = 0.07$). There was no effect of takeover number, and no interactions.

Table 3. Markov-chain analysis of the location of drivers' first fixation away from the Arrows task after a TOR.

	Console	Driver Lap	Driver Mirror	Instrument Cluster	Off Road	On Road	Passenger Mirror	Rear Mirror
Mann-Whitney U	11,751	12,508	12,525	10,430	12,449	11,225	12,450	12,525
Wilcoxon W	23,076	26,536	23,850	24,458	26,477	22,550	23,775	23,850
Z	−2.04	−0.11	0.00	−2.98	−0.23	−2.01	−0.95	0.00
p	0.04	0.91	1.00	<0.001	0.81	0.04	0.34	1.00
Average Probability (Light-band HMI)	10%	0%	0%	50%	5%	35%	0%	0%
Average Probability (Auditory HMI)	5%	0%	0%	65%	6%	24%	0%	0%

Table 4. Statistical results for takeover analysis.

Effect	HMI Type			Event Number			HMI Type X Event Number		
	$F(df1, df2)$	p	η_p^2	$F(df1, df2)$	p	η_p^2	$F(df1, df2)$	p	η_p^2
Hands on wheel time	3.443 (1,29)	0.074	0.106	1.553 (4,29)	0.19	0.051	1.265 (4,116)	0.228	0.042
Automation disengagement time	0.364 (1,38)	0.55	0.010	0.284 (4,38)	0.88	0.007	0.302 (4,152)	0.867	0.008

**Figure 7.** Time taken from TOR to (A) both hands on wheel and (B) automation disengagement for each HMI. Centre lines show the medians; box limits indicate the 25th and 75th percentiles as determined by R software; whiskers extend 1.5.

There was also no effect of HMI type on automation disengagement time, where the means for the Light-band HMI ($M = 3.46$ s, $SD = 0.15$) and Auditory HMI ($M = 3.36$ s, $SD = 0.14$) conditions were similar (Table 4, Figure 7). There was no effect of takeover number and no interactions.

3.3. Perception of Safety and Trust

Participants were asked to provide a perceived trust rating on a 5-point scale (Strongly disagree to Strongly agree) before the experiment and after each drive. The statement they were asked to rate was 'I trust/trusted that the vehicle will drive safely, while I do/did the Arrows task'. A one-way ANOVA was conducted to compare differences in the three perceived trust ratings for the baseline (i.e., pre-experiment) and post-drive ratings (with the Light-band and Auditory HMIs). There was no significant difference between the conditions ($F(2,78) = 2.86$, $p = 0.084$, $\eta_p^2 = 0.068$; Figure 8).

Participants were also asked to rate 'I felt safe while doing the Arrows task during automated driving' on a 5-point scale (Strongly disagree to Strongly agree, Figure 8). We compared responses using a Kruskal–Wallis test. Overall, participants reported feeling safe while doing the Arrows task during automated driving, but there was no difference between the Light-band and Auditory HMI conditions [$Z(1, 40) = 0.37$, $p = 0.713$].

At the end of the second drive, we also asked drivers whether there was anything apart from takeover requests that interrupted their engagement in the Arrows task while automation was on. 76% of participants ($n = 31$) reported that either a lead vehicle en-

tering the lane or their vehicle braking interrupted their engagement in the Arrows task, confirming that the kinematic cue of the braking lead vehicle was felt by these drivers.

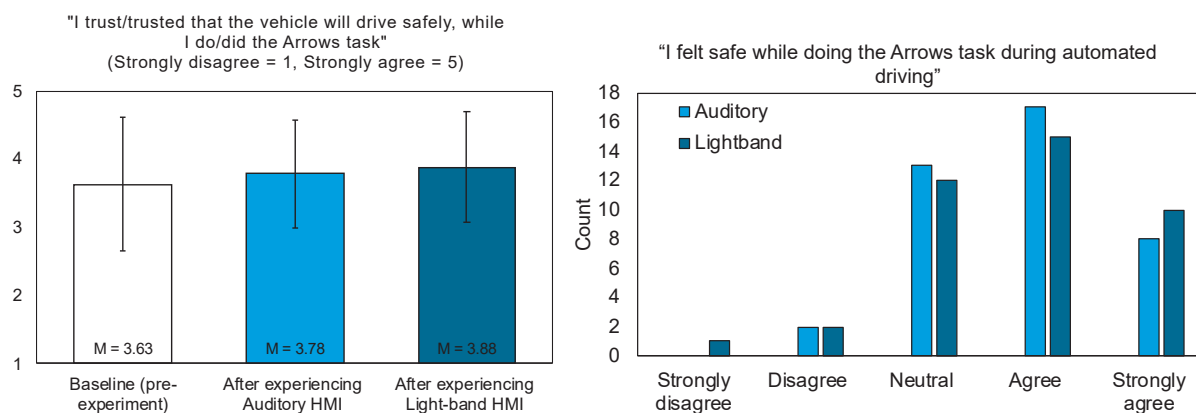


Figure 8. Mean scores of participants' responses to the statement "I trust/trusted that the vehicle will drive safely, while I do/did the Arrows task" (Left) and distribution of participants' responses to the statement "I felt safe while doing the Arrows task during automated driving" (Right). Note: The Auditory HMI was only used during takeover requests.

4. Discussion

Results showed that there were no major differences for the two HMIs in terms of the pattern of drivers' eye movements during the cut-in events, with a similar level of glances away from the NDRT and towards the forward roadway for both HMI conditions. This result suggests that the same degree of trust was bestowed to the automated system, regardless of the HMI condition. Regardless of HMI, the highest number of fixations to the road centre was seen during the third cut-in event (hard brake), which suggests that the kinematic cue was noticeable during this condition, with less reliance on the HMI. Overall, these results go against our initial hypothesis, which predicted that the presence of a peripheral Light-band, indicating that the automated system was engaged and fully functional, would put drivers more at ease, reducing the need to look up during the cut-in events to check automation status. Results from the subjective trust scales also suggest that the two HMIs were trusted equally.

In terms of the distribution of eye gaze at TOR, for both HMI conditions, visual attention was distributed between the road centre (On Road) and the Instrument Cluster areas (with the latter stating clearly whether drivers were in control of the vehicle). Therefore, for both HMI conditions, drivers initially prioritized glances to the Instrument Cluster over the On Road area, after the takeover alert, to confirm their role in the takeover. Similar behaviour has been found by [28–31], who suggested that drivers look at the dash area to receive information about system status, and their own role, before taking an action at TOR. However, our Markov-chain analysis showed that this behaviour was more pronounced when drivers only received an auditory takeover alert, possibly because the auditory alert provided no contextual information and, therefore, while it removed drivers' attention away from the Arrows task, it did not provide any further information about automation status. In contrast, the different colours and patterns of the Light-band HMI provided drivers with additional cues about the automation status, encouraging more drivers to glance On Road first, in favour of the Instrument Cluster. It must be noted that neither HMI concept (Light-band or Auditory) are a substitute for a standard cluster HMI, as, in both cases, drivers still relied on the instrument cluster to check the status of the automated system.

Previous studies investigating gaze patterns during transition of control [3] have found that drivers who were looking at the forward roadway during a TOR were less likely to crash into a lead vehicle, compared to those who were looking to other areas, including the Instrument Cluster. Therefore, the use of supportive information from such peripherally

presented Light-bands may be beneficial for directing drivers' attention to more safety-relevant areas during a TOR, such as the forward roadway (as also observed in [24]). However, further work needs to be done in this context to understand the value of these ambient messages for more critical takeovers, and for a wider range of scenarios, including those which require a response to other areas of the road and not just the forward roadway.

Regarding drivers' performance during the takeover request, there was no significant difference in time to resume control between the two HMI conditions. Further work is required in this context to establish the value of each HMI in assisting drivers for more critical TORs and scenarios, such as those that may lead to a crash if response is too slow. It is important to note that the delay seen in drivers' automation disengagement, which was after they placed their hands on the wheel, reflects the time taken to meet the three criteria of the driver disengagement algorithm. These results demonstrate that it is possible to elicit efficient takeovers that are also safe by using driver monitoring technologies to ensure drivers are following a series of actions while actively disengaging automation. However, it is equally important to support drivers who take a long time to resume control. In the current study, some drivers took longer than 10 s to disengage automation. Video analysis confirmed that, in these cases, drivers struggled to meet the three criteria of the driver disengagement algorithm. Therefore, more research is needed to understand the optimal parameters for such a disengagement algorithm to ensure that drivers are not unnecessarily hindered from resuming control. At the same time, this technique shows some value in the use of such algorithms, for avoiding unintended takeovers by drivers, while also ensuring that the driver is indeed ready and alert for a takeover.

5. Conclusions

The results of this study showed that driver behaviour and response to a Light-band takeover request was similar to that of an auditory alert. Drivers' level of engagement in the NDRT was also similar for the two takeover HMIs, and eye-tracking data showed that the same proportion of drivers looked at the road ahead during the harsh deceleration event (caused by the lead vehicle), regardless of HMI condition. Drivers' trust levels for the ADS were also quite similar for each drive. These results are somewhat in contrast to our original hypothesis, as we assumed that the presence of the ambient Light-band, which confirmed the automation status, would put drivers more at ease, increasing their trust in the system and precluding them from looking up at the driving scene when the L3 vehicle produced a harsh deceleration. Further work is warranted in this context to establish if more prolonged experience with such ambient lights increases drivers' trust and sustains their engagement in NDRTs. On the other hand, since the Light-band was at least as informative as the Auditory HMI, the use of this sort of stimulus in the vehicle may help reduce the use of the number of sound-based alerts and messages in vehicles.

It is acknowledged that simple light-band displays may not be suitable for conveying complex system-related information to drivers, which is in slight contrast to the conclusions of [24]. However, this type of ambient stimulus may be effective for attracting drivers' peripheral attention during NDRT engagement in L3 vehicles, after which more detailed information about system status and driver responsibility can be given by other displays, such as dash-based HMIs.

In terms of study limitations, the current study only used non-critical takeover scenarios after relatively short periods of automation. Therefore, future research should consider how response to these HMIs differs for more critical scenarios. Future research should also evaluate the value of such peripheral ambient displays for a range of messages and more strategic information, such as system-induced changes and events. Finally, this study investigated the value of a visual and auditory stimulus for takeovers in isolation. Future work should consider how the combination of these stimuli will affect driver takeover response and behaviour in such L3 driving conditions.

Author Contributions: Conceptualization, T.L., Y.M.L., R.M. and N.M.; Methodology, R.C.G., T.L. and N.M.; Software, J.K. and M.L.; Formal analysis, R.C.G.; Investigation, R.C.G. and T.L.; Resources,

J.K. and M.L.; Writing—original draft, R.C.G. and T.L.; Writing—review & editing, R.C.G., Y.M.L. and N.M.; Supervision, Y.M.L. and N.M.; Project administration, N.M. All authors have read and agreed to the published version of the manuscript.

Data Availability Statement: Not applicable.

Acknowledgments: This research was developed in partnership with Toyota Motors Europe (TME). We would like to thank the institution for their financial support, as well as their input on the development of the experiment, that made this study possible.

Conflicts of Interest: The authors declare no conflict of interest.

References

- SAE. Summary of SAE International's Levels of Driving Automation for On-Road Vehicles. In Proceedings of the 8th International Conference on Automotive User Interfaces and Interactive Vehicular Applications, Ann Arbor, MI, USA, 24–26 October 2021; pp. 237–244. [CrossRef]
- Group, M.-B. Mercedes-Benz Backs Redundancy for Safe Conditionally Automated Driving | Mercedes-Benz Group > Innovation > Product Innovation > Autonomous Driving. 2022. Available online: <https://group.mercedes-benz.com/innovation/product-innovation/autonomous-driving/redundancy-drive-pilot.html> (accessed on 12 April 2023).
- Louw, T.; Madigan, R.; Carsten, O.; Merat, N. Were they in the loop during automated driving? Links between visual attention and crash potential. *Inj. Prev. J. Int. Soc. Child Adolesc. Inj. Prev.* **2016**, *23*, 281–286. [CrossRef] [PubMed]
- Unece. GRAV-13-36 Safety Requirements for Automated Driving Systems. 2022. Available online: <https://unece.org/sites/default/files/2022-05/GRVA-13-36e.pdf> (accessed on 12 April 2023).
- Endsley, M.R. Toward a Theory of Situation Awareness in Dynamic Systems. *Hum. Factors Ergon. Soc.* **1995**, *37*, 32–64. [CrossRef]
- Carrasco, M. Visual attention: The past 25 years. *Vis. Res.* **2011**, *51*, 1484–1525. [CrossRef] [PubMed]
- Svärd, M.; Bärgrman, J.; Victor, T. Detection and response to critical lead vehicle deceleration events with peripheral vision: Glance response times are independent of visual eccentricity. *Accid. Anal. Prev.* **2021**, *150*, 105853. [CrossRef] [PubMed]
- Lamble, D.; Laakso, M.; Summala, H.; Laakso, M.A.; Tra HI, E.S. Detection thresholds in car following situations and peripheral vision: Implications for positioning of visually demanding in-car displays. *Ergonomics* **2010**, *42*, 807–815. [CrossRef]
- Lee, J.D. Trust and the teleology of technology Commentary on Hancock (2019) Some pitfalls in the promises of automated and autonomous vehicles. *Ergonomics* **2019**, *62*, 500–501. [CrossRef] [PubMed]
- Parasuraman, R.; Sheridan, T.B.; Wickens, C.D. A Model for Types and Levels of Human Interaction with Automation. *Syst. Hum.* **2000**, *30*, 286–297. [CrossRef] [PubMed]
- Cohen-Lazry, G.; Katzman, N.; Borowsky, A.; Oron-Gilad, T. Directional tactile alerts for take-over requests in highly-automated driving. *Transp. Res. Part F Traffic Psychol. Behav.* **2019**, *65*, 217–226. [CrossRef]
- Feierle, A.; Schlichtherle, F.; Bengler, K. Augmented Reality Head-Up Display: A Visual Support During Malfunctions in Partially Automated Driving? *IEEE Trans. Intell. Transp. Syst.* **2022**, *23*, 4853–4865. [CrossRef]
- Nadri, C.; Ko, S.; Diggs, C.; Winters, M.; Sreehari, V.K.; Jeon, M. Novel Auditory Displays in Highly Automated Vehicles: Sonification Improves Driver Situation Awareness, Perceived Workload, and Overall Experience. *Proc. Hum. Factors Ergon. Soc. Annu. Meet.* **2021**, *65*, 586–590. [CrossRef]
- Tang, Q.; Guo, G.; Zhang, Z.; Zhang, B.; Wu, Y. Olfactory Facilitation of Takeover Performance in Highly Automated Driving. *Hum. Factors* **2020**, *63*, 553–564. [CrossRef] [PubMed]
- Nikolic, M.I.; Sarter, N.B. Peripheral Visual Feedback: A Powerful Means of Supporting Effective Attention Allocation in Event-Driven. *Data-Rich Environ.* **2016**, *43*, 30–38. [CrossRef]
- Danielsson, L.; Lind, H.; Bekiaris, E.; Gemou, M.; Amditis, A.; Miglietta, M.; Stålberg, P. HMI principles for lateral safe applications. In *Universal Access in Human-Computer Interaction, Proceedings of the International Conference on Universal Access in Human-Computer Interaction, Beijing, China, 22–27 July 2007*; Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics); Springer: Berlin/Heidelberg, Germany, 2007; Volume 4555, pp. 330–338. [CrossRef]
- Kunze, A.; Summerskill, S.J.; Marshall, R.; Filtner, A.J. Conveying Uncertainties Using Peripheral Awareness Displays in the Context of Automated Driving. In Proceedings of the 11th International Conference on Automotive User Interfaces and Interactive Vehicular Applications, Utrecht, The Netherlands, 21–25 September 2019; pp. 329–341. [CrossRef]
- Meschtscherjakov, A.; Döttlinger, C.; Rödel, C.; Tscheligi, M. Chase Light: Ambient LED Stripes to Control Driving Speed. In Proceedings of the 7th International Conference on Automotive User Interfaces and Interactive Vehicular Applications, Nottingham, UK, 1–3 September 2015; pp. 212–219. [CrossRef]
- Van Huysduynen, H.H.; Terken, J.; Meschtscherjakov, A.; Eggen, B.; Tscheligi, M. Ambient light and its influence on driving experience. In Proceedings of the 9th International Conference on Automotive User Interfaces and Interactive Vehicular Applications, Oldenburg, Germany, 24–27 September 2017; pp. 293–301.

20. Schmidt, G.J.; Rittger, L. Guiding driver visual attention with LEDs. In Proceedings of the AutomotiveUI 2017-9th International ACM Conference on Automotive User Interfaces and Interactive Vehicular Applications, Oldenburg, Germany, 24–27 September 2017; pp. 279–286. [CrossRef]
21. Trösterer, S.; Wuchse, M.; Döttlinger, C.; Meschtscherjakov, A.; Tscheligi, M. Light my way: Visualizing shared gaze in the car. In Proceedings of the 7th International Conference on Automotive User Interfaces and Interactive Vehicular Applications, Nottingham, UK, 1–3 September 2015; pp. 196–203.
22. Langlois, S. ADAS HMI using peripheral vision. In Proceedings of the 5th International Conference on Automotive User Interfaces and Interactive Vehicular Applications, AutomotiveUI, Eindhoven, The Netherlands, 28–30 October 2013; pp. 74–81. [CrossRef]
23. Borojeni, S.S.; Chuang, L.; Heuten, W.; Boll, S. Assisting drivers with ambient take-over requests in highly automated driving. In Proceedings of the 8th International Conference on Automotive User Interfaces and Interactive Vehicular Applications, Ann Arbor, MI, USA, 24–26 October 2016; pp. 237–244.
24. Dziennus, M.; Kelsch, J.; Schieben, A. Ambient light based interaction concept for an integrative driver assistance system—A driving simulator study. In Proceedings of the Human Factors and Ergonomics Society Europe Chapter 2015 Annual Conference, Groningen, The Netherlands, 14–16 October 2015; Available online: <http://elib.dlr.de/99076/> (accessed on 13 April 2023).
25. Jamson, A.H.; Merat, N. Surrogate in-vehicle information systems and driver behaviour: Effects of visual and cognitive load in simulated rural driving. *Transp. Res. Part F Traffic Psychol. Behav.* **2004**, *8*, 79–96. [CrossRef]
26. Metz, B.; Rosener, C.; Louw, T.; Aittoniemi, E.; Bjorvatn, A.; Worle, J.; Weber, H.; Torrao, G.A.; Silla, A.; Innamaa, S.; et al. L3Pilot Deliverable D3.3—Evaluation and Methods. Available online: https://l3pilot.eu/fileadmin/user_upload/Downloads/Deliverables/Update_07102021/L3Pilot-SP3-D3.3_Evaluation_Methods-v1.0_for_website.pdf (accessed on 12 April 2023).
27. Salvucci, D.D.; Goldberg, J.H. Identifying fixations and saccades in eye-tracking protocols. In Proceedings of the 2000 Symposium on Eye Tracking Research & Applications, Palm Beach Gardens, FL, USA, 6–8 November 2000; pp. 71–78. [CrossRef]
28. Gonçalves, R.; Louw, T.; Madigan, R.; Merat, N. Using Markov chains to understand the sequence of drivers’ gaze transitions during lane-changes in automated driving. In Proceedings of the 10th International Driving Symposium on Human Factors in Driver Assessment, Training, and Vehicle Design, Santa Fe, NM, USA, 24–27 June 2019; Volume 10.
29. Gonçalves Rafael, C.; Louw, T.L.; Quaresma, M.; Madigan, R.; Merat, N. The effect of motor control requirements on drivers’ eye-gaze pattern during automated driving. *Accid. Anal. Prev.* **2020**, *148*, 105788. [CrossRef] [PubMed]
30. Goncalves, R.C.; Louw, T.L.; Madigan, R.; Quaresma, M.; Romano, R.; Merat, N. The effect of information from dash-based human-machine interfaces on drivers’ gaze patterns and lane-change manoeuvres after conditionally automated driving. *Accid. Anal. Prev.* **2022**, *174*, 106726. [CrossRef] [PubMed]
31. Louw, T.; Kuo, J.; Romano, R.; Radhakrishnan, V.; Lenné, M.G.; Merat, N. Engaging in NDRTs affects drivers’ responses and glance patterns after silent automation failures. *Transp. Res. Part F Traffic Psychol. Behav.* **2019**, *62*, 870–882. [CrossRef]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

Article

Transparency Assessment on Level 2 Automated Vehicle HMIs

Yuan-Cheng Liu ^{1,*}, Nikol Figalová ² and Klaus Bengler ¹

¹ Chair of Ergonomics, Technical University of Munich, Boltzmannstr. 15, 85748 Garching, Germany

² Clinical and Health Psychology, University of Ulm, Albert-Einstein-Allee 41, 89069 Ulm, Germany

* Correspondence: yuancheng.liu@tum.de; Tel.: +49-162-215-3303

Abstract: The responsibility and role of human drivers during automated driving might change dynamically. In such cases, human-machine interface (HMI) transparency becomes crucial to facilitate driving safety, as the states of the automated vehicle have to be communicated correctly and efficiently. However, there is no standardized transparency assessment method to evaluate the understanding of human drivers toward the HMI. In this study, we defined functional transparency (FT) and, based on this definition, proposed a transparency assessment method as a preliminary step toward the objective measurement for HMI understanding. The proposed method was verified in an online survey where HMIs of different vehicle manufacturers were adopted and their transparencies assessed. Even though no significant result was found among HMI designs, FT was found to be significantly higher for participants more experienced with SAE Level 2 automated vehicles, suggesting that more experienced users understand the HMIs better. Further identification tests revealed that more icons in BMW's and VW's HMI designs were correctly used to evaluate the state of longitudinal and lateral control. This study provides a novel method for assessing transparency and minimizing confusion during automated driving, which could greatly assist the HMI design process in the future.

Keywords: automated driving; human-machine interface; transparency; assessment method

1. Introduction

Automated vehicles are considered a revolutionary technology that could relieve human drivers from tedious and long-distance drives. In recent years, with the ability to execute both lateral and longitudinal controls, Level 2 driving automation [1] has been commercially available and become more and more prevalent. However, the system's safety could be compromised easily when the role transitions of human drivers and the capabilities and limitations of the automated systems are not well understood. The first fatal accident involving an automated vehicle gives us essential insight into this issue [2]. Despite being initially concluded as a “driver error”, some argued that it is more likely a “designer error” and is possibly owing to the lack of clear boundaries to allocate the responsibilities during driving between human and automated systems [3].

The duty allocation between human and automated systems should be transparent. In the definition of SAE Level 2 automated vehicle (L2 AV), human drivers are required to supervise the Automated Driving System (ADS) and be ready to intervene and perform the remaining driving tasks not performed by the ADS when the ADS is engaged [1]. Hence, to guarantee driving safety when Level 2 ADS is engaged, human drivers should be well informed of the current ADS status. In other words, L2 AVs should continuously provide drivers with the necessary information so that a correct and effortless understanding of AVs' capabilities and safe take-over maneuvers can be achieved [4–6].

Before fully autonomous vehicles are generally adopted, human drivers must rely heavily on the Human-Machine Interface (HMI) to learn about the system state and take the corresponding action when a system boundary is reached. However, there has not been a standardized evaluation procedure or training session to guarantee that human drivers understand the HMI correctly. Furthermore, studies show that a significant number of ADS

drivers did not receive any information regarding the ADS, and the situation worsens for used-car owners [7]. Without correct information, proper training, or enough transparency regarding the ADS, the system capabilities and limitations would be unclear to drivers and could easily lead to more confusion and misuses [8].

Hence, the HMI design must be transparent to support unerring understanding of the automation system status. Naujoks et al. [9] proposed an AV HMI design guideline which aggregates HMI design recommendations from experts and uses heuristic verification as the evaluation method. However, the transparency of the HMI to human drivers is not addressed. It is also pointed out that a standardized test protocol for validation is not yet available [9]. Similarly, in the fields of robotics and vehicle automation, transparency for human-machine interaction has been emphasized [10–13], but a systematic and standardized evaluation method for transparency does not exist.

In this study, a transparency assessment method is proposed as a preliminary step toward a standardized validation protocol. The proposed method is aimed to facilitate a more efficient HMI design process and guarantee driving safety by evaluating the transparency of the HMI. Possible suggestions for a more transparent HMI design were also identified during the course to help increase the transparency of future HMI designs.

Transparency

Transparency of the HMI design has been studied across various ADS levels, and it is considered a basis for trust and acceptance [11,13,14]. Human drivers must develop a correct understanding of the ADS, which in turn supports driving safety. Maarten Schraagen et al. [15] evaluated the effects of providing transparency, post-hoc explanations, or both using videos of Level 2 ADS in different driving conditions. Results showed increased trust, satisfaction, and situational awareness when a moderate level of transparency is provided. Körber et al. [14] investigated whether explaining the take-over request in Level 3 ADS would result in different transparency, trust, or acceptance of the HMI. The study was carried out in a driving simulator. The subjective evaluation of system understanding did suggest that the HMI with explanations could increase transparency. However, further research is necessary to validate the actual improvement of system understanding.

Moving to Level 4 ADS, Pokam et al. [10] defined and divided the transparency into two levels: Robot-to-Human transparency and Robot-of-Human transparency. The former categorizes the information that a robotic system should convey to a human. Contrarily, the latter represents information about the awareness and understanding of the human that the system receives and presents to the human. The authors of the study were interested in the effect of HMI transparency on situational awareness, discomfort feelings, and participants' preferences. Five different HMI designs with varying levels of transparency were evaluated in the driving simulator. It was found that the transparent HMI provides a better understanding of the surrounding, and the participants would prefer higher transparency. In contrast to the study by Korber et al., who measured transparency by rating scales, Pokam et al. defined subjectively different transparency levels as the combinations of information provided on the HMI.

In another study on Level 4 ADS, the system transparency information was treated as a constant across different HMIs [11]. The authors of this study compared four types of interfaces and two types of transparency information (presentation of hazard and intended driving path) using the Wizard-of-Oz paradigm [16]. With three seven-point scale questions [17], the results showed increased system transparency in all three interfaces compared to the baseline one.

In the field of robotics, Chen et al. [18] proposed the situation awareness-based agent transparency (SAT) model, based on Endsley's situation awareness model [19], and categorized three levels of transparency based on the information type. In the first level, information about the current state and goals of the automation is provided; followed by the information regarding the reasoning behind the action in the second level; and finally, human users are supplied with predictive information in the third level. Yang et al. [20]

applied the definition of transparency in SAT to investigate the effect of automation transparency on human operators' trust. Two types of alarms with different transparency levels were designed. The traditional one (less transparent) uses a binary alarm, while the other one (more transparent) uses a likelihood alarm to provide additional information regarding the confidence and urgency level.

However, the definitions of transparency in the literature mentioned above are either subjective or merely used to categorize the information types, which do not disclose the driver's (or user's) understanding and efforts applied to the HMI. Additionally, a systematic and standardized evaluation method for estimating the HMI's transparency is not addressed. Without a standardized evaluation method, researchers could only identify a more transparent HMI design through comparison and were not able to differentiate in detail which components in the HMI design are critical to transparency. It would be even more challenging to estimate the effect of transparency on other measurements, such as trust and acceptance, when there is no standardized evaluation method. Moreover, a transparent HMI design should not solely be a combination of information topics since the information needed to understand the system differs across automation levels, traffic situations and driver characteristics [4,21–23]. In other words, giving the same HMI to different participants under different scenarios would result in different “functional transparency” for each individual. Providing more information (e.g., a higher transparency level in SAT model) is not necessarily what human drivers need. This was demonstrated by Carsten and Martens [4], who concluded that human drivers with more trust in the system tend to prefer HMI with less information. The differences in the information needed also make it more urgent to have a transparency assessment method, so the critical information leading to better understanding could be identified and applied to enable transparent HMI design.

Hence we argue that a standardized assessment method for functional transparency should exist so that the information needs under various conditions and driver characteristics can be identified efficiently and facilitate a transparent HMI design. Furthermore, the exact requirement for HMIs to facilitate correct understanding and minimum efforts could be identified by adopting the transparency assessment method. Here we outline the objectives of this study:

- A standardized and robust transparency assessment method would be proposed.
- Verification of the proposed method would be conducted using commercially available HMI designs.
- Information critical to HMI designs' functional transparency would be identified using the proposed method.

Based on the objectives, we defined the research questions and hypotheses as follows:

- Q1: How sensitive is the proposed transparency assessment method when evaluating different HMI designs and ADS experiences?
 - H1a: There is significant difference in functional transparency among different HMI designs.
 - H1b: There is a significant difference in functional transparency among participants with different ADS experiences.
- Q2: How does the proposed functional transparency relate to self-reported transparency?
 - H2 The higher the functional transparency, the higher the self-reported transparency.
- Q3: How is the information used by participants with different levels of functional transparency?
 - H3: Participants with different levels of functional transparency use different information sources when estimating system states.

To further corroborate the internal and external validity of the proposed assessment method, we evaluate whether different ADS experiences affect the understanding of HMIs

in Hypothesis 1b to obtain internal validity. Although the proposed method uses only objective data, self-reported data is also included in this study and compared to the proposed method in Hypothesis 2 to establish external validity.

2. Materials and Methods

In the literature on human-machine interaction, the transparency level of the HMI is often manipulated as the amount of information provided to humans [10,14,24]. However, as described in the previous section, the information must vary across participants and scenarios. The HMI transparency level in the literature does not reflect human drivers' actual understanding of the environment and might thus diminish safety during automated driving. Hence, we developed a transparency assessment method to evaluate human drivers' true understanding of the environment facilitated by the HMI. In this section, we first define the transparency we attempt to evaluate. Then the study design comparing different HMIs is introduced. Finally, the analysis of the collected data is presented.

2.1. Definition of Transparency

During the interaction between humans and automation, shared goals can only be achieved when there is "a harmonization of control strategies of both actors towards a common control strategy" [25]. This suggests that the ADS should be functional, transparent, and understandable to human drivers in the context of human drivers and automated vehicles. The term functional transparency (FT) is used to be distinguished from transparency in the literature. Note that the transparency in the literature solely represents the information provided by the HMI, while the FT is considered as the resulting understandability of the HMI after the interaction with humans. To evaluate and gain deeper insights, here we list the three basic requirements for the FT.

The first and fundamental one is that the HMI should enable a correct understanding of the ADS states (i.e., minimum mode confusion). As pointed out in the HMI guidelines of Naujoks et al. [9], the HMI should inform human drivers of the current ADS mode and the system state changes. For Level 2 and Level 3 ADS, the responsibility of human drivers during automated driving is constantly changing. The role of human drivers could be passengers during Level 3 drivings and suddenly become drivers when the system reaches its boundary. Hence, clear indications of the system states are indispensable for human drivers to avoid critical situations and to increase the FT.

The second requirement is that human drivers should be well informed of what automated functions could be used and how they should be adopted. Cao et al. [26] emphasizes the importance of "user awareness", which represents the user's understanding of "the available and possible automated driving modes, of the currently active mode and transitions among driving modes". ADS with higher FT should permit correct activation and transition among ADS modes.

The last requirement for FT is also stressed in the literature, where researchers argue that HMI should be efficient and easy to understand (i.e., minimizing the workload) and without confusion so that drivers can stay focused on the road and reduce the response time [4,26].

Combining all three requirements, we define FT as:

how easy it is for users to understand and respond to ADS correctly

More than just a definition is required to develop a standardized transparency assessment method. Quantifiable measurements would be indispensable to make the assessment efficient during evaluation and analysis. When evaluating user experience, self-report measurement is a standard method to scale the constructs like usability, acceptance, and trust [27–29]. However, self-report measures only account for participants' preferences, which might not reflect how easy or successful the interaction with the system is [30].

A novel way to formulate the FT is proposed to approach this issue by including the concept of understandability. In computer science, it is defined as "the capability of the

software product to enable the user to understand whether the software is suitable, and how it can be used for particular tasks and conditions of use” [31]. By adopting the method of estimating code understandability [32], the formulation for FT is

$$T_{functional} = \begin{cases} 0, & \text{if "No" is answered} \\ AU(1 - \frac{TNPU}{TNPU_{max}}), & \text{otherwise} \end{cases} \quad (1)$$

where AU represents the actual understandability, which is acquired through verification questions about the states of the HMI. Since AU is calculated by the percentage of correct answers, it would be a number between 0 and 1. $TNPU$ stands for the time needed for perceived understandability, which is the time required by participants to state whether they understand the HMI or not. $TNPU_{max}$ is the maximum $TNPU$ measured within the targeted HMI cluster and is used to normalize the $TNPU$.

With the proposed formula, the FT of the HMI could be easily estimated from simple experimental setups.

2.2. Study Design

The FTs of a series of HMI images of Level 2 ADS from different brands and scenarios were evaluated to verify the proposed transparency assessment method. Besides the difference in HMI brands, different levels of experience in Level 2 ADS were also considered. The study was carried out on the online survey tool LimeSurvey.

2.2.1. HMI Designs

In this study, HMI designs from BMW (Bayerische Motoren Werke AG, Germany) 3 Series, Tesla (Tesla, Inc, USA.) Model 3 and VW (Volkswagen AG, Germany) Passat were employed. The HMI images used were from the driver’s perspective, where the instrument clusters were shown for the images of BMW and VW, while the touchscreen next to the driver was used for Tesla’s images. The HMI designs on the market were chosen for their distinct design concept, where VW keeps the traditional layout (i.e., speedometer and tachometer) and puts system status on the bottom with relatively smaller icons (Figure 1). On the other hand, Tesla provides detailed information for ego and surrounding vehicles and takes traditional meters away (Figure 2). The HMI design for BMW combines the above two (Figure 3). Furthermore, using existing HMIs could make the results of this study a valuable foundation for future studies concerning various user experiences and mental models of different HMI designs.



Figure 1. Example VW HMI design.

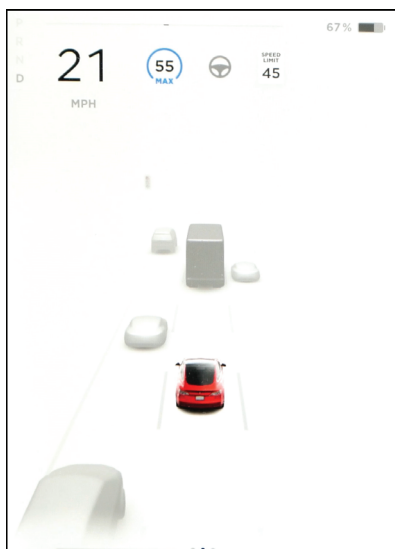


Figure 2. Example Tesla HMI design.



Figure 3. Example BMW HMI design.

Besides different brands, various HMI images under different scenarios were also considered. On Level 2 ADS, longitudinal and lateral control was supported by adaptive cruise control (ACC) and lane-keeping assistance (LKA). Whether these sub-systems are engaged or not, icons on the HMI designs would have different effects, resulting in various scenarios. Together with “whether the front vehicle is detected” and “if the Level 2 ADS could be activated”, all possible HMI images for different scenarios are listed in Table 1. Note that similar or even the same icons might have different meanings across different HMI designs, so the availability of sub-systems is used to avoid confusion. With 11 different scenarios for each brand, a total of 33 HMI images are adopted in this study.

Table 1. HMI scenarios under different conditions.

Scenarios	Is ACC Available?	Is LKA Available? *	Is Front/Side Vehicle Visible?	Is There Warning Signal?
Nothing activated	Yes	No	No	(None)
	Yes	No	Yes	(None)
	Yes	Yes	No	(None)
	Yes	Yes	Yes	(None)
Only ACC activated	(activated)	No	No	(None)
	(activated)	No	Yes	(None)
	(activated)	Yes	No	(None)
	(activated)	Yes	Yes	(None)
ACC and LKA activated (Level 2)	(activated)	(activated)	No	No
	(activated)	(activated)	Yes	No
	(activated)	(activated)	(None)	Yes

* Note: No icon on BMW’s HMI design indicates if the LKA is available. Hence, for the design, this column becomes “Is LKA on standby?”.

2.2.2. Transparency Assessment Test

A transparency assessment test (TRASS) was employed after each HMI image was examined to estimate the FT. The critical questions to assess FT that are included in the TRASS should depend on the scenario the researchers would like to test. In this study, we mainly focused on the HMI of SAE Level 2 automated vehicles, in which the role of the driving automation system by definition is “Performs part of the DDT by executing both the lateral and the longitudinal vehicle motion control subtasks” [1]. Hence the states of these two sub-systems (longitudinal and lateral control systems) are critical to users and should be clearly transmitted. Together with experts’ perspectives and literature [4,8,9,33], the following questions are critical for drivers to acquire a correct understanding of the Level 2 ADS:

1. Is the driving assistance system carrying out longitudinal control?
2. Is the driving assistance system carrying out lateral control?
3. Is the front vehicle detected by the driving assistance system?
4. Is the lane marking detected by the driving assistance system?
5. Can you activate the automated driving assistance (which performs both longitudinal and lateral controls automatically)

Each formal TRASS briefly introduced information regarding Level 2 ADS and its sub-systems (ACC and LKA). Participants were then instructed to answer questions regarding the ADS states based on the upcoming HMI image. After the image was shown, Participants were asked to choose either “Yes, I understand” or “No, I do not understand” based on whether they felt confident answering the HMI image’s ADS states and were required to make the decision as fast as possible. Then, five questions regarding ADS states mentioned above were asked with three options: “Yes”, “No” and “Uncertain”.

The answers to these questions allow us to estimate the actual understanding of the participant regarding the ADS states, which is the AU in Equation (1). On the other hand, the time used to comprehend the HMI, which is the TNPU in Equation (1), is calculated by the time needed by participants to choose either “Yes, I understand” or “No, I do not understand”, depending on whether they understood the HMI image and considered themselves capable of answering the questions mentioned above.

2.2.3. Self-Reported Transparency Test

A self-reported question regarding perceived transparency was asked on a 5-point Likert scale. The question referred to whether the information provided by the HMI was easy to understand (“With the information provided by the HMI image shown, do you agree that ‘this HMI is easy to understand’”), which was later compared to the proposed transparency assessment method.

2.2.4. Information Used Test

The HMI image from each brand having its icons labeled with numbers was shown to the participants. Participants were then asked to identify the icons they used to answer the questions in TRASS: whether ADS is carrying out longitudinal or lateral control, whether front vehicle or lane marking is detected, and whether Level 2 ADS is activated. Both functional and irrelevant icons are labeled with numbers in each HMI image.

2.2.5. Procedure

In total, 33 individuals participated in the 2×3 mixed design. The within-subject factor was three different brands of HMI designs: (1) BMW, (2) Tesla, and (3) VW. Experience in ADS was the between-subject factor in the two groups. Regarding the question “How often have you used the following driving assistance systems in the past 12 months: cruise control, adaptive cruise control, lane-keeping assistance, and automated driving assistance?” those who answered “used sometimes” or “used regularly” are categorized as “experienced” ($n = 16$), and the rest (“rarely used”, “known but never used” and “unknown”)

as “novice” ($n = 17$). Regarding the experiences of certain vehicle brands, among those characterized as “experienced” ($n = 16$), nine of them are experienced in VW’s ADS, four in BMW’s ADS, and four in Tesla’s ADS. Participants were balanced for gender (18 males and 15 females) and age ($M = 29.48$, $SD = 6.00$), and had been driving for 3 years or more ($M = 10.61$, $SD = 6.47$).

After collecting demographic data, calibration tests were conducted to mitigate possible internet delays and individual reaction time differences. The reaction time was later used for calibration in the following analysis. Participants were informed to click either “Yes, I understand” or “No, I do not understand” (randomly assigned) on a dummy figure after pressing the next button. Then, the dummy figure was shown with both options below, which is the same procedure as in TRASS. Practice tests were also presented before the formal TRASS to familiarise the experimental process. The same design and layout of formal TRASS were introduced in the practice test, but still, the section for the HMI image was replaced with a blank figure to avoid any bias.

During the formal TRASS, ten tests were executed for each participant and took around 30 to 40 mins to complete. From 33 HMI images with different HMI brands and scenarios, ten images were randomly chosen for each participant. The process described earlier for TRASS was followed, and a self-reported transparency test was conducted at the end of each TRASS. Finally, the information used test was performed for each brand at the end of the survey.

2.3. Analysis

2.3.1. Transparency Assessment Test

Using the proposed Equation (1), the functional transparency (FT) for each TRASS was calculated with the reaction time (TNPU) and the actual understanding (AU) collected during the test. To determine whether the HMI brands and Level 2 ADS experiences have any effect on transparency, we performed the linear mixed effect analysis with lme4 [34] in the R-4.2.1 programming environment [35]. HMI brand and Level 2 ADS experience were treated as fixed effects in the model, with their interaction term also considered. For random effects, intercepts for participants and scenarios were set, but no by-participant or by-scenario random slopes for either fixed effect could be added as the model could not converge. The representation for the final maximal structure [36] model is:

$$FT \sim \text{HMI brands} * \text{Level 2 ADS experience} + (1 | \text{participant}) + (1 | \text{scenario}) \quad (2)$$

No apparent deviations from homoscedasticity or normality were found with a visual inspection of residual plots. Likelihood ratio tests were applied to models with and without the targeted effect. Further pairwise comparisons were conducted using emmeans [37] also in the R programming environment, where Kenward–Roger degrees-of-freedom approximation and Tukey adjusted p -value were applied.

2.3.2. Self-Reported Transparency Test

The responses were collected and compared to the proposed FT. Their relationship would be determined using Spearman’s correlation.

2.3.3. Information Used Test

Results of icons used to answer each question in the TRASS were collected. Participants with different levels of FT (higher and lower) and their responses to different HMI designs (BMW, Tesla, VW) were analyzed using mixed ANOVA in JASP 0.14.1.0. The post hoc Holm-Bonferroni test was also carried out for comparisons among groups.

3. Results

3.1. Transparency Assessment Test

Table 2 shows the means and standard deviation of FT and self-reported transparency given the HMI designs and ADS experience levels. The data reveal that Self-reported transparency is higher than FT across all conditions. Furthermore, with Equation (2), the corresponding coefficients could be determined. Table 3 shows estimates and standard error of fixed effects, and 95% confidence intervals (abbreviated as 95% Conf. Int.) for the estimates. We are 95% confident that participants experienced in ADS have 0.027 to 0.22 higher FT than novice ones. Finally, Using the linear mixed-effect model with likelihood ratio test, the results indicate that three different HMI designs have no significant effect on FT, $\chi^2(2) = 1.18, p = 0.56$. However, the effect of ADS experience levels on FT was found significant, $\chi^2(1) = 9.64, p = 0.02$, where experienced participants had 0.12 ± 0.048 higher FT than novice participants. No significant interaction effect between HMI designs and ADS experience levels was found, $\chi^2(2) = 2.33, p = 0.33$.

Table 2. Mean and standard deviation (SD) of measured FT and self-reported transparency with different HMI designs and ADS experience levels.

HMI Design	ADS Experience Level	Functional Transparency Mean (SD)	Self-Reported Transparency Mean (SD)	N
BMW	experienced	0.45 (0.26)	0.67 (0.23)	52
	novice	0.32 (0.24)	0.71 (0.21)	54
Tesla	experienced	0.41 (0.24)	0.56 (0.24)	51
	novice	0.34 (0.24)	0.59 (0.23)	60
VW	experienced	0.42 (0.21)	0.62 (0.20)	59
	novice	0.29 (0.21)	0.58 (0.19)	54

Note: The experience level here is regarding the general ADS experience, and not specifically for a certain HMI design.

Table 3. Estimates of fixed effects for the mixed effect model.

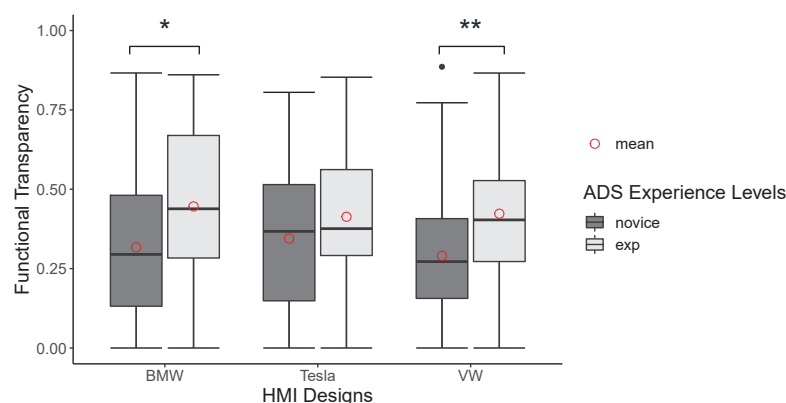
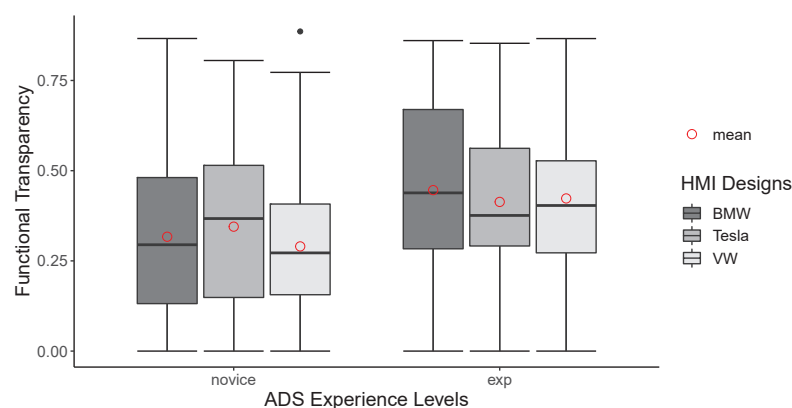
	Estimate	Std. Error	df	t Value	Sig.	95% Conf. Int.	
						Lower Bound	Upper Bound
Intercept	0.45	0.040	65.11	11.24	<0.0005	0.37	0.52
ADS experience: exp.	0	0	.	.	.	.	.
ADS experience: novice	−0.12	0.048	122.01	−2.54	0.012	−0.22	−0.027
HMI design: BMW	0	0	.	.	.	.	.
HMI design: Tesla	−0.035	0.043	309.08	−0.80	0.43	−0.12	0.051
HMI design: VW	−0.018	0.041	302.56	−0.44	0.66	−0.10	0.063

Note: Fixed effects of interactions are not listed.

Further comparisons are shown in Figures 4 and 5. In Figure 4, FT from the same HMI design were grouped, where participants with different experience levels were compared. The result indicates that, more experienced participants showed significantly higher FT in BMW HMI, $t(127) = 2.50, p = 0.014$, and VW HMI, $t(118) = 3.06, p = 0.003$. However, the effect of different ADS experiences levels on Tesla HMI design was insignificant, $t(121) = 1.26, p = 0.21$. In Figure 5, given the same ADS experience level, no significant effect was found across comparisons between different HMI designs. Results of all the pairwise comparisons are listed in Table 4.

Table 4. Results of all pairwise comparisons.

Given Variable	Comparison	Estimate	SE	DOF	t-Value	p-Value
HMI design: BMW	exp - novice	0.12	0.049	127	2.50	0.014 **
HMI design: Tesla	exp - novice	0.06	0.048	121	1.26	0.21
HMI design: VW	exp - novice	0.14	0.047	118	3.06	0.003 ***
ADS experience level: exp	BMW - Tesla	0.035	0.044	313	0.79	0.71
	BMW - VW	0.018	0.042	307	0.43	0.90
	Tesla - VW	−0.017	0.042	305	−0.39	0.92
ADS experience level: novice	BMW - Tesla	−0.027	0.042	318	−0.66	0.79
	BMW - VW	0.040	0.043	320	0.94	0.62
	Tesla - VW	0.068	0.041	305	1.65	0.23

*** $p < 0.01$, ** $p < 0.05$.**Figure 4.** Comparing levels of ADS experience given HMI designs (exp stands for experienced in ADS experience levels). ** $p < 0.05$, * $p < 0.1$.**Figure 5.** Comparing HMI designs given ADS experience levels (exp stands for experienced in ADS experience levels).

3.2. Self-Reported Transparency Test

The correlation between the proposed FT and self-reported transparency was calculated using Spearman's correlation (r_s). The self-reported transparency was normalized between 0 and 1, in the same range as the proposed FT. The result suggests that there was a weak monotonic relationship, $r_s = 0.25$, $p < 0.0001$, between the objectively measured FT and the self-reported transparency.

3.3. Information Used Test

To understand the differences in the information used, we divided the participants into high FT ($n = 16$) and low FT ($n = 17$) groups based on their average FT throughout the total of 10 TRASSes, where the median was chosen as the threshold ($Mdn = 0.39$).

The HMI designs are shown in Figures 6–8, having their icons labeled with a number. The corresponding icons chosen by participants to answer TRASS questions are shown as bar charts in Figures 9–11, with each subplot representing one of the TRASS questions: questions regarding longitudinal control (Long), lateral control (Lat), front vehicle (FV), lane marking (LM), and Level 2 ADS availability (Ava).



Figure 6. BMW HMI design and corresponding icon number.

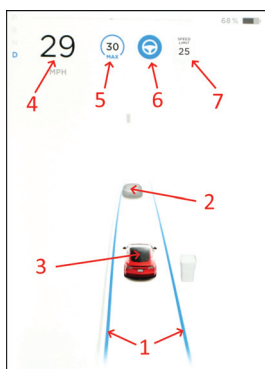


Figure 7. Tesla HMI design and corresponding icon number.



Figure 8. VW HMI design and corresponding icon number.

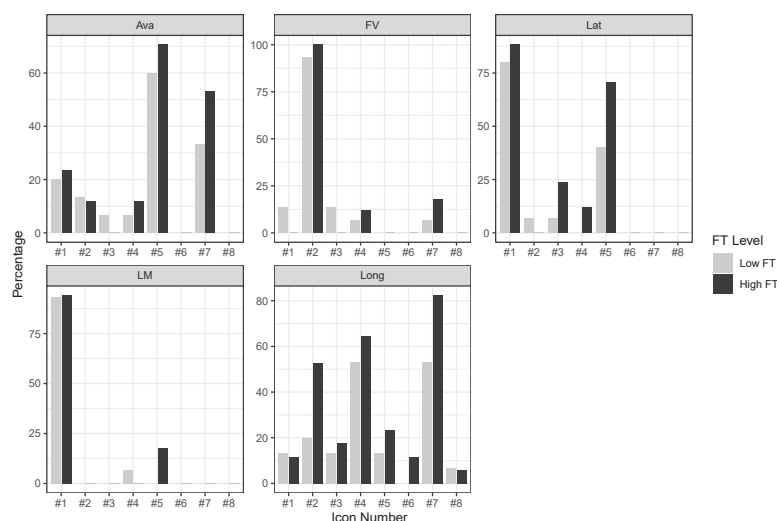


Figure 9. Percentages of BMW icons used to answer TRASS questions from participants with high and low FT.

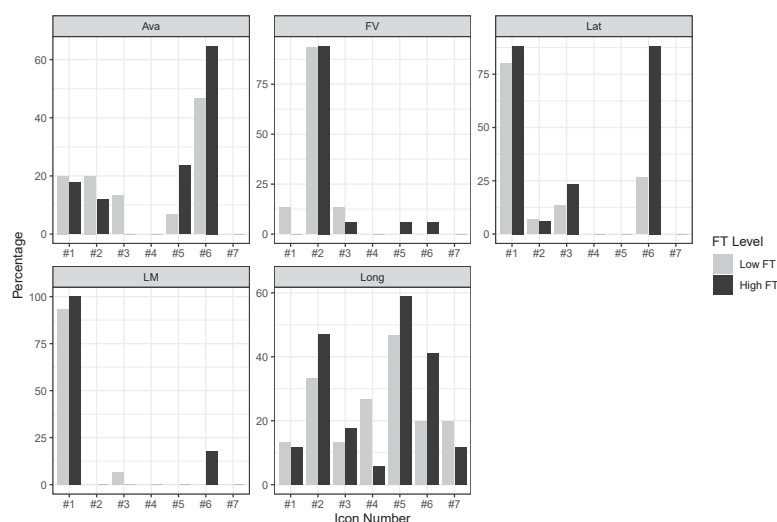


Figure 10. Percentages of Tesla HMI icons used to answer TRASS questions from participants with high and low FT.

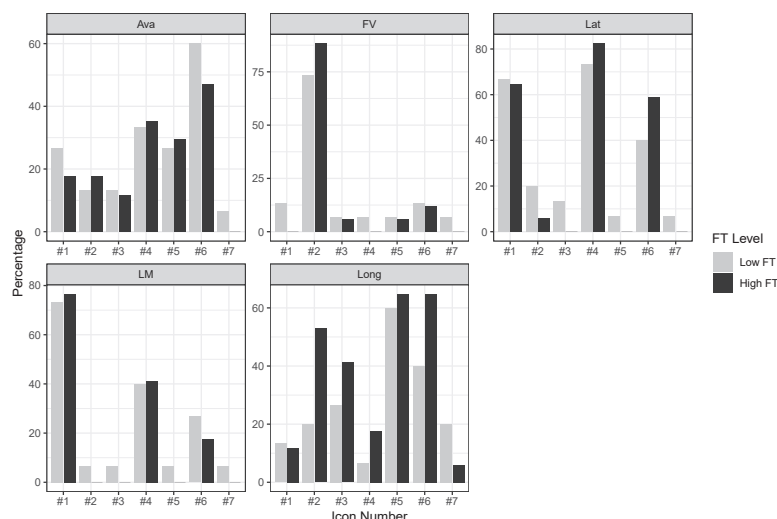


Figure 11. Percentages of VW HMI icons used to answer TRASS questions from participants with high and low FT.

As shown in Table 5, no valid icon could be used in BMW and VW's HMI designs to determine whether automated longitudinal and lateral controls could be activated. For Tesla, icon #6 is used to indicate the availability of level 2 ADS.

To determine whether front vehicle is detected, icon #2 is used in all HMI designs (Figures 6–8), and there is an extra icon for BMW (icon #7 in Figure 6). There was no significant difference on valid icon used for FT levels, $F(1, 31) = 1.48, p = 0.23$, and HMI designs, $F(2, 62) = 2.84, p = 0.067$, and there was also no interaction between these factors, $F(2, 62) = 0.50, p = 0.61$. There were generally low false icon selection rates in answering questions regarding the detection of front vehicles for all three designs. Still, there was a relatively higher portion of participants choosing icon #6 in VW's design (Figure 8), where the small vehicle icon is shown regardless of whether the front vehicle is detected or not.

To answer the question concerning whether lateral control is activated, icon #1 and #5 are used on BMW's design (Figure 6), and icon #1 and #6 for Tesla and VW's (Figures 7 and 8). No significant effect was found on different FT levels, $F(1, 31) = 1.00, p = 0.33$, but the effect of HMI design was found significant, $F(2, 62) = 4.31, p = 0.018$, where more participants chose at least one valid icon on BMW's design comparing to VW's, $t(32) = 2.54, p = 0.041$, and same applies to Tesla's comparing to VW's, $t(32) = 2.54, p = 0.041$, but no difference

between BMW's and Tesla's design, $t(32) = 0.00, p = 1.00$. In Figure 11, icon #4 on VW (Figure 8) was chosen falsely by participants, and it indicates the activation of "lane assist", which would only intervene when the vehicle is about to cross the lateral boundary (lane marking).

In all HMI designs, lane marking detection was indicated by icon #1 (Figures 6–8), and additional icons were used on VW's design (#4 and #6 in Figure 8). The effect of FT level was not significant, $F(1, 31) = 0.26, p = 0.61$, and was the same for HMI designs, $F(2, 62) = 0.21, p = 0.81$, and for the interaction as well, $F(2, 62) = 0.21, p = 0.81$.

The question regarding if longitudinal control is activated could be answered by icon #2, #4 and #7 on BMW's, #5 on Tesla's, and #2, #3, #5 and #6 on VW's design, where no significant effect was found on different FT level $F(1, 31) = 0.73, p = 0.40$, but the effect of HMI design was significant, $F(2, 62) = 8.66, p < 0.001$, as more participants chose at least one valid icon on the HMI design of BMW than Tesla, $t(32) = 3.61, p = 0.002$, and same for VW's design comparing to Tesla's, $t(32) = 3.61, p = 0.002$, but not between BMW's design and VW's, $t(32) = 0.00, p = 1.00$. No interaction effect was found between the two factors, $F(2, 62) = 0.034, p = 0.97$. Although no effect was found for FT levels on the number of participants choosing at least one valid icon, higher percentages of valid icon selections were observed across all HMI designs, where for each valid icon, more participants with high FT selected it than participants with low FT.

Table 5. Valid and false icons on HMI designs concerning TRASS question categories.

Question	HMI Designs	Valid Icons	False Icons
Ava	BMW	Unknown	Unknown
	Tesla	#6	#1, #2, #3, #4, #5, #7
	VW	Unknown	Unknown
FV	BMW	#2, #7	#1, #3, #4, #5, #6, #8
	Tesla	#2	#1, #3, #4, #5, #6, #7
	VW	#2	#1, #3, #4, #5, #6, #7
Lat	BMW	#1, #5	#2, #3, #4, #6, #7, #8
	Tesla	#1, #6	#2, #3, #4, #5, #7
	VW	#1, #6	#2, #3, #4, #5, #7
LM	BMW	#1	#2, #3, #4, #5, #6, #7, #8
	Tesla	#1	#2, #3, #4, #5, #6, #7
	VW	#1, #4, #6	#2, #3, #5, #7
Long	BMW	#2, #4, #7	#1, #3, #5, #6, #8
	Tesla	#5	#1, #2, #3, #4, #6, #7
	VW	#2, #3, #5, #6	#1, #4, #7

Note: For BMW and VW, level 2 ADS could be engaged without guaranteed longitudinal and lateral control. Hence, no valid icon could be used to answer the question "Ava".

4. Discussion

4.1. Summary

HMI is the bridge that allows humans to understand the intentions and capabilities of ADS. At the same time, transparency of the HMI is critical and fundamental for the ADS to be understood with minimum effort [4]. In this study, a preliminary transparency assessment method was proposed, integrating the measurement of understandability toward mode awareness, and using time as the workload indicator. With the proposed method, the functional transparency of the static level 2 ADS system could be evaluated with minimum effort.

The results using the proposed transparency assessment method did not confirm Hypothesis 1a, but support Hypothesis 1b. During the verification test of the transparency assessment method, no significant difference was found among the three different HMI designs. This could be explained by the generally low FTs measured, suggesting that participants had difficulties understanding these three HMI designs. When the proposed

method is properly utilized and incorporated into the HMI design process, we could then more efficiently develop a more understandable HMI design and overcome this problem. For instance, critical elements that help increase HMI understandability could be efficiently identified by observing what information users with high FT use to understand it. And the opposite could also be done to identify potentially misleading information. On the other hand, the effect of ADS experience levels on FT was found significant and was significantly more prominent in the HMI designs of BMW and VW. The result supports Hypothesis 1b and establishes internal validity. This significant gap in understanding the HMI designs between experienced and novice users also suggests that human drivers require some levels of training to understand the HMI [33,38,39]. Apart from general ADS experience, it might also be interesting to look into how a specific brand's experience interacts with the HMI designs of some other brands on the FT. In the feedback section of the survey, some participants mentioned that they made specific suggestions based on their experiences in L2 AV. Since the design, icons, and logic behind the HMI designs are different, further studies investigating such interaction could be valuable in the future.

The subjective transparency estimated with one item Likert scale showed a weak positive correlation with FT. This result supports Hypothesis 2 and establishes external validity. The self-reported measurement was considered as the perceived transparency of participants in contrast to the proposed FT, which is the resulting understanding after the interaction. From the result, it appears that what participants thought they understood was detached from what they did. However, a certain level of monotonicity still exists between the two measurements, which gives an insight for future studies on the relationship between perceived transparency (what one thinks one knows) and functional, or true, transparency (what one actually knows).

What and how information should be conveyed through HMI has been an essential topic in automated vehicle research [4,9,21]. In this study, we evaluated the information used to understand level 2 ADS statuses (i.e., longitudinal control status, lateral control status, front vehicle detection status, lane marking detection status, and Level 2 ADS availability status). Overall, participants with higher FT levels did not choose icons more correctly. Still, in evaluating specific system statuses, they selected more valid icons, which suggests that participants with higher FT relied on multiple information sources when estimating these system statuses. And this result supports Hypothesis 3. On the other hand, HMI designs had no significant impact on icon identification across all system statuses, but it was found significant when evaluating longitudinal and lateral status. By looking closer, e.g., when assessing the longitudinal status, those with a higher percentage of at least one valid icon selection are HMI designs with a higher number of valid icons (BMW and VW's designs). This information redundancy in status indication might seem to be helpful, but it could also lead to confusion and wrong icon selection [40]. Again using the longitudinal status estimation as an example, the redundant icon for lane detection and "lane assist" (only intervene when the boundary is reached) in VW's HMI design (#4 in Figure 8) was mistaken as the indication for the lateral control. Hence, to achieve transparent HMI, it is more critical to provide information with quality (correct information given the level of automation and the scenario) instead of quantity (merely stacking the information). And the proposed method would be a suitable design tool to help identify critical information for transparent HMIs.

4.2. Limitations and Future Works

In this study, the differences in understandability of the adopted HMI designs are limited, which might also be the culprit for failing to confirm Hypothesis 1a. However, we also identify some elements of the HMI designs that have an impact on FT. With these elements, we could create and adopt HMI designs with more significant differences in understandability in future research.

The information needed for the HMI differs across levels of automation and scenarios [41]. This study focused on the level 2 HMI designs, where human users are still responsible for

longitudinal and lateral controls and driving environments. However, in a higher level of automation, human users are no longer in the loop under specific conditions and are allowed to conduct non-driving-related tasks. These differences in responsibility would also significantly impact information needed for FT, thus affecting the questions asked during the transparency assessment test. It is crucial first to ascertain the critical questions for human users to be capable of safely operating the automation under various levels. Furthermore, to apply the transparency assessment method more efficiently, a non-intrusive way of estimating mode awareness [42] could be applied and used to replace the AU.

To further validate the proposed transparency assessment method, simulator or test-track studies are required. In this study, the transparency assessment test was based on HMI images under different scenarios. However, the interaction between participants and the automation was limited. For instance, with only images, participants wouldn't be able to experience the transition between different modes (i.e., activation, deactivation, take-over, etc.) and the corresponding reactions from the vehicle. Plus, as automation users are prone to learn automation by trial and error [7], understanding how transparency changes throughout the interaction would also be beneficial for future HMI designs.

One finding in this study is that participants with different levels of level 2 ADS experience did differ in their understanding of the automated system. But since the HMI designs and the logic behind the system's activation or deactivation diverge and might contradict one another, a more precise distinction on ADS experience would be essential (e.g., familiarity across different automation brands). With this variable, additional system information or tailored training procedure could provide better transparency for the HMI.

This study was a preliminary attempt to assess transparency, and the relationships between the proposed metric and other psychometrics (e.g., trust, acceptance, mental workload) remain unknown. Shedding light on these correlations would help make the proposed transparency assessment method more robust and assist in comparing it with the results in the literature.

4.3. Conclusions

In this study, we proposed and verified a standardized transparency assessment method, which can be used to estimate the understandability of the HMI design. We further established this method's internal and external validity by confirming that the effect of ADS experiences was significant on functional transparency and that a positive correlation between self-reported understandability and functional transparency existed. The proposed method can also help identify critical elements in HMI designs that can significantly impact the understandability of the HMI.

Author Contributions: Conceptualization, Y.-C.L.; methodology, Y.-C.L.; software, Y.-C.L.; validation, Y.-C.L.; formal analysis, Y.-C.L.; investigation, Y.-C.L.; resources, Y.-C.L. and K.B.; data curation, Y.-C.L.; writing—original draft preparation, Y.-C.L.; writing—review and editing, Y.-C.L. and N.F.; visualization, Y.-C.L.; supervision, K.B.; project administration, K.B.; funding acquisition, K.B. All authors have read and agreed to the published version of the manuscript.

Funding: This project has received funding from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No. 860410.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki and approved by the Ethics Committee of the Technical University of Munich (377/21 S-KH) for studies involving humans.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The survey and collected data are accessible at: https://github.com/mobydickhm1851/transparencyAssessmentMethod_openAccessData.git (accessed on 29 September 2022).

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

ACC	adaptive cruise control
ADS	automated driving system
AU	actual understandability
BMW	Bayerische motoren werke AG
FT	Functional Transparency
HMI	human-machine interface
L2 AV	SAE level 2 automated vehicle
LKA	lane-keeping assistance
SAT	situation awareness-based agent transparency
TRASS	transparency assessment test
TNPU	time needed for perceived understandability
VW	Volkswagen AG

References

1. SAE International. *Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles* (SAE Standard J3016, Report No. J3016-202104); Technical Report; SAE International: Warrendale, PA, USA, 2021.
2. National Highway Traffic Safety Administration. ODI Resume. Technical Report. 2017. Available online: <https://static.nhtsa.gov/odi/inv/2016/INCLA-PE16007-7876.PDF> (accessed on 29 September 2022).
3. Banks, V.A.; Plant, K.L.; Stanton, N.A. Driver error or designer error: Using the Perceptual Cycle Model to explore the circumstances surrounding the fatal Tesla crash on 7th May 2016. *Saf. Sci.* **2018**, *108*, 278–285. [CrossRef]
4. Carsten, O.; Martens, M.H. How can humans understand their automated cars? HMI principles, problems and solutions. *Cogn. Technol. Work.* **2019**, *21*, 3–20. [CrossRef]
5. Rezvani, T.; Driggs-Campbell, K.; Sadigh, D.; Sastry, S.S.; Seshia, S.A.; Bajcsy, R. Towards trustworthy automation: User interfaces that convey internal and external awareness. In Proceedings of the 2016 IEEE 19th International conference on intelligent transportation systems (ITSC), Rio de Janeiro, Brazil, 1–4 November 2016; pp. 682–688.
6. Russell, S.M.; Blanco, M.; Atwood, J.; Schaudt, W.A.; Fitchett, V.; Tidwell, S. *Naturalistic Study of Level 2 Driving Automation Functions*; Technical Report; Department of Transportation, National Highway Traffic Safety: Washington, DC, USA, 2018.
7. Boelhouwer, A.; Van den Beukel, A.P.; Van der Voort, M.C.; Hottentot, C.; De Wit, R.Q.; Martens, M.H. How are car buyers and car sellers currently informed about ADAS? An investigation among drivers and car sellers in the Netherlands. *Transp. Res. Interdiscip. Perspect.* **2020**, *4*, 100103. [CrossRef]
8. Banks, V.A.; Eriksson, A.; O'Donoghue, J.; Stanton, N.A. Is partially automated driving a bad idea? Observations from an on-road study. *Appl. Ergon.* **2018**, *68*, 138–145. [CrossRef] [PubMed]
9. Naujoks, F.; Wiedemann, K.; Schömig, N.; Hergeth, S.; Keinath, A. Towards guidelines and verification methods for automated vehicle HMIs. *Transp. Res. Part F Traffic Psychol. Behav.* **2019**, *60*, 121–136. [CrossRef]
10. Pokam, R.; Debernard, S.; Chauvin, C.; Langlois, S. Principles of transparency for autonomous vehicles: First results of an experiment with an augmented reality human-machine interface. *Cogn. Technol. Work.* **2019**, *21*, 643–656. [CrossRef]
11. Oliveira, L.; Burns, C.; Luton, J.; Iyer, S.; Birrell, S. The influence of system transparency on trust: Evaluating interfaces in a highly automated vehicle. *Transp. Res. Part F Traffic Psychol. Behav.* **2020**, *72*, 280–296. [CrossRef]
12. Chen, J.Y.; Lakhmani, S.G.; Stowers, K.; Selkowitz, A.R.; Wright, J.L.; Barnes, M. Situation awareness-based agent transparency and human-autonomy teaming effectiveness. *Theor. Issues Ergon. Sci.* **2018**, *19*, 259–282. [CrossRef]
13. Ososky, S.; Sanders, T.; Jentsch, F.; Hancock, P.; Chen, J.Y. Determinants of system transparency and its influence on trust in and reliance on unmanned robotic systems. In *Proceedings of the Unmanned Systems Technology XVI*; International Society for Optics and Photonics: Bellingham, WA, USA, 2014; Volume 9084, p. 90840E.
14. Körber, M.; Prasch, L.; Bengler, K. Why do I have to drive now? Post hoc explanations of takeover requests. *Hum. Factors* **2018**, *60*, 305–323. [CrossRef]
15. Maarten Schraagen, J.; Kerwien Lopez, S.; Schneider, C.; Schneider, V.; Tönjes, S.; Wiechmann, E. The role of transparency and explainability in automated systems. In Proceedings of the Human Factors and Ergonomics Society Annual Meeting, Baltimore, MD, USA, 3–8 October 2021; SAGE Publications Sage CA: Los Angeles, CA, USA, 2021; Volume 65, pp. 27–31.
16. Bengler, K.; Omozik, K.; Müller, A.I. The Renaissance of Wizard of Oz (WoOz): Using the WoOz methodology to prototype automated vehicles. In Proceedings of the Human Factors and Ergonomics Society Europe Chapter, Nantes, France, 28 October–1 November 2019; pp. 63–72. Available online: https://www.researchgate.net/profile/Kamil-Omozik/publication/346659448_The_Renaissance_of_Wizard_of_Oz_WoOz_-_Using_the_WoOz_methodology_to_prototype_automated_vehicles/links/5fcd24ef9251c00f8588cbf/The-Renaissance-of-Wizard-of-Oz-WoOz-Using-the-WoOz-methodology-to-prototype-automated-vehicles.pdf (accessed on 29 September 2022).

17. Choi, J.K.; Ji, Y.G. Investigating the importance of trust on adopting an autonomous vehicle. *Int. J.-Hum.-Comput. Interact.* **2015**, *31*, 692–702. [CrossRef]
18. Chen, J.Y.; Procci, K.; Boyce, M.; Wright, J.; Garcia, A.; Barnes, M. *Situation Awareness-Based Agent Transparency*; Technical Report; Army Research Lab Aberdeen Proving Ground MD Human Research and Engineering: Aberdeen Proving Ground, MD, USA, 2014.
19. Endsley, M.R. Toward a theory of situation awareness in dynamic systems. In *Situational Awareness*; Routledge: London, UK, 2017; pp. 9–42.
20. Yang, X.J.; Unhelkar, V.V.; Li, K.; Shah, J.A. Evaluating effects of user experience and system transparency on trust in automation. In Proceedings of the 2017 12th ACM/IEEE International Conference on Human-Robot Interaction (HRI 2017), Vienna, Austria, 6–9 March 2017; pp. 408–416.
21. Bengler, K.; Rettenmaier, M.; Fritz, N.; Feierle, A. From HMI to HMIs: Towards an HMI framework for automated driving. *Information* **2020**, *11*, 61. [CrossRef]
22. Feierle, A.; Danner, S.; Steininger, S.; Bengler, K. Information needs and visual attention during urban, highly automated driving—An investigation of potential influencing factors. *Information* **2020**, *11*, 62. [CrossRef]
23. Bhaskara, A.; Skinner, M.; Loft, S. Agent transparency: A review of current theory and evidence. *IEEE Trans.-Hum.-Mach. Syst.* **2020**, *50*, 215–224. [CrossRef]
24. Akash, K.; Polson, K.; Reid, T.; Jain, N. Improving human-machine collaboration through transparency-based feedback—Part I: Human trust and workload model. *IEAC-PapersOnLine* **2019**, *51*, 315–321. [CrossRef]
25. Flemisch, F.; Schieben, A.; Kelsch, J.; Löper, C. Automation spectrum, inner/outer compatibility and other potentially useful human factors concepts for assistance and automation. In *Human Factors for Assistance and Automation*; Shaker Publishing: Duren, Germany, 2008.
26. Cao, Y.; Griffon, T.; Fahrenkrog, F. Code of Practice for the Development of Automated Driving Functions; Technical Report, L3Pilot Deliverable D2.3; version 1.1; 2021. Available online: https://www.eucar.be/wp-content/uploads/2022/06/EUCAR_CoP-ADF.pdf (accessed on 29 September 2022).
27. Albers, D.; Radlmayr, J.; Loew, A.; Hergeth, S.; Naujoks, F.; Keinath, A.; Bengler, K. Usability evaluation—Advances in experimental design in the context of automated driving human-machine interfaces. *Information* **2020**, *11*, 240. [CrossRef]
28. Jian, J.Y.; Bisantz, A.M.; Drury, C.G. Foundations for an empirically determined scale of trust in automated systems. *Int. J. Cogn. Ergon.* **2000**, *4*, 53–71. [CrossRef]
29. Venkatesh, V.; Morris, M.G.; Davis, G.B.; Davis, F.D. User acceptance of information technology: Toward a unified view. *MIS Q.* **2003**, *27*, 425–478. [CrossRef]
30. Forster, Y.; Hergeth, S.; Naujoks, F.; Krems, J.F. How usability can save the day—methodological considerations for making automated driving a success story. In Proceedings of the 10th International Conference on Automotive User Interfaces and Interactive Vehicular Applications, Toronto, ON, Canada, 23–25 September 2018; pp. 278–290.
31. ISO/IEC 9126; Software Engineering—Product Quality. ISO/IEC: Geneva, Switzerland, 2001.
32. Scalabrino, S.; Bavota, G.; Vendome, C.; Poshyvanyk, D.; Oliveto, R. Automatically assessing code understandability. *IEEE Trans. Softw. Eng.* **2019**, *47*, 595–613. [CrossRef]
33. Mueller, A.S.; Cicchino, J.B.; Singer, J.; Jenness, J.W. Effects of training and display content on Level 2 driving automation interface usability. *Transp. Res. Part F Traffic Psychol. Behav.* **2020**, *69*, 61–71. [CrossRef]
34. Bates, D.; Mächler, M.; Bolker, B.; Walker, S. Fitting Linear Mixed-Effects Models Using lme4. *J. Stat. Softw.* **2015**, *67*, 1–48. doi: 10.18637/jss.v067.i01. [CrossRef]
35. R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2021.
36. Barr, D.J.; Levy, R.; Scheepers, C.; Tily, H.J. Random effects structure for confirmatory hypothesis testing: Keep it maximal. *J. Mem. Lang.* **2013**, *68*, 255–278. [CrossRef]
37. Lenth, R.V. *Emmeans: Estimated Marginal Means, aka Least-Squares Means*; R package version 1.7.0; R Foundation for Statistical Computing: Vienna, Austria, 2021. .
38. Forster, Y.; Hergeth, S.; Naujoks, F.; Krems, J.; Keinath, A. User education in automated driving: Owner’s manual and interactive tutorial support mental model formation and human-automation interaction. *Information* **2019**, *10*, 143. [CrossRef]
39. Boos, A.; Emmermann, B.; Biebl, B.; Feldhütter, A.; Fröhlich, M.; Bengler, K. Information Depth in a Video Tutorial on the Intended Use of Automated Driving. In *Proceedings of the Congress of the International Ergonomics Association*; Springer: Berlin/Heidelberg, Germany, 2021; pp. 575–582.
40. Bauerfeind, K.; Stephan, A.; Hartwich, F.; Othersen, I.; Hinzmann, S.; Bendewald, L. Analysis of potentials of an HMI-concept concerning conditional automated driving for system-inexperienced vs. system-experienced users. In Proceedings of the Human Factors and Ergonomics Society Europe, Rome, Italy, 28–30 September 2017; pp. 67–77. Available online: https://www.researchgate.net/profile/Kassandra-Bauerfeind/publication/339900783_Analysis_of_potentials_of_an_HMI-concept_concerning_conditional_automated_driving_for_system-inexperienced_vs_system-experienced_users/links/5e6b682c458515e55576ac14/Analysis-of-potentials-of-an-HMI-concept-concerning-conditional-automated-driving-for-system-inexperienced-vs-system-experienced-users.pdf (accessed on 29 September 2022).

41. Beggiato, M.; Hartwich, F.; Schleinitz, K.; Krems, J.; Othersen, I.; Petermann-Stock, I. What would drivers like to know during automated driving? Information needs at different levels of automation. In Proceedings of the 7. Tagung Fahrerassistenzsysteme, Munich, Germany, 25–26 November 2015.
42. Forster, Y.; Geisel, V.; Hergeth, S.; Naujoks, F.; Keinath, A. Engagement in non-driving related tasks as a non-intrusive measure for mode awareness: A simulator study. *Information* **2020**, *11*, 239. [CrossRef]

Article

A Cognitive Model to Anticipate Variations of Situation Awareness and Attention for the Takeover in Highly Automated Driving

Marlene Susanne Lisa Scharfe-Scherf ^{1,2,*}, Sebastian Wiese ² and Nele Russwinkel ²

¹ Robert Bosch GmbH, Robert-Bosch-Allee 1, 74232 Abstatt, Germany

² Department of Psychology and Ergonomics, Technical University Berlin, 10587 Berlin, Germany

* Correspondence: marlene-susanne-lisa.scharfe@de.bosch.com

Abstract: The development of highly automated driving requires dynamic approaches that anticipate the cognitive state of the driver. In this paper, a cognitive model is developed that simulates a spectrum of cognitive processing and the development of situation awareness and attention guidance in different takeover situations. In order to adapt cognitive assistance systems according to individuals in different situations, it is necessary to understand and simulate dynamic processes that are performed during a takeover. To validly represent cognitive processing in a dynamic environment, the model covers different strategies of cognitive and visual processes during the takeover. To simulate the visual processing in detail, a new module for the visual attention within different traffic environments is used. The model starts with a non-driving-related task, attends the takeover request, makes an action decision and executes the corresponding action. It is evaluated against empirical data in six different driving scenarios, including three maneuvers. The interaction with different dynamic traffic scenarios that vary in their complexity is additionally represented within the model. Predictions show variances in reaction times. Furthermore, a spectrum of driving behavior in certain situations is represented and how situation awareness is gained during the takeover process. Based on such a cognitive model, an automated system could classify the driver's takeover readiness, derive the expected takeover quality and adapt the cognitive assistance for takeovers accordingly to increase safety.

Keywords: takeover; highly automated driving; automation; situation awareness; adaptive automation; human–automation interaction; cognitive architectures; decision making; user studies

1. Introduction

The development of highly automated driving is a central topic of current transportation research [1]. Still, the next generations of highly automated vehicles (Levels 3 and 4) will need the driver as fallback level [2,3]. Highly automated driving solutions thus require dynamic approaches including the anticipation of the cognitive driver state within dynamic situations and need to be adapted in a way that control could be handed over intuitively between driver and vehicle. In case the driver has to resume control, a takeover request is triggered. However, the assumption that a human driver stays alert and is always ready to takeover control within seconds is fallacious [1]. The switch between a non-driving-related task (NDRT) and the main driving task is crucial during the takeover. The driver is required to build up situation awareness according to the situation that unfolds in order to make an action decision [4,5]. Only when the driver has collected enough information of the surrounding and is in the loop [6], safe actions in a dynamic situation are possible. This process requires manual control skills as well as cognitive skills [7]. A drivers' takeover readiness when resuming control after an automated driving section may be influenced by various factors [8], inter alia cognitive and environmental factors. Within this paper, the focus lies on cognitive mechanisms and the process to build up situation awareness.

Driver behavior thereby is determined by complex interactions with other traffic participants [9] and the system itself. To increase safety and comfort, it has to be identified where humans best fit into the system [1,10].

It has been shown [11] that the usability and comprehensibility in highly automated driving can be increased when using adaptive systems. This shows that it is highly important to consider the individual driver when developing advanced driver assistance systems. For such an individual adaption in highly automated driving, the need to simulate human agents arises [12]. Thus, it becomes important to determine the takeover readiness of each individual driver and develop adaptive cognitive assistance systems to support the driver accordingly.

Such cognitive assistance systems serve the purpose of supporting the human driver during a takeover task. Such a support could be, for example, a pre-deceleration before the takeover, the best maneuver indication, or the suppression of irrelevant information in a complex situation. In the long run, these systems do not drive the vehicle, but anticipate the driver and enable the driver to takeover better and more safely. To develop such a system, cognitive processes during a takeover should be simulated and anticipated to understand the cognitive status of situation awareness within a particular situation. To illustrate cognitive processes and the development of situation awareness, this paper uses cognitive modeling to represent a spectrum of driving decisions in dependence of the current dynamic situation during a takeover.

1.1. Cognitive Modeling in Highly Automated Driving

The use of cognitive driver models within advanced driver assistance systems allows anticipating the driver in context to all traffic participants in the close vicinity. A prediction of other traffic participants is used by [9] to adapt the advanced driver assistance system reaction and consequently increase safety and comfort. Nevertheless, more importantly, time dynamics of driver perception, scene interpretation and decision making [12] need to be considered during the takeover to adapt assistance systems accordingly and increase safety.

A first approach to include the driver into an advanced driver assistance system is done by [8]. They propose an adaptive advanced driver assistance system to keep the driver in-the-loop or support in takeover situations through supplementary features such as gaze guidance or increased deceleration. These are based on the complexity of the traffic situation, the current secondary task of the driver, and the gazes on the road [8]. Still, this approach does not simulate and predict the cognitive processes that are undergone during a takeover in different situations.

Other authors [13] proposed a driver model that focuses on the multitasking nature of the driving task. Within their model, general driving in different situations is modeled. When it comes to highly automated driving, the execution of different tasks in parallel is relevant to consider. However, their approach only models the driving task per se, but does not include the takeover in highly automated driving.

The same applies for other approaches [14,15] that implement cognitive models for the driving task. The cognitive model by [15] provides a useful approach to implement the general driving task within a cognitive architecture. He implemented the driving task based on several cognitive processes that are applied during the driving task to perceive the surrounding and perform motoric actions. However, this model does not include the takeover process, the whole dynamic driving environment, and the built-up of situation awareness.

An ACT-R (Adaptive Control of Thought-Rational; [16]) cognitive model for the takeover in highly automated driving [17] shows that such a model is able to represent cognitive processes during a takeover. Different complexities of situations can already be handled very well by that model, but no spectrum of different behaviors is included. This is addressed within this paper.

The aim of the present paper is to show how a spectrum of driving decisions can be integrated into such a modeling approach. Depending on how the dynamic situation

unfolds and what kind of action and attention decisions are made. These two aspects will influence the build-up of the individual situation awareness. The model follows such a situated cognition approach and tries to capture the variety of driving decisions in takeover situations that would follow under reasonable considerations. The representation of such differences can be modeled using ACT-R. Some authors [18], for example, developed a cognitive model to simulate automated driving using reinforcement learning for adaptations on automated vehicle operations. This adaptation process is situated and very relevant for automated driving. However, related to takeover maneuvers, complex decision strategies and the build up of situation awareness are not the focus of that work. Furthermore, another study [19] followed a situated cognition account on how a situation unfolds and what kind of behavior would be reasonable regarding task and context. When behavior deviates from model expectations, further explanations (e.g., via neurophysiological information) are explored by the model. Another approach [19] uses a form of model-tracing to represent individual situated pilots' behavior. Nevertheless, this work is more focused on auditory attention than on visual.

Such complex cognitive driver models that incorporate driving differences need to be integrated into embedded systems [9] to support a safe and comfortable takeover. It is a method for higher-level assessment of the situation at hand that can be described as situation awareness [12].

1.2. Situation Awareness

Situation awareness as defined by [20] includes three main stages to gain an understanding of a situation. In the first stage, the perception phase, objects in the surrounding environment are perceived. The perception phase is especially important for the takeover as it includes the visual perception of relevant vehicles in the surrounding traffic environment. This is followed by forming an understanding of the current situation in the comprehension phase [20]. In this paper, the phase is understood as the process when information about the situation and the task are integrated to gain a deeper understanding of the whole situation. After this, it is important that the driver is able to project the future situation with the altered state of other vehicles in the surrounding traffic (depending on speed, lane change indications, etc.), called the projection phase [20].

Overall, situation awareness can be understood as interrelated perceptual decisions about the world [12]. These are cognitively integrated to base decisions on.

1.3. SEEV Theory

Based on the SEEV theory, visual attention distribution in workplace scenarios can be predicted [21] and whether unexpected stimuli are perceived [22]. One approach [23] assessed the plausibility of SEEV in a driving scenario. SEEV (Salience Effort Expectancy Value) uses four parameters to describe attention-attracting properties. These are a combination of the bottom-up factors, salience and effort, and the top-down factors, expectancy and value. The workplace is subdivided into areas of interest (AOI). Value corresponds to the task relevance of an AOI and expectancy to how often this information changes. This indicates how often an AOI needs to be sampled in order to get the relevant information. For a driver, the view through the front window is, for example, a highly relevant area of interest. Nevertheless, a passenger doing a non-driving-related task (NDRT) will find his tablet display, for example, more important. Thus, the passenger will less frequently look through the front window (low expectancy, low value). However, this would still be more often than through the rear window, as it is closer to the current point of fixation, needing less effort to look there.

By understanding driver cognition and the state of situation awareness during the takeover, cooperative systems in highly automated driving can be developed. Such cooperative systems adjust both-sided (driver–system and system–driver) and can support the human driver directly and on an individualized level. A cognitive model that integrates situation awareness and SEEV enables a better prediction of cognitive processes. Especially to predict the process of building up situation awareness after a takeover request, the fusion of cognitive modeling and the SEEV theory is highly important. It allows the usage of AOIs in combination with human cognitive processes. Based on such a cognitive model, a better approximation to vehicle system dynamics can be developed and the cooperation between vehicle and driver enhanced. The goal in this paper is to show that the present cognitive model is able to (a) represent cognitive processes during a takeover, (b) interact with dynamic traffic environments, (c) use AOIs for visual perception based on the SEEV theory, (d) predict deviations that represent a spectrum of different behaviors in dependence of the current situation, (e) represent situation awareness during a takeover, and (f) match empirical data.

2. Materials and Methods

To simulate human cognition and represent corresponding cognitive processes, ACT-R [24] is used. The architecture ACT-R explains how multiple modules that represent cognitive processing units such as memory and auditory/visual perception are integrated to produce behavior in complex tasks [16]. Models within such an architecture can generate novel predictions and helps explain human behavior [25]. To represent the takeover in driving, the standard ACT-R modules are used. The procedural system is used in this work for the takeover process, as productions enable the most adaptive response from any state of the modules [26]. The original ACT-R visual module is replaced by a new vision module called SEEV-VM (Salience Effort Expectancy Value-Vision Module [27]). In contrast to the standard module that identifies objects in the visual field, the SEEV-VM uses AOIs that more closely represent human visual attention. The following sections describe the model environment, the new SEEV-VM module, and how the cognitive model is built up.

2.1. Graphical User Interface

The cognitive model interacts with a graphical user interface (GUI). In this GUI, different traffic scenarios are represented. Two visualizations exist, the in-car view (Figure 1) and the bird's-eye view (Figure 2).

The in-car view (Figure 1) shows the simulation and the AOIs that the model interacts with. This view is synchronized (y-position dependency) to the bird's-eye view [28]. The bird's-eye view is for visualization and control of the surrounding traffic by the modeler (Figure 2). Figure 2 shows the setup of the surrounding traffic environment when the takeover is requested. The vehicles in the GUI move different pixels per time-frame, resembling the different speed that the traffic environment in the empirical study has. The traffic environment resembles a six-lane highway (three lanes per direction), with the ego-vehicle driving on the center lane. The vehicles in the GUI move several pixels per time-frame, resembling the different speed that the traffic environment in the empirical study has.

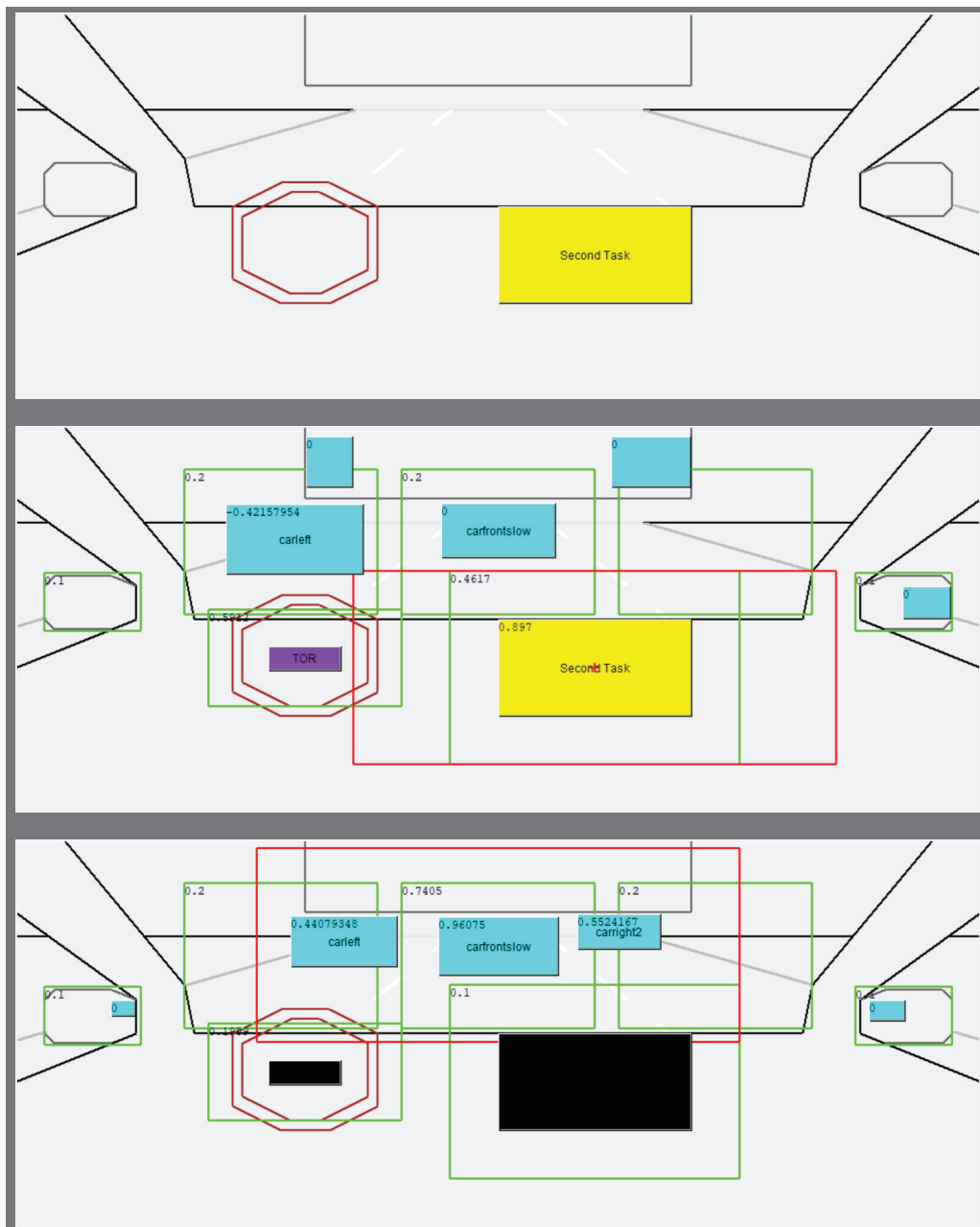


Figure 1. Original screenshot of the graphical user interface that the cognitive model interacts with. The screenshot represents the areas of interest (green frames), the field of view (red frames), and vehicles in the traffic environment (solid blue/green squares). Depicted example scenes show from top: the lane graphics and the NDRT, a complex traffic scenario with TOR (takeover request; violet solid square), and NDRT active (yellow solid square), traffic scenario with NDRT and TOR deactivated (black solid squares; source: own figure).

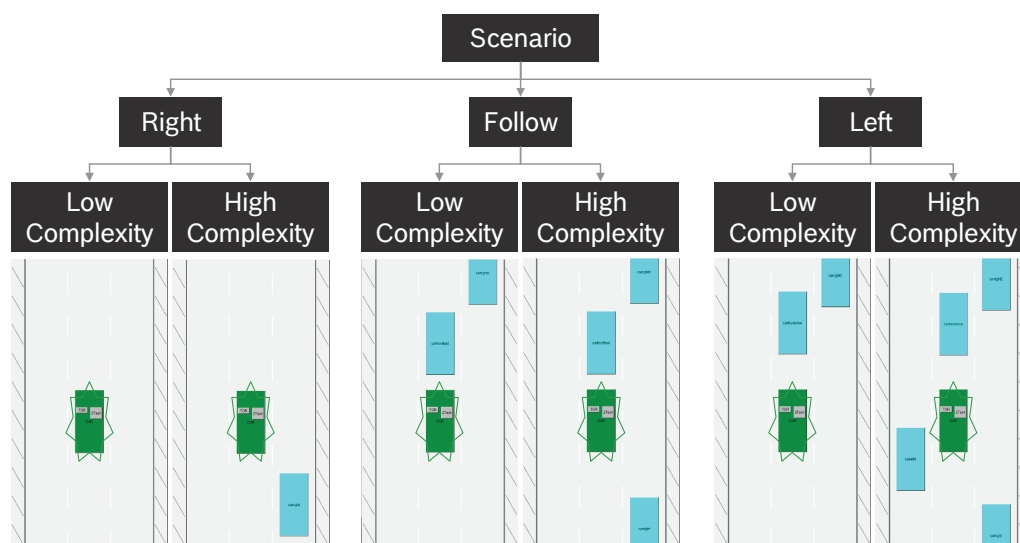


Figure 2. Representation of traffic scenarios in the graphical user interface representing the ego-vehicle (star-green) and relevant vehicles in the surrounding traffic environment at the takeover point (blue/green; source: own figure).

2.2. Scenarios

Every scenario starts with an NDRT until the takeover request, located behind the steering wheel, appears. Traffic is set up to make three different maneuvers necessary (lane-change left, right, and car following).

Each maneuver is set up in a high or low complex surrounding traffic environment. This results in overall six different scenarios. Figure 2 gives an overview of the scenarios and the corresponding traffic in the bird's-eye view. In order to trigger the corresponding maneuver, the speed of the vehicles differs between the different scenarios. In the following condition, for example, the leading vehicle has a higher speed and is thus followed, whereas in the left condition with low complexity, the leading vehicle is significantly slower and triggers a lane change to the left.

2.3. Cognitive Model

The cognitive model represents a spectrum of cognitive processes during a takeover in dynamic traffic environments that builds up a representation of situation awareness. The model captures processes upon the situation as it unfolds and includes decision-points where several options are possible and chosen by chance. The model uses the ACT-R default parameters in all cases. No additional parameter fitting is done. The default production specified action time is 50 ms. Furthermore, at this stage of the model, learning is not integrated.

Within the model, a bandwidth of different driving strategies is implemented that the model applies during a takeover task. Some drivers, for example, look over the shoulder before changing a lane. Others prefer using the mirrors, others do both. Furthermore, drivers differ in their caution when taking over the driving task. Some drivers, for example, are more cautious (more braking) than others (acceleration). Based on the surrounding traffic, the model perceives the situation and makes corresponding action decisions. For the perception of the blind spot, a motoric movement is used, indicating the view over the shoulder. No AIOs are set for the blind spot as it is a yes/no perception without visual tracking. The model represents different decision points where several options are eligible with the same probability. This results in a divergent rather than a deterministic model that produces a high spectrum of behavioral representations. This kind of model includes all kinds of reasonable decisions considering the current situation and the information that has already been collected. This model could at a later point be used for model

tracing. In the model tracing approach (e.g., [29]), actions of the user are traced and the underlying model provides means to figure out the appropriate productions that lead to the actions. Thus, gaze and actions drive the tracing process and provide hypothesis about the current cognitive state and representation that explain the behavior. With this information, appropriate assistance or intervention can be given whenever necessary. At this point, the model is required to show a variety of behaviors that is in accordance with the situation and the sequence of actions to see whether this shows reasonable performance compared to participants' performance. In addition, the model builds up situation awareness that is represented in a chunk. The current state of situation awareness influences decisions at decision points where premises apply based on the dynamics of the traffic situation.

The following sections describe the visual guidance of the scene perception, the representation of situation awareness within the model, and the corresponding decision flow.

2.4. Visual Guidance

Visual perception is an integral part of the cognitive model to build up situation awareness. Three requirements to model visual guidance are identified for this model. First, the detection of unexpected, not explicitly requested visual features such as alarm signals. Second, prediction-based gaze control, and third, the realistic simulation of eye movements in visually rich and highly dynamic scenes.

In this approach, the SEEV vision module (SEEV-VM; [27]) is used as a replacement for the default ACT-R vision module. The following properties of the original SEEV theory are implemented in this vision module. Top-down gaze control in the form of areas of interest (AOI). These are defined by the modeler. However, the original SEEV theory does not provide an approach to predict visual guidance in dynamic scenes, as realized here. Visual guidance here is composed of salience, effort, and relevance. The equation below shows how to compute visual guidance.

$$\text{guidance value} = \text{relevance} + \text{salience} - \text{effort} \quad (1)$$

In the following section, salience, relevance and effort are explained in detail.

Salience: Visual salience in general is considered as the physical attractiveness of an object. It is based on the contrast of an object to the close surrounding. E.g., a flashing red light in an otherwise indistinguishable gray space is highly salient. It can be calculated, but this would require a very extensive simulation of the environment.

Relevance: Every AOI has a relevance that is a combination of value and expectancy (which are the factors in the original account). The value reflects the task-related importance of expected information at the corresponding location. Expectancy describes here the amount of relevant objects and the frequency of change within an AOI. In the original SEEV theory, an AOI contains information of every object within this AOI. In the SEEV-VM implementation, objects lie within AOIs and are attended and encoded individually. Therefore, relevance changes dynamically. If an object within an AOI is attended, the relevance of this AOI decreases by the consumption value. Accordingly relevance increases over time by the refresh rate. Both values correspond to the expectancy value of the original SEEV theory and have to be set by the modeler.

Effort: Saccadic effort is calculated based on the spatial distance of a target object to the current point of fixation. Effort is high when objects are far away.

Within the GUI, scenarios are composed of objects that are defined by the modeler. The modeler describes objects by assigning salience values and contextual meaning (e.g., car, lane marking) to them. The SEEV-VM approach frees the modeler of designing the visual perception processes in detail but uses an indirect mechanism based on the guidance value Formula (1).

By using SEEV-VM, the model occasionally checks the vehicle interface while primarily conducting an NDRT. This is realized by two AOIs. One AOI for the NDRT display and one for the takeover request display (Figure 1). Once the takeover request signal is activated, it has a high salience value. This increases the guidance value of the takeover request signal.

In this way, a bottom-up process can trigger a top-down process that initiates a task switch. Following this task switch, relevance values of AOIs are readjusted.

2.5. Situation Awareness Representation

The process to build up situation awareness within this model is a top-down process. In contrast to [15], this model includes a more complex approach to build up situation awareness that is more goal-directed rather than general maintenance. Furthermore, the takeover in highly automated driving requires to reconstitute situation awareness. It is dependent on the current situation and is built-up progressively. The model starts after the takeover request is triggered. Using the SEEV-VM, the AOI of the front lane has the highest relevance in the beginning. The first goal, after the takeover request has been attended, is to build up situation awareness. After the front lane is perceived, the front part of situation awareness is built-up and the information stored in a mental representation is called a chunk. This chunk is located in the imaginal buffer. It is able to store information (in the slots) about the position of vehicles on the three lanes, representing current situation awareness. Based on the information that is stored in this chunk, decisions are made. Depending on the current state of situation awareness, different AOIs rise in relevance, depending on the current goal. Thus, if there is no upcoming vehicle, for example, the right lane becomes an AOI with a high relevance to check whether it is free (obligation to drive on the right). Based on the information about the situation of the perceived AOI, situation awareness is updated and the goal adapted. The information that the model perceives in the dynamic situations is matched to conditions of different production rules.

2.6. Decision Flow

In order to provide an understanding of how such a production rule looks like, Figure 3 provides an exemplary production rule. Only productions that fit to the information in the buffers (these are specified in the upper half) can fire and changes, e.g., the goal, the imaginal buffer or requests to the visual-organization (control concept of the SEEV-VM) are conducted.

(p right-SA-car		
=goal>		
state		decide
TOR		t
=imaginal>		
isa		sa
perceivedTOR		t
perceivedR		nil
carright		99
=visual-object-buffer>		
type		CAR
relative-position		2
?visual-organization>		
state		free
==>		
+visual-organization>		
cmd		update
aoi-name		right-aoi
relevance		0
refresh		0
=imaginal>		
isa		sa
carright		t
perceivedR		t
=goal>		
state		decide)

Figure 3. Exemplary production rule that fires when a car on the right oncoming lane is perceived while situation awareness is build up (99 = dummy variable; t = true; TOR = takeover request; R = right; aoi = area of interest; sa = situation awareness; source: own figure).

The production fires if the current goal is to make a decision and a takeover request has been perceived. This information is stored in the imaginal buffer together with the information that the right lane has not been perceived yet (nil). Therefore, the value for a vehicle on the right is still the dummy (99). This indicates that no car on the right has been perceived yet, neither has the right been attended. Furthermore, a visual object has been perceived of the type CAR on the right oncoming lane (Position 2) and the visual organization has to be free. If all this is the case, then ($\Rightarrow$) the right AOI's relevance is set to 0, as it has just been attended. Furthermore, the imaginal buffer stores the information that the right lane has been perceived and that there is a car driving on the front oncoming lane (t). The goal is still to make an action decision. In case the conditions match a certain production, as described above, this production will be carried out. As in this version of the model no utilities are used, one of several possible productions is chosen by chance. Figure 4 represents the whole decision flow. First, the model sets the pre-defined relevant AOIs for NDRT, HMI, the three lanes (left, oncoming/front, right), and left and right mirror. As the model represents the takeover, it starts with an NDRT (Figure 4). The NDRT of the model equals the automated mode when the driver engages into an NDRT. The model initially attends the NDRT that is active (NDRT).

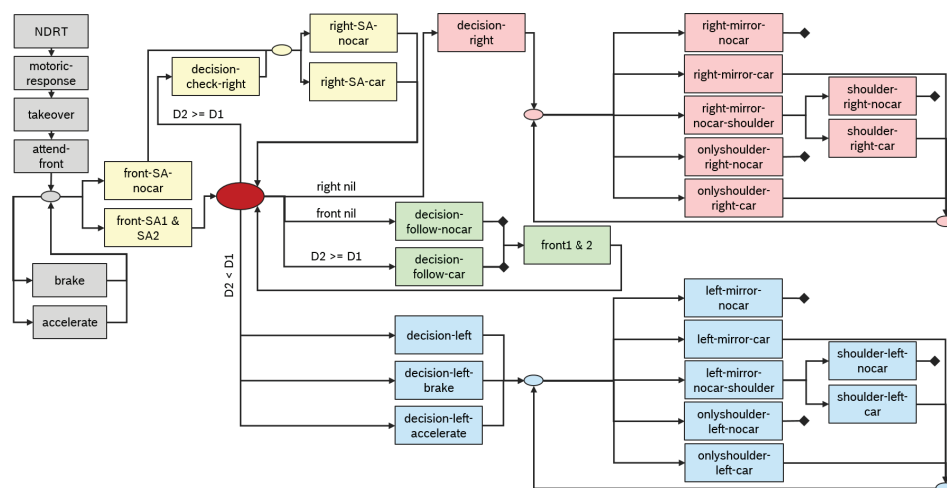


Figure 4. Structure of the Cognitive Model. Boxes with productions that represent the procedural system. Arrows show possible steps in the process. Ovals mark process points, where the goal in the goal buffer is changed. Colors mark different steps in the process to build up situation awareness (SA; gray = takeover and first motoric reactions, yellow = building of first SA to base a decision on, blue = lane change right, green = following, red = lane change left; D = distance; source: own figure).

As soon as the takeover request appears, the model responds with first automated reactions (gray boxes; Figure 4). The takeover request is recognized, the model simulates moving the hands to the steering wheel (*motoric-response*), and the model takes over the driving task (*takeover*). The front is attended (*attend-front*) and another motoric reaction is executed. This motoric longitudinal driving reaction can either be an acceleration (*accelerate*) or a deceleration (*brake*). As the ACT-R architecture does not provide movements of the feet, the default ACT-R movement of the hands is used as substitute for a motoric action. Next, the situation awareness (SA) of the current traffic situation has to be built up (yellow boxes; Figure 4). If no upcoming vehicle is present (*front-SA-nocar*), the right side is attended and it is decided whether there is a car (*right-SA-car*) driving on the right lane or not (*right-SA-nocar*). In cases, where an upcoming vehicle is present, this vehicle is attended twice. The two distance points (D; Figure 4) are contrasted. Based on this information, the model goes back into the central decision stage (red oval; Figure 4) where it takes an action decision based on the collected information. If the distance increases or stays the same, the right side is attended as described above. If the distance decreases, the model decides to

overtake the leading vehicle with a lane-change to the left. Thus, it starts checking the left lane for overtaking vehicles.

2.6.1. Maneuver Decision Lane-Change to the Right

Based on the obligation to drive on the right, a decision to make a lane change to the right (blue boxes; Figure 4) is made when the right lane is free (*decision-right*). This is the case when no car is detected on the right lane and right equals nil. In case a car on the right is detected, right would not equal nil and a different decision is made. As soon as the lane-change right decision is made, the lane-change maneuver is prepared. Three different behavioral options to make sure no car is driving in the right back are possible. The first option is to check only the mirror. If no vehicle is detected in the right mirror (*right-mirror-nocar*), the lane change is executed. In case there is a vehicle in the mirror (*right-mirror-car*), a new decision between the three behavioral options is made. The second optional behavior is to check the mirror and if there is no car, look over the shoulder before executing the lane change (*right-mirror-nocar-shoulder*). When there is still no car detected with the shoulder check (*shoulder-right-nocar*), the lane change is executed. If a car is recognized due to the over-shoulder check (*shoulder-right-car*), a new behavioral decision for the right back area is made. The third behavioral option is to look only over the shoulder. Again, if no car is in the field of view, the lane change is executed (*onlyshoulder-right-nocar*) and if there is a car detected, a new behavioral decision is made (*onlyshoulder-right-car*).

2.6.2. Maneuver Decision Car Following

Figure 4 represents car-following productions with green boxes. If no upcoming vehicle is present but the right lane occupied (*right-SA-car*), the decision to stay on the current lane is made (*decision-follow-nocar*). If an upcoming vehicle is present, the car in front is attended twice and distances are compared (*front-SA1 & SA2*). In case the distance (D; Figure 4) to the car in front increases or stays the same, the right lane is checked for a vehicle (*decision-check-right*). The above-described procedure for a lane change to the right is performed, if no car is detected on the right (*right-SA-nocar*). In case there is a car present on the right lane (*right-SA-car*) and the distance (D; Figure 4) to the car in front increases or stays the same, car following is decided (*decision-follow-car*). The front is attended again (*front1 & 2*) to update the status of the vehicle in front.

2.6.3. Maneuver Decision Lane-Change to the Left

In case the leading vehicle is driving slower than the ego-vehicle, a decision to overtake is made (red boxes; Figure 4). Three options are possible when deciding to do a lane change to the left. The lane change can be executed without changing the current speed (*decision-left*), slower (*decision-left-brake*) or faster (*decision-left-accelerate*). All three options are followed by attending the left back. As in the condition for the right maneuver, three behavioral patterns are possible. Those are: first, checking the mirror only (*left-mirror-car & left-mirror-nocar*). Second, checking the mirror (*left-mirror-nocar-shoulder*) and if there is no car, rechecking the blind spot with the over-shoulder view (*shoulder-left-car & shoulder-left-nocar*). Third, using the over-shoulder view only (*onlyshoulder-left-car & onlyshoulder-left-nocar*). There can either be a vehicle in the attended area (*...-car*) or not (*...-nocar*). If there is no vehicle detected, the lane change is executed. If a vehicle is detected in the relevant area, a new behavioral decision to attend the left back is made.

2.7. Empirical Data

In order to validate predictions of the cognitive model, these have been compared to a driving simulator study in Figure 5. As driving simulators are identified as the prevalent testing environment in the field of usability assessments of HMIs and an efficient and risk-free alternative to real driving experiments [30], this method was chosen to validate the cognitive model. The study was conducted in spring 2019 and the test design, methods, and technical functionality were tested in a pre-study. The study was conducted

with $N = 20$ participants (13 male, 7 female) with a mean age of 26.2 years ($SD = 2.69$). Most participants drive on average 30 min per day, mostly on highways and indicate a moderate driving style [31,32]. Each participant ran through every scenario three times. Thus, empirical data represent $N = 60$ samples per scenario. Commonly, it is proposed to run a model 100 times [33] with $N = 20$ participants. However, based on the different scenarios, the model is run through each scenario 60 times to have comparable sample sizes. The implementation of the study was approved by the ethics committee of the TU Berlin in April 2019 and Robert Bosch GmbH.

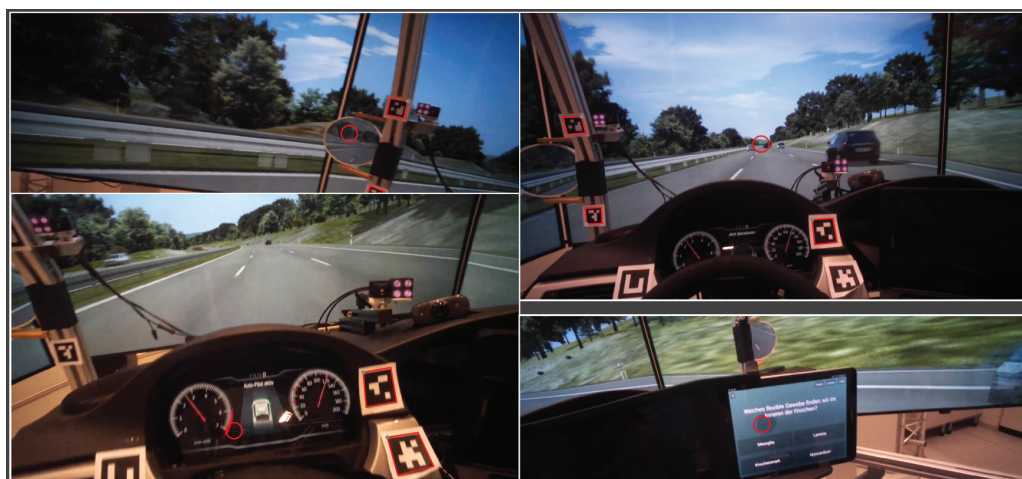


Figure 5. Depiction of the driving simulator study. Source: [31].

3. Results

The model is able to interact with different dynamic environments and make the most common maneuver decisions (lane change left, right, and following; [34]) as shown in Figure 4. The specific results of these maneuver decisions are described in the sections below.

The model is run through each scenario 60 times ($N = 60$) with parameters set to initial values each time. The vehicles in the GUI move different amounts of pixels per time-frame. This resembles the different speeds within the traffic environment of the empirical study in which the ego-vehicle drove at a speed of 120 km/h when the takeover was triggered. Results show that the model is able to make maneuver decisions in different scenarios. Prediction times show variations depending on the maneuver and the complexity of the surrounding traffic environment (Figure 6). The rule firing times vary due to a different selection of production rules depending on the driver and the traffic environment. Figure 6 represents these differences in execution times by indicating the points NDRT, takeover, decision, and execution of an action. The buildup of situation awareness mainly happens between the NDRT and the takeover, but it is an ongoing process and essential to make and execute an action in a complex environment.

In the following, results of the three different scenarios are represented in tables in separate sections. The column *Events* of the tables (Tables 1–3) indicates how often a production is fired. Productions can be fired more than once in a case. Braking, for example, can be fired more often, indicating a stronger deceleration. The tables start with the listing of the first action. Before this, the takeover request appears and is perceived. In the end, model predictions are compared to empirical data. Reaction times that are compared to each other always correspond to the initiation of an action on the timeline of the takeover. Based on these action initiations, cognitive model predictions and empirical data are compared.

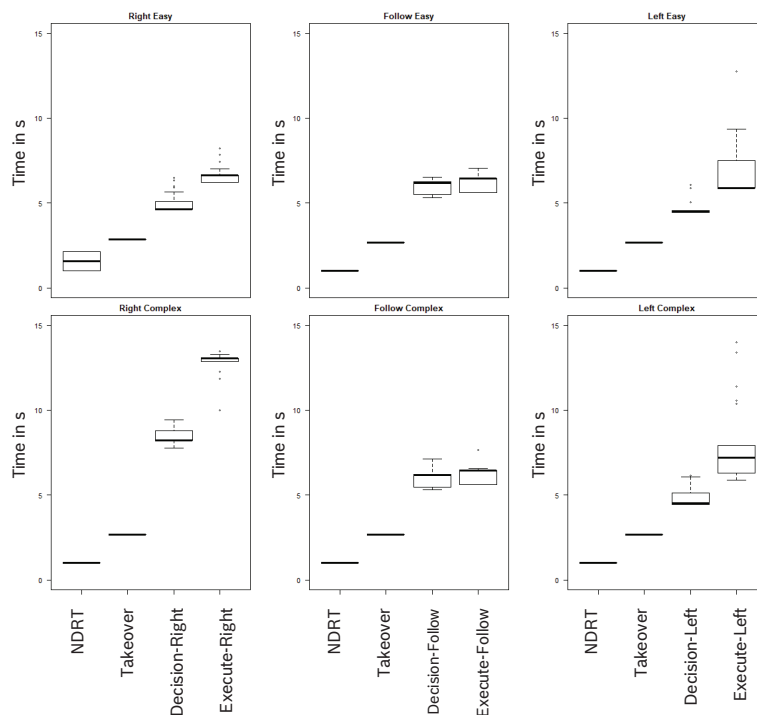


Figure 6. Box-whisker plot of model predictions for each task in each condition. The y-axis shows the timeline in seconds (s) including the NDRT, the takeover, the corresponding action decision, and the action execution (source: own figure).

Table 1. Predicted time statistics of the cognitive model with a lane-change right maneuver. Representation of minimum (Min), mean, and maximum (Max) values over $N = 60$ trials. Wrong decisions are in lower case letters (source: own table).

Production	Events	Predicted Time			
		Min	Mean	Max	SD
Right Low Complexity					
ACCELERATE	74	3.07 s	3.49 s	4.86 s	0.51
BRAKE	77	3.07 s	3.4 s	34.86 s	0.37
FRONT-SA-NOCARFRONT	60	3.36 s	3.62 s	4.98 s	0.44
RIGHT-SA-NOCAR	60	4.53 s	4.83 s	6.43 s	0.49
DECISION-RIGHT	60	4.58 s	4.92 s	6.48 s	0.61
RIGHT-MIRROR-NOCAR	19	6.16 s	6.38 s	6.97 s	0.31
RIGHT-MIRROR-NOCAR-SHOULDER	19	6.16 s	6.52 s	7.78 s	0.55
ONLYSHOULDER-RIGHT-NOCAR	22	6.16 s	6.46 s	7.78 s	0.44
SHOULDER-RIGHT-NOCAR	19	6.57 s	6.92 s	8.19 s	0.56
EXECUTE-RIGHT	60	6.21 s	6.63 s	8.24 s	0.5
Right High Complexity					
ACCELERATE	154	2.87 s	3.64 s	7.75 s	0.88
BRAKE	146	2.87 s	3.58 s	6.98 s	0.78
FRONT-SA-NOCARFRONT	1	7.8 s	7.8 s	7.8 s	-
FRONT-SA-NOCARFRONT-PERCEIVED	59	3.42 s	3.92 s	6.27 s	0.62
RIGHT-SA-NOCAR	46	7.73 s	8.41 s	9.39 s	0.43
DECISION-RIGHT	46	7.78 s	8.49 s	9.44 s	0.44
RIGHT-MIRROR-NOCAR	16	9.95 s	12.68 s	13.02 s	0.75
RIGHT-MIRROR-NOCAR-SHOULDER	16	11.38 s	12.87 s	13.23 s	0.43
ONLYSHOULDER-RIGHT-NOCAR	14	9.95 s	12.73 s	13.23 s	0.82
SHOULDER-RIGHT-NOCAR	16	11.79 s	13.1 s	13.43 s	0.39
EXECUTE-RIGHT	46	10 s	12.89 s	13.48 s	0.69
decision-follow-nocar	14	10.66 s	11.55 s	13.02 s	0.71
execute-follow	14	10.71 s	11.6 s	13.06 s	0.71

Right Maneuver Scenarios: The first scenario that the model interacts with is the condition in which a lane change to the right is the best maneuver decision. Table 1 shows that in the complex condition, a maneuver decision is made 3.57 s (42%) later and the execution 6.26 s (48.6%) later. In addition, the model accelerates more often in the complex condition where a vehicle is present in the right back of the ego-vehicle. In the complex scenario, the model makes the wrong action decision to stay on the center lane in 14 out of 60 cases although the right oncoming lane is free.

Follow Maneuver Scenarios: In scenarios in which the best maneuver decision is car following, the model does not show significant differences between complexities. Model predictions do only show minimal difference in the time until an action decision is made ($M_{D.Low} = 5.94$ s; $M_{D.High} = 5.95$ s; Table 2). In both situations, only the leading and the right vehicles are attended. This result shows that only if vehicles in the surrounding traffic environment are relevant, they have an impact on the cognitive processes of the driver. In both car-following cases, the model makes a wrong decision. This is a lane change to the right ($M_{D.Wrong} = 5.67/5.81$ s; Table 2) although the right lane is occupied.

Table 2. Predicted time statistics of the cognitive model with a follow maneuver. Representation of minimum (Min), mean, and maximum (Max) values over $N = 60$ trials. Wrong decisions are in lower case letters (source: own table).

Production	Events	Predicted Time			
		Min	Mean	Max	SD
Follow Low Complexity					
ACCELERATE	84	2.87 s	3.11 s	3.37 s	0.22
BRAKE	96	2.87 s	3.12 s	3.37 s	0.2
FRONT-SA1	60	3.42 s	3.44 s	3.47 s	0.03
FRONT-SA2	60	4.81 s	5.28 s	5.62 s	0.4
DECISION-CHECK-RIGHT	376	4.86 s	5.55 s	6.37 s	0.42
RIGHT-SA-CAR	59	5.26 s	5.84 s	6.42 s	0.42
DECISION-FOLLOW-CAR	59	5.31 s	5.94 s	6.52 s	0.43
FRONT1	59	5.62 s	6.13 s	7.04 s	0.42
decision-right	1	5.76 s	5.76 s	5.76 s	-
execute-right	1	6.49 s	6.49 s	6.49 s	-
Follow High Complexity					
ACCELERATE	87	2.87 s	3.13 s	3.37 s	0.2
BRAKE	93	2.87 s	3.01 s	3.37 s	0.21
FRONT-SA1	60	3.42 s	3.44 s	3.47 s	0.02
FRONT-SA2	60	4.82 s	5.3 s	5.62 s	0.4
DECISION-CHECK-RIGHT	388	4.86 s	5.57 s	7.02s	0.45
RIGHT-SA-CAR	59	5.26 s	5.86 s	67.07 s	0.43
DECISION-FOLLOW-CAR	59	5.31 s	5.95 s	7.12 s	0.43
FRONT1	59	5.62 s	6.16 s	7.65 s	0.44
decision-right	1	5.81 s	5.81 s	5.81 s	-
execute-right	1	7.51 s	7.51 s	7.51 s	-

Left Maneuver Scenarios: The third maneuver option is a lane change to the left. Results show that in both scenarios, the action decision is taken at almost the same time ($M_{D.Low} = 4.79$ s; $M_{D.High} = 4.77$ s). This is due to the fact that the decision is based on the slower leading vehicle. For the action execution on the other hand, the vehicles on the left lane are relevant. Hence, 0.47 s (6%) more are needed to execute the action in the complex scenario (Table 3). In the easy scenario, the model makes 14 decisions for a lane change to the right. In the complex scenario, 12 times a lane change to the right is executed. This decision is suboptimal in the given traffic situation and thus considered as a wrong decision ($M_{D.Wrong} \approx 3.7$ s; Table 3).

Table 3. Predicted time statistics of the cognitive model with a lane change left maneuver. Representation of minimum (Min), mean, and maximum (Max) values over $N = 60$ trials. Wrong decisions are in lower case letters (source: own table).

Production	Events	Predicted Time			
		Min	Mean	Max	SD
Left Low Complexity					
ACCELERATE	105	2.87 s	3.18 s	3.65 s	0.23
BRAKE	93	2.87 s	3.15 s	3.65 s	0.26
FRONT-SA1	46	3.6 s	3.64 s	3.7 s	0.05
FRONT-SA2	46	4.41 s	4.83 s	5.83 s	0.64
DECISION-LEFT	14	4.46 s	4.74 s	6.08 s	0.55
DECISION-LEFT-ACCELERATE	12	4.46 s	4.96 s	5.98 s	0.7
DECISION-LEFT-BRAKE	20	4.46 s	5 s	5.98 s	0.69
LEFT-MIRROR-NOCAR	17	5.84 s	6.73 s	9.31 s	1.37
LEFT-MIRROR-NOCAR-SHOULDER	12	5.84 s	7.1 s	9.31 s	1.49
ONLYSHOULDER-LEFT-NOCAR	17	5.84 s	6.34 s	9.31 s	1.18
SHOULDER-LEFT-NOCAR	12	6.66 s	8.88 s	12.7 s	2.38
EXECUTE-LEFT	46	5.89 s	7.2 s	12.75 s	1.91
decision right	14	3.7 s	3.71 s	3.85 s	0.04
execute right	14	5.07 s	5.14 s	5.28 s	0.1
Left High Complexity					
ACCELERATE	94	2.87 s	3.15 s	3.65 s	0.24
BRAKE	103	2.87 s	3.17 s	3.65 s	0.25
FRONT-SA1	48	3.6 s	3.64 s	3.7 s	0.04
FRONT-SA2	48	4.41 s	4.79 s	6.03 s	0.61
DECISION-LEFT	16	4.46 s	4.86 s	6.13 s	0.66
DECISION-LEFT-ACCELERATE	15	4.46 s	4.72 s	6.13 s	0.47
DECISION-LEFT-BRAKE	17	4.46 s	5 s	6.18 s	0.68
LEFT-MIRROR-CAR	21	5.22 s	5.65 s	8.48 s	0.98
ONLYSHOULDER-LEFT-CAR	23	5.22 s	5.65 s	8.48 s	0.94
LEFT-MIRROR-NOCAR	23	5.84 s	6.95 s	13.36 s	1.78
LEFT-MIRROR-NOCAR-SHOULDER	19	5.84 s	7.6 s	13.36 s	2.37
ONLYSHOULDER-LEFT-NOCAR	9	5.84 s	7.35 s	13.36 s	2.38
SHOULDER-LEFT-NOCAR	16	6.65 s	8.74 s	13.97 s	2.36
EXECUTE-LEFT	48	5.89 s	7.67 s	14.02 s	2.21
decision-right	12	3.7 s	3.7 s	3.8 s	0.03
execute right	12	7.46 s	7.63 s	7.87 s	0.21

Comparison to Empirical Data

Predictions of the model in the scenarios are compared to a simulator study with $N = 20$ participants [31]. Each participant ran through every scenario three times. Thus, empirical data represent $N = 60$ samples per scenario.

In the empirical simulator study, participants took a decision in the same scenarios as the model did. In this paper, empirical decision times until the first action decision is made (including errors; between subject) are used, rather than times until participants take the right action decision as in [31]. This is of high interest, as the model is able to predict errors. Results of the traffic simulator study and model predictions are shown in Table 4 and Figure 7. The comparison between predicted and empirical values below shows that model predictions are already very close to empirical data in most conditions, but need to be improved for right complex and left complex scenarios. Figure 7 shows model predictions and empirical data in the different conditions for the main takeover points. The gray (empirical) and white (model) boxes represent the standard deviations of the data that are additionally listed in Table 4. Mean values of model predictions are in general close to the mean values of the empirical data. However, the standard deviations in empirical data are much higher than the model data. In the right easy scenario, participants took between 0.73 and 12.42 s to take an action decision. The model, in contrast, represents driver differences that deviate between 4.58 and 6.48 s. Furthermore, empirical data show

higher deviations in the time to take over the driving task. At this stage of the takeover, model predictions do not show any variation. In the right complex and the follow scenarios, participants are faster than the model. In the other three conditions on the other hand, the model is faster. Overall, predictions of right easy and left complex scenarios are mostly different from empirical data. Here, the model is faster than participants. Especially in the left complex conditions, standard deviations of empirical data are very high (3.11). In contrast, model predictions do not show higher standard deviations in the left complex scenario than in the other conditions. The highest standard deviation that the model is able to represent is in the right complex scenario (1.4). In empirical data, the complex conditions show higher standard deviations than the easy conditions. This difference is also found in model predictions except for the left scenarios. Here, the model predicts less differences between drivers in the complex scenarios than in the easy ones.

Table 4. Action decision (including wrong decision) times of empirical data vs. model predictions. Representation of minimum (Min), mean, maximum (Max) values, standard deviation (SD) for right (R), follow (F), and left (L) scenarios in easy (E) and complex (C) conditions (source: scharfeTRF).

Scenario	Empirical Data				Model Predictions			
	Min	Mean	Max	SD	Min	Mean	Max	SD
R E	0.73 s	5.06 s	12.42 s	2.07	4.58 s	4.92 s	6.48 s	0.5
R C	2.1 s	5.34 s	12.63 s	2.23	7.78 s	9.2 s	13.02 s	1.4
F E	2.82 s	5.09 s	9.7 s	1.79	5.31 s	5.94 s	6.52 s	0.42
F C	1.5 s	5.11 s	23.54 s	3.13	5.31 s	5.95 s	7.12 s	0.43
L E	2.73 s	4.84 s	10.25 s	1.73	3.7 s	4.63 s	6.08 s	0.76
L C	2.5 s	5.69 s	21.1 s	3.11	3.7 s	4.63 s	6.18 s	0.72

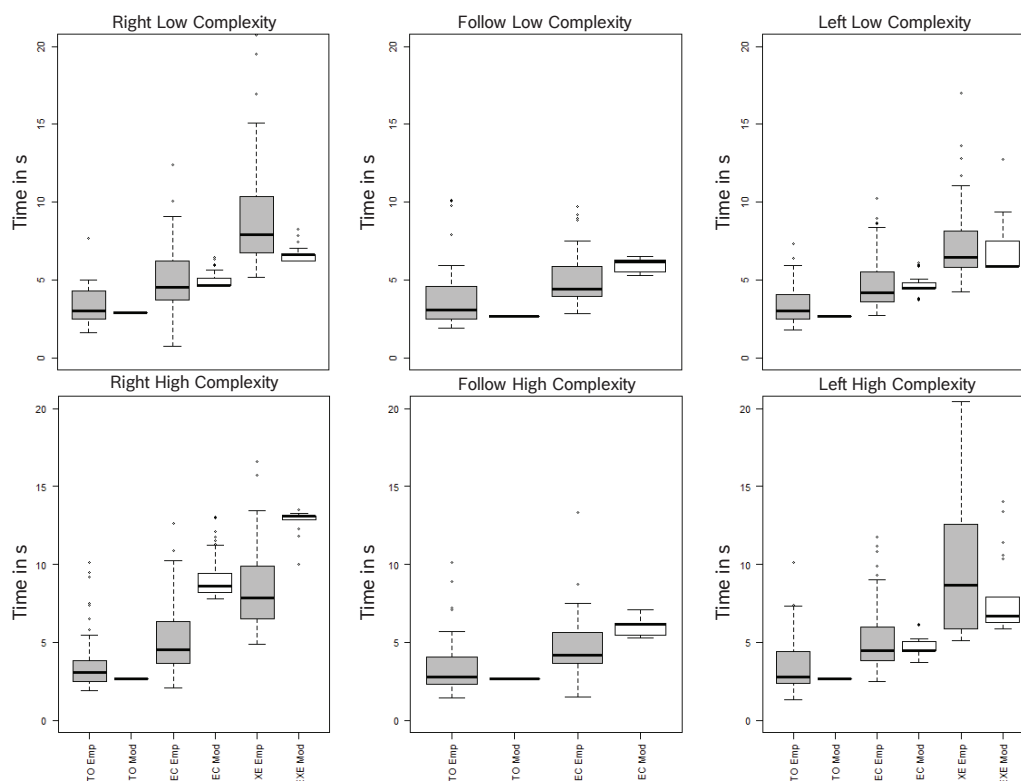


Figure 7. Box-Whisker plot of empirical data (Emp, gray) and model predictions (Mod, white) for the timeline in seconds (s) on the y-axis of the takeover (TO) and the time of the action decision (DEC). In right and left maneuver cases, the time of the action execution (EXE) is also included (source: own table).

Moreover, the model and participants make wrong decisions in the different scenarios. The model makes 42 wrong decisions in $N = 360$ (60 per scenario) takeovers. Participants of the empirical study made 13 wrong decisions in $N = 360$ (60 per scenario) takeover tasks. Wrong model decisions are mostly based on vehicles that have not been detected or too much acceleration/deceleration in the beginning. In the empirical study, wrong decisions are mostly based on overwhelmed drivers that decide to brake strongly (although the task is to keep the speed) and make a lane change to the right or follow the leading vehicle that is far slower than the ego-vehicle at the time of the takeover.

4. Discussion

The current model provides concrete operationalization of situation awareness, representing a plausible model of perceptual decisions. It is able to simulate cognitive processes during the takeover in highly automated driving from a non-driving-related task until the action execution. Interestingly, the model takes more time in the right complex maneuver scenario than in the left maneuver scenarios. As left maneuvers are usually more complex than a lane change to the right, lane changes to the left are assumed to take more time than to the right, due to more validation steps that are required to take over. Within the right complex maneuver, however, the vehicle in the right back had to be tracked and the speed calculated to make sure that it is safe to perform the lane change. As vehicles driving on the right are the slowest, the tracking of the vehicle in the right back takes longer until a safe distance to change the lane is reached. In the left maneuvers, the model only has to perceive that the right lane is occupied and the leading vehicle is slower. Overtaking vehicles in the left complex scenario also have to be perceived, but they overtake faster, and thus, the lane change can be executed.

Altogether, the model provides the basis to develop cognitive assistance systems for highly automated driving in the following way. As early warnings help distracted drivers to react more quickly [35], it is especially important to identify the takeover readiness and derive relevant adaptations. Those could be, for example, handing over step by step, visualizing the best maneuver option or indicating important aspects of the traffic environment to bring the driver back into-the-loop and enable a safe takeover. In combination with information of the vehicle sensors and eye tracking or physiological data of the driver, this cognitive model could directly interact with assistance systems by providing information about the current cognitive state of the driver.

Figure 8 shows how sensory data of the vehicle and wearables can be combined with the cognitive model to provide a situational adaption of assistance systems in highly automated driving. The integration of eye tracking into cognitive model predictions also allows the integration of the situation awareness status of the driver. Thus, a driver who is partly or completely in-the-loop would have already perceived certain objects of the traffic environment. These objects are stored in the declarative module. This represents the amount to which situation awareness is already created. This highlights the benefit of integrating a cognitive model into the development of adaptive assistance systems in highly automated driving. The model is able to represent a spectrum of different driving behaviors, resulting in deviating time that is needed to make an action decision and execute the action. In addition, the current state of situation awareness at any stage of the takeover process can be represented.

Nevertheless, the model shows limitations. At the moment, productions in the model check static conditions regularly to gather information about the dynamics of the situation. Thus, the model uses several productions to perceive changes. In a next step, productions that are able to trace a moving object could be used for enhancing the model. Another aspect that should be considered in future development is learning. The integration of learning into the model would help represent differences in experience levels between drivers. Furthermore, the model makes a decision and sticks to it. Thus, as soon as a decision is made, the model cannot come back to an earlier decision point. More looping options to change a decision while taking over the driving task have to be implemented.

Moreover, more speed variations of the traffic should be tested and compared to further data of empirical investigations. There might be great differences due to speed deltas that have an influence on the reaction times, as it is in our case with the right complex maneuver. Another aspect, the model does not address yet, are further driver differences in gaze guidance and motoric reactions. Furthermore, the projection stage of situation awareness is not integrated into the model yet. By integrating these points, the model might be able to better predict the deviations that are found in empirical data and represent driver differences for the whole takeover process more closely.

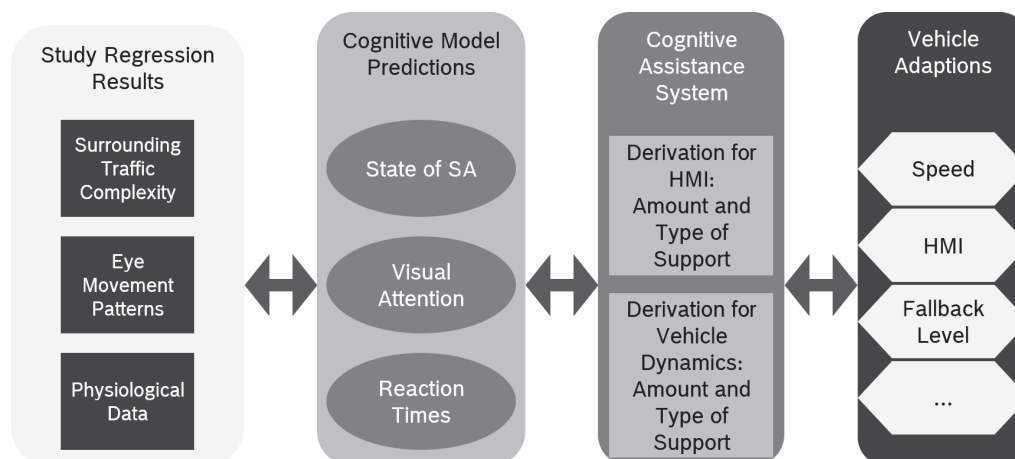


Figure 8. Usage of the cognitive model to support situational adaptations of adaptive assistance systems in highly automated driving (source: own figure).

To better predict individual subjects' patterns of performance, the authors of [36] proposed a computational model that accounts for differences in working memory capacity. The integration of this approach into the presented cognitive model could be a first step to further predict individual differences. In combination with that, different forms of learning should be integrated into the cognitive model of the takeover task, as [37] did for a fault-finding task to represent how drivers adapt to the new forms of driving. In the future, more flexible approaches that address the above-mentioned limitations have to be analyzed to better represent driver differences. Especially when dealing with the visual process during a takeover, saccades and fixation times become relevant when predicting the individual takeover process. In a next step, eye-tracking results of a simulator study will be compared to the visual traces of the cognitive model to improve predictions that represent individual variations and analyze errors. Better individualization can result in a cognitive assistance system that is able to guide the driver through a takeover individually and increase safety. Information that the driver needs in a certain situation can be displayed individually and vehicle dynamics can be adapted to hand over the driving task dynamically and adapt the given takeover time individually.

5. Conclusions

In order to enhance the usage of such a cognitive model in highly automated driving, some limitations have to be considered. First, the presented model runs only through six scenarios. More scenarios should be tested and compared to a larger amount of empirical data. Second, the productions of the current model check static conditions and compare them to collect information about the dynamics of a situation, but do not actively trace an object. Third, the model has not yet been connected to a driving simulator. These aspects should be met in future research. Furthermore, the model is only able to represent some of the differences between drivers. These are mainly speed adaption and safety habits. Other differences between driver takeovers still have to be investigated that lead to the higher standard deviations that are found in empirical data. However, the model is able to represent different cognitive processes during a takeover in dynamic environments. The

usage of AOIs enhances the simulation of the visual perception. Furthermore, a spectrum of different driving decisions in different situations can be represented. It is able to illustrate the current state of situation awareness in different takeover situations that are found within empirical studies. However, time predictions are close but not perfectly in line with times that are found in empirical data. The model can be used to predict different driving strategies and integrate vehicle and participant data into adaptive assistance systems. By knowing, for example, that a driver is overwhelmed by the situation and has not yet collected enough information of the surrounding traffic environment, the automated system could adapt correspondingly. In such a situation, the system would decelerate, visually guide the driver, and support the maneuver. Using cognitive modeling, assistance systems are able to integrate the cognitive state of the driver on hand and adapt to the situation and the present cognitive state of the driver. Based on such a cognitive model, an automated system could classify the driver's takeover readiness, derive the expected takeover quality [8] and adapt the cognitive assistance for non-critical takeovers accordingly.

Author Contributions: Conceptualization, M.S.L.S.-S. and N.R.; methodology, M.S.L.S.-S.; software, M.S.L.S.-S.; validation, M.S.L.S.-S.; formal analysis, M.S.L.S.-S.; investigation, M.S.L.S.-S.; resources, M.S.L.S.-S. and S.W.; data curation, M.S.L.S.-S.; writing—original draft preparation, M.S.L.S.-S.; writing—review and editing, S.W. and N.R.; visualization, M.S.L.S.-S.; supervision, N.R.; project administration, M.S.L.S.-S. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki, and approved by the or Ethics Committee of TU Berlin (Eingangsdatum des Antrages: 20 March 2019, Antragsnummer: SF0120190320, Datum der Beschlussfassung: 1 April 2019).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: Not applicable.

Acknowledgments: We wish to acknowledge the help provided by Kathrin Zeeb and Michael Schulz at Robert Bosch GmbH, our department at TU Berlin and Klaus Bengler from TU Munich as second supervisor.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Sheridan, T.B. Human-robot interaction: Status and challenges. *Hum. Factors* **2016**, *58*, 525–532. [CrossRef] [PubMed]
2. Louw, T.; Merat, N.; Jamson, A. Engaging with Highly Automated Driving: To Be or Not to Be in the Loop? In Proceedings of the 8th International Driving Symposium on Human Factors in Driver Assessment, Training and Vehicle Design, Salt Lake City, UT, USA, 22–25 June 2015. [CrossRef]
3. SAE-International. Surface Vehicle Recommended Practice. Taxonomy and Definitions for Terms Related to On-Road Motor Vehicle Automated Driving Systems. In Proceedings of the SAE International, Hong Kong, China, 4 May 2016.
4. Kerschbaum, P.; Lorenz, L.; Bengler, K. A transforming steering wheel for highly automated cars. In Proceedings of the 2015 IEEE Intelligent Vehicles Symposium (IV), Seoul, Korea, 1–28 June 2015; pp. 1287–1292.
5. Zeeb, K.; Buchner, A.; Schrauf, M. What determines the take-over time? An integrated model approach of driver take-over after automated driving. *Accid. Anal. Prev.* **2015**, *78*, 212–221. [CrossRef] [PubMed]
6. Merat, N.; Seppelt, B.; Louw, T.; Engström, J.; Lee, J.D.; Johansson, E.; Green, C.A.; Katagaki, S.; Monk, C.; Itoh, M.; et al. The “Out-of-the-Loop” concept in automated driving: Proposed definition, measures and implications. *Cogn. Technol. Work.* **2019**, *21*, 87–98. [CrossRef]
7. Bainbridge, L. Ironies of automation. In *Analysis, Design and Evaluation of Man–Machine Systems*; Elsevier: Amsterdam, The Netherlands, 1983; pp. 129–135.
8. Braunagel, C.; Rosenstiel, W.; Kasneci, E. Ready for Take-Over? A New Driver Assistance System for an Automated Classification of Driver Take-Over Readiness. *IEEE Intell. Transp. Syst. Mag.* **2017**, *9*, 10–22. [CrossRef]
9. Klimke, J.; Themann, P.; Klas, C.; Eckstein, L. Definition of an embedded driver model for driving behavior prediction within the DESERVE platform. In Proceedings of the 2014 International Conference on Embedded Computer Systems: Architectures, Modeling, and Simulation (SAMOS XIV), Samos, Greece, 14–17 July 2014; pp. 343–350. [CrossRef]
10. Gold, C.; Happee, R.; Bengler, K. Modeling take-over performance in level 3 conditionally automated vehicles. *Accid. Anal. Prev.* **2018**, *116*, 3–13. [CrossRef]

11. Naujoks, F.; Forster, Y.; Wiedemann, K.; Neukum, A. A human-machine interface for cooperative highly automated driving. In *Advances in Human Aspects of Transportation*; Springer: New York, NY, USA, 2017; pp. 585–595.
12. Markkula, G.; Boer, E.; Romano, R.; Merat, N. Sustained sensorimotor control as intermittent decisions about prediction errors: Computational framework and application to ground vehicle steering. *Biol. Cybern.* **2018**, *112*, 181–207. [CrossRef]
13. Aasman, J.; Michon, J.A. Multitasking in driving. In *Soar: A Cognitive Architecture in Perspective*; Springer: New York, NY, USA, 1992; pp. 169–198.
14. Cao, J.; Wang, H. Vision navigation for driver cognitive model in act-r architecture. In Proceedings of the 9th IEEE International Conference on Cognitive Informatics (ICCI'10), Beijing, China, 7–9 July 2010; pp. 946–951. [CrossRef]
15. Salvucci, D. Modeling driver behavior in a cognitive architecture. *Hum. Factors* **2006**, *48*, 362–380. [CrossRef]
16. Anderson, J.; Bothell, D.; Byrne, M.D.; Douglass, S.; Lebiere, C.; Qin, Y. An integrated theory of the mind. *Psychol. Rev.* **2004**, *111*, 1036. [CrossRef]
17. Scharfe, M.S.L.; Russwinkel, N. Towards a Cognitive Model of the Takeover in Highly Automated Driving for the Improvement of Human Machine Interaction. In Proceedings of the International Conference on Cognitive Modelling, Montreal, QC, Canada, 30 July 2019; pp. 210–216.
18. Morita, J.; Miwa, K.; Maehigashi, A.; Terai, H.; Kojima, K.; Ritter, F. Cognitive Modeling of Automation Adaptation in a Time Critical Task. *Front. Psychol.* **2020**, *11*, 2149. [CrossRef]
19. Klaproth, O.W.; Halbrügge, M.; Krol, L.R.; Vernaleken, C.; Zander, T.O.; Russwinkel, N. A Neuroadaptive Cognitive Model for Dealing With Uncertainty in Tracing Pilots' Cognitive State. *Top. Cogn. Sci.* **2020**, *12*, 1012–1029. [CrossRef]
20. Endsley, M.R. Toward a theory of situation awareness in dynamic systems. *Hum. Factors* **1995**, *37*, 32–64. [CrossRef]
21. Wickens, C. *Noticing Events in the Visual Workplace: The SEEV and NSEEV Models*; Cambridge University Press: Cambridge, UK, 2015; pp. 749–768. [CrossRef]
22. Wickens, C.; McCarley, J.; Steelman-Allen, K. NT-SEEV: A model of attention capture and noticing on the flight deck. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*; SAGE Publications: Los Angeles, CA, USA, 2009; Volume 53, pp. 769–773.
23. Horrey, W.J.; Wickens, C.D.; Consalus, K.P. Modeling drivers' visual attention allocation while interacting with in-vehicle technologies. *J. Exp. Psychol. Appl.* **2006**, *12*, 67. [CrossRef]
24. Anderson, J.; Matessa, M.; Lebiere, C. act-r: A theory of higher level cognition and its relation to visual attention. *Hum.-Comput. Interact.* **1997**, *12*, 439–462. [CrossRef]
25. Ritter, F.; Tehranchi, F.; Oury, J. ACT-R: A cognitive architecture for modeling cognition. *Wiley Interdiscip. Rev. Cogn. Sci.* **2019**, *10*, e1488. [CrossRef] [PubMed]
26. Anderson, J. *How Can the Human Mind Occur in the Physical Universe?* Oxford University Press: Oxford, UK, 2007. [CrossRef]
27. Wiese, S.; Lotz, A.; Russwinkel, N. SEEV-VM: ACT-R Visual Module based on SEEV theory. In Proceedings of the 17th International Conference on Cognitive Modeling, Montreal, QC, Canada, 20 July 2019; pp. 301–307.
28. Rauschke, J. Development, Implementation and Validation of a Simulation Environment for the Cognitive Modelling of the Driver-Vehicle-Interaction in Highly Automated Driving. Practical Training Thesis, Robert Bosch GmbH and Cooperative State University Baden-Württemberg, Stuttgart, Germany, 2019.
29. Fu, W.T.; Bothell, D.; Douglass, S.; Haimson, C.; Sohn, M.H.; Anderson, J. Toward a real-time model-based training system. *Interact. Comput.* **2006**, *18*, 1215–1241. [CrossRef]
30. Albers, D.; Radlmayr, J.; Loew, A.; Hergeth, S.; Naujoks, F.; Keinath, A.; Bengler, K. Usability evaluation—Advances in experimental design in the context of automated driving human-machine interfaces. *Information* **2020**, *11*, 240. [CrossRef]
31. Scharfe-Scherf, M.S.L.; Russwinkel, N. Familiarity and Complexity during a Takeover in Highly Automated Driving. *Int. J. Intell. Transp. Syst. Res.* **2021**, *19*, 525–538. [CrossRef]
32. Scharfe, M.S.L.; Zeeb, K.; Russwinkel, N. The Impact of Situational Complexity and Familiarity on Takeover Quality in Uncritical Highly Automated Driving Scenarios. *Information* **2020**, *11*, 115. [CrossRef]
33. Ritter, F.; Schoelles, M.; Quigley, K.; Klein, L. Determining the number of model runs: Treating cognitive models as theories by not sampling their behavior. *Hum. Loop Simul. Methods Pract.* **2011**, *11*, 97–116.
34. Li, G.; Li, S.E.; Jia, L.; Wang, W.; Cheng, B.; Chen, F. Driving maneuvers analysis using naturalistic highway driving data. In Proceedings of the 2015 IEEE 18th International Conference on Intelligent Transportation Systems, Gran Canaria, Spain, 15–18 September 2015; pp. 1761–1766.
35. Lee, J.D.; McGehee, D.V.; Brown, T.L.; Reyes, M.L. Collision warning timing, driver distraction, and driver response to imminent rear-end collisions in a high-fidelity driving simulator. *Hum. Factors* **2002**, *44*, 314–334. [CrossRef]
36. Daily, L.Z.; Lovett, M.C.; Reder, L.M. Modeling individual differences in working memory performance: A source activation account. *Cogn. Sci.* **2001**, *25*, 315–353. [CrossRef] [PubMed]
37. Ritter, F.; Bibby, P.A. Modeling how, when, and what is learned in a simple fault-finding task. *Cogn. Sci.* **2008**, *32*, 862–892. [CrossRef] [PubMed]

Article

What Attracts the Driver's Eye? Attention as a Function of Task and Events

Yke Bauke Eisma *, Dirk J. Eijssen and Joost C. F. de Winter *

Department of Cognitive Robotics, Faculty of Mechanical, Maritime and Materials Engineering,
Delft University of Technology, 2628 CD Delft, The Netherlands; eijssen.dirk@gmail.com

* Correspondence: y.b.eisma@tudelft.nl (Y.B.E.); j.c.f.dewinter@tudelft.nl (J.C.F.d.W.)

Abstract: This study explores how drivers of an automated vehicle distribute their attention as a function of environmental events and driving task instructions. Twenty participants were asked to monitor pre-recorded videos of a simulated driving trip while their eye movements were recorded using an eye-tracker. The results showed that eye movements are strongly situation-dependent, with areas of interest (windshield, mirrors, and dashboard) attracting attention when events (e.g., passing vehicles) occurred in those areas. Furthermore, the task instructions provided to participants (i.e., speed monitoring or hazard monitoring) affected their attention distribution in an interpretable manner. It is concluded that eye movements while supervising an automated vehicle are strongly ‘top-down’, i.e., based on an expected value. The results are discussed in the context of the development of driver availability monitoring systems.

Keywords: driving; eye-tracking; attention

1. Introduction

Driver availability monitoring systems (DAMS) are finding their way into consumer vehicles and will soon become mandatory for newly sold cars [1]. DAMS do not only have utility in manual driving but also in automated driving [2,3]. In particular, SAE Level 2 automated driving requires drivers to focus on the driving task, which can be verified using eye- and head-tracking [4,5] and hands-on-wheel detection technology [6,7].

One point of importance in the development of DAMS is *what* the driver should look at. Lappi et al. [8] noted that a vital regularity in eye movements in driving is that drivers’ “gaze is focused on task-relevant objects and locations . . . rather than the visually most salient ones . . .”. Research into hazard perception shows that more experienced drivers are better able to detect and respond to hazards and precursors of hazards [9,10]. Additionally, more experienced drivers have a larger horizontal variance of gaze and are less likely to look at the road center, which may be explained by their superior skills in identifying hazards or their higher spare mental capacity [11–14]. In short, a DAMS that merely measures whether the driver looks towards the forward roadway may not be optimal as drivers acquire situational awareness by perceiving various elements of the task environment. Furthermore, safe driving involves more than looking at relevant features. For example, previous research into highway hypnosis [15,16] and the looked-but-failed-to-see phenomenon [17] showed that looking does not imply that the driver comprehends the situation.

Another factor concerns the driving task itself. It can be expected that a driver’s attention distribution depends on whether the driver is driving manually or whether he or she is tasked to supervise an automated driving system. Previous research indicates that drivers of automated vehicles are less likely to glance at the forward roadway than manual drivers, which can be explained by the visual demands associated with manual lane-keeping [18–23].

When drivers supervise their automated vehicles, they still need to be situationally aware in order to be able to make an appropriate evasive maneuver after reclaiming control [24]. For example, the supervisor of the (SAE-2-level) automated vehicle may have to keep track of the speed of the vehicle and will need to be vigilant to other vehicles in the vicinity as these vehicles may have to be avoided after taking over control [25]. The extent to which drivers of automated vehicles should focus on the speedometer or hazards in the driving environment depends on the expected capabilities of the automated driving system. That is, relatively unreliable automated vehicles may require a higher level of alertness than others. In summary, there is no single definition of driving, but the goals the driver pursues may depend on the tasks the driver has. These statements resonate with classical research by Yarbus [26], which demonstrated that the task given to an observer affects eye movements to a great extent, although replication studies suggest that this finding is still contentious (see [27–30], for replication studies with mixed success).

The current study, conducted with a desktop-based eye tracker, aimed to explore how participants distribute attention while monitoring an automated car on a highway. Attentional distribution was investigated in three tasks: no extra task, an additional speed monitoring task, and an additional hazard detection task. Participants viewed the same non-interactive videos for the three tasks, allowing us to examine viewing behavior in a controlled and contextualized manner—that is, we examined how events in the environment, such as the proximity of other vehicles, affect viewing behavior. We end this work with a discussion of our findings, limitations, and recommendations for the development of DAMS.

2. Methods

2.1. Participants

Twenty students (seventeen males and three females) from the Delft University of Technology participated in the experiment. The participants had a mean age of 24.8 years ($SD = 2.1$) and were all in possession of a driver's license. They held their license for an average of 5.8 years ($SD = 3.2$). The participants' yearly driving distance was 3530 km per year ($SD = 2621$). All participants provided written informed consent before partaking in the experiment.

2.2. Experimental Setup

The experiment was performed on a 24-inch monitor with a resolution of 1920×1080 pixels (display area 531×298 mm). A head support was used to minimize participant head movements. The monitor was positioned at a distance of approximately 0.9 m from the head support; see Figure 1 for a photo of the experimental setup. Eye-tracking was performed using the SR Research EyeLink 1000 Plus eye-tracker, which recorded the right eye.



Figure 1. Experimental setup. Shown from back to front is the monitor for displaying the videos, the eye-tracker camera with infrared illuminator, the keyboard used for pressing the spacebar and arrow keys, and the head support.

2.3. Stimuli

Videos were created in the Euro Truck Simulator 2, a virtual environment that offers realistically looking roads, environments, and road users (e.g., cars and trucks). The driving speed and lane position of other road users was simulated by the AI of the Euro Truck Simulator.

A total of three 120-s videos were used. In each of the pre-recorded videos, the experimenter drove the ego-vehicle by means of the keyboard. In all three videos, the ego-vehicle drove on a two-lane highway with a speed that varied between 80 and 130 km/h. The videos had a frame rate of 60 fps.

- In Video 1, the ego-vehicle overtook six vehicles and was overtaken by one vehicle. The speed of the ego-vehicle crossed the 100 km/h mark six times, and the vehicle changed lanes six times.
- In Video 2, the ego-vehicle overtook two vehicles and was overtaken by five vehicles. The speed of the ego-vehicle crossed the 100 km/h mark 13 times, and the vehicle changed lanes four times.
- In Video 3, the ego-vehicle overtook eight vehicles and was overtaken by two vehicles. The 100 km/h mark was crossed six times, and six lane changes occurred.

2.4. Experimental Task and Procedures

Participants were presented with instructions in written form on the computer monitor. It was mentioned that participants should act as if they were in a vehicle that drove itself and which they had to monitor. It was also mentioned that they should be able to take over control at any time. It was mentioned that the experiment comprised three different tasks:

1. Baseline: You do not have any secondary tasks to perform.
2. Speed Task: You should shortly press the spacebar every time the speed dial crosses the 100 km/h mark. This works two ways around, i.e., when the car is braking or accelerating. As soon as the dial is on the 100 km/h mark, you should press; as soon as the pointer has passed the mark, you should release the spacebar.
3. Hazard Task: You will have to indicate whether a car or truck is driving in the lane next to you. As long as this is the case, you have to press the arrow key. Press the right arrow key when someone is driving on your right side and the left arrow key when someone is driving on your left. When there is nothing on your left or right, you have to release the arrow key.

Although the Speed Task and Hazard Task can reliably be performed by automation technology (namely through cruise control and blind-spot monitoring systems, respectively), supervisors of automated vehicles will still need to perform these tasks in order to stay situationally aware.

Next, the participants were shown a short (13 s) example video to give them an idea of the driving environment. Then, calibration of the eye-tracker was performed. Before each trial, the task for the upcoming trial was shown on the monitor. It was repeated that the primary task was to monitor the self-driving car, to stay aware of the state of the vehicles and traffic signs, and to be able to take over steering at any time. Furthermore, before the Speed Task trial, it was mentioned that the secondary task was to press the spacebar every time the speed crossed 100 km/h. Before the Hazard Task trial, it was mentioned that the secondary task was to hold the left/right arrow key as long as a vehicle was driving next to them.

Each participant performed eighteen 120-s trials. These trials comprised three different videos and three different task instructions, as specified above. Furthermore, half of the videos contained gaze-contingent visual feedback in the form of a red outline of the front windshield, speed dial, left mirror, or right mirror (for a similar approach, see [31]). The gaze-contingent feedback was based on an algorithm that kept track of the participants' previous eye movements and accordingly aimed to guide their eye movements to relevant areas. The order of the 18 trials was randomized and different for each participant.

For this paper, results from nine of eighteen trials were used. The nine videos with gaze-contingent feedback were not analyzed because the onset and offset of the gaze-contingent feedback showed large individual differences, and preliminary analyses did not provide insights that contribute to the present paper's purpose.

2.5. Data Processing

First, data gaps (e.g., due to blinks) in the horizontal and vertical gaze coordinates were filled through linear interpolation. Next, the horizontal and vertical gaze coordinates were filtered using a median filter with a time interval of 100 ms.

The attention distribution was determined by analyzing whether a participant's gaze point fell within one of four areas of interest (AOIs): windshield, left mirror, right mirror, and dashboard (note that the rear-view mirror was not included in the animated video clips of this experiment). The AOIs for the left mirror, right mirror, and dashboard were pragmatically defined as rectangular shapes. All other gaze points were labeled as the windshield. Because of the aforementioned linear interpolation, there were no missing data—that is, for each frame, the percentage of participants looking at the left mirror, right mirror, dashboard, and windshield adds to 100%.

The number of participants who had their eyes in one of these four AOIs was counted for each time sample and moving-average-filtered with a time interval of 500 ms. Subsequently, the participants' attention to the AOIs was graphically associated with events in the scenario, such as ego-vehicle lane changes, ego-vehicle speed, and vehicle size in the mirrors.

The driving simulator we used for creating the videos did not record the state of other vehicles (nor the view the ego-driver had of those vehicles). Additionally, future automated vehicles will likely rely on cameras to obtain information about surrounding traffic [32] and to replace the outside mirrors with screens [33]. Therefore, the size (~closeness) of vehicles visible in the left and right mirrors was obtained using a pre-trained YoloV4 model [34] (obtained from [35]). YoloV4 can detect 80 different object types. From these, we extracted the cars, trucks, and buses. The width of the object in the mirror that yielded the highest classification score was used as an index of vehicle size. Only objects yielding classification scores greater than 0.8 were taken into consideration, to exclude false-positive and low-quality detections (such as opposing traffic on the other side of the road).

3. Results

Figure 2 shows the heatmap for the three task conditions. It can be seen that the attention of participants was highly concentrated on either the left mirror, right mirror, the road ahead, or the speedometer. Participants also occasionally looked at the navigation display, which was ignored in the present analysis and grouped with the dashboard AOI. In summary, the heatmaps confirm that the use of four AOIs (windshield, left mirror, right mirror, dashboard) is appropriate.

Figure 3 shows the attention distribution across the four AOIs for Videos 1, 2, and 3 in the Baseline condition, i.e., without a secondary task. It is noticeable that the attention distribution strongly fluctuated with elapsed time. In an attempt to elucidate the causes of these fluctuations, Figures 4–7 show the moments in Video 3 when a maximal proportion of participants focused their attention on the windshield ($t = 94.1$ s), left mirror ($t = 24.8$ s), right mirror ($t = 84.0$ s), and dashboard ($t = 17.0$ s), respectively.

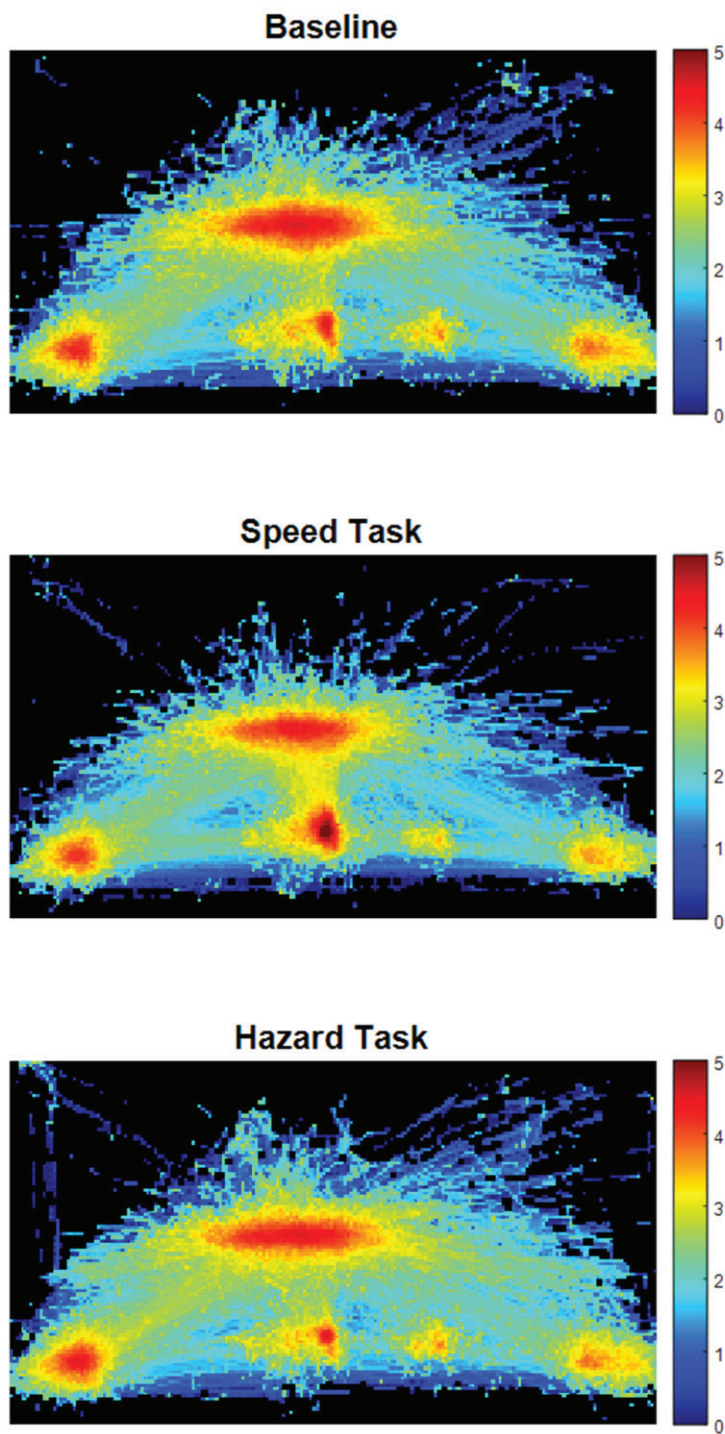


Figure 2. Heatmap of gaze points for the three videos aggregated for the Baseline Task (**top**), Speed Task (**middle**), and Hazard Task (**bottom**). The heatmap was divided into 10×10 -pixel squares and normalized. The colors are on a base-10 logarithmic scale. For example, a dark red color (value = 5) means there was 100 times more attention to that region than a yellow color (value = 3).

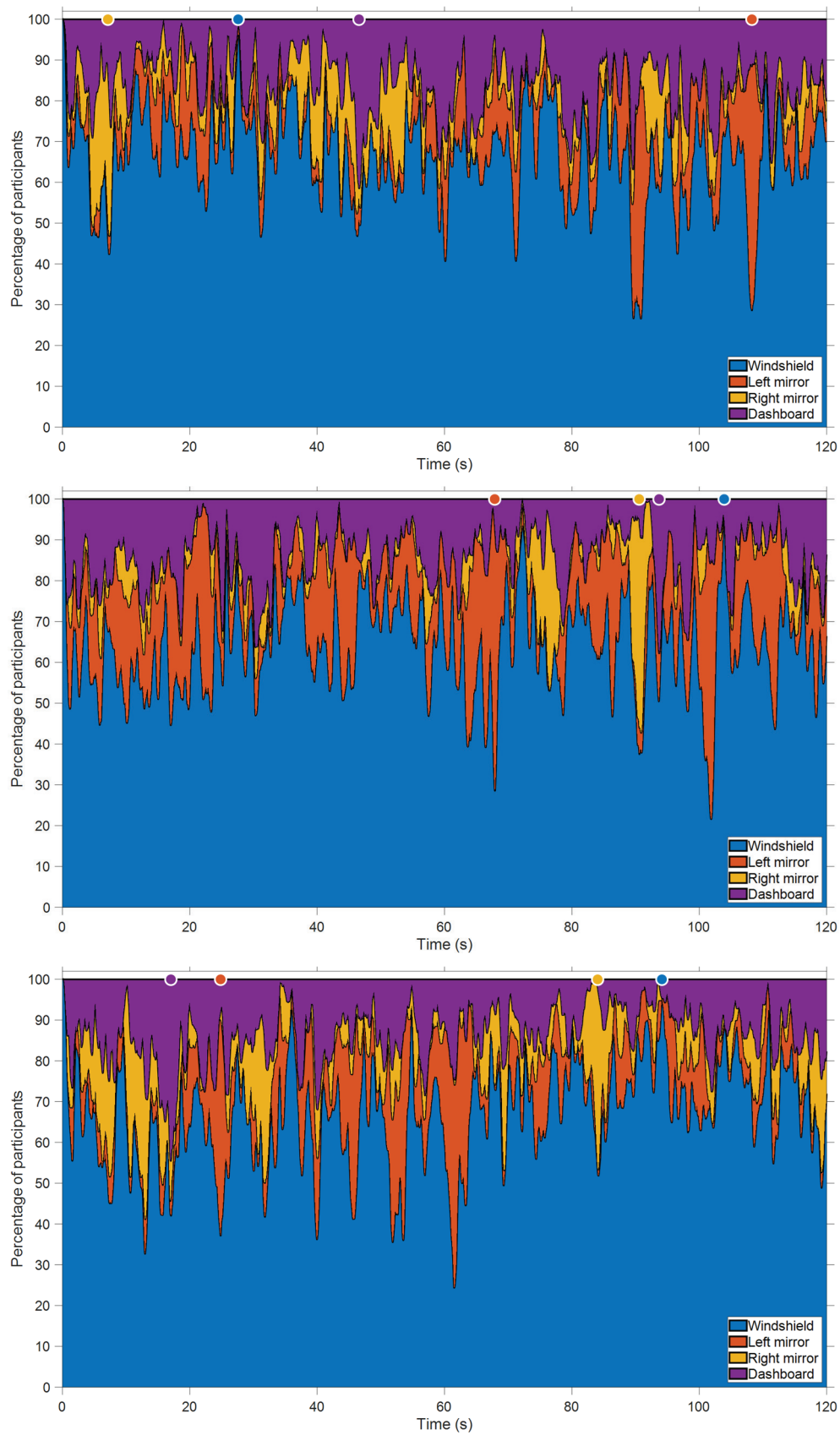


Figure 3. Attention distribution across four areas of interest (AOIs) for Video 1 (**top**), Video 2 (**middle**), and Video 3 (**bottom**) in the Baseline task. The colored circular markers at the top of each figure represent the moments where the maximal number of participants glanced at the AOI.

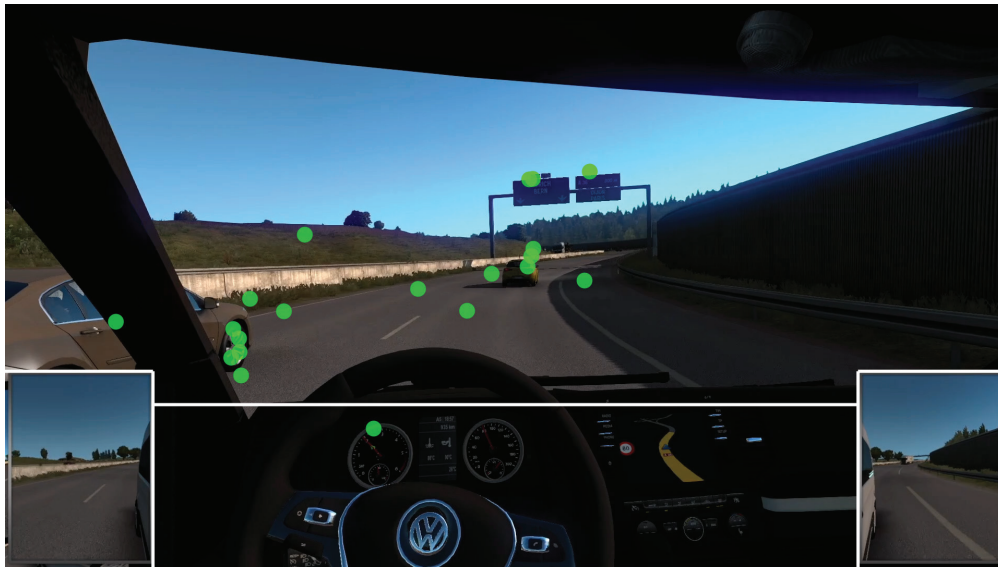


Figure 4. Moment in Video 3 where most participants attended to the windshield (Baseline Task). The green markers represent the gaze coordinates of individual participants. White lines are drawn around the AOIs.



Figure 5. Moment in Video 3 where most participants attended to the left mirror (Baseline Task). The green markers represent the gaze coordinates of individual participants. White lines are drawn around the AOIs.



Figure 6. Moment in Video 3 where most participants attended to the right mirror (Baseline Task). The green markers represent the gaze coordinates of individual participants. White lines are drawn around the AOIs.



Figure 7. Moment in Video 3 where most participants attended to the dashboard (Baseline Task). The green markers represent the gaze coordinates of individual participants. White lines are drawn around the AOIs.

- The windshield attracted attention when events (e.g., car in front, curve, road signs) were happening in that AOI (Figure 4).
- The left mirror attracted attention right before a lane change to the left lane and while a vehicle was visible in the left mirror (Figure 5). The participants may have anticipated that a lane change was about to happen based on the proximity of other vehicles as well as the turn indicator usage visible on the dashboard.
- The right mirror attracted attention right before changing lanes to the right after having overtaken another vehicle (Figure 6).
- Finally, the speedometer attracted attention in otherwise uneventful conditions while the road was relatively empty (Figure 7).

Figure 8 illustrates that vehicles in the mirror attracted attention—that is, when another vehicle was driving in the left lane to overtake the ego-vehicle, participants were increasingly likely to focus on the left mirror as this vehicle got closer.

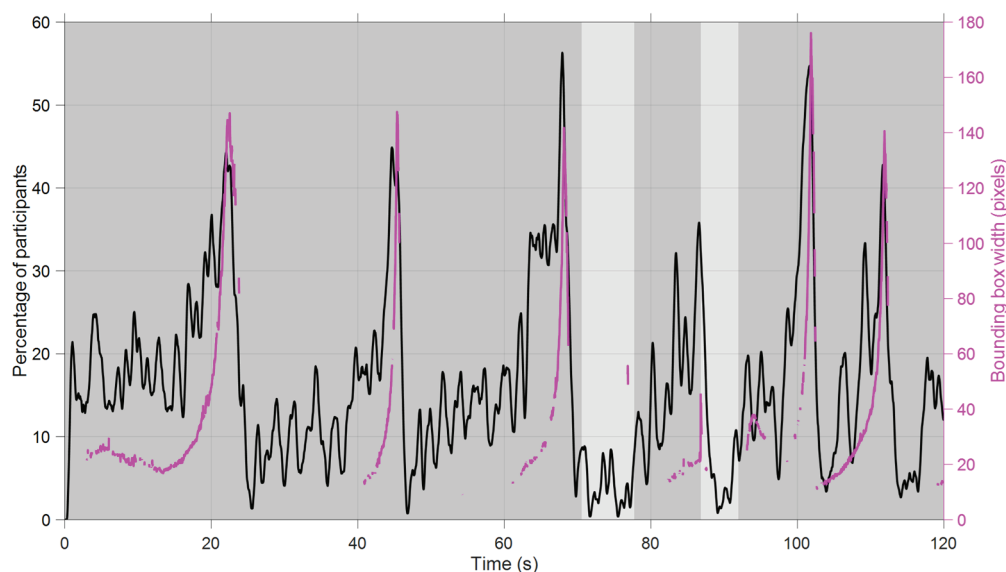


Figure 8. Left axis (black): Percentage of participants with their eyes on the left mirror (Video 2; three tasks averaged). Right axis (magenta): Width of the bounding box of the detected vehicle in the left mirror. The shaded background indicates whether the ego-vehicle was driving on the right lane (dark gray) or left lane (light gray).

The effects of the three task-instruction conditions are shown in Table 1. The Speed Task caused attention to be focused on the dashboard (~40% vs. ~14% in the Baseline Task). In the Hazard Task, attention was directed away from the dashboard and towards the left mirror (13–21% vs. 10–17% in the Baseline Task).

Table 1. Attention distribution to the four Areas of Interest (AOIs) for the nine trials.

Video Number	Task	Windshield (% of Time)	Left Mirror (% of Time)	Right Mirror (% of Time)	Dashboard (% of Time)
1	Baseline	67	10	8	15
2	Baseline	65	17	5	14
3	Baseline	68	11	8	13
1	Speed Task	50	8	5	37
2	Speed Task	43	12	3	43
3	Speed Task	46	8	5	42
1	Hazard Task	67	13	8	11
2	Hazard Task	62	21	5	12
3	Hazard Task	72	13	8	7

The focus on the dashboard during the Speed Task fluctuated over time, with the peaks in attention to the dashboard corresponding to moments when the speed was close to 100 km/h (Figures 9 and 10). In the same vein, in the Hazard Task, participants were relatively likely to attend to the mirrors when objects were visible in those mirrors (see Figure 11 for an illustration). Thus, this again shows that attention distribution is a function of the situation and task.

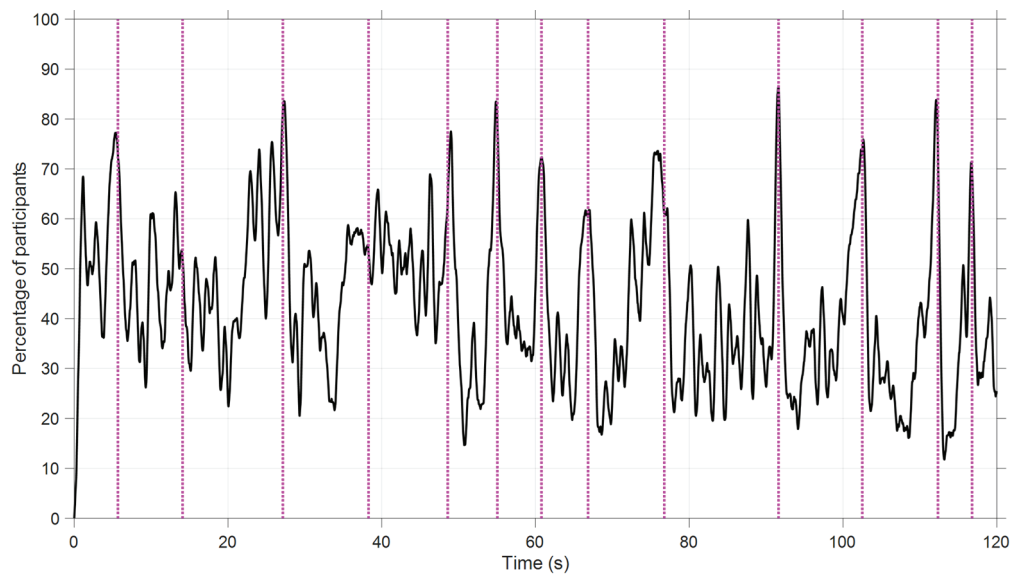


Figure 9. Percentage of participants with their eyes on the dashboard (Video 2, Speed Task). The dotted vertical lines represent the moments of 100 km/h threshold crossings.

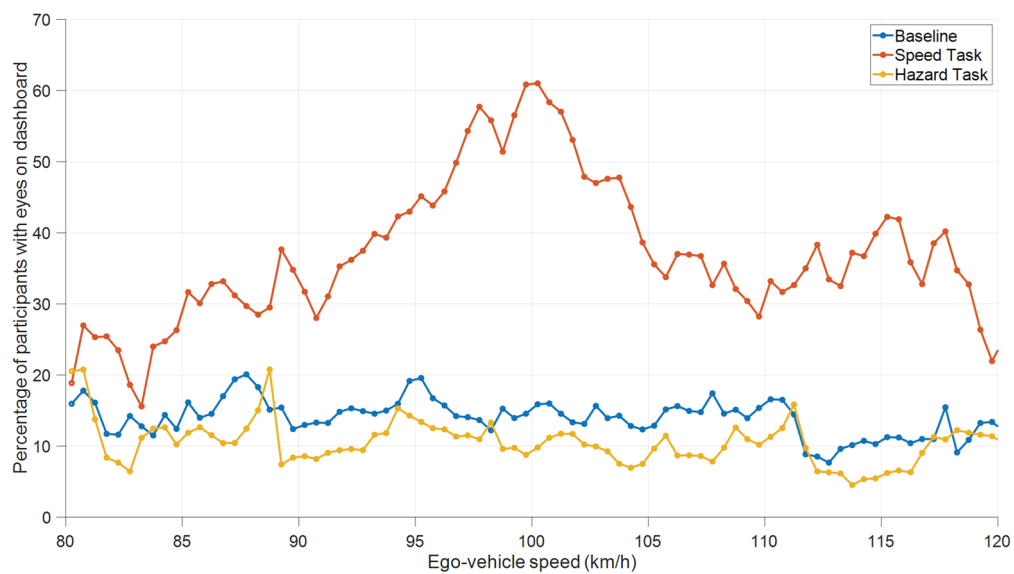


Figure 10. Percentage of participants with their eyes on the dashboard as a function of vehicle speed in the Baseline Task, Speed Task, and Hazard Task (Videos 1–3 combined, first 5 s of each video excluded).

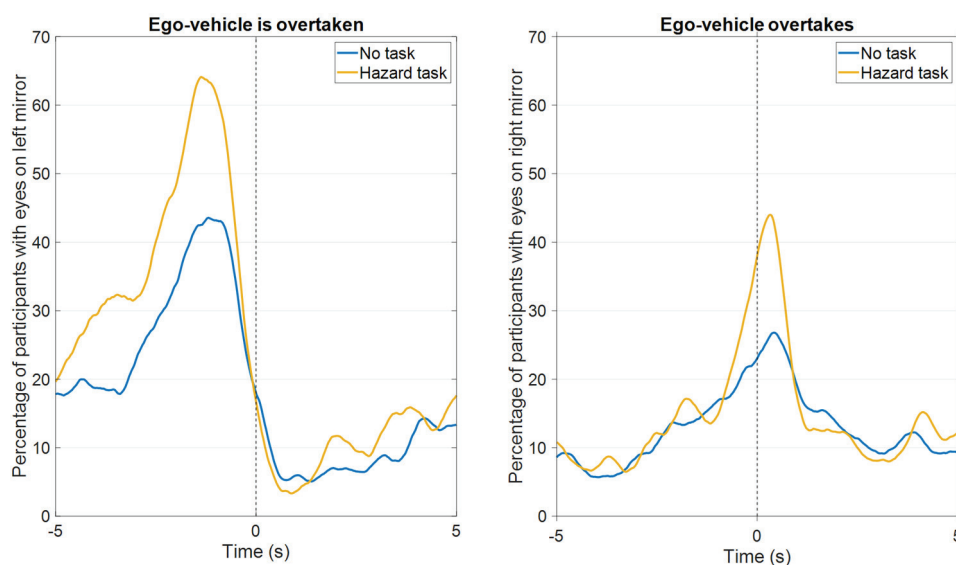


Figure 11. Percentage of participants with their eyes on the left mirror when another vehicle overtook the ego-vehicle in the left lane (left panel) and with their eyes on the right mirror when the ego-vehicle overtook another vehicle in the right lane (right panel) (Videos 1–3 combined, 8 and 15 overtakes, respectively). The vertical line at $t = 0$ s represents the first moment when part of the overtaking car became visible in the left side window (**left panel**) or the last moment when part of the overtaken car was visible in the front window (**right panel**).

4. Discussion

This study examined participants' viewing behavior when presented with videos depicting a simulated automated car. The use of videos enabled us to evaluate participants' viewing behavior for three different task instructions under identical traffic conditions for all participants.

The results indicated that viewing behavior in driving is strongly top-down. More precisely, top-down attention refers to attention that is governed by task-related expectancies and the value of task components (i.e., expected value), as opposed to bottom-up attention, which is governed by salient stimuli and head/eye-movement effort [36]. In particular, the task given to drivers determined their attention distribution (speedometer, hazards), and the drivers' eyes were directed towards task-relevant elements (e.g., overtaking cars) in a highly time-varying manner.

A relatively novel aspect of our work is that it focused on eye movements in relation to tasks and events in the environment. In comparison, research about eye movements in driving seems most developed in regard to curve driving on relatively uneventful roads (see [37] for a review). For example, there is research that reports that on curved roads, drivers focus on the tangent point [38,39], a future path [40,41], or the vanishing point [42]. That said, there is a growing body of research that examines eye movements in relation to the driving context [43] and specific driving events, such as overtaking and lane-changing [44,45], approaching road crossings and roundabouts [46], processing feedback and advice on the dashboard [47], responding to warnings [48,49], and reclaiming control from an automated vehicle (e.g., [50–52]). Especially noteworthy is research by Kircher and Ahlstrom [53], which provide similar figures as our Figure 11 but obtained in real traffic with the help of manual annotations of driving maneuvers. The participants' eye movements in our study seemed aimed at *predicting* the behaviors of other road users, a finding consistent with earlier hazard perception research (e.g., [54]). For example, viewing behavior was forward-facing when events occurred that could be relevant to the driver, such as vehicles approaching or possibly planning to perform a lateral maneuver.

The present study also revealed indications of task prioritization. For example, participants were more inclined to look at less critical elements such as the speedometer when

no hazards were present in the vicinity. In the Speed Task, participants looked at the speedometer considerably more often than in the Baseline task, but still in a functional manner—that is, attention seemed to be focused on the speedometer when there was a high probability that the threshold speed would be exceeded. This latter finding is in line with Senders’s [55] notion of conditional sampling and with Eisma et al. [56], who observed similar results when participants had to detect threshold crossings while watching an instrument panel of moving dials.

The current study revealed a congruence between environmental events and attention distribution. Future research could use more formal models, such as the Salience, Effort, Expectancy, Value (SEEV) model, which has been used before in various transportation research studies. Horrey et al. [57], for example, studied the SEEV model in a driving simulator, where participants had to maintain lane position while performing an in-vehicle technology task. Their results showed a high correlation between predicted and observed percentage dwell time. Other studies also showed high correlations between SEEV-predictive and observed dwell time percentages in various types of human–machine interaction tasks [58–63].

A limitation of this work is the small horizontal field of view of 30 degrees, which prevents the use of peripheral vision—an important component of driving [64,65]. In comparison, the human visual field is about 200 degrees [66], meaning that drivers could infer moving objects from, e.g., the side windows [67]. The participants also could not turn their heads—something that drivers normally do [68]. At the same time, there is ample research that claims valid results for movie clips. For example, there are studies on hazard perception (with and without eye-tracking) that use a small field of view and provide discriminative validity between beginner and expert drivers [69,70]. Additionally, our findings for lane changes resemble the findings of an on-road study by Kircher and Ahlstrom [53]. More classically, Hughes and Cole [71] claimed that “visual information presented by the movie film is sufficient to generate attentive processes characteristic of driving” (p. 377).

The use of short 2-min highway-driving trips only and the lack of a rear-view mirror are other limitations of the present study. Additionally, the participants were young engineering students, who can be expected to have fast response times and strong visual-spatial ability [72,73]. The current work could be replicated with older drivers. Finally, no sound, motion, or tactile cues were produced during the experiment. Future research into predicting eye movements in driving should be conducted in real vehicles (e.g., [43,74]).

What do the current findings mean for the development of DAMS? One interpretation of the results is that drivers’ attention should be distributed. For example, a DAMS could assess whether drivers look in the left or right mirrors when another car is passing or when they pass another car. This observation corresponds to earlier work on a gaze-based DAMS that was built on the assumption “that glances to the mirror and the speedometer are necessary for safe driving” [75] (p. 967). However, their DAMS was not contextualized; that is, it did not consider whether meaningful objects were present in the front view or mirrors. The present results may hold promise for object recognition in combination with eye-tracking. It is expected that future cars and trucks will be equipped with camera systems as a replacement for exterior mirrors in cars and trucks [76]. Object recognition on these camera images (such as using YoloV4 in this study) may prove useful input for DAMS.

Author Contributions: Conceptualization, Y.B.E., D.J.E. and J.C.F.d.W.; data curation, J.C.F.d.W. and Y.B.E.; formal analysis, J.C.F.d.W. and D.J.E.; funding acquisition, J.C.F.d.W.; investigation, D.J.E.; methodology, Y.B.E., D.J.E. and J.C.F.d.W.; resources, Y.B.E., D.J.E. and J.C.F.d.W.; software, Y.B.E. and D.J.E.; supervision, Y.B.E. and J.C.F.d.W.; validation, J.C.F.d.W.; visualization, J.C.F.d.W.; writing—original draft preparation, J.C.F.d.W., Y.B.E. and D.J.E.; writing—review and editing, J.C.F.d.W. and Y.B.E. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki, and approved by the Human Research Ethics Committee of the Delft University of Technology (approval number 1479, 22 March 2021).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: Raw data, videos, and a MATLAB script that reproduces all figures can be found at <https://doi.org/10.4121/20254071> (accessed on 8 May 2022).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. European Commission. Regulation (EU) 2019/2144. Type-approval requirements for motor vehicles and their trailers 2019, and systems, components and separate technical units intended for such vehicles, as regards their general safety and the protection of vehicle occupants and vulnerable road users. *Off. J. Eur. Union* **2019**, *L 325*, 1–40.
2. Cabrall, C.D.D.; Happee, R.; De Winter, J.C.F. From Mackworth’s clock to the open road: A literature review on driver vigilance task operationalization. *Transp. Res. Part F Traffic Psychol. Behav.* **2016**, *40*, 169–189. [CrossRef]
3. Hecht, T.; Feldhütter, A.; Radlmayr, J.; Nakano, Y.; Miki, Y.; Henle, C.; Bengler, K. A review of driver state monitoring systems in the context of automated driving. In Proceedings of the 20th Congress of the International Ergonomics Association, Florence, Italy, 26–30 August 2018; Bagnara, S., Tartaglia, R., Albolino, S., Alexander, T., Fujita, Y., Eds.; Springer: Cham, Switzerland, 2018; pp. 398–408. [CrossRef]
4. Cabrall, C.D.D.; Janssen, N.; De Winter, J.C.F. Adaptive automation: Automatically (dis)engaging automation during visually distracted driving. *PeerJ. Comput. Sci.* **2018**, *4*, e166. [CrossRef] [PubMed]
5. Cadillac. CT6 (Owner’s Manual). 2020. Available online: <https://www.cadillac.com/content/dam/cadillac/na/us/english/index/ownership/technology/supercruise/pdfs/2020-cad-ct6-owners-manual.pdf> (accessed on 1 May 2022).
6. Tesla. About Autopilot. 2022. Available online: https://www.tesla.com/ownersmanual/models/en_us/GUID-EDA77281-42DC-4618-98A9-CC62378E0EC2.html (accessed on 17 June 2022).
7. Volvo. Pilot Assist. 2020. Available online: <https://www.volvocars.com/en-th/support/manuals/v60/2018w46/driver-support/pilot-assist/pilot-assist> (accessed on 1 May 2022).
8. Lappi, O.; Rinkkala, P.; Pekkanen, J. Systematic observation of an expert driver’s gaze strategy—An on-road case study. *Front. Psychol.* **2017**, *8*, 620. [CrossRef] [PubMed]
9. Garay-Vega, L.; Fisher, A.L. Can novice drivers recognize foreshadowing risks as easily as experienced drivers? In Proceedings of the Third Driving Assessment Conference, Rockport, ME, USA, 27–30 June 2005; Volume 3, pp. 471–477. [CrossRef]
10. Malone, S.; Brünken, R. The role of ecological validity in hazard perception assessment. *Transp. Res. Part F Traffic Psychol. Behav.* **2016**, *40*, 91–103. [CrossRef]
11. Falkmer, T.; Gregersen, N.P. A comparison of eye movement behavior of inexperienced and experienced drivers in real traffic environments. *Optom. Vis. Sci.* **2005**, *82*, 732–739. [CrossRef] [PubMed]
12. Mourant, R.R.; Rockwell, T.H. Strategies of visual search by novice and experienced drivers. *Hum. Factors* **1972**, *14*, 325–335. [CrossRef]
13. Underwood, G.; Crundall, D.; Chapman, P. Driving simulator validation with hazard perception. *Transp. Res. Part F Traffic Psychol. Behav.* **2011**, *14*, 435–446. [CrossRef]
14. Van Leeuwen, P.M.; Happee, R.; De Winter, J.C.F. Changes of driving performance and gaze behavior of novice drivers during a 30-min simulator-based training. *Procedia Manuf.* **2015**, *3*, 3325–3332. [CrossRef]
15. Cerezuela, G.P.; Tejero, P.; Chóliz, M.; Chisvert, M.; Monteagudo, M.J. Wertheim’s hypothesis on ‘highway hypnosis’: Empirical evidence from a study on motorway and conventional road driving. *Accid. Anal. Prev.* **2004**, *36*, 1045–1054. [CrossRef] [PubMed]
16. Wertheim, A.H. Explaining highway hypnosis: Experimental evidence for the role of eye movements. *Accid. Anal. Prev.* **1978**, *10*, 111–129. [CrossRef]
17. Herslund, M.B.; Jørgensen, N.O. Looked-but-failed-to-see-errors in traffic. *Accid. Anal. Prev.* **2003**, *35*, 885–891. [CrossRef]
18. Damböck, D.; Weißgerber, T.; Kienle, M.; Bengler, K. Requirements for cooperative vehicle guidance. In Proceedings of the 16th International IEEE Annual Conference on Intelligent Transportation Systems, Hague, The Netherlands, 6–9 October 2013; pp. 1656–1661. [CrossRef]
19. Goncalves, R.C.; Louw, T.L.; Quaresma, M.; Madigan, R.; Merat, N. The effect of motor control requirements on drivers’ eye-gaze pattern during automated driving. *Accid. Anal. Prev.* **2020**, *148*, 105788. [CrossRef]
20. Louw, T.; Merat, N. Are you in the loop? Using gaze dispersion to understand driver visual attention during vehicle automation. *Transp. Res. Part C Emerg. Technol.* **2017**, *76*, 35–50. [CrossRef]
21. Mackenzie, A.K.; Harris, J.M. Eye movements and hazard perception in active and passive driving. *Vis. Cogn.* **2015**, *23*, 736–757. [CrossRef]
22. Miyajima, C.; Yamazaki, S.; Bando, T.; Hitomi, K.; Terai, H.; Okuda, H.; Hirayama, T.; Egawa, M.; Suzuki, T.; Takeda, K. Analyzing driver gaze behavior and consistency of decision making during automated driving. In Proceedings of the 2015 IEEE Intelligent Vehicles Symposium, Seoul, Korea, 28 June 2015; pp. 1293–1298. [CrossRef]

23. Navarro, J.; Lappi, O.; Osiurak, F.; Hernout, E.; Gabaude, C.; Reynaud, E. Dynamic scan paths investigations under manual and highly automated driving. *Sci. Rep.* **2021**, *11*, 3776. [CrossRef]
24. Lu, Z.; Happee, R.; de Winter, J.C. Take over! A video-clip study measuring attention, situation awareness, and decision-making in the face of an impending hazard. *Transp. Res. Part F Traffic Psychol. Behav.* **2020**, *72*, 211–225. [CrossRef]
25. Gold, C.; Körber, M.; Lechner, D.; Bengler, K. Taking over control from highly automated vehicles in complex traffic situations: The role of traffic density. *Hum. Factors* **2016**, *58*, 642–652. [CrossRef] [PubMed]
26. Yarbus, A.J. *Eye Movements and Vision*; Plenum Press: New York, NY, USA, 1967.
27. Borji, A.; Itti, L. Defending Yarbus: Eye movements reveal observers' task. *J. Vis.* **2014**, *14*, 29. [CrossRef]
28. DeAngelus, M.; Pelz, J.B. Top-down control of eye movements: Yarbus revisited. *Vis. Cogn.* **2009**, *17*, 790–811. [CrossRef]
29. Greene, M.R.; Liu, T.; Wolfe, J.M. Reconsidering Yarbus: A failure to predict observers' task from eye movement patterns. *Vis. Res.* **2012**, *62*, 1–8. [CrossRef]
30. Tatler, B.W.; Wade, N.J.; Kwan, H.; Findlay, J.M.; Velichkovsky, B.M. Yarbus, eye movements, and vision. *i-Perception* **2010**, *1*, 7–27. [CrossRef]
31. Pomarjanschi, L.; Dorr, M.; Rasche, C.; Barth, E. Safer driving with gaze guidance. In *Bio-Inspired Models of Network, Information, and Computing Systems*; Suzuki, J., Nakano, T., Eds.; Springer: Berlin/Heidelberg, Germany, 2010; pp. 581–586. [CrossRef]
32. Yeong, D.J.; Velasco-Hernandez, G.; Barry, J.; Walsh, J. Sensor and sensor fusion technology in autonomous vehicles: A review. *Sensors* **2021**, *21*, 2140. [CrossRef] [PubMed]
33. Large, D.R.; Crundall, E.; Burnett, G.; Harvey, C.; Konstantopoulos, P. Driving without wings: The effect of different digital mirror locations on the visual behaviour, performance and opinions of drivers. *Appl. Ergon.* **2016**, *55*, 138–148. [CrossRef] [PubMed]
34. Bochkovskiy, A.; Wang, C.Y.; Liao, H.Y.M. Yolov4: Optimal speed and accuracy of object detection. *arXiv* **2020**, arXiv:2004.10934.
35. sbairagy-MW. Pretrained-Deeplabv3plus. 2021. Available online: <https://github.com/matlab-deep-learning/pretrained-deeplabv3plus> (accessed on 1 May 2022).
36. Wickens, C.D.; McCarley, J.S. Visual attention control, scanning, and information sampling. In *Applied Attention Theory*; CRC Press: Boca Raton, FL, USA, 2008; pp. 41–61.
37. Lappi, O. Gaze strategies in driving—An ecological approach. *Front. Psychol.* **2022**, *13*, 821440. [CrossRef] [PubMed]
38. Kandil, F.I.; Rotter, A.; Lappe, M. Driving is smoother and more stable when using the tangent point. *J. Vis.* **2009**, *9*, 11. [CrossRef]
39. Land, M.F.; Lee, D.N. Where we look when we steer. *Nature* **1994**, *369*, 742–744. [CrossRef]
40. Wann, J.P.; Swapp, D.K. Why you should look where you are going. *Nat. Neurosci.* **2000**, *3*, 647–648. [CrossRef]
41. Wilkie, R.M.; Kountouriotis, G.K.; Merat, N.; Wann, J.P. Using vision to control locomotion: Looking where you want to go. *Exp. Brain Res.* **2010**, *204*, 539–547. [CrossRef]
42. Deng, T.; Yang, K.; Li, Y.; Yan, H. Where does the driver look? Top-down-based saliency detection in a traffic driving environment. *IEEE Trans. Intell. Transp. Syst.* **2016**, *17*, 2051–2062. [CrossRef]
43. Palazzi, A.; Abati, D.; Solera, F.; Cucchiara, R. Predicting the driver's focus of attention: The DR (eye) VE project. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *41*, 1720–1733. [CrossRef]
44. Lethaus, F.; Rataj, J. Do eye movements reflect driving manoeuvres? *IET Intell. Transp. Syst.* **2007**, *1*, 199–204. [CrossRef]
45. Martin, S.; Trivedi, M.M. Gaze fixations and dynamics for behavior modeling and prediction of on-road driving maneuvers. In *Proceedings of the 2017 IEEE Intelligent Vehicles Symposium, Los Angeles, CA, USA, 11–14 June 2017*; pp. 1541–1545. [CrossRef]
46. Abbasi, J.A.; Mullins, D.; Ringelstein, N.; Reilhac, P.; Jones, E.; Glavin, M. An analysis of driver gaze behaviour at roundabouts. *IEEE Trans. Intell. Transp. Syst.* **2021**. [CrossRef]
47. Ahlstrom, C.; Kircher, K. Changes in glance behaviour when using a visual eco-driving system—A field study. *Appl. Ergon.* **2017**, *58*, 414–423. [CrossRef]
48. Lehtonen, E.; Malhotra, N.; Starkey, N.J.; Charlton, S.G. Speedometer monitoring when driving with a speed warning system. *Eur. Transp. Res. Rev.* **2020**, *12*, 16. [CrossRef]
49. Morando, A.; Victor, T.; Dozza, M. Drivers anticipate lead-vehicle conflicts during automated longitudinal control: Sensory cues capture driver attention and promote appropriate and timely responses. *Accid. Anal. Prev.* **2016**, *97*, 206–219. [CrossRef]
50. Louw, T.; Kuo, J.; Romano, R.; Radhakrishnan, V.; Lenné, M.G.; Merat, N. Engaging in NDRTs affects drivers' responses and glance patterns after silent automation failures. *Transp. Res. Part F Traffic Psychol. Behav.* **2019**, *62*, 870–882. [CrossRef]
51. Petermeijer, S.M.; Cieler, S.; De Winter, J.C.F. Comparing spatially static and dynamic vibrotactile take-over requests in the driver seat. *Accid. Anal. Prev.* **2017**, *99*, 218–227. [CrossRef]
52. Goncalves, R.C.; Louw, T.L.; Madigan, R.; Quaresma, M.; Romano, R.; Merat, N. The effect of information from dash-based human-machine interfaces on drivers' gaze patterns and lane-change manoeuvres after conditionally automated driving. *Accid. Anal. Prev.* **2022**, *174*, 106726. [CrossRef]
53. Kircher, K.; Ahlstrom, C. Evaluation of methods for the assessment of attention while driving. *Accid. Anal. Prev.* **2018**, *114*, 40–47. [CrossRef] [PubMed]
54. Lee, S.E.; Klauer, S.G.; Olsen, E.C.; Simons-Morton, B.G.; Dingus, T.A.; Ramsey, D.J.; Ouimet, M.C. Detection of road hazards by novice teen and experienced adult drivers. *Transp. Res. Rec.* **2008**, *2078*, 26–32. [CrossRef] [PubMed]
55. Senders, J.W. Visual Sampling Processes. Ph.D. Thesis, Catholic University, Tilburg, The Netherlands, 1983.
56. Eisma, Y.B.; Cabral, C.D.D.; De Winter, J.C.F. Visual sampling processes revisited: Replicating and extending Senders (1983) using modern eye-tracking equipment. *IEEE Trans. Hum. Mach. Syst.* **2018**, *48*, 526–540. [CrossRef]

57. Horrey, W.J.; Wickens, C.D.; Consalus, K.P. Modeling drivers' visual attention allocation while interacting with in-vehicle technologies. *J. Exp. Psychol. Appl.* **2006**, *12*, 67–78. [CrossRef]
58. Bos, A.J.; Ruscio, D.; Cassavaugh, N.D.; Lach, J.; Gunaratne, P.; Backs, R.W. Comparison of novice and experienced drivers using the SEEV model to predict attention allocation at intersections during simulated driving. In Proceedings of the Eighth International Driving Symposium on Human Factors in Driver Assessment, Training and Vehicle Design, Salt Lake City, UT, USA, 22–25 June 2015; pp. 120–126.
59. Cassavaugh, N.D.; Bos, A.; McDonald, C.; Gunaratne, P.; Backs, R.W. Assessment of the SEEV model to predict attention allocation at intersections during simulated driving. In Proceedings of the Seventh International Driving Symposium on Human Factors in Driver Assessment, Training and Vehicle Design, Bolton Landing, NY, USA, 17–20 June 2013; pp. 334–340.
60. Steelman-Allen, K.S.; McCarley, J.S.; Wickens, C.; Sebok, A.; Bzostek, J. N-SEEV: A computational model of attention and noticing. In Proceedings of the Human Factors and Ergonomics Society Annual Meeting, San Antonio, TX, USA, 19–23 October 2009; Volume 53, pp. 774–778. [CrossRef]
61. Wickens, C.D.; Helleberg, J.; Goh, J.; Xu, X.; Horrey, W.J. *Pilot Task Management: Testing an Attentional Expected Value Model of Visual Scanning*; Technical Report No. ARL-01-14/NASA-01-7; Aviation Research Lab, Institute of Aviation: Savoy, IL, USA, 2001.
62. Wickens, C.D.; Sebok, A.; Bagnall, T.; Kamienski, J. Modeling of situation awareness supported by advanced flight deck displays. In Proceedings of the Human Factors and Ergonomics Society Annual Meeting, Baltimore, MD, USA, 1–5 October 2007; Volume 51, pp. 794–798. [CrossRef]
63. Wortelen, B.; Lüdtke, A.; Baumann, M. Simulating attention distribution of a driver model: How to relate expectancy and task value? In Proceedings of the 12th International Conference on Cognitive Modeling, Ottawa, ON, Canada, 11–14 July 2013.
64. Kotseruba, I.; Tsotsos, J.K. Behavioral research and practical models of drivers' attention. *arXiv* **2021**, arXiv:2104.05677.
65. Wolfe, B.; Dobres, J.; Rosenholtz, R.; Reimer, B. More than the useful field: Considering peripheral vision in driving. *Appl. Ergon.* **2017**, *65*, 316–325. [CrossRef]
66. Strasburger, H. Seven myths on crowding and peripheral vision. *i-Perception* **2020**, *11*, 1–45. [CrossRef]
67. Bickerdt, J.; Wendland, H.; Geisler, D.; Sonnenberg, J.; Kasneci, E. Beyond the tracked line of sight—Evaluation of the peripheral usable field of view in a simulator setting. *J. Eye Mov. Res.* **2021**, *12*. [CrossRef]
68. Fridman, L.; Lee, J.; Reimer, B.; Victor, T. 'Owl' and 'Lizard': Patterns of head pose and eye pose in driver gaze classification. *IET Comput. Vis.* **2016**, *10*, 308–314. [CrossRef]
69. Scialfa, C.T.; Deschênes, M.C.; FERENCE, J.; Boone, J.; Horswill, M.S.; Wetton, M. A hazard perception test for novice drivers. *Accid. Anal. Prev.* **2011**, *43*, 204–208. [CrossRef]
70. Vlaskveld, W.P. A comparative study of two desktop hazard perception tasks suitable for mass testing in which scores are not based on response latencies. *Transp. Res. Part F Traffic Psychol. Behav.* **2014**, *22*, 218–231. [CrossRef]
71. Hughes, P.K.; Cole, B.L. What attracts attention when driving? *Ergonomics* **1986**, *29*, 377–391. [CrossRef] [PubMed]
72. Salthouse, T.A. When does age-related cognitive decline begin? *Neurobiol. Aging* **2009**, *30*, 507–514. [CrossRef] [PubMed]
73. Wai, J.; Lubinski, D.; Benbow, C.P. Spatial ability for STEM domains: Aligning over 50 years of cumulative psychological knowledge solidifies its importance. *J. Educ. Psychol.* **2009**, *101*, 817–835. [CrossRef]
74. Fridman, L.; Langhans, P.; Lee, J.; Reimer, B. Driver gaze region estimation without use of eye movement. *IEEE Intell. Syst.* **2016**, *31*, 49–56. [CrossRef]
75. Ahlstrom, C.; Kircher, K.; Kircher, A. A gaze-based driver distraction warning system and its effect on visual behavior. *IEEE Trans. Intell. Transp. Syst.* **2013**, *14*, 965–973. [CrossRef]
76. Schmidt, E.A.; Hoffmann, H.; Krautscheid, R.; Bierbach, M.; Frey, A.; Gail, J.; Lotz-Keens, C. Camera-monitor systems as a replacement for exterior mirrors in cars and trucks. In *Handbook of Camera Monitor Systems*; Terzis, A., Ed.; Springer: Cham, Switzerland, 2016; pp. 369–435. [CrossRef]

S.A.D.E.—A Standardized, Scenario-Based Method for the Real-Time Assessment of Driver Interaction with Partially Automated Driving Systems

Nadja Schömig ^{1,*}, Katharina Wiedemann ¹, André Wiggerich ² and Alexandra Neukum ¹

¹ Würzburg Institute for Traffic Sciences GmbH (WIVW), Robert-Bosch-Str. 4, 97209 Veitshoechheim, Germany

² Federal Highway Research Institute, 51427 Bergisch-Gladbach, Germany

* Correspondence: schoemig@wivw.de

Abstract: Vehicles equipped with so-called partially automated driving functions are becoming more and more common nowadays. The special feature of this automation level is that the driver is relieved of the execution of the lateral and longitudinal driving task, although they must still monitor the driving environment and the automated system. The method presented in this paper should enable the assessment of the usability and safety of such systems in a standardized manner. It is designed to capture a driver's interaction with a system via the human-machine interface in specific scenarios in user studies. It evaluates several observable aspects of this interaction in real time and codes inadequate behavior in the categories "system operation", "driving behavior" and "monitoring behavior". A generic rating regarding the overall handling of the scenario is derived from these criteria. The method can be used with the assistance of a tablet application called the S.A.D.E. app (Standardized Application for Automated Driving Evaluation). Initial studies using driving simulators show promising results regarding its ability to detect problems related to a system or HMI, with some future challenges remaining open.

Keywords: evaluation; HMI; tablet application; partially automated driving; Level 2 automation; driver monitoring; system operation; driving performance

1. Background

Vehicles equipped with so-called partially automated driving functions (so-called Level 2 systems or L2 systems; see SAE definition, 2021 [1]) are becoming more and more common nowadays. Such systems are able to execute parts of the operational driving task, i.e., lateral vehicle control via steering and longitudinal vehicle control via acceleration and deceleration. These features are dedicated to being used in specific operational design domains, such as highways; however, they are also available on other road types. L2 systems require immediate driver intervention when the system reaches its limits. In this case, the system either switches off without any explicit take-over request or it remains active because its sensors have not detected the reason for necessary driver intervention (e.g., an obstacle). Therefore, the driver must continue to perform the task of object and event detection while using such systems [1]. Most of the serial systems today are hands-on systems requiring drivers to leave their hands on the steering wheel. If such a system detects that the driver takes their hands off the steering wheel for a certain period of time, it will request the driver put their hands back via a warning.

Although L2 systems are designed to increase driving comfort, some specific problems might arise during their usage. While on the one hand drivers are partly released from continuous driving tasks (steering and keeping speed/distance), they still remain responsible for intervening at system limits. In highly reliable systems, this might require the detection of and the reaction to very rare events after long periods of inactivity while the driver was simply monitoring the road and the system. Such a task is very difficult for

human operators to handle (see Bainbridge's *Ironies of Automation*, 1980 [2]). In addition, an inadequate assumption about a system's abilities and limits, i.e., an inadequate mental model, might result in overtrust and potentially safety-critical situations in the case of a non-response or the delayed or inadequate response of the operator to a system limit. A prominent example is the brand Tesla labeling its L2 feature "Autopilot" which could be associated with comparably higher automation levels in the aircraft domain (see, for example, the Tesla crash with a truck in May 2016 in Florida [3]). However, more data from real roads and from crash databases are still missing, which are needed to estimate the real impact of such potential problems.

Empirical research on L2 systems in simulators, however, has been conducted extensively. Many studies show that drivers accept and trust such systems [4,5]. When using L2 systems in traffic jams, drivers feel relieved of the driving task. However, the monitoring task which still has to be performed is experienced as rather strenuous [6,7]. It has also been observed that drivers increasingly turn their attention to non-driving-related secondary activities when driving with such systems [8]. With regard to responsiveness at system limits, some studies suggest that drivers are able to react relatively well if the system limits are announced by warnings and the driver is fully attentive [5,9,10]. However, it is typical that L2 systems are not able to predict system limits and therefore they cannot be announced in advance. Reacting to such sudden, critical situations is far more difficult and leads to greater problems with the driver's reaction [7,11,12]. The better the HMI (human-machine interface) prepares the driver for such situations and assists the driver in responding to them, the more likely it is that safety-critical consequences can be avoided [13].

In order to objectively evaluate the effects of such driver assistance systems and their HMI design in terms of usability and driving safety as efficiently as possible and in the most standardized way, suitable methods are needed that can be optimally used as a kind of "live" rating during interaction with the system. In contrast to L2 systems, some approaches were already developed for L3 systems (conditionally automated driving), e.g., by Naujoks and colleagues. They developed a method for the expert assessment of the controllability of take-over situations in L3 driving, the so-called TOC-Rating [14]. Raters are trained to use a coding sheet which includes several observation criteria to generate a global controllability rating from a video of the take-over situation. However, this method is very detailed as it is dedicated for use in subsequent data analyses and focuses exclusively on controllability but not on the usability aspects of L3 HMIs. Another method developed by [15] deals with the verification of minimum design requirements for L3 system HMIs as proposed by the National Highway Traffic Safety Administration (NHTSA, [16]). This test procedure contains a set of relevant use cases that should be tested as well as a set of evaluation criteria based on drivers' observations in the intuitive interaction with the system and on drivers' subjective reports evaluating the understandability of system mode indicators for L3 systems.

For Level 2 systems, such tools for the assessment of usability and safety do not exist so far. A new method should be able to evaluate whether a driver understands L2 system functionality and the resulting responsibilities of the driving task, whether they adequately interact with the system and are able to react to system limits effectively and safely. This requires the observation of the direct interaction of a user with a system via the HMI. Most suitable for this approach are user studies with a sample of real users who have no prior experience. In order to evaluate these interaction-related aspects of a safe human-machine interaction with partially automated driving functions, the method should fulfill the following requirements, which were estimated as meaningful by the authors' team based on an evaluation of already existing methods:

- The method should be able to identify problems related to HMI aspects as well as to system functionality aspects;
- The method should allow both a global analysis as well as a very differentiated analysis of the interaction with the system in the investigated scenarios;

- The method should be adaptable to different test environments (e.g., driving simulation, test track or real-world driving);
- The method should allow a quick and efficient data assessment and analysis;
- The method should enable a standardized procedure with regard to the testing conditions, test scenarios and evaluation criteria;
- The method should allow an objective evaluation according to clear rules;
- The method should allow the assessment of driver interaction and experiences with the system directly in the situation, i.e., in real time;
- The method should be unobtrusive (i.e., it should not disturb the user).

The following chapter describes the development and application of such a standardized scenario-based method for the real-time assessment of driver interaction with partially automated driving systems called S.A.D.E. (Standardized Application for Automated Driving Evaluation) which is intended to be employed in user studies. This method was developed within a project funded by the German Federal Highway Research Institute (German BAST: Bundesanstalt für Straßenwesen) among other instruments for the evaluation of L2 HMIs (see [17] for a detailed overview of the project contents and outputs).

2. Method Description

2.1. Evaluation Criteria—Which Behavior Should Be Observed?

The first step of the method development was to define the criteria for the evaluation of the human–machine interaction of users with partially automated systems: The human–machine interface (HMI) must be able to communicate a correct understanding of the driver’s responsibilities while using the system and must be able to create an adequate awareness of the system status and the surrounding situation (mode and situation awareness). The driver must be able to efficiently use the system (which means the operation of the system with low mental and physical effort and without impairments in driving performance) and to adequately and safely handle situations in which they need to take over vehicle control. Aspects such as the efficiency (as part of usability) and safety of the interaction are therefore within the scope of evaluation. Collecting these aspects of objectively observable behavior allows the inference of underlying implicit aspects, such as comprehension of the system’s logic, the level of workload in system operation, comprehension of system outputs, the level of mode and situation awareness and knowledge about the driver’s responsibilities.

In order to evaluate the effort and resulting workload in system usage, the driver’s system operations must be assessed. This must include the observation of driver-initiated transitions between system modes (e.g., activation or deactivation of the system or its subfunctions). Errors, imprecisions or unnecessary actions during system usage point to a high complexity of the system logic or operational logic and potential difficulties in the comprehensibility of necessary operational steps to achieve a certain goal (e.g., in the case of activation, is the system already active or still in standby mode?).

In addition, the observation of the driver’s behavior in the case of system-initiated transitions at system limits or system failures helps to assess whether a driver has enough system and situation awareness to detect system limits and to adequately choose necessary interventions. This assessment can be achieved by observing whether the driver adequately deactivates the system or its subfunctions in a timely manner in cases of system limits or system failures or whether the driver adequately reacts by taking over control without impairments in driving safety. Driving errors or driver endangerment of themselves or other road users indicate that the driver has insufficient system and situation awareness and is therefore surprised and excessively challenged by system limits and failures.

While driving in certain system states (especially while driving with active L2 system) it must be determined whether the driver is aware of their responsibility to monitor the driving environment and the system. This can be observed in the degree to which the driver shows adequate monitoring behavior, both with regard to their visual attention

distribution for the observation of the surrounding traffic and for checking the system status as well as to the adequate motoric involvement in the driving task, i.e., keeping hands on the wheel and cooperating with the system on steering tasks if needed.

Taken together, it is necessary to observe a driver's system operations and driving behavior as well as driver monitoring behavior to obtain a comprehensive understanding of the assumed underlying implicit aspects named in the chapter above. In addition to the observation of the driver's behavior, the driver's subjective comprehension of the system and the HMI should also be measured via scenario-specific questions. This allows the assessment of subjective statements about the perceived level of certainty with regard to interaction with the system (e.g., knowing what to do, understanding what the system is indicating, understanding why the system behaves in a certain way). In contrast to a more generic system understanding, which can be assessed in a follow-up survey after testing, the driver is asked to evaluate specific HMI outputs and system behaviors directly in the respective scenarios.

2.1.1. Observation Categories of Driver Behavior

To achieve a standardized, objective and efficient observation, driver behavior is coded using three main categories:

- System operation;
- Driving behavior;
- Monitoring behavior.

To increase the efficiency of the observation, not all behavior is rated, but only inadequate behavior is coded and protocolled. The level or type of inadequateness is further divided into category-specific errors or problems. These error types were derived from the long-term experience of the authors' team from numerous studies about such systems (simulator studies and real driving studies with relation to the evaluation of take-over scenarios, see, for example, [14]) and resemble the most frequent problems with L2 systems. Error definitions are shown in Table 1.

Table 1. Error definitions for observation categories.

Category	Observable Error/Problem	Description/Example
System operation (especially driver-initiated operations)	Noticed nothing	Driver does not notice changes in system state
	Uncertain/delayed operation	Driver shows uncertainties in system operation, e.g., searches for a certain button; Driver takes a long time to perform an action
	Inadequate operation	Driver shows inadequate system operation, e.g., activates the system in situations where it should not be used
	Operation error	Driver initiates an incorrect operation, e.g., driver presses wrong button, driver presses correct button but not firmly enough, driver wants to activate the system when it is not possible/available
	Support by experimenter in operation required	Driver is not able to execute the expected action until a certain defined point in time so that the experimenter must give support to reach the designated system mode

Table 1. Cont.

Category	Observable Error/Problem	Description/Example
Driving behavior (especially relevant to driver-initiated system operations and system-initiated transitions)	No reaction	Driver does not show any reaction in a given situation which would require one
	Reaction delayed	Driver reacts to an event with a clear delay, e.g., by a braking maneuver
	Reaction too strong	Driver reacts too strongly to an event, e.g., with oversteering
	Lane exceedance	The vehicle crosses the lane marking with one wheel
	Poor lane keeping	The vehicle visibly swerves within the lane to the right and/or to the left
	Insufficient securing behavior	Driver does not execute a control glance in the mirror in the case of a lane change
	Endangerment	Safety distance below 1 s to the front/side or behind/towards other vehicles
	Collision	Vehicle collides with another traffic participant or a stationary obstacle
Monitoring behavior	Uncertainties in hands-on behavior	Driver shows clear uncertainties as to whether hands should be left on the steering wheel or not, takes them away repeatedly or rests them too weakly on the wheel
	Not attentive enough	Driver shows clear signs of inattention, e.g., no control glances to HMI for longer time intervals, direction of attention towards NDRT (non-driving-related task)
	Hands-off warning	The hands-off warning was triggered by the system
	Stage of hands-off warning	The maximum stage of the hands-off warning was reached within a scenario

2.1.2. Subjective Driver Evaluation

In addition to observational categories, drivers' subjective evaluations should be assessed in certain test scenarios (especially system- and driver-initiated transitions). With this intention, several interview questions are included in the method. The driver is asked to answer them directly after having experienced the transitions. The following questions are used to assess the subjective evaluation of the HMI by the driver:

- Comprehensibility of the required driver action: Does the driver know what to do in a certain situation, e.g., in order to activate the system, to deactivate it or to adequately react to a system limit?
- Understandability of system behavior: Does the driver understand why the system behaves in a certain way in a situation, e.g., when the lateral control is switched off?
- Comprehensibility of system outputs: Does the driver understand what the visual system indicators or acoustic signals mean?
- Perceived situation criticality: How critical does the driver perceive a certain situation as a result of the combination of the objective demands of the situation and the required reaction?

In addition, there might be further suitable questions depending on the specific study question that could be included in the test protocol, e.g., perceived safety when driving with the system or perceived system trust. In terms of participant response, a seven-point verbalized, bipolar rating scale is recommended with rating values from -3 to $+3$ which can be flexibly adapted to the specific questions. According to the literature review in [18], five to seven categories achieve the best measures with regard to reliability, validity and differentiation. In addition, this reduced number can be easily assessed in real time during the drive and easily learned. All values are verbalized to increase reliability and validity (according to [18]). The additional numerical values enhance the rapid answering of the questions during the drive. The scale should be mounted at a location in the vehicle where the drivers can assess its content with a short glance. For the question regarding the perceived criticality of the situation, the 11-point rating scale of [19] can be used.

2.1.3. Global Rating of Scenario Handling

In addition, the experimenter is requested to give a global rating per scenario, which describes how well the driver has handled the overall driving scenario (see Figure 1). The proposed rating scale was adapted from the so-called Fitness-to-Drive-Scale (FtD-Scale; [20,21], on the basis of [19]), which was originally used as a rating for driver fitness in manual driving. This scale and the categories were adapted to the context of automated driving for this test protocol.



Figure 1. Experimenter global rating of the driver's handling of a use case in a specified scenario. The cross marks an exemplary rating.

The experimenter should derive the rating from the observed errors or problems during interactions with the system in the specific test scenario. Table 2 summarizes the recommendations for global ratings dependent on the observed errors or problems. Depending on the type of scenario, different observations can lead to a classification of the rating into various categories. For example, a rating of 10 can be given if either the driver is not able to bring the system into a designated state without the support of the experimenter or if the driver caused an accident in the case of a system limit. This requires a separate interpretation of the rating per scenario type (e.g., system operation scenarios separated from take-over scenarios). The definitions of the severity of the errors are recommendations regarding which category the experimenter should assign an observed behavior. When refining the rating within a certain category, the rater should take the consequences of the erroneous behavior into account.

Table 2. Recommendations for global ratings depending on the observed errors or problems.

Error/Problem	Verbal Category	Numeric Category
- System operation: no successful operation, even after support by the experimenter - Driving behavior: occurrence of a collision with another vehicle or obstacle - Monitoring behavior: system switching off due to repeated non-reaction to hands-off warnings (last warning stage of the system)	Scenario not handled successfully	10

Table 2. Cont.

Error/Problem	Verbal Category	Numeric Category
System operation: driver did not notice that a transition to a lower level of automation was required; support required, resulting in successful operation	Not acceptable problems	9
Driving behavior: no or delayed reaction with the consequence of endangerment of self or other road users		8
Monitoring behavior: more than one hands-off warning or hands-off warning at stage 2		7
System operation: operation error or inadequate operation	Error-prone, but acceptable	6
Driving behavior: delayed or too strong reaction, e.g., with regard to braking, lane exceedance; insufficient securing		5
Monitoring behavior: clear signs of inattentive behavior, at least one hands-off warning at stage 1		4
System operation: uncertain/delayed operation	good	3
Driving behavior: poor lane keeping		2
Monitoring behavior: uncertainties in hands-on behavior		1
No errors/problems	perfect	0

2.1.4. Rater Training

To ensure that the method achieves high objectivity, reliability and validity, training the users on the tool beforehand in the application of the method is recommended. This should include a description of the defined categories, some examples of behaviors that should be rated as errors or problems in the different categories and some rules for the achievement of the global rating. If several raters are to be deployed within a study, a discussion round should be included before the start of the study to clarify unclear definitions and to gather a common understanding of the categories and the coded behaviors. The result should be sufficiently high interrater reliability. Training material in the form of a PowerPoint presentation is available on request from the authors' team.

2.2. Definition of Relevant Test Scenarios

The following system states and transitions between these states were selected as the most relevant states while driving with L2 systems by the authors' team. They should be used as relevant test scenarios where the driver's interaction with the system should be observed based on the above defined categories in order to assess the HMI's usability and safety:

- System activation by the driver;
- System deactivation by the driver;
- Longer driving with active L2 system;
- Driver-initiated lane change;
- Temporary standby mode of lateral control;
- System limit and/or system malfunction (system limits can be both detectable and predictable as well as not detectable and not predictable; e.g., in longitudinal control: sensors are not able to detect a stationary vehicle or any other obstacle; in lateral control: system is not able to apply the necessary steering torque to manage a situation, e.g., a sharp bend).

These more general use cases must be transferred into testable scenarios for evaluation in a driving simulator or on a test track. Driver-initiated activation and deactivation processes as well as longer phases of driving with an active L2 system and lane changes can simply be instructed by the experimenter. Depending on the specific system, a temporary standby mode of the lateral control can be achieved by the absence of lane markings. Longitudinal system limits can be realized by placing an obstacle in the ego-vehicle's lane. Realistic lateral system limits that are not manageable by the system could be construction sites or sharp bends during real driving conditions.

2.3. Implementation of the Method in a Tablet App

The assessment method described can be used as a paper-pencil tool. In order to economize the process of data entry, a tablet application called S.A.D.E. (Standardized Application for Automated Driving Evaluation; see Figure 2) was developed. The app was programmed for Windows tablets with WIVW simulation software SILAB® (v6.0). The experimenter is able to protocol the occurrence of defined errors in system operation, driving and monitoring behavior via the tablet separately for each scenario in real time during the scenario (this is performed in the left area of the app). In addition, it is possible to rate the driver's handling of a scenario globally based on the observed errors (performed in the lower area of the app) and to conduct a standardized interview during the driving course (right area of the app). The upper control buttons on the right side can be used to start, pause and stop the evaluation and to switch to previous or subsequent test scenarios. If the application is used within driving simulator studies programmed with the simulation software SILAB®, it is possible to automatically count some of the errors (e.g., endangerments, collisions, lane exceedances and hands-off warning events). The application additionally integrates specific features such as acoustic feedback at the start of a new scenario or the upcoming end of the current scenario, a visual enhancement of relevant observation categories or survey questions for a specific scenario. The data are recorded and collected in a data sheet that also includes the data of the participant and study-related information and that can be easily imported into established statistical analyzing tools.

S.A.D.E. is active / scenario ID 11		Scenario 1_Activation (long. and lat)				▶ ■ X	
S.A.D.E. is active / scenario ID 11	noticed nothing	uncertain/delayed	inadequate		Instruction: Please activate the complete system, i.e. lateral and longitudinal control!		
	operation error	support required					
	no reaction	reaction delayed	reaction too strong	poor lane keeping	1. How clear was it what do to? ? -3 -2 X 0 1 2 3		
	lane exceedance	insufficient securing	endangerment	collision	2. How clear was system's reaction? X -3 -2 -1 0 1 2 3		
	uncertain hands-on	not attentive	hands-off warning	Stage of HOR	3. How clear were the system outputs? X -3 -2 -1 0 1 2 3		
					4. How safe did you feel during the ride? X -3 -2 -1 0 1 2 3		
evaluation of scenario handling					5. How critical was the situation? ? 0 1 2 3 X 5 6 7 8 9 10 X 0 1 2 3 4 5 6 7 8 9 10		

Figure 2. Tablet surface of the S.A.D.E. application.

2.4. Application of the Method

In simulator studies, the S.A.D.E. app should be used in the following way: After starting the test course, S.A.D.E. automatically starts (when using SILAB[®]). The experimenter can enter data that are specific to the participant (e.g., demographics) and the test trial. After that, the driver completes the test course in a predefined sequence of scenarios. During the respective scenarios, such as a system limit or a certain instruction such as to activate the L2 system, the experimenter observes driver behavior and directly protocols the observed errors or problems defined in the app. In doing so, the experimenter must have the opportunity to monitor the driver's system operations at the steering wheel as well as the driver's glance behavior. In the best instance, two cameras are fixed at proper positions in the vehicle mockup to record the driver's face and the steering wheel with the control elements as well as the driver's hand position. If possible, the driving scenery should be observable on a monitor from a bird's eye view for the experimenter to assess the driving performance and lane-keeping quality. After having recorded the observed driving errors and problems, the driver can be asked the predefined questions. After the scenario, the experimenter is expected to give the global rating of the driver handling the scenario. It is recommended to extensively train the experimenters of a user study beforehand in the usage of the application.

If the app is used in a simulator study, it is possible to make entries into the tablet only during the current scenario. After reaching the end of a scenario, the next scenario starts, all inputs are nulled and new entries can be made. Therefore, it is important that the length of a scenario is designed in a way so that the conduction of the observation and the survey are smoothly possible (i.e., there should be enough time between subsequent scenarios or simulation should be paused after each scenario). If the experimenter does not manage to enter all observations or answers from the driver, they can pause the app at the end and go through all scenarios again to make corrections before they finally stop the application.

3. Results of an Explorative Study in the WIVW Driving Simulator

An initial explorative study to test the feasibility of the developed method was performed in the motion-based driving simulator of the WIVW GmbH. The goal of the study was to evaluate whether HMI design aspects defined as possibly problematic within a previously conducted expert evaluation (using a checklist method including guidelines about the design of the human-machine interface of L2 automated vehicles, see [17]) influence real users' observable behavior and their subjective experience during interaction with the L2 system assessed by the S.A.D.E. app.

Two HMI variants A and B were investigated as part of the study. Their HMI design differed in several dimensions:

- System operation: operation logic regarding the activation of the longitudinal vehicle control (one-step vs. two-step activation);
- Control elements: labeling consistent vs. not consistent with the user manual;
- Visual indicators for active lateral vehicle control: with vs. without additional symbol of a steering wheel and text;
- Visual contrast: high vs. low contrast between foreground and background;
- Warning concept in situations with predictable system limits: presence vs. absence of a visual and acoustic warning.

These dimensions were derived from an expert checklist including several generic design guidelines derived from the literature (for more information about this checklist and its development, see [17]). HMI variant A was designed to achieve a high compliance with these guidelines (therefore named "highly compliant" HMI variant A), whereas HMI variant B was designed to achieve a low compliance with these guidelines (therefore named "low-compliant" variant B). Figure 3 shows three exemplary system outputs and how they differed between variant A and B in various scenarios.

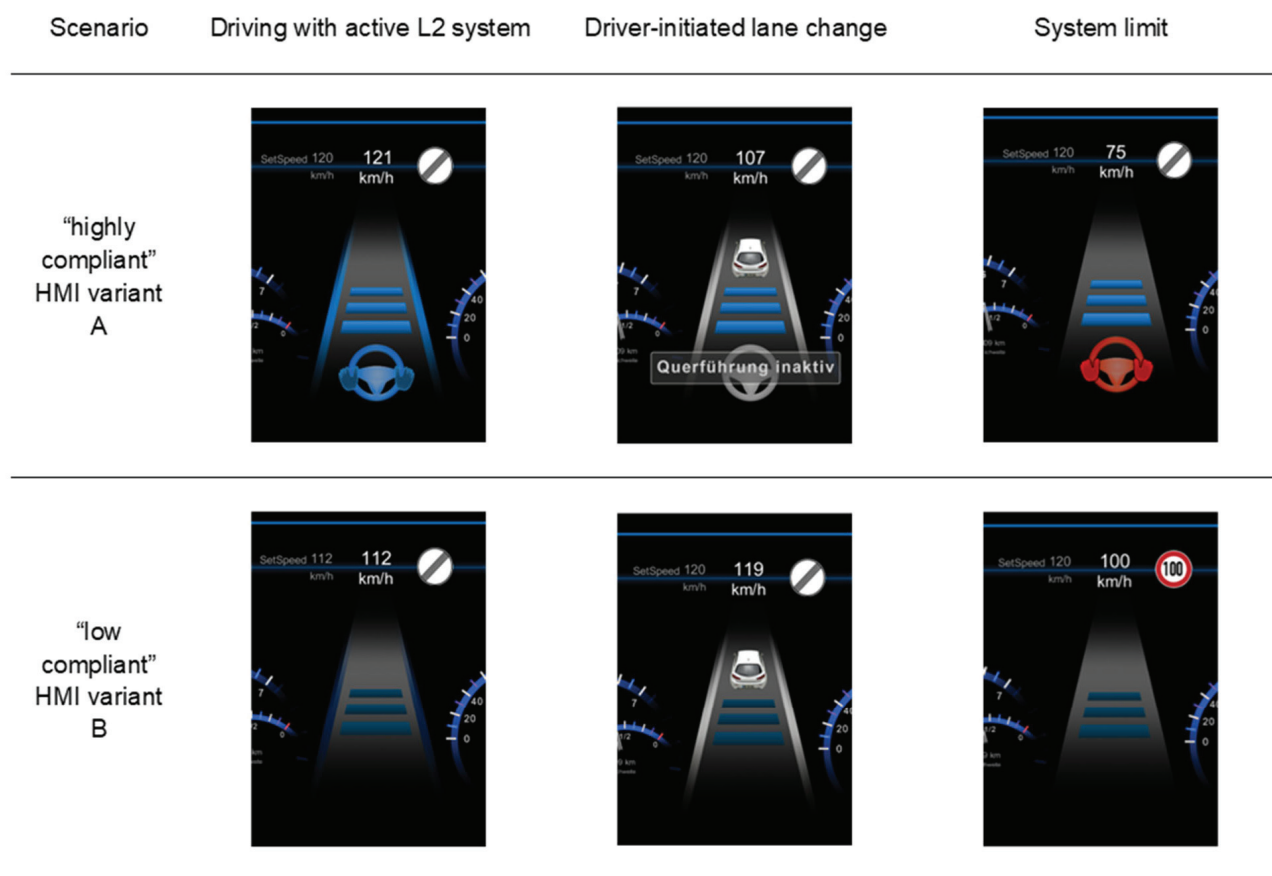


Figure 3. Exemplary scenarios in which the HMI variants differed with regard to system outputs.

The test course in the driving simulation consisted of different driving scenarios specified in the chapter “Definition of Relevant Test Scenarios”. The test course was a two-lane highway with moderate traffic and a speed limit of 120 km/h. The participants were instructed to drive in the right lane. A sharp bend as well as a stationary vehicle in the ego-vehicle’s lane (scenario “obstacle”) served as critical use cases reaching the system limits of the L2 system. In the first scenario, “sharp bend”, the necessary steering torque to manage the strong curvature could not be fully applied by the L2 system, resulting in a deactivation of lateral control. If the driver did not intervene timely enough, the vehicle would leave the lane towards the left and finally collide with the guardrail. The second scenario, “obstacle”, consisted of a broken-down vehicle placed behind a hilltop on the right lane. The situation is announced to the participant by a reflective triangle in front of the hilltop (as in the case of a broken-down vehicle in real life). As the simulated sensors of the L2 system were not able to detect this stationary object, the L2 system would collide with the broken-down vehicle if the driver did not intervene by braking and/or steering. The test course took about 20 min to complete.

N = 24 drivers (11 female) recruited from the WIVW test driver panel participated in the study. The subjects were on average 41.2 years old (SD = 14.3 years). The experimental factor HMI-variant (A vs. B) varied between the subjects so that half of the drivers experienced HMI variant A, while the other half experienced variant B. Both groups had equal preconditions regarding their prior experience with assistance systems and automated driving systems from other simulator studies that the participants took part in. They did not differ in terms of driver age and driving experience.

All the test drivers experienced the following procedure: During a short manual drive, the drivers were able to get used to driving in the simulator again. After that, they read a short user manual which was adapted to the two HMI variants in terms of system operation and HMI displays. Afterwards, drivers received general instructions regarding the simulator drive and the usage of the rating scales. This was followed by the 20 min test drive. During the test drive, the participants experienced the different test scenarios and answered the survey questions directly after each scenario. The experimenter recorded the problems observed during system operation, driving behavior and monitoring behavior using the tablet app. A post-survey questionnaire (not included in the app) after the test drive contained questions on various HMI design aspects. In total, the study took approximately one hour. In addition to the observed behavioral data from the tablet, continuous driving data from the simulation software were analyzed for the two critical scenarios (minimum time-to-collision TTC to the broken-down vehicle in the scenario “obstacle”, and maximum deviation from lane center during the event “sharp bend”). Only one rater evaluated all the trips, meaning that no interrater reliability could be calculated.

In brief, the study showed the following results:

- The designed differences in the two HMI variants affected drivers’ behavior and subjective experiences of the system assessed via the S.A.D.E. app only to some extent. For most of the analyses, only a tendency towards the significant effects of the HMI variant was found. On the one hand, this could be due to the small sample size. On the other hand, it could be possible that some design issues did not affect driver behavior in such a significant way that real problems occurred which could be detected by the tool.
- The following results were observed on the level of single observational categories (effects with p -values < 0.15 are defined as tendentially significant, effects with p -values < 0.05 are defined as significant):
 - Effect of different warning strategies (with vs. without visual–acoustic warning) in scenario “sharp bend”: In HMI variant B (without the warning), a tendency towards worse lane-keeping behavior was observed (i.e., a higher frequency of problems in the category “driving behavior” was coded; $p = 0.132$). The subjective evaluation of the drivers revealed greater problems in system understanding for HMI variant B ($p = 0.000$).
 - Effect of differences in system operation in scenario “first system activation” and scenario “deactivation”: The more complex system activation and deactivation in HMI variant B resulted in a higher frequency of problems in the category “system operation” for HMI variant B (especially more frequently coded events in the category “support required by the experimenter” in scenario “first system activation”: $p = 0.019$; a tendency towards this effect in scenario “system deactivation”: $p = 0.140$). In addition, there was a tendency towards a higher perceived subjective difficulty for system activation in variant B ($p = 0.061$).
 - Effect of differences in visual contrast and visual indicators for the active lateral vehicle control subtask in the scenario “standby mode of lateral control”: The lower distinctiveness of system states in HMI variant B did not lead to observable differences in driving performance. However, the drivers from HMI variant B subjectively reported a tendency towards greater problems in identifying the system status based on the HMI output ($p = 0.093$).

- Global rating of the experimenter per scenario: The global rating of the experimenter differed in the scenarios “first system activation” (tendentially significant; $p = 0.111$), “second system activation” (significant; $p = 0.024$), “system deactivation” (tendentially significant; $p = 0.064$) and “obstacle” (statistically significant; $p = 0.041041$). No differences were found in the scenarios “lane change”, “standby mode of lateral control”, “sharp bend” and “active driving with the system”.
- HMI differences in the scenario “obstacle”: In contrast to expectations, a tendency towards worse driving behavior was observed for drivers in the group with HMI variant A, i.e., a higher frequency of drivers produced endangerments in terms of too low a minimum distance from the obstacle ($p = 0.102$). System behavior and HMI outputs did not differ between the HMI variants in this scenario. One possible explanation for this result could be that the highly compliant HMI variant A resulted in overtrust in the system, leading to the impression of reliable system performance. As a result, drivers may have taken longer to realize that the system would not be able to handle the situation. However, this interpretation can currently only remain on a hypothetical level. Additional questions regarding driver trust would have helped to support this explanation.
- The identified trends in the observed driving behavior based on the categorial evaluation in the S.A.D.E. app corresponded with the analysis of measured continuous driving data:
 - The tendentially significant more frequent lane exceedances (observed and coded via the S.A.D.E. app; $p = 0.132$) corresponded with tendentially significant lower ratings of vehicle handling (rated by the experimenter via the S.A.D.E. app, $p = 0.145$) and with tendentially significant higher maximum measured lateral deviations in the scenario “sharp bend” (measured via the simulation software; see Figure 4 above for a comparison of the measures in the scenario “sharp bend”, $p = 0.058$).
 - The significantly more frequent delayed braking reactions and endangerments (observed and coded with the S.A.D.E. app, $p = 0.012$) corresponded with a significantly lower rating of situation handling (rated by the experimenter via the S.A.D.E. app, $p = 0.024$) and with tendentially significant smaller minimum time-to-collision values in the scenario “obstacle” (indicating more critical scenarios; measured via the simulation software; see Figure 4 below for a comparison of the measures in the scenario “obstacle”, $p = 0.058$).
 - These results indicate that categorial observation can partly replace the very time-intensive and resource-intensive analysis of time-based measures without losing too much information. This is an advantage of the method when used in studies with real vehicles, since it is usually time-consuming and costly to collect the necessary continuous driving data for an evaluation.

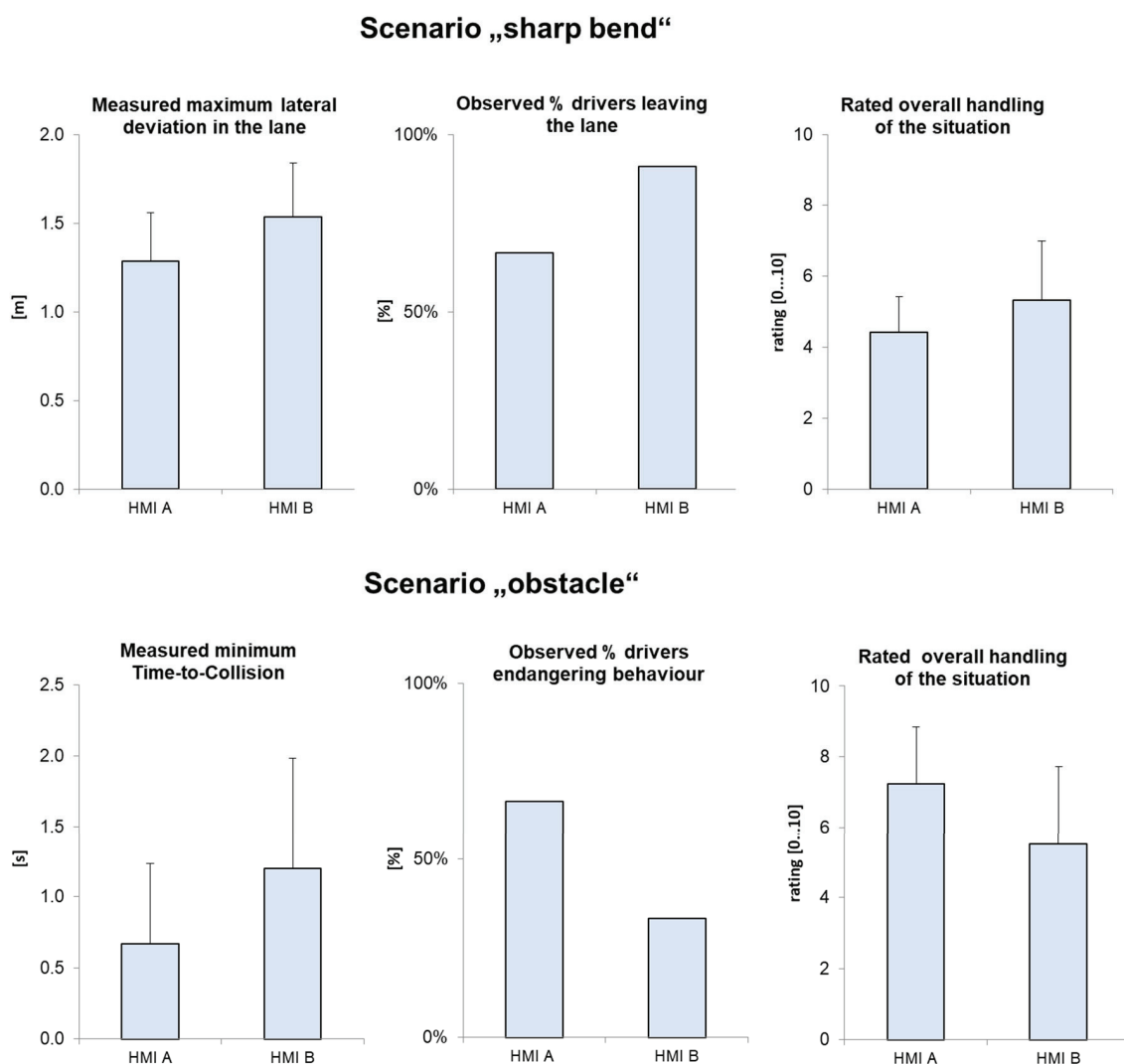


Figure 4. Comparison of measures for the two scenarios “sharp bend” (above: maximum lateral lane deviation; observed % of drivers leaving the lane; overall handling of the situation) and “obstacle” (below: minimum time-to-collision; observed % drivers showing endangering behavior; overall handling of the situation).

4. Results from the First Application of the Method in the BAST Driving Simulator

The method was also applied in a simulator study conducted at the Federal Highway Research Institute [17]. The authors explored the effects of different levels of quality of the lateral vehicle control of a partially automated system on drivers’ interactions with the system. N = 56 test drivers participated in the study. The participants performed a 30 min test drive using a prototypical L2 system on a two-lane highway. The system status was indicated in a cluster display. Lateral control quality was manipulated between the subjects by two factors. The first factor was the variability of the lateral control. The systems differed in the extent that the system swerved within the lane, requiring the driver to steer cooperatively with the system without overriding it. There was a condition with “low variability”, i.e., completely lane centered without any deviations, and another condition with “high variability”, i.e., including deviations from the lane center to the right and to the left without leaving the lane. The second factor was the reliability of the lateral control (i.e., how often the lateral control completely fails for a longer period of time; resulting in the condition “high reliability” without any failures of lateral control vs. “low reliability” including system fails of 3–4 s every 30 s).

During the drive, the subjects experienced two scenarios including a system limit or a system malfunction in which they had to intervene to prevent a safety-critical situation (within a 12 min interval). The scenarios were the two that were already described and used in the WIVW driving simulator study (see Section 3): The first scenario was the scenario “sharp bend” in which lateral control was deactivated without any warning to the driver. For the participants, this deactivation could only be noticed from the change in the visual status indicator in the cluster HMI. The second scenario was the scenario “obstacle” where the system did not detect the stationary vehicle behind the hilltop. Again, the drivers received no warning. As the system itself was not able to detect this limit, the visual status indication did not change in the HMI.

Drivers’ interactions with the system and especially drivers’ reactions to the system limits were assessed using both objective driving data measured via the simulation software SILAB[®] as well as observed driver behavior via the S.A.D.E. app. Experimenters conducting the rating were kept blind to the experimental condition the subjects were assigned to.

The results of the study showed that a high reliability of the L2 system’s lateral control (in terms of stable active system status without intermittent fails) worsened driver intervention performance at critical system limits both in the scenario “sharp bend” (measured by tendentially higher lateral deviations) and in the scenario “obstacle” (measured by tendentially lower minimum TTC). This result occurred despite the fact that drivers had been informed about the system limits beforehand. The factor variability in lateral guidance had no effects on driver intervention behavior in the study. Furthermore, the effects could also be observed in the experimenters’ ratings via the S.A.D.E. app. Correlations between the measures were quite high: The correlation between the SDLP and observer rating was $r = 0.80$ ($p < 0.01$), and the correlation between minimum TTC and observer rating was $r = -0.56$ ($p < 0.01$).

The interrater reliability was calculated separately for the five studied scenarios with the Spearman rank order correlation. It reached values between $r_s = 0.602$ (scenario “second activation”) and $r_s = 0.742$ (scenario “obstacle”).

5. Summary

Taking the results from these initial simulator studies, the developed method is a promising tool for the evaluation of human–machine interaction in partially automated driving in the context of user studies. It showed a high although not fully perfect correlation with objective continuous driving data. The method has the advantage that it can be used in different test environments, such as a driving simulator, test track or in field studies. Instead of an elaborate analysis of continuous driving data, which requires complex measurement equipment, especially in real vehicles, the focus lies on a standardized observation of driver behavior in real time during system use. It was shown that statements can be made about driving behavior, system operation and monitoring performance, which can provide valuable information about the quality of an L2 system and its HMI in terms of usability and safety. The method can be used as a paper–pencil tool applying the proposed observational categories and the rules for the derivation of the global rating for each investigated scenario. The observation, logging and analysis of the data can be facilitated by the tablet-based S.A.D.E. application, but this is not mandatory.

By specifying a defined test procedure, test scenarios and error categories, a standardized application of the method is ensured. The global rating procedure is based on clearly defined rules, which makes the method objective. Therefore, the tool seems well suited for use in different institutions, such as universities, OEMs and testing institutions for consumer protection purposes.

By using the method in the context of user studies, it can be evaluated whether assessments of the degree of fulfillment of design guidelines based on the literature actually have practical relevance to driver interaction with a system and its HMI. The inclusion of subjective assessments by drivers can provide additional clues as to how well an HMI is able to convey an appropriate understanding of driver responsibilities and subjectively

perceived system functionalities, which is the basis of maintaining an appropriate level of mode and situation awareness.

The implementation of realistic test scenarios allows an estimation of system- and HMI-related aspects of driving safety.

6. Discussion of Limitations and Future Challenges

The current version of the tool nevertheless has some limitations and future challenges. The observation categories used with regard to system operation, driving performance and driver monitoring were defined based on the authors' expertise. The validation of the method in the two simulator studies focused predominantly on the two categories of system operation and driving behavior. It was found that the observation criteria defined in these categories were largely suitable for identifying the problems arising from inappropriate HMI design, which are serious enough that they also manifest in observable behavior. In this respect, the method seems to be valid for the detection of major problems in interaction with an existing system in a fast and efficient way. However, the method cannot accomplish the detection of relatively small differences in system design. These differences may need to be evaluated during the development process of a system. Otherwise, more detailed analyses, e.g., of gaze behavior or reaction times, are recommended.

The objectivity and reliability of the method will strongly depend on the degree to which a thorough rater training (about the general method, but also for a specific study) takes place. However, the objectivity of the method will always be lower than when a quantitative analysis of behavioral and vehicle data is performed. In addition, depending on environmental conditions, some behaviors may be difficult to observe (e.g., lane keeping if a bird's eye view is not available in a real traffic experiment). Likewise, the requirements for observing various problems simultaneously may overload the rater, causing them to overlook individual aspects.

Conversely, the method has the potential to observe and identify certain more complex cause–effect relationships that would not be apparent in the analysis of measured objective data. For example, problems with system activation are usually expressed as a prolonged reaction time until successful system activation. Whether this occurs due to a complex search for the correct operation element or due to an initially incorrect system operation can be identified far more quickly by observation.

Furthermore, there is a clear need for the optimization of the precise definition of the evaluation criteria for appropriate monitoring behavior of the drivers. In the short-term studies reported in this paper, problems in monitoring tended to be neglected or did not occur (the drivers were highly focused, were not distracted by any additional secondary activities and always had their hands on the steering wheel). The operationalization of the error category “not attentive enough” is currently still relatively vaguely defined. In the future, this will require the use of more clearly defined criteria, e.g., gaze behavior and a meaningful approach on the basis of which an attentive and situation-aware driver can be recognized (both based on data from eye tracking measurement but also purely from observation).

Kircher and Ahlström [22] list expert judgments as one possible evaluation method for the assessment of the minimum required attention to the driving task aside from other evaluation methods, such as the usage of eye tracking sensors to directly measure drivers' glance directions and eye blinks. They cite a publication [23] which includes recommendations for eye scanning rules. The named examples are: “keep the eyes moving”, “scan the entire traffic scene”, “center the gaze on the travel path” and “look at mirrors and instruments” [22]. However, they found that the recommendations of driving instructors in a laboratory condition regarding how attention should be distributed on a highway were not reliable for the assessment of driver attentional distribution in traffic if the situational context was unknown.

What the method cannot provide is a detailed explanatory model of why certain problems occur during system interaction, since only behavior itself is observed. If conclusions

are to be drawn about the underlying psychological mechanisms, it is advisable to supplement the method with other methods that can go deeper into the analysis of psychological processes, such as precise analyses of gaze and reaction times or the addition of specific surveys of drivers, e.g., about their mental model, system trust and acceptance, etc.

Lastly, the use of the method has so far been limited to simulators. Initial tests in real traffic had mainly focused on the applicability of the tool. Currently, the method is also used and validated in real vehicles for the evaluation of driver–vehicle interaction in L2 systems by BASt. The results will be made available to the public and used to further develop the method.

Author Contributions: N.S. and K.W.: project administration, study conceptualization, methodology and validation; N.S.: writing—original draft preparation; A.N.: supervision and funding acquisition; A.W.: conceptualization and validation. All authors have read and agreed to the published version of the manuscript.

Funding: This work was funded by the Federal Highway Research Institute within a project with the overall goal of developing evaluation methods for the assessment of human–machine interaction for partially automated driving functions (Grant Number: FE82.0708).

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. J3016_202104; SAE On-Road Automated Vehicle Standards Committee. Taxonomy and definitions for terms related to on-road motor vehicle automated driving systems. SAE International: Warrendale, PA, USA, 2021.
2. Bainbridge, L. Ironies of automation. In *Analysis, Design and Evaluation of Man–Machine Systems*; Elsevier: Amsterdam, The Netherlands, 1983; pp. 129–135.
3. National Transportation Safety Board. *Collision between a Car Operating with Automated Vehicle Control Systems and a Tractor–Semitrailer Truck near Williston, Florida, May 7, 2016*; Highway Accident Report NTSB/HAR-17/02; National Transportation Safety Board: Washington, DC, USA, 2017.
4. Schaller, T.; Schiehlen, J.; Gradenegger, B. Stauassistentz–Unterstützung des Fahrers in der Quer- und Längsführung: Systementwicklung und Kundenakzeptanz. In *3 Tagung Aktive Sicherheit durch Fahrerassistenz*; Lehrstuhl für Fahrzeugtechnik, Technische Universität: München, Germany, 2008.
5. Kircher, K.; Larsson, A.; Hultgren, J.A. Tactical driving behavior with different levels of automation. *IEEE Trans. Intell. Transp. Syst.* **2014**, *1*, 158–167. [CrossRef]
6. Begiatto, M.; Hartwich, F.; Schleinitz, K.; Krems, J.; Othersen, I.; Petermann-Stock, I. What would drivers like to know during automated driving? Information needs at different levels of automation. In *7. Tagung Fahrerassistenzsysteme*; Lehrstuhl für Fahrzeugtechnik, Technische Universität: München, Germany, 2015.
7. Large, D.R.; Banks, V.A.; Burnett, G.; Baverstock, S.; Skrypchuk, L. Exploring the behaviour of distracted drivers during different levels of automation in driving. In Proceedings of the 5th International Conference on Driver Distraction and Inattention (DDI2017), Paris, France, 20–22 March 2017; pp. 20–22.
8. Llaneras, R.E.; Salinger, J.; Green, C.A. Human Factors Issues Associated with Limited Ability Autonomous Driving Systems: Drivers’ Allocation of Visual Attention to the Forward Roadway. In Proceedings of the 7th International Driving Symposium on Human Factors in Driver Assessment, Training, and Vehicle Design, Bolton, NY, USA, 17–20 June 2013.
9. Naujoks, F.; Purucker, C.; Neukum, A. Secondary task engagement and vehicle automation—Comparing the effects of different automation levels in an on-road experiment. *Transp. Res. Part F* **2016**, *38*, 67–82. [CrossRef]
10. Kleen, A.T. Beherrschbarkeit von teilautomatisierten Eingriffen in die Fahrzeugführung. Ph.D. Thesis, Technische Universität Braunschweig, Braunschweig, Germany, 2014.
11. Louw, T.; Kuo, J.; Romano, R.; Radhakrishnan, V.; Lenné, M.G.; Merat, N. Engaging in NDRTs affects drivers’ responses and glance patterns after silent automation failures. *Transp. Res. Part F Traffic Psychol. Behav.* **2019**, *62*, 870–882. [CrossRef]
12. Stanton, N.A.; Young, M.S.; Walker, G.H.; Turner, H.; Randle, S. Automating the driver’s control tasks. *Int. J. Cogn. Ergon.* **2001**, *5*, 221–236. [CrossRef]
13. Wulf, F.; Zeeb, K.; Rimini-Döring, M.; Arnon, M.; Gauterin, F. Effects of human-machine interaction mechanisms on situation awareness in partly automated driving. In Proceedings of the 16th International IEEE Conference on Intelligent Transportation Systems (ITSC 2013), The Hague, The Netherlands, 6–9 October 2013; pp. 2012–2019.
14. Naujoks, F.; Wiedemann, K.; Schömig, N.; Jarosch, O.; Gold, C. Expert-based controllability assessment of control transitions from automated to manual driving. *MethodsX* **2018**, *5*, 579–592. [CrossRef] [PubMed]

15. Naujoks, F.; Hergeth, S.; Wiedemann, K.; Schömig, N.; Forster, Y.; Keinath, A. Test procedure for evaluating the human–machine interface of vehicles with automated driving systems. *Traffic Inj. Prev.* **2019**, *20*, 146–151. [CrossRef] [PubMed]
16. National Highway Traffic Safety Administration. *Federal Automated Vehicles Policy 2.0*. Washington, DC: National Highway Traffic Safety Administration (NHTSA), Department of Transportation (DOT); National Highway Traffic Safety Administration: Washington, DC, USA, 2017.
17. Wiggerich, A.; Hoffmann, H.; Schömig, N.; Wiedemann, K.; Segler, K. Bewertung der Sicherheit der Mensch-Maschine-Interaktion teilautomatisierter Fahrfunktionen (Level 2). In Proceedings of the Beitrag auf dem 13, Uni-DAS e.V. Workshop Fahrerassistenz und automatisiertes Fahren, digital conference, 16–17 July 2020.
18. Menold, N.; Bogner, K. Gestaltung von Ratingskalen in Fragebögen. *Mannh. GESIS—Leibniz-Inst. Für Soz. (SDM Surv. Guidel.)* **2015**. [CrossRef]
19. Neukum, A.; Lübbecke, T.; Krüger, H.-P.; Mayser, C.; Steinle, J. ACC-Stop&Go: Fahrerverhalten an funktionalen Systemgrenzen. In Proceedings of the Workshop Fahrerassistenzsysteme—FAS2008, Walting, Germany, 2–4 April 2008; Maurer, M., Stiller, C., Eds.; fmrt Karlsruhe: Walting, Germany, 2008; pp. 141–150.
20. Kaussner, Y. Assessment of driver fitness: An alcohol calibration study in a high-fidelity simulation. In Proceedings of the Fit to Drive 7th International Traffic Expert Congress, Berlin, Germany, 25–26 April 2013.
21. Kenntner-Mabiala, R.; Kaussner, Y.; Jagiellowicz-Kaufmann, M.; Hoffmann, S.; Krüger, H.-P. Driving performance under alcohol in simulated representative driving tasks: An alcohol calibration study for impairments related to medicinal drugs. *J. Clin. Psychopharmacol.* **2015**, *35*, 134–142. [CrossRef] [PubMed]
22. Kircher, K.; Ahlstrom, C. Evaluation of methods for the assessment of attention while driving. *Accid. Anal. Prev.* **2018**, *114*, 40–47. [CrossRef] [PubMed]
23. Zwahlen, H.T. Eye scanning rules for drivers: How do they compare with actual observed eye scanning behaviour. *Transp. Res. Rec. J. Transp. Res. Board* **1991**, *1403*, 14–22.

MDPI AG
Grosspeteranlage 5
4052 Basel
Switzerland
Tel.: +41 61 683 77 34

Information Editorial Office
E-mail: information@mdpi.com
www.mdpi.com/journal/information



Disclaimer/Publisher's Note: The title and front matter of this reprint are at the discretion of the Guest Editors. The publisher is not responsible for their content or any associated concerns. The statements, opinions and data contained in all individual articles are solely those of the individual Editors and contributors and not of MDPI. MDPI disclaims responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.



Academic Open
Access Publishing

mdpi.com

ISBN 978-3-7258-4638-2