TUM School of Engineering and Design
Technische Universität München

Transparency Assessment of Automated Vehicle Human-Machine Interface

Yuan-Cheng Liu

Vollständiger Abdruck der von der TUM School of Engineering and Design der Technischen Universität München zur Erlangung eines

Doktors der Ingenieurwissenschaften (Dr.-Ing.)

genehmigten Dissertation.

Vorsitz:
Prof. Dr.-Ing. Hartmut Spliethoff

Prüfende der Dissertation:

1. Prof. Dr. Klaus Bengler
2. Assistant Prof. Dr.ir. Riender Happee

Die Dissertation wurde am 19.07.2024 bei der Technischen Universität München eingereicht und durch die TUM School of Engineering and Design am 24.10.2024 angenommen.

Acknowledgement

For Pris, Nikol, Jurek, and all the people.

Abstract

Before fully autonomous vehicles are generally adopted, continuous interactions are required between human users and automated driving systems during automated driving. Since the control of the vehicle is shared among the human users and the automation, it is then of great importance for the two agents to be capable of sharing current states, motivations, and intentions. The Human-Machine Interface (HMI) works as the only media that could transmit critical information from the automated driving system to users regarding the states of the automated driving system. However, misunderstanding and confusion remain major issues and might lead to dangerous situations, compromising the safety of human users. To minimize the risk of accidents and improve the automated driving experiences of human users, a correct understanding of the HMI should be guaranteed by the automated system.

To address the issue, a standardized and objective transparency assessment method was developed in Article 2. Using the metrics that evaluate understandability and effort to understand the automated system simultaneously, the degree to which the AV HMI is understood by human users could be objectively estimated. To further adapt the proposed transparency assessment method to a more dynamic environment, electroencephalogram (EEG) was adopted and analyzed in Article 2 to investigate how sensitive the psychophysiological measure is and whether it is suitable to be incorporated into the proposed transparency assessment method. Furthermore, electrocardiography (ECG) and electrodermal activity (EDA) were also used in a driving simulator to investigate the effects of various HMI designs on the proposed transparency assessment method in Article 3.

The results showed the capability of the proposed transparency assessment method to identify significant differences among different HMI designs, and the method was also able to be utilized to determine critical information topics on the HMI design for a better understanding of the automated driving system. The sensitivities of various psychophysiological workload measures were also identified and compared to subjective workload measures. It was also indicated that with the adoption of the proposed TRASS method, the design and evaluation process for the HMI could be improved by the inclusion of objective measures for understandability and workload when trying to understand the automated driving system. This not only provides future researchers with a powerful assessment tool to evaluate the HMI design, but it could also be beneficial and help to modify the HMI designs based on this standardized objective metric. However, more research is still required to increase ecological validity across all complex driving scenarios. Still, this research provides a firm basis for future studies regarding the HMI design process, training program, as well as safety regulations of the automated driving system.

Zusammenfassung

Bevor vollautonome Fahrzeuge allgemein eingeführt werden, ist eine kontinuierliche Interaktion zwischen menschlichen Nutzern und automatisierten Fahrsystemen während des automatisierten Fahrens erforderlich. Da sich die menschlichen Nutzer und die Automatisierung die Kontrolle über das Fahrzeug teilen, ist es von großer Bedeutung, dass die beiden Agenten in der Lage sind, aktuelle Zustände, Motivationen und Absichten auszutauschen. Der Human-Machine Interface (HMI) funktioniert als einziges Medium, das kritische Informationen über den Zustand des automatisierten Fahrsystems an die Benutzer übertragen kann. Missverständnisse und Verwirrung sind jedoch nach wie vor ein großes Problem und können zu gefährlichen Situationen führen, die die Sicherheit der menschlichen Nutzer gefährden. Um das Unfallrisiko zu minimieren und das automatisierte Fahren für die menschlichen Nutzer zu verbessern, sollte ein korrektes Verständnis des HMI gewährleistet sein.

Um dieses Problem anzugehen, wurde in Artikel 2 eine standardisierte und objektive Methode zur Bewertung der Transparenz entwickelt. Mit Hilfe der Metriken, die gleichzeitig die Verständlichkeit und den Aufwand für das Verstehen des automatisierten Systems bewerten, konnte der Grad des Verständnisses der AV HMI durch menschliche Benutzer objektiv abgeschätzt werden. Zur weiteren Anpassung der vorgeschlagenen Transparenzbewertungsmethode an eine dynamischere Umgebung wurde Electroencephalogram (EEG) in Artikel 2 übernommen und analysiert, um zu untersuchen, wie empfindlich das psychophysiologische Maß ist und ob es in die vorgeschlagene Transparenzbewertungsmethode aufgenommen werden kann. Darüber hinaus wurden Electrocardiography (ECG) und Electrodermal Activity (EDA) auch in einem Fahrsimulator verwendet, um die Auswirkungen verschiedener HMI-Designs auf die vorgeschlagene Transparenzbewertungsmethode in Artikel 3 zu untersuchen.

Die Ergebnisse zeigten, dass die vorgeschlagene Transparenzbewertungsmethode in der Lage ist, signifikante Unterschiede zwischen verschiedenen HMI-Designs zu identifizieren, und die Methode konnte auch dazu verwendet werden, kritische Informationsthemen über das HMI-Design für ein besseres Verständnis des automatisierten Fahrsystems zu bestimmen. Die Empfindlichkeit verschiedener psychophysiologischer Belastungsmessungen wurde ebenfalls ermittelt und mit subjektiven Belastungsmessungen verglichen. Es wurde auch darauf hingewiesen, dass mit der Annahme der vorgeschlagenen TRASS-Methode der Entwurfs- und Bewertungsprozess für das HMI durch die Einbeziehung objektiver Maße für Verständlichkeit und Arbeitsbelastung beim Versuch, das automatisierte Fahrsystem zu verstehen, verbessert werden könnte. Dies gibt zukünftigen Forschern nicht nur ein leistungsfähiges Bewertungsinstrument zur Beurteilung des Designs von HMI an die Hand, sondern könnte auch von Vorteil sein und dazu beitragen, die HMI-Designs auf der Grundlage dieser standardisierten objektiven Metrik zu ändern.

Es sind jedoch noch weitere Forschungen erforderlich, um die ökologische Validität für alle komplexen Fahrszenarien zu erhöhen. Dennoch bietet diese Forschung eine solide Grundlage für zukünftige Studien in Bezug auf den Entwurfsprozess, das Trainingsprogramm und die Sicherheitsvorschriften für automatisierte Fahrsysteme.

Contents

List of Figures

List of Tables

Acronyms

ACC	Adaptive Cruise Control.
ADS	Automated Driving System.
AV	Automated Vehicle.
ECG	Electrocardiography.
EDA	Electrodermal Activity.
EEG	Electroencephalogram.
FT	Functional Transparency.
GM	General Motors.
GPS	Global Positioning System.
HMI	Human-Machine Interface.
HRV	Heart Rate Variability.
IMU	Inertial Measurementunit.
LIDAR	Light Detection And Ranging.
NASA-TLX	Nasa-Task Load Index.
NDRT	Non-Driving Related Task.
RADAR	Radio Detection and Ranging.
RMSSD	Root Mean Square of Successive R-R Intervals.
SAT	Situation Awareness-based Agent Transparency.
SCL	Skin Conductance Level.
SCR	Skin Conductance Response.
TRASS	Transparency Assessment.

1 Introduction

He who knows the answer to this
knows all things.

Les Misérables
Victor Hugo

In 1769, Richard Arkwright built his first cotton-spinning factory with a fully automated spinning mill (Fitton & Wadsworth, 1958), which set the wheel of the industrial revolution in the 18th century in motion. From then on, and with the aid of electrification, computerization, and robotics, automated technologies have been extensively integrated into our daily lives, especially in the transportation and automobile manufacturing sectors. From traffic infrastructures to public transportation and to Automated Vehicle (AV) that are widely adopted, automation has become part of our everyday experiences.

The history of automation on vehicles goes way back to the early 20th century. One well-known example was Futurama, whereGeneral Motors (GM) put the idea of remotely-controlled automated vehicles into the future city (Miller, 2020). In this conceptual city, vehicles on the highways were radio-controlled and followed the electromagnetic field embedded under the surface of the road. People at the time believed it to be a promising solution for the increasingly complex highway system, which allowed a substantially safer journey. Exploration of different sensory solutions for AV navigation keeps pushing this technology into the public eye.

Dickmanns, Mysliwetz, and Christians (1990) introduced computer vision to the self-driving car and reached almost 100 km per hour. Back then, hardware limitations were significant, and thus, sophisticated algorithms were required, and only single sensor input was possible. However, as the computing power of the processors advances, multi-sensory self-driving vehicles become possible and allow more reliable controls. In the 2005 DARPA Grand Challenge, the winner, Stanley, incorporated Light Detection And Ranging (LIDAR), Radio Detection and Ranging (RADAR), Inertial Measurementunit (IMU), Global Positioning System (GPS), and camera, and self-navigated a 142-mile long course in 6 hours 53 minutes and 58 seconds (Thrun et al., 2006). Till today, these technologies have been widely adopted in automated vehicles, including those commercially available models.

Despite the blooming of automated technologies in recent years, AVs nowadays are still at a distance from the dream of fully autonomous cars. In SAE levels of driving automation, six levels of driving automation are defined, where the lowest level of level zero represents the vehicle with very limited warning or assistance provided, and the highest level of level five represents the fully autonomous vehicle that can drive everywhere in all conditions. The wide adoption of automated and autonomous vehicles is believed to benefit society by alleviating congestion, reducing energy used and associated emissions, and improving safety (Brown, Gonder, & Repac, 2014; Winkle, 2016; C. D. Yang & Fisher, 2021), still, limitations and potential concerns still persist and require a closer look.

One of the challenges raised is regarding the use of Human-Machine Interface (HMI) in the AV. HMI is often the only interface allowing human users to understand the state of the automation, and thus, is considered one of the critical parts during the interaction. However, instead of improving the efficiency and performance while using them, AV HMIs tend to cause confusion and misuse (Kaleefathullah et al., 2022; Wilson, Yang, Roady, Kuo, & Lenné, 2020). For instance, in the study of Wilson et al. (2020), divers in Level 2 automation sometimes were unaware of the disengagement of the automated system, which can result in significant risks. In another study, drivers who transit from highly automated driving to manual driving tend to perform risky maneuvers with higher speeds owing to the misconception of the automation status (Calvi, D'Amico, Ciampoli, & Ferrante, 2020).

Hancock et al. (2020) highlighted several challenges raised by the advancement of automated levels, including the complication and confusion for human users' roles. When under a low automation level, human users are drivers using driving assistance provided by the vehicle. As the level goes up, human users become supervisors, monitoring the more powerful assistance system and taking back control whenever needed. Transitions between these levels alter the duties of human users, making the role dynamic. It is thus critical to ensure that the AV HMI is well comprehended to minimize risks and facilitate driving safety.

To guarantee a safe transition and efficient cooperation between humans and automation, suggestions and guidelines have been discussed upon (Beggiato et al., 2015; Debernard, Chauvin, Pokam, & Langlois, 2016; Naujoks, Wiedemann, Schömig, Hergeth, & Keinath, 2019; Richardson, Lehmer, Lienkamp, & Michel, 2018). However, there has been minimal coverage on whether the HMI design is transparent or understandable to users and, more importantly, how to measure the corresponding transparency of the HMI. Motivated by this insufficiency, this thesis aims to define and develop a transparency assessment method for automated vehicle human-machine interfaces. This thesis is structured as follows: In Chapter 2, definitions of transparency and the state-of-the-art assessment methods for automated vehicle HMI will be discussed. Then, the criticality and operationalization of understandability will be reviewed and discussed in Chapter 3 together with workload measures, which are also considered critical during the AV HMI evaluation process. Following are the summary and objectives, which will be given in Chapter 4. Motivated by the challenges and gaps in AV HMI evaluation, a standardized and objective transparency assessment method was proposed in Chapter 5. To expand the use scenarios and adapt to real-time measures, the application of an electroencephalogram (EEG) for workload measurements in a simulator environment was studied (Chapter 6). The proposed metric was further explored in Chapter 7 for its fidelity in a more dynamic environment using both electrocardiography (ECG) and electrodermal activity (EDA). Results from these studies relating to the proposed transparency assessment method and potential future research were discussed in Chapter 8.

2 Transparency of Human-Machine Interface and Evaluation Strategies

Not being heard is no reason for silence.

Les Misérables
Victor Hugo

Transparency is not a new word in the field of human-computer interaction. It has been used to describe the amount and type of information provided by the automated system (Bhaskara et al., 2021; J. Y. Chen et al., 2018; Körber, Baseler, & Bengler, 2018; Kraus, Scholz, Stiegemeier, & Baumann, 2020; Kunze, Summerskill, Marshall, & Filtness, 2019; J. Lee, Abe, Sato, & Itoh, 2020; Maarten Schraagen et al., 2021; Oliveira, Burns, Luton, Iyer, & Birrell, 2020). However, the quantity of information does not compensate for the quality. For example, extra information on the Human-Machine Interface (HMI) is less preferable for novice automated vehicle users compared to more experienced ones (Kraft, Naujoks, Wörle, & Neukum, 2018). Beggiato et al. (2015) also found that users with more trust in the automated system tend to demand less information on the HMI. Hence, it is evident that using solely the amount of information displayed on the HMI does not correlate to the extent of users' comprehension of the automated system. A construct representing the resulting comprehension of the automated system through the HMI is required. Later in the Chapter, current HMI evaluation methods are examined and discussed. Multiple constructs have been deployed during the HMI design process to enable a safer and more satisfying experience during human-AV interactions. However, the outlook appears unfavorable for bridging the gaps in measuring the "true transparency" of the automated system.

2.1 Automated Vehicle and Human-Machine Interface Transparency

According to the Oxford English Dictionary, the definition of transparency is:

> The quality of being easy to perceive or detect. (*transparency, n.*, n.d.)

A transparent object allows one to see through, while a transparent behavior represents actions that are easily comprehended by others and create clear communications. For either scenario, having transparency promotes detailed knowledge for the recipient and, what is more preferable, should guarantee a profound understanding of the information provided from the perspective of the recipient.

In the definition of the work of J. Y. Chen et al. (2014), transparency is defined as the ability of the interface to provide users "about an intelligent agent's intent, performance, future plans, and reasoning process" (p. 2). The agent here can refer to intelligent agents or autonomous robots that are teaming or interacting with human users. In the Situation Awareness-based Agent Transparency (SAT) model, there are three transparency levels associated with the information needs to facilitate the human-agent teamwork (J. Y. Chen et al., 2014). The SAT model was later expanded to allow bidirectional communications for even more complex or dynamic situations (J. Y. Chen et al., 2018). Generally, higher subjective trust was found with a higher level of transparency. Regarding workload, however, inconclusive results were found across studies. In the study of Selkowitz, Larios, Lakhmani, and Chen (2016), no differences in workloads across interfaces with various transparency levels were found, indicating effective information usage in certain situations is possible without additional effort for human users. Meanwhile, J. Y. Chen and Barnes (2012) found that depending on the increased levels of transparency, no differences in perceived workload were found, but the performances of participants were negatively affected in high transparency conditions. However, Helldin, Ohlander, Falkman, and Riveiro (2014) reported a higher workload during the interaction with a high-transparency scenario.

A similar concept of transparency has also been adopted in the research field of automated vehicles. Maarten Schraagen et al. (2021) adopted the SAT model by J. Y. Chen et al. (2014) and applied it to the navigation system of the automated vehicle. The navigation system and the planned routes themselves were used as the first level of transparency, while colored images of the front vehicle, where the color changed from green to yellow to red when the distance got shorter to the front one, represented the second level of transparency. Finally, the future state of the automated vehicle was shown and was the highest transparency level. The study showed that adding transparency as a factor could increase trust and acceptance, but the situation awareness was decreased.

In another study in the area of automated vehicles, Kunze et al. (2019) used uncertainty display as the means to increase the transparency level. They found out that with the additional uncertainty information, increases in task-solving rate, situation awareness response, and subjective trust were significant. Similar effects were found on the indices for take-over performances, where the minimum time-to-collision was significantly higher, and the maximum lateral acceleration was significantly lower for the group with uncertainty display. However, the results of subjective workload and post-experiment interviews indicated that the additional uncertainty information increases the need for participants to glance at the display and thus increases the mental demand for the uncertainty display group. This also made participants feel uneasy and frustrated when using the uncertainty display.

Transparency has also been communicated as the factor of the information quantity in the field of automated vehicles. J. Lee et al. (2020) toggled system transparency between "detailed" and "less", where in the detailed scenario, more prior information and examples of specific situations were shown to the participants. Kraft, Maag, Cruz, Baumann, and Neukum (2020) applied additional explanation messages on HMIs for the failures of the system and resulted in a higher level of acceptance throughout the scenarios. In other situations, no effect of the additional explanatory messages on trust and understanding except for certain occasions(Kraft et al., 2020). This again demonstrates that using solely transparency does not guarantee the understanding of the users and thus might put driving safety at risk.

To facilitate safe driving when interacting with automated vehicles, it is necessary and critical to enable clear communication from the HMI and allow users to have a full understanding of the automated system (Beggiato et al., 2015; Bengler, Rettenmaier, Fritz, & Feierle, 2020). Defining transparency as different levels of information and the amount of information provided only distinguished the quantity and the type of the information displayed, and thus, it was not sufficient to evaluate whether the HMI design on the AV was easily understandable. Moreover, if only the quantity of information is considered and users are only provided with information of a higher level of transparency, users might be overwhelmed and thus not produce the best possible outcome (Maarten Schraagen et al., 2021). To evaluate the Human-Machine Interface of automated systems in a more comprehensive manner, more constructs than only types of information should be considered to facilitate an understandable and safe human-automation interaction. The next section will briefly introduce and discuss state-of-the-art AV HMI evaluation methods.

2.2 Human-Machine Interface Evaluation Methods

Multiple constructs during human interaction with AV have been studied, such as usability (Albers et al., 2020; Voinescu, Morgan, Alford, & Caleb-Solly, 2020; Walch, Mühl, Baumann, & Weber, 2018), trust and acceptance(Adnan, Nordin, bin Bahruddin, & Ali, 2018; Ayoub, Zhou, Bao, & Yang, 2019; Körber et al., 2018), comfort(Beggiato, Hartwich, & Krems, 2019; Bellem, Thiel, Schrauf, & Krems, 2018; Peng et al., 2022), etc. Studies regarding usability measured subjective perceptions and objective reactions from human users during their interaction with HMI to evaluate how well users enjoy and react to the HMI.

The concept of usability and usability tests has been widely adopted in many fields of studies, including web page design (Kokil & Scott, 2017; Kous, Pušnik, Heričko, & Polančič, 2020), mobile devices (Pensabe-Rodriguez, Lopez-Dominguez, Hernandez-Velazquez, Dominguez-Isidro, & De-la Calleja, 2020; Weichbroth, 2020), or Virtual Reality applications (Ebnali, Lamb, Fathi, & Hulme, 2021; Voinescu et al., 2020) so that the subjective or objective user experiences could be estimated. The implementation is also common when evaluating human-AV interaction. As summarized in the work of Albers et al. (2020), many studies used usability testing to evaluate the Human-Machine Interface (HMI), however, with different definition and operationalization variables. Among the definitions of usability, the definition from ISO 9241-11:2018 has been one of the most used definitions (DIN EN ISO 9241-11, 2018). According to ISO 9241-11:2018, usability consists of attributes of effectiveness, efficiency, and user satisfaction. Effectiveness is often measured by the completion rate of certain tasks, while efficiency could be measured by time-based or overall averaged efficiency. Compared to the other two, user satisfaction is measured subjectively and usually with the System Usability Scale (SUS) (Brooke, 1996). In the operationalization of usability measurement, the three constructs do not always appear together, as there is no standardized experimental procedure explicitly defined in ISO 9241-11. In many cases, subjective SUS was used to measure user satisfaction as the operationalization of the usability test (Richardson et al., 2018; Voinescu et al., 2020).

Meanwhile, other subjective evaluations, such as trust, were also often estimated during the interaction between humans and AV to determine the level of influence on the willingness to use automation (Ekman, Johansson, & Sochor, 2017; Gold, Körber, Hohenberger, Lechner, & Bengler, 2015; J. D. Lee & See, 2004; Ma, Morris, Herriotts, & Birrell, 2021; Maarten Schraagen et al., 2021). Trust is usually characterized as a belief, attitude, intention, or behavior. J. D. Lee and See (2004) defined it simply as "the attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability" (p. 51). Highly automated vehicle experience was shown to increase the self-reported trust in automation (Gold et al., 2015), where automated vehicles with high visual feedback also increased subjective trust rating in automation (Ma et al., 2021). The appropriate level of trust in automation is believed to be pivotal when it comes to the adaptation of the automated vehicle both now and in the future (Walker et al., 2023).

Trust could be estimated by self-reported measures, behavioral measures, or psychophysiological measures (Gold et al., 2015; Kohn, de Visser, Wiese, Lee, & Shaw, 2021; Walker, Wang, Martens, & Verwey, 2019). The latter two are favored in a more dynamic setting where real-time changes in trust are desired, while most studies rely on self-reported measures. A combination of behavioral and psychophysiological measures is also developed to increase the reliability of measuring trust objectively Walker et al. (2019).

Previous paragraphs cover some of the constructs used widely during the development stages of automated vehicle HMI. However, little has been addressed regarding the essence of the HMI design, which is to allow human users to understand complex autonomous systems easily. Not understanding the automation in use might pose a threat to driving safety, which might further develop and lead to catastrophic accidents. This idea of being "transparent" tends to be incorporated in other constructs rather than be evaluated on its own. Hence, the author argues that there should exist such a construct that can answer the question of "How well do human users understand the human-machine interface?" or, to be more precise, "How transparent is the human-machine interface?". Here, the concept of transparency in the previous quoted sentence that the author tries to deliver is the combination of the resulting understanding and effort used by the human user after the interaction with the automated system, not the amount or levels of information provided by it according to the SAT model. In order to distinguish this concept from the transparency in SAT model (J. Y. Chen et al., 2014), this new construct, which is of the highest relevance to a safe interaction between human and automated vehicles, is defined as Functional Transparency (FT). FT represents a scale that reflects the easiness and correctness of users to understand and react to an AV HMI (Hoc, Young, & Blosseville, 2009; Passchier, 2021; Reilhac, Hottelart, Diederichs, & Nowakowski, 2017; Rezaei & Caulfield, 2021; Seppelt & Lee, 2019).

The concept of the intended construct FT is to bridge the gap in measuring whether the HMI is understood correctly with minimum effort. If this construct could be more concrete and reliable, the efficiency during AV HMI design and validation phases could be largely improved, as this concept could provide the ultimate outcome of users' understanding and efforts used, which are considered to be some of the most critical factors for safety during human-AV interaction. Despite the similarities between FT and effectiveness and efficiency, where effectiveness could be interpreted as

The concept of FT in AV HMI differs from usability in that no reaction from the human user is required during the interaction, and thus, the measurement of the FT could be conducted in an earlier stage of the design process. Some of the attributes in FT might seem to be overlapping with that of usability. However, FT should be treated as a prerequisite for a product or system to be usable. It describes the extent to which the measured item, either a function, system, or structure, is understandable with minimum effort and thus could potentially be usable in more diverse scenarios. In other words, the understanding of the HMI showing the take-over request, for instance, is a prerequisite for prompt and accurate reaction to take over control.

Several subjective transparency measures have been studied to evaluate the self-reported or perceived transparency of the automation system. With a three-item subjective evaluation, Choi and Ji (2015) attempted to measure the construct "system operation transparency", and determine its relationship with trust. Here, the transparency was measured by the questions "I believe that autonomous vehicle acts consistently and its behavior can be forecast.", "I believe that I can form a mental model and predict the future behavior of the autonomous vehicle.", and "I believe that I can predict what autonomous vehicle will act in a particular way.", from which the mental models of participants to the automation system was evaluated. Later in the study of Oliveira et al. (2020), the three 7-point scale questions were used to evaluate the perceived transparency of participants among different HMI settings. Questions used in this study were adapted from the work of Choi and Ji (2015), while the scenarios were extended according to different HMI designs.

The subjective evaluation of transparency, despite the ambiguity regarding the functions of HMI, could be used to provide a clue about users' subjective understanding of the automated system, and the results could also be acquired for a more "subjectively understandable" HMI design. However, misunderstanding of the system should also be noted and prevented, as it could be the culprit of serious accidents (Serter, Beul, Lang, & Schmidt, 2017; Smyth, Ulahannan, Florek, Shaw, & Mansfield, 2021; Wörle & Metz, 2023). The complacency of human users tends to overwhelm users during human-AV interaction. But we should also note here that it is the responsibility of AV HMIs to convey correct and understandable information to human users. Besides design principles, design guidelines, and expert opinions, there should exist a standardized evaluation method to gauge whether the minimum system requirements are reached.

Hence, the author argues that such a method should exist to answer the question, "How well and effortlessly do users really understand the HMI to ensure driving safety?". Merely subjective 'perceived understanding" of the AV HMI might increase the trust and acceptance of the automation but would not guarantee driving safety. Thus, an objective evaluation method should exist in such a way that the "true understanding" of the automated system can be assured, and minimize the error rate and misuse during the human-AV interaction.

Besides the correctness of the understanding of the AV HMI, keeping the process of understanding effortless is another topic to be considered. In other words, it would not be sufficient to only evaluate situation awareness, for instance, during automated driving, since an HMI allowing high situation awareness might, at the same time, significantly increase the workload, which is not desirable HMI design since it negatively influences driving safety by impairing performance and potentially increasing the error rate (Di Flumeri et al., 2018; Loeches De La Fuente et al., 2019). Hence, a standardized methodology to evaluate an AV HMI design should exist to guarantee that the necessary level of situation and mode awareness is reached while not sacrificing the mental workload of the human user.

In Chapter 3, an in-depth discussion regarding understandability and workload will be discussed. The chapter will also highlight the mental workload during automated driving as well as the subjective and objective measurement methods.

3 Understandability and Objective Workload Measures

> If I speak, I am condemned. If I stay silent, I am damned!
>
> *Les Misérables*
> *Victor Hugo*

In Chapter 2, we have come to the conclusion that a standardized evaluation method for Functional Transparency (FT) of the Automated Vehicle (AV) Human-Machine Interface (HMI) is urgently required for safer driving during the interaction with automated systems. The driving assistance system alleviates the attention required (e.g., SAE Level 2) or even allows users to shift the role from driving to monitoring (e.g., SAE Level 3 and higher) during the automated driving. However, the design of the HMI, serving as the primary element during human-AV interaction, is not always understandable, which might impact users' understanding of the automation system and its capabilities. Another critical factor for automated driving safety is the workload. If the AV HMI is complicated and difficult to understand, the workload of the users during the interaction would increase and lead to misunderstanding of the HMI and situation. A high workload could also lead to delayed responses and impair the situation awareness of the users, which increases the risk during automated driving and influences the trust and acceptance in the automated system. Hence, it is crucial to design and estimate an HMI in a way that clear and necessary information is provided without resulting in a high workload during the interaction. In this chapter, the importance of understandability and its operationalization will be reviewed in detail. Different workload measures will also be compared and discussed.

3.1 Understandability

Common methods to operationalize the psychological construct like FT is the inclusion of measures that could probe the understanding or mental model of the users through survey, behavioral or performance outcomes, as well as psychophysiological measures. One of the main concepts that FT attempted to capture is the degree to which the measured HMI is able to transfer the critical information to users, and this requires the HMI to be understandable. In the field of Computer Science, system or code understandability has been widely studied Bansiya and Davis (2002); Capiluppi, Morisio, and

Lago (2004); Lin and Wu (2006); Misra and Akman (2008); Scalabrino et al. (2019); Srinivasulu, Sridhar, and Mohapatra (2014).

As the complexities of the code structure or system go high, the understandability and maintainability will be deprecated. Misra and Akman (2008) compared different cognitive complexity measures and determined that the cognitive weight complexity measure stood out to meet the basic requirements proposed by Weyuker (1988), which listed nine properties to be satisfied. In the cognitive weight complexity measure, each control structure component was evaluated and weighted before the summation into the system complexity measurement.

The understandability of a code system could also be decomposed and deemed as the total effect of these smaller factors. Capiluppi et al. (2004) defined a method to estimate the understandability of open-source projects, where the projects were decomposed into factors such as macro-module (i.e., the first order grouping of the source code), micro-model (i.e., files in the whole project), and the size of the code. The researchers later defined the indistinctness, overlooked content, and micro-module size. These attributes were calculated, and the unuderstandabilities of 19 projects were evaluated, and found that the understandability of these projects tend to increase during their life cycles.

Understandability is also a multifaceted and complex construct that might be affected by the characteristics of the person interacting with it. Similar to code readability, the subjectivity of understandability makes personal taste, personal habit, and previous experience and mental model of the developer critical factors (Storey, 2005). Scalabrino et al. (2019) correlated existing metrics and developed a model to estimate understandability by measuring the perceived and actual understandability of participants. The perceived understanding could be represented by a binary *true* or *false* answers, depending on whether developers perceived themselves that the code was understandable. As for the actual understandability, it was operationalized by answering questions about the code they perceived as understandable.

Despite the ideal of understandability being majorly discussed in the field of Computer Science, this construct could be feasible for the measurement of Functional Transparency (FT). With FT, we are interested in how well users comprehend the information transmitted by the automation system, and what we really want to know, is the ”true” understanding of the system state that is learned by the users, as this would be of the utmost criticality to the driving safety. In many cases, as we see in Chapter 1, users perceived themselves and believed that they understood the states of the automated driving system, while the understanding was incorrect or not comprehensively, and neglected their responsibilities to monitor while the system is engaged. This discrepancy between the ”true” and ”perceived” understanding of the system will haunt driving system automation until the penetration rate of the fully autonomous vehicle reaches 100 percent.

However, if there exists such a method to allow researchers or HMI designers to assess this gap between the ”true” and ”perceived” understanding of the given automated driving system, the driving safety while the automation is engaged could be substantially increased, and the design process could also be more efficient. In Computer Science, the true understanding of the code can be estimated with a question set regarding the function and the targeted alternations the code intends to achieve. Similar comparisons could also be seen in the Human Factor of vehicle automation. Since the main purpose of the existence of HMI is to help human users understand the states of the automated driving system correctly, the questions to estimate the "true" understanding could, as a consequence, be those regarding the current states of the automated vehicle, for example, the longitudinal and lateral states (Beggiato et al., 2015; Bengler et al., 2020; Naujoks et al., 2019). This result in "true" understanding could be analogized to the objective outcome, as it is measured objectively with true references (i.e., answers), but it does not necessarily deny the importance of the subjective ones. The subjective understanding measures could offer a hint about the mental model of users to the system faced in a certain scenario. The consistency between the results of objective and subjective understanding (i.e., "true" and "perceived" understanding) could be regarded as the expected and desired result in the case of HMI design of the automated vehicle. The inconsistency between the results of objective and subjective understanding, on the other hand, give us a "false negative", which provides us with valuable information on what information or information type could be potentially confusing to the users, and prone to erroneous or dangerous behaviours.

3.2 Workload Estimation

Human workload is arguably one of the most important topics in the field of Human Factors and Ergonomics (Longo, Wickens, Hancock, & Hancock, 2022; Nygren, 1992; Van Acker, Parmentier, Vlerick, & Saldien, 2018; Wickens, 2008). Starting to be recognized for its importance in aviation research, the method and its adaptation have been extended to the use of general driving scenarios (Brookhuis, van Driel, Hof, van Arem, & Hoedemaeker, 2009; Cantin, Lavallière, Simoneau, & Teasdale, 2009; De Waard & Brookhuis, 1996; Paxion, Galy, & Berthelon, 2014; Recarte & Nunes, 2003) and now in the field of automation (Bueno et al., 2016; Kaduk, Roberts, & Stanton, 2021; Y.-c. Liu et al., 2023; Radhakrishnan et al., 2023; Stapel, Mullakkal-Babu, & Happee, 2019). Mental workload refers to the requirements of a certain task imposed on the cognitive, perceptual, and motor efforts required by the person performing such task. It is known to be a complex and multifaceted construct that could be influenced by various factors, such as the nature of tasks, the state and capabilities of participants, interacting scenarios, etc. The primary objective of using such a metric is to optimize the performance and experience of users and attempt to mitigate potential negative influences like stress, frustration, and mental overload to minimize errors during the tasks.

In the oncoming trend of wide adoption of automated vehicles, human users will be eligible to savor the journey with the automation system, which is expected to alleviate their workload during the ride. One of the difficulties faced during the process of vehicle "automationalization" lies in the fact that there would be a transition from the automated driving system to human users when the system limit is reached. This huge wall will continue to hinder the use of automation as long as the full self-driving vehicle is unavailable. Such a transition from simply monitoring the automated system, where the workload should be lower, to perceiving the environment, understanding the states of the automated system, and controlling the vehicle, where the workload should be higher, is of great importance to driving safety.

Researchers have been trying to bridge the gap from various perspectives, and one of them is through the research of the Human-Machine Interface (HMI), as it works as the only communicating tunnel between the automated system and the human users. Hecht, Kratzert, and Bengler (2020) studies the effect of transition frequencies on various constructs, including workload, of highly automated vehicles in the driving simulator, and different "predictive" HMI were designed and used. Significant differences in workload among different transition frequencies were found, while the effect of the HMI concepts was not significant. Now, we understand that various task frequencies or types during the interaction would have an effect on the workload, but we should also note here that the relationship could be mutual. Different levels of workload could also affect performance during driving. In the studies of Bueno et al. (2016), this time, the workload was the independent variable, which was manipulated to investigate its effect on performance during the transitions. Although the effect of different workload levels was not significant, the negative effects of such differences on driving performance were discovered.

To increase the fidelity of the research regarding workload and the use of automated driving systems, studies were also carried out in the real-driving or test-track setting. In the study of Stapel et al. (2019), a series of on-road experiments were carried out to identify the impact of automated driving systems as well as user experience in automated systems on the workload. The results showed that experienced automation users perceived lower self-reported workload, while the non-experienced participants found little difference. However, the objective evaluation of the workload revealed that higher workload loads were actually experienced when interacting with the automated system, especially during complex driving scenarios. These opposite results not only showed the potential issues of the possibilities to increase the workload despite the use of the automated driving system but also emphasized the importance of estimating with both subjective and objective measures for a more comprehensive view.

Figalová et al. (2024) also adopted both subjective and objective workload measures to identify the effect of different levels of the automated driving systems on the workload measures. Subjectively, participants in the study perceived no difference between manual driving and the SAE Level 2 automated driving system, which might be owing to the fact that the participants were novice users of the automated system and, hence, did not immediately delegate the control duty to the automation system. This could also be a potential issue of the Functional Transparency (FT) of the system used, which is not transparent and hence could not alleviate the workload of the users. On the other hand, the objective workload measure with EEG showed the lack of attention allocation on the road for both SAE Level 2 and SAE Level 3 automated driving systems, suggesting a miscommunication between the HMI and human users, resulting in the negligence of carry out the duty of being supervisors.

With the inclusion of both subjective and objective workload measures, we have more perspectives from various angles to evaluate the automated driving system and the HMI and utilize the results to enable a safer and easier automated driving experience. In the following sections, subjective and objective workload measures commonly used in the field of vehicle automation will be discussed in detail.

3.2.1 Subjective Workload Measures

The subjective workload often refers to the mental and physical demands of an individual while performing the task. Depending on the complexity of the task, corresponding mental and physical resources have to be allocated by the individual. However, since these resources are limited, individuals might experience higher effort and strain when this demand in certain perception modalities rises, resulting in an increased workload. Given that effort and strain are qualities experienced by the subject itself, the subjective rating should work as a direct and sensitive measure for workload estimation. Nasa-Task Load Index (NASA-TLX) is a multidimensional scale designed to estimate the workload while performing a certain task, and it is often used immediately after conducting the task to mitigate the effect of short-term memory loss. This popular technique assesses workload using six sub-scales, which are mental demand, physical demand, temporal demand, performance, effort, and frustration level.

The workload experienced by participants while interacting with various systems has been investigated using the NASA-TLX, including warning systems and automation systems, and mostly regarding visual and auditory displays (Hart, 2006). The generation of automation technologies is aiming at its power to reduce human operation workload, so it is not surprising that the workload during the interaction with automated vehicles would draw interest.

W. Chen, Sawaragi, and Horiguchi (2019) investigated the relationship between workload and different levels of automated driving systems in a driving simulator while performing designated secondary tasks. The driving scenario was simply a loop on a highway, and after each trial, participants were requested to evaluate their workload with NASA-TLX. As expected, the subjective workload measured with NASA-TLX during driving with the lowest level (i.e., without adaptive cruise control or lane-keeping assist) was significantly lower. However, similar subjective workload measurements were shown between SAE Level 1 (i.e., with adaptive cruise control only) and SAE Level 2 (i.e., with both adaptive cruise control and lane-keeping assist) automated driving systems.

In order to understand the performance of manual driving before and after the usage of the automated driving system, Kaduk et al. (2021) conducted a simulator study and used NASA-TLX to estimate the workload during each task. The performance during manual driving was estimated with factors such as lane position, steering angle, heading angle, and longitudinal speed. The results showed that the manual driving performance was significantly lower after the usage of automation compared to that before the automation usage. The subjective workload measurements with NASA-TLX also indicated that a higher workload was required after manual driving, while the demand was lower after the engagement of the automated driving system.

One of the perks of the automation systems is that human users are allowed to engage in Non-Driving Related Task (NDRT) during highly automated driving. However, users are required to take back control whenever the limit of the automated system is reached.

In this regard, Yoon and Ji (2019) conducted a driving simulator study to investigate the effects of different NDRT on the takeover performance during highly automated driving. Despite no difference being found in the reaction time to take back control among various NDRT, the results on subjective workload demonstrated a significant effect of these tasks, where taking back control while watching video would induce significantly higher subjective workload. Data of the NASA-TLX was collected after each session, where participants were asked to perform a certain type of NDRT. Authors of this study also further investigated the sub-scales of NASA-TLX comparing to the overall workload scores.

Besides the application in the simulator environment, NASA-TLX could also be adopted in real driving scenarios. Heikoop, de Winter, van Arem, and Stanton (2019) measured the workload and stress of human users during partially automated driving in an on-road study. Generally, among participants, the overall workload demand required while engaging in automated driving with real traffic was low, and the mean workload was lower compared to studies carried out in the simulated environment. Due to the real driving setting, the NASA-TLX was provided not directly after the driving and tasks during the motorway but was given after stopped in a nearby parking lot.

Subjective workload measures offer a direct measurement of the efforts experienced by participants during the task. It is easy to be deployed and robust with different scenarios and driving settings. However, one of the biggest drawbacks is that it can not be measured continuously. What human drivers and automation users are facing during driving is a dynamic environment. The rises and falls in workload should be constantly changing due to this nature. The perceptions of the driver or automation user are constantly updated upon receiving new information from the environment, which makes the subjective workload measure unable to capture a comprehensive view during the dynamic change. Objective workload measures, on the other hand, provide continuous workload evaluation throughout the study, providing another perspective of the participants' workload.

3.2.2 Objective Workload Measures

Objective workload measures, compared to subjective workload measures, are quantifiable traits that could be captured objectively without input from the participants. In the driving scenarios, the most commonly used objective workload measures could be categorized into two categories, which are involuntary response, typically with psychophysiological measures (Figalová et al., 2024; Y.-C. Liu, Figalova, Baumann, & Bengler, 2023; Y.-c. Liu et al., 2023; Lohani, Payne, & Strayer, 2019; Radhakrishnan et al., 2023; Shakouri, Ikuma, Aghazadeh, & Nahmens, 2018), and voluntary response, typically with reaction time or performance data (Melnicuk, Thompson, Jennings, & Birrell, 2021; Pouliou, Kehagia, Bekiaris, Pitsiava-Latinopoulou, & Poulios, 2022; Pouliou, Kehagia, Poulios, Pitsiava-Latinopoulou, & Bekiaris, 2023). Voluntary response or control movement offers a direct correspondence between the action and the workload, while involuntary physiological responses, on the other hand, are usually difficult to interpret as the signals are continuous, and some come with the nature of being delayed responses.

However, the psychophysiological workload measures have been well established in the driving scenarios, providing accurate and real-time information regarding the change in the workload of participants, as we can see in Section 3.2.2.1.

3.2.2.1 Psychophysiological Workload Measures

Psychophysiological indicators such as heart rate, skin conductance, and brain waves are able to be associated with the increased demand for mental task efforts and are found to be capable of increasing the accuracy in detecting mental workload when such data are utilized (S. Yang, Hosseiny, Susindar, & Ferris, 2016). Compared to voluntary responses like task performance and reaction time, psychophysiological measures are less affected by undesired or uncorrelated constructs and individual preferences, and thus, the results tend to be more representative of the targeted task workload estimation (Shakouri et al., 2018).

The brain controls how individuals think, move, and feel by sending and receiving chemical and electrical signals. Hence, capturing where and when these signals happen could provide us with valuable information about the perceptions, emotions, and even thoughts of humans. To achieve this, Electroencephalogram (EEG) records both the oscillatory and aperiodic electrical activity in the brain. By connecting electrodes to specific areas on the skull, EEG could provide brain activity data with high temporal resolution, allowing real-time electrical activities for mental workload assessment.

The variation in workload during driving will be reflected in the alternations of certain frequency bands in the EEG recording, and these changes between workload and EEG signals have been reliably associated with (Käthner, Wriessnegger, Müller-Putz, Kübler, & Halder, 2014; Zander et al., 2017). The workload intrigued by the task affects the alpha and theta power bands in different directions. During over-arousal states during driving, such as when driving contexts or environments are complicated, the increase in workload is reflected on these power bands in the way where the alpha power decreases and the theta power increases. The influences of workload on these power bands could also be evaluated jointly. Borghini et al. (2015) used ratios of frontal theta and parietal alpha power spectral density to evaluate the workload of helicopter pilots during simulator operation.

In the real driving context, the use of EEG has also been widely adopted to estimate the workload of automated system users. To investigate the effect of rising cognitive demand during certain driving scenarios and to provide prompt information to the drivers, Borghini et al. (2012) introduced an EEG-based workload index to identify different mental workload levels during various driving tasks. Results showed an increase in the power spectra of the theta band and a simultaneous decrease in the power spectra of the alpha band during difficult drive conditions, indicating the feasibility of adopting EEG in the context of driving scenario to identify differences in workload, which also has the potential to assess the mental states of driving in real-time during the driving task. Similar results had also been found during the engagement of Adaptive Cruise Control (ACC). In a recent study, Acerra et al. (2019) had found the increase in mental workload using the measurement of EEG, which is in accordance with the increase in the perception-reaction time.

With high temporal resolution, EEG offers real-time data for the brain's electrical activities during driving scenarios, enabling the dynamic recording of drivers' workload, which could help gain more insights compared to the post-event subjective workload evaluation methods. Besides EEG, other psychophysiological measures like Electrocardiography (ECG) based measures have also been widely adopted and explored in measuring the workload of human users during driving scenarios, which is considered a reliable measure in estimating driver's status (Heine et al., 2017; Mulder, de Waard, & Brookhuis, 2004; Reimer, Mehler, Pohlmeyer, Coughlin, & Dusek, 2006; S. Yang et al., 2016). With steady heart beats, typical ECG pattern is composed of P, Q, R, S waves (also known as QRS complex), which could be used to identify the rate and rhythm of the heart. Heart rate is generally considered as in beat-per-minute, or bpm, and would rise with the cognitive load of the task Heine et al. (2017); Mulder et al. (2004). The differences in age also have effects on the heart rate pattern change during increased task load, where younger drivers (19-23 years old) had higher amounts of increases when facing tasks with higher cognitive loads (Reimer et al., 2006).

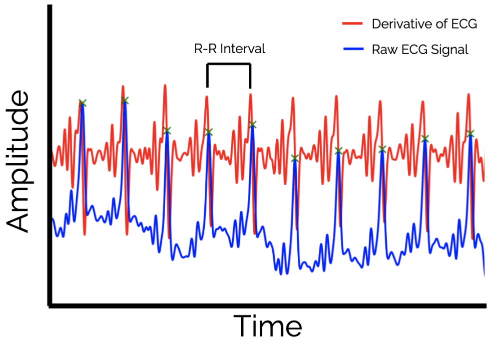


Figure 3.1: Illustration of the raw ECG signal and processed R-R interval.

Another effective way of using the ECG data is to analyze the variability of heart rate intervals within a certain time duration, ranging from short (one to five minutes) to long (up to 24 hours) duration. Metrics derived from such analyses are referred to as Heart Rate Variability (HRV), and could roughly be characterized into time domain and frequency domain. Time domain analyses such as standard deviation of normal R-R intervals (SDNN) and root mean square of successive R-R intervals (RMSSD) calculate the standard deviation and root mean square of the variations of R-R intervals during the designated time intervals (R-R interval is as shown in the Fig. 3.1). Frequency domain measures, on the other hand, analyze the power of different frequency bands of heart rate by transforming the heart beats intervals from time to frequency domain using Fourier transformation. The powers of the lower frequency band (0.04–0.15 Hz) and higher frequency band (0.15–0.40 Hz) are calculated as low-frequency (LF) power and high-frequency power (HF) and are used independently to estimate the workload during driving scenarios (Heine et al., 2017; Reimer et al., 2006), and sometimes jointly using the ration of LF and HF (Shakouri et al., 2018; Tjolleng et al., 2017).

Heart Rate Variability (HRV) was concluded to be sensitive to increased workload and raised awareness of the situation faced (Jasper, Sibley, & Coyne, 2016; Lohani et al., 2019). In the driving simulator setting, a higher workload was identified and reflected on the RMSSD and SDNN, confirming the validity of the usage of HRV analysis in detecting differences in workload using ECG signals (Heine et al., 2017). A similar trend of the HRV metrics (SDNN and LF/HF ratio) was found between simulator studies and on-road studies, where the same car following scenarios and the same equipment were applied, showing the validity of the ECG usage among different driving settings (Heikoop et al., 2019).

Another common way to estimate real-time workload is through the usage of Electrodermal Activity (EDA), where electrical resistance is measured by applying a constant and small current through a pair of electrodes fixed on a certain section of skin. This measure could provide a continuous state of skin conductance derived from the skin resistance, which alters according to the workload levels of participants. As we can see in Fig. 3.2, after the stimulus is introduced, a short latency period follows before the rising phase of the SCR. The magnitude of the SCR rising from the baseline is calculated as the SCR amplitude, which is associated with many physiological responses, such as mental efforts during unexpected events (Dawson, Schell, & Filion, 2007).

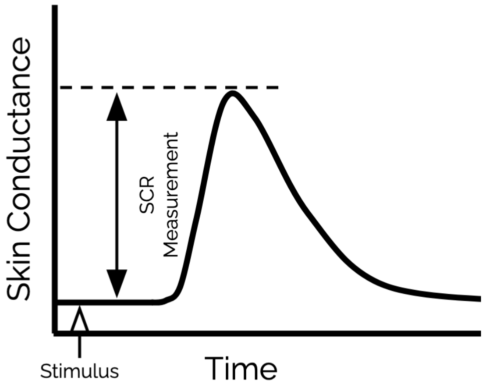


Figure 3.2: Illustration of the EDA signal and the skin conductance response (SCR).

In the driving context, SCR and SCL measures have been shown to increase with stress and workload while driving in simulated environments (Foy & Chapman, 2018; Radhakrishnan et al., 2022). Multiple factors, including car following scenarios (automated or manual), levels of automation engaged (SAE Level 2 or SAE Level 3), or distance to the front vehicle, were manipulated to estimate the difference in the driver's workload level. Results show significant effects of all the factors (car following type, automation level, and distance to front vehicle) using the metric derived from Electrodermal Activity (EDA), suggesting that the EDA signal is sensitive to the difference in workload in the driving context (Radhakrishnan et al., 2023).

Several metrics could be derived from EDA signals to estimate the workload of users during driving. Perello-March, Burns, Woodman, Elliott, and Birrell (2021) used SCR counts, SCR amplitude, and SCR magnitude to verify whether scenarios with varying traffic complexities would have effects on the physiological responses. The effects between high-complexity urban traffic scenario and low-complexity urban traffic scenario with Non-Driving Related Task (NDRT) were shown to be equivalent using SCR counts, while the highest values among all the experimental conditions were found among SCR amplitude and magnitude values.

3.2.2.2 Reaction Time

Reaction time has been associated closely with driving safety in real traffic settings as well as during automated driving (W. Chen, Sawaragi, & Hiraoka, 2022; Dozza, 2013; Müller, Fernandes-Estrela, Hetfleisch, Zecha, & Abendroth, 2021; Pawar & Velaga, 2020; Shi & Bengler, 2022). Users of the automation have to react in a timely manner to avoid critical situations, especially when the automation technologies are not fully developed to handle various scenarios (Dargahi Nobari, Albers, Bartsch, Braun, & Bertram, 2022; Dixit, Chand, & Nair, 2016; Morales-Alvarez, Marouf, Tadjine, & Olaverri-Monreal, 2021; Paula et al., 2023). A study of Dixit et al. (2016) has also pointed out factors that are highly correlated with the reaction time, including the disengagement types, roadway layouts, and autonomous miles traveled (i.e., automated driving experience). The cognitive model of users during driving is often with the following stages: sensory perception, decision-making process, and motor reaction (Dargahi Nobari et al., 2022). Researches also reveal a high correlation between workload and reaction time in normal driving settings as well as simulator studies (Cantin et al., 2009; Dozza, 2013; Makishita & Matsunaga, 2008; Pawar & Velaga, 2020).

Longer reaction time is often associated with higher workload during the operation. In some cases, the increased workload induced by certain events was applied to evaluate the effects of workload on the reaction time. In one of the examples, researchers used mental calculations to manipulate workload and found that the reaction time increased significantly with the increased workload in both scenarios where participants were sitting in a static vehicle and driving in simulated environments (Makishita & Matsunaga, 2008). Effects across each age group were also found to be significant (young drivers: 20–29 years old; middle-aged drivers: 41-54 years old; elderly drivers: 61–64 years old.)

In other cases, the reaction time was used as the probe to estimate the workload in different driving settings. Various road layouts in the urban as well as rural sections often come with different complexities, which induce momentary workload during driving. To estimate how these stimuli, together with the effects of age, may affect the momentary workload, a study was carried out in the simulated environment (Cantin et al., 2009). The results revealed that higher reaction time to the auditory stimulus, which suggested a higher momentary workload, was found during more complex driving scenarios such as the intersections. The results also revealed that the error rates during driving were proportional to the complexities of the driving layouts, which is highly influenced by the increased temporary workload.

Visual signals were also applied in studies as stimuli to initiate the process of reaction time estimation (Chowdhury et al., 2020; Droździel, Tarkowski, Rybicka, & Wrona, 2020). A red signal was designed to indicate the initiation of moving the right foot from the gas pedal to the braking pedal while estimating the distance of the brake to correlate to the reaction time from perceiving the signal to the finish of the movement (Droździel et al., 2020). Chowdhury et al. (2020) used the alternated visual signal as the trigger to test the reaction time the participants required to press the space bar, where the results of the Electroencephalogram (EEG) signal were used to construct a predictive model.

In the automated vehicle setting, similar relationships and effects of workload on the reaction time were also discovered, where visual stimuli also play a significant part as one of the Non-Driving Related Task (NDRT) in the experiments to induce the rising of workload, which is often used in studies to manipulate differences in workload (Müller et al., 2021; Yoon & Ji, 2019). Müller et al. (2021) found differences in workload among the NDRT selected during automated driving, where each was associated with different physiological modalities. They also found a high correlation between workload and response time, where NDRT connected with a higher workload, i.e., reading and texting, leading to longer reaction time, and the consistent results were found across subjective workload measures, psychophysiological measures, and performance measures. Similar results were again found in the work of Yoon and Ji (2019), where the visual stimuli were found to be positively correlated with the take-over time in a highly automated driving setting. The study also indicated that the subjective workload significantly rose after taking back control from the automated vehicle after highly automated driving and negatively influenced driving performance.

4 Summary and Objectives

During the interaction between human users and Automated Vehicle (AV), Human-Machine Interface (HMI) functions as the only media between the two agents (i.e., human users and automation system), making it of utmost criticality to ensure the understanding of the automation system so that the driving safety could be guaranteed, as detailed discussed in Chapter 1. There have been multiple HMI design guidelines existing to provide heuristic suggestions in improving user experiences when engaged in automation systems. Several HMI evaluation methods focusing on usability, trust and acceptance, and comfort, as described in Chapter 2, do provide some insights into these constructs during automated driving. However, there has not been a standardized and systematic assessment method for automated driving systems that could provide objective metrics and evaluation regarding whether human users could understand the automated system with minimum effort.

To bridge this gap, a novel concept for Functional Transparency (FT) was delivered in Chapter 2, where the definition of this construct was discussed in detail to facilitate a standardized and more efficient HMI evaluation method. Given the definition of FT, the understandability of the automated system and HMI would be one of the most important factors to evaluate the FT. Despite being a complex and multifaceted construct that is often evaluated subjectively, objective evaluation methods were considered robust in understanding code systems in the field of computer science. In Chapter 3, understandability evaluation methods were discussed in depth, followed by comprehensive research in workload estimations, as workload represents yet another crucial factor in the definition of FT, which has to be minimized to guarantee minimum effort when interacting with the automated driving system.

In the article discussed in Chapter 5, a transparency assessment method for Level 2 (SAE Level 2) automated vehicles was proposed and verified. This provides a strong basis for further research in estimating the Functional Transparency (FT) of automated driving system, which could be further utilized to improve the efficiency and standardize the HMI design and evaluation processes. The proposed transparency assessment method was then validated in a simulated environment, as discussed in Chapter 6, where the inclusion of psychophysiological measure (i.e., the inclusion of Electroencephalogram (EEG)) was performed to evaluate the sensitivity and applicability of such objective measures in evaluating automated vehicle HMI. Aiming to propose a systematic and standardized transparency assessment method, in Chapter 7, experiments with inclusions of more physiological workload measures were conducted to explore satisfactory objective workload measures to enable the standardization and efficiency of the proposed transparency assessment method.

Motivated by the challenges faced and the research questions, this thesis aims to examine and achieve the following objectives :

- Develop a standardized and objective transparency assessment method.
- Validate the proposed transparency assessment method in more dynamic scenarios (i.e., driving simulator studies).
- Identify psychophysiological measures that are sensitive in estimating workload when evaluating Automated Vehicle (AV) Human-Machine Interface (HMI).

5 Article 1: “Transparency Assessment on Level 2 Automated Vehicle HMIs“

Liu Y.-C., Figalová N., & Bengler K. (2022). Transparency Assessment on Level 2 Automated Vehicle HMIs. *Information, 13*(10):489.

In order to address the need for human users to understand the automated system they are interacting with, a standardized assessment method for automated vehicle transparency is inevitable. Motivated by this, the definition of such a construct has to be first determined.

A novel approach to evaluating the understandability of Human-Machine Interface (HMI) in automated vehicles was proposed, where the definition went beyond the traditional transparency definition. The legacy definition of transparency only defines it as the quantity and levels of information provided and neglects the true user comprehension of the automated system, which is sensitive to user characteristics and driving scenarios. To genuinely reflect the true understanding of human users of the automated system, the proposed Functional Transparency (FT) was defined as ”how easy it is for users to understand the automated driving system correctly”. The proposed FT included the understandability and workload estimation to objectively consider the true comprehension of human users toward the current states and the efforts consumed when interacting with the automated driving system.

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

Figure 5.1: Definition of Functional Transparency developed in Article 1 Y.-C. Liu et al. (2022). AU stands for actual understandability; $TNPU$ stands for time needed for perceived understanding.

The transparency (functional transparency) assessment method was verified using an online survey, where various HMI designs available on the market were adopted to investigate their effects, as well as user experiences in automated driving systems on the resulting functional transparency. The results suggested the applicability of the proposed method in evaluating the understanding of users toward the automated driving system and found the effects of user experiences to be significant, which was aligned with the study of Mueller, Cicchino, Singer, and Jenness (2020). The results also showed the potential of identifying the critical information items on the instrument clusters on the HMI designs that were pivotal in affecting the understanding of the automated driving system of human users, which could, in return, facilitate correct understanding and minimize user efforts.

By focusing on functional transparency and developing a standardized assessment method, the proposed method found a solid basis for future studies in the development and adoption of automated driving system technologies, especially regarding the topic of AV HMI designs and automation training and educational programs, to guarantee that the users are well informed and understand the provided HMI, and ultimately contribute to safer and more efficient automated driving.

6 Article 2: “Human-Machine Interface Evaluation Using EEG in Driving Simulator“

Liu, Y.-C., Figalová, N., Baumann, M., & Bengler, K. (2023). Human-machine interface evaluation using eeg in driving simulator. In *2023 IEEE intelligent vehicles symposium (iv)* (pp. 1–6).

Building on the previous study, this experiment aimed to validate the proposed transparency assessment method in a driving simulator to explore the applicability of the method in the environment with higher fidelity. In the previous study, Functional Transparency (FT) was validated with an online experiment, where the workload was estimated by the time to understanding. This evaluation method could provide objective measurement on how easily the Human-Machine Interface (HMI) designs were understood; however, it is difficult to be applied in an environment requiring higher fidelity since that real-time data would be required to estimate the workload in a more dynamic condition.

To observe real-time and continuous workload measurement, Electroencephalogram (EEG) is a common psychophysiological measure used to estimate driver workload in driving scenarios (Lohani et al., 2019). Three HMI designs were also developed based on the results from the previous study (Y.-C. Liu et al., 2022) to intrigue differences in FT and estimate the effect of FT to workload evaluation using EEG.

Results from the study showed that the Trans HMI design (Transparency HMI design, derived from a previous study and expected to be understood with the least effort among the HMI designs) brought about the highest alpha power and lowest theta power, aligned with the literature to have the lowest workload to understand, despite the fact that the difference was found not statistically significant. Subjective measurements, which were NASA-TLX test and subjective transparency questionnaires, showed the same trend where the Trans HMI design had the lowest averaged NASA-TLX score and highest subjective transparency score, suggesting that the trans HMI design was the most understandable one with least workload to interact with. The differences from both subjective evaluation methods (i.e., NASA-TLX and subjective transparency test) were significant.

These findings suggest that the Trans HMI design could be more effective in conveying information clearly and efficiently to users with minimum effort, and the differences found among the HMI designs indicate the potential that these HMI designs developed could be applied in future studies to evaluate the effectiveness of other psychophysiological measures. Despite the spectral power of EEG signals showing no significant result, consistent results on the workload measurements of HMI designs were found. Further research is needed to identify a more sensitive, robust, and objective real-time measure with continuous data.

7 Article 3: “Workload Assessment of Human-Machine Interface: A Simulator Study with Psychophysiological Measures“

Liu, Y.-C., Figalová, N., Pichen, J., Hock, P., Baumann, M., & Bengler, K. (2023). Workload assessment of human-machine interface: A simulator study with psychophysiological measures. *Human Systems Engineering and Design (IHSED 2023): Future Trends and Applications, 112* (112).

Continued to the last study and in order to facilitate a standardized assessment method for Automated Vehicle (AV) Human-Machine Interface (HMI) in a higher fidelity environment, Electrocardiography (ECG) and Electrodermal Activity (EDA) were adopted for the fact that they play a crucial role in evaluating workload in driving context due to their sensitivity and accuracy in detecting changes in cognitive states (Foy & Chapman, 2018; Heine et al., 2017).

Results from both ECG and EDA metrics were effective in identifying significant differences in objective workload among different HMI designs in a simulator study. More specifically, the Root Mean Square of Successive R-R Intervals (RMSSD) derived from the Heart Rate Variability (HRV) was found to be the highest on the Trand HMI design, indicating that the HMI design demanded the lowest workload among the three HMI designs. Similarly, the Skin Conductance Response (SCR) derived from the EDA can also capture the dynamic changes in workload, showing significantly lower SCR values on the Trans HMI design.

Subjective evaluation methods for workload were also applied and found to align with the objective psychophysiological measures. Significant differences were found among HMI designs, where both NASA-TLX scores and self-reported subjective transparency scores results indicated that the Trans HMI design demanded the lowest workload during the interaction, followed by the Trans-fog HMI design, and lastly, the Fog HMI design. These results also aligned with the previous study where the Trans HMI design was found to be the most transparent and required minimum workload to be understood.

The results demonstrated that both ECG and EDA are valuable tools for objectively measuring mental workload resulting from monitoring and interacting with different HMI designs. These measurements of the physiological responses offer a more accurate and continuous assessment of workload compared to traditional subjective methods, where the workload is challenging to detect using subjective or behavioral measures alone. The findings could facilitate a more robust transparency assessment method in dynamic automated driving scenarios, which have the potential to contribute to future developments of a more transparent and understandable HMI design and a more effective and systematic HMI design process.

8 General Discussion and Future Works

In the following chapter, general discussions regarding the proposed Transparency Assessment (TRASS) will be provided, including the estimation of the understandability in Section 8.1, the workload estimation in Section 8.2, and finally, the potential contribution of the proposed TRASS will be discussed in Section 8.3.

8.1 Understanding of the automated vehicle human-machine interface

Transparency of automated driving systems has been shown to be a critical topic in the field, as discussed in Chapter 2. However, there has yet to be a standardized and systematic assessment method to estimate the understanding of human users of the automated driving system. In Article 1, a novel Transparency Assessment (TRASS) was proposed and further validated in Article 2 and Article 3, providing a more efficient path to evaluate an Automated Vehicle (AV) Human-Machine Interface (HMI).

One of the main factors influencing the TRASS value is the understandability that the human user possesses toward the automated driving system. Hence, the assessment of such a construct was derived from the fundamental function or objective of AV, which is to transport humans from point A to point B safely. To achieve this, the teaming and duties allocation between the automated system and human users has to be correct and clear, i.e., transparent, as discussed in detail in Chapter 1 and Chapter 2.

The information transferred through the HMI is critical during automated driving before they are fully autonomous. In the traditional definition of transparency, provided with more information or information with a higher level of transparency is considered more transparent. However, too much or even too little information could contribute to a false perception of the system state or an unnecessarily higher workload during the interaction (Beggiato et al., 2015; Kraft et al., 2018).

Hence, the proposed TRASS focuses not on the quantity of the information but the quality and the resulting perception and understanding of the AV HMI after the interaction, which is defined as the Functional Transparency (FT). The word "functional" represents the functional part of the information that could contribute to the transparency and understanding of the automated driving system through AV HMI. To facilitate safe driving, the evaluation standard for the correct understanding of the automated driving system is thus a correct understanding of the system states and the surrounding agents. In Article 1, the following questions regarding critical information to obtain current automated driving system states for driving safety were asked:

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?

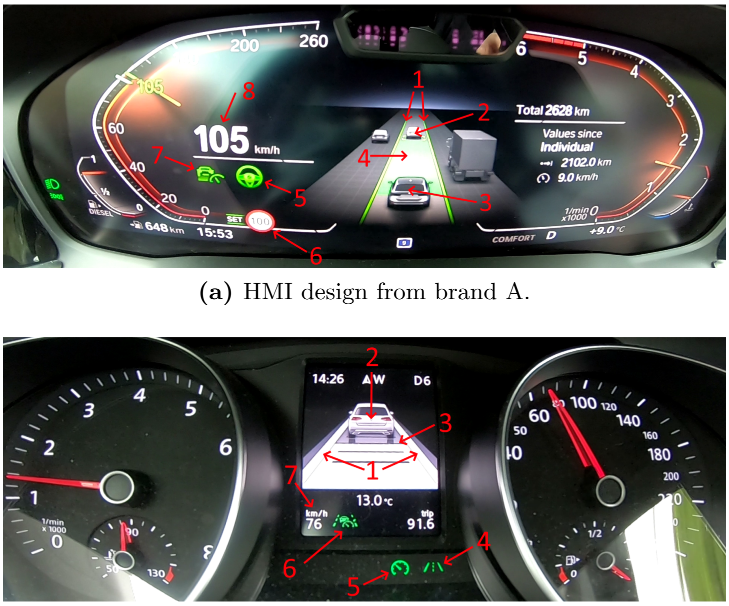
5. Can you activate the automated driving assistance (which performs both longitudinal and lateral controls automatically)

The indication of the front vehicle and lane marking detection are important as they showcase the availability of automated system engagement, providing drivers with enough situation awareness and future projection of the possible outcome when the automated driving system is activated. The potentially dangerous maneuvers could occur if human users possess the wrong conception and understanding of the automated driving system. In the simulator study conducted in Article 2, large deviations from the center line were shown when participants falsely believed the automated driving system was activated, but in fact, it was not and was in standby mode, resulting in zero input to the steering wheel when correction maneuver should be performed. This lateral deviation was considered highly correlated to fatigues and risky maneuver (Ting, Hwang, Doong, & Jeng, 2008) and was considered a safety factor during driving scenarios (Suh, Park, Park, & Chon, 2006).

In order to answer these questions demonstrated to evaluate the understanding of the user toward the HMI interacting with, information icons on the interface design are used to help the user judge the state of the automated driving system. However, not all icons on the HMI provided clear and understandable information about the system. In one section of the questionnaire, participants were surveyed about which information icons they used to understand the state of the automated driving system and answer understandability questions, as illustrated in Fig. 8.1. The results identified that redundant icons or indications might, in fact, confuse the automation users. For instance, the icon labeled with the number five in Fig. 8.1b will be illuminated when the Adaptive Cruise Control (ACC) was activated and was kept on when the SAE Level 2 automated driving assistance was engaged. Some participants misunderstood this icon as an indication of whether the automated driving assistance was activated, while others reported that this icon confused them when they tried to ascertain if the automated driving assistance was activated and thought that they failed to activate it and were still in ACC mode.

The findings could be used to assist future researchers in identifying information icons and designs that could potentially lead to misuse and confusion. Combining this with the proposed TRASS, models could be constructed by utilizing the corresponding TRASS scores to each information icon, and help to identify critical information icons required for safe driving in automated driving system.

The understandability questions could also be adapted to different levels of automated driving systems, as the critical information needed for safety among these scenarios might be slightly different (Beggiato et al., 2015; Lu, Happee, Cabrall, Kyriakidis, & De Winter, 2016). For example, different automated initiated take-over requests in higher levels of



(a) HMI design from brand A.

(b) HMI design from brand B.

Figure 8.1: Illustration of icon identification with HMI designs from various brands, where the red numbers are indications of icons demonstrating different states of automated driving systems. (Y.-C. Liu et al., 2022)

automated driving systems might have various purposes. In some scenarios, the take-over request was initiated to raise the attention of the human users to regain their situation awareness and bring them back into the loop to minimize the risk (Gonçalves & Bengler, 2015). In other scenarios, take-over requests were initiated in critical scenarios where the situation was out of the capabilities of the automated driving system to cope with. Hence, input from human users was mandatory. Different information and indications should then be provided so that human users' trust and perceived safety would not be compromised.

In conclusion, the understandability questions the proposed TRASS used could be adopted to help identify users' true understanding of the automated driving system. The proposed method probed the degree to which users understand the current state of the automated system, which was quantified and normalized to provide us with an objective measurement of their true understanding of the system. Ordinal scales were often used to estimate a rough scale of how much workload the participants experienced. However, the proposed method aimed to provide continuous and real-time data by adopting psychophysiological measures that measure users' workload while interacting with the automated HMI system. The quantified workload measurements and users' true understanding provide us with a more general overview of the effectiveness and ease of use of the automated system. Furthermore, it also provides insights into the critical information icons that are truly beneficial to obtaining correct understanding and minimizing misuses and confusion, which might increase workload during the interaction. Nevertheless, a few improvements are still necessary to obtain more accurate and efficient measurements. The questions to a true understanding of the system could be further improved to probe it in real-time, which could be better combined with the real-time data from the psychophysiological measures. Various driving scenarios and automated system levels should also be explored to understand the effectiveness of various psychophysiological measures and aid us in getting closer to the "absolute value" of automated vehicle transparency. Even though more studies are required to extend the understanding assessment to more levels of automated driving systems, the understandability evaluation method provides the possibility to construct a standardized evaluation method for AV HMI transparency.

8.2 Efforts to understand the automated vehicle human-machine interface

For the work and experiment conducted in Article 1, the workload evaluation method used to estimate the easiness for human users to correctly understand the HMI design was the time to understand. This method was shown to be an effective and robust measure when used to evaluate code understandability in the field of computer science, as described in Chapter 3. One limitation of using the reaction time when targeting the interaction with monitoring or interacting with visual HMI was that only averaged and discrete results could be gathered. Since the reaction time was collected during an online study during the verification of the proposed TRASS (Y.-C. Liu et al., 2022), the evaluation conditions were relatively static and suitable for the workload to be measured with such a method. However, in the cases where the interaction is dynamic, continuous and real-time data would be necessary since the scenario, state of the automation as well as the state of human users change rapidly. Discrete data still provide some insight when the situation is relative static, otherwise the resulting evaluation would be the averaged results over certain period of time, which might increase the uncertainty and noises and further away from the desired metrics.

Considering this gap and aiming to increase the ecological validity of the proposed TRASS, studies were conducted to explore various psychophysiological measures for their sensitivity during the HMI evaluation process. Psychophysiological measures have the advantage of being more robust in estimating workload as what they measure are involuntary controls, which essentially avoid undesired constructs initiated by the participants. Results of Electroencephalogram (EEG) reported in Article 2 did not conclude the effect of different HMI designs resulting in significant workload differences using alpha and theta band waves (Y.-C. Liu et al., 2023).

However, consistent results were found where Trans HMI was found to result in the lowest workload when interacting with the interface. This finding was in line with the previous study (Y.-C. Liu et al., 2022) as the information traits used to design the Trans HMI design was found to be more transparent and understandable. Since this is the first study using EEG power spectral analysis to estimate the effects of AV HMI designs on workload during the interaction, little references could be used as comparisons to identify the factors. One possible reason for the non-significant result might be the small sample size of twelve participants and the fact that data from 18 trials per participant were collected. Also, the inclusion of event-related potentials analysis with EEG signals should also be considered.

During the driving, multiple resources are engaged to perform the task. Using power spectrum analysis might include the workload induced by tasks other than interacting with the AV HMI to understand the automated driving system. The analysis of event-related potentials, on the other hand, could narrow the window of data collection to those considered crucial to the construct desired to be measured. In the study, the trigger points could be identified by corresponding the eye movements of participants. Whenever the gaze of the participant was on the HMI design, we could categorize such time points as points where the monitoring or interaction with the automated driving system was initiated. Then, corresponding event-related potentials from the EEG signals could be identified and analyzed. In the experiment in Article 2, the trigger points could be identified, but the data quantity was not large enough for the event-related potential analysis. More research should be conducted to validate the sensitivity of EEG to the workload while interacting with the HMI of the automated driving systems and to identify the effects of other interaction methods on the workload estimation with the neural activities.

Comparing to the non-significant results using EEG power spectral analysis, results from measures using Electrocardiography (ECG) and Electrodermal Activity (EDA), reported significant differences on workload when interacting with the same HMI designs during same experimental settings and scenarios. This result, on the one hand, demonstrated some aspect of workload measured with ECG and EDA were identified to be sensitive in such scenarios, but on the other hand, raised the concern in regards to the discrete results among workload measures. As suggested by Conti-Kufner (2017), however, the fact that EEG is sensitive to other psychological effects than workload might conceal the effect of cognitive load and result in the lack of differences among AV HMI. Literature has also shown that when measuring workload, there is no universal standard or one-size-fits-all solutions (Boumann, Hamann, Biella, Carstengerdes, & Sammito, 2023; Matthews, Reinerman-Jones, Barber, & Abich IV, 2015).

As discussed in Chapter 3, the workload is not directly observable, and there is no general workload construct (Matthews et al., 2015) applicable to all aspects of the workload. A multiple-measures approach is strongly recommended, but the challenge lies in the determination of valid measures for the desired aspect of the workload. For instance, heart rate signal from ECG data is found to be more sensitive to emotional arousal, while the Heart Rate Variability (HRV), also derived from the ECG data, is more sensitive to cognitive demand (Lohani et al., 2019; Matthews et al., 2015). During the simulated flight, EEG was found to be more sensitive than ECG when the change in workload was applied (Dussault, Jouanin, Philippe, & Guezennec, 2005). Individual differences in automatic arousal may also cause the difference in workload measures, as some neural or physiological responses are more sensitive than others.

As a preliminary series of studies in identifying the sensitivity of various psychophysiological workload measures during human AV HMI interaction, the objective was to construct a standardized and systematic framework for transparency estimation of the automated driving system. The proposed TRASS was shown to be effective to the extent where differences in Functional Transparency (FT) were identified among various HMI designs. With the application of the TRASS, the finding, in return, helped the process of HMI design. However, to allow the method to be more robust against various scenarios, what should be considered is to identify the overloaded resources during the interaction and further determine the load-sensitive process. Another limitation is that when estimating workload with psychophysiological measures, the wearing comfort and intrusiveness of the measures used were not taken into account, which might have a negative impact on the measures. Despite baseline data being collected and these undesired effects being minimized, more research should be conducted to fully understand the size of these effects and possible factors contributing to the aspect of workload measured.

A similar approach should also be taken in investigating the transparency of other interacting resources, for example, when the automated driving system communicates the intents and states of the system through other media such as audio or haptic feedback, or even the combinations of all the interfaces (Bengler et al., 2020). The sensitivity of the psychophysiological measures might be different when acquiring the workload measurements of these different information media. For instance, given randomly visual and auditory tasks, the peak amplitude of visual modality was found to be significantly higher than that of auditory one on p300 wave (Mazaheri & Picton, 2005). Hence, further experiments should be conducted for each specific sensor source and modality to achieve a more exhaustive overview and measures during the interaction between automated driving systems and human users.

In conclusion, these preliminary studies have identified several subjective and objective psychophysiological measures capable of finding differences in effort required to understand the HMI designs. However, further studies are still required to validate the workload assessments regarding the aspects of workload measured, as well as how to include multiple workload measures approach to take into account different aspects of workload that influence the efforts required to understand the HMI design.

8.3 Conclusions and contributions to the future research

The objectives of this thesis have been studied and achieved, although further research is still required to identify the most sensitive psychophysiological measures under different scenarios and use cases. Regardless, a standardized and objective transparency assessment method has been developed. By applying the proposed method, the degree to which the AV HMI is understood by the human users could be objectively estimated using the metrics that evaluate understandability and effort to understand the automated system simultaneously. The proposed Transparency Assessment (TRASS) method was further verified and validated in online and driving simulator experiments. Sensitivities of multiple psychophysiological measures were also explored to expand the fidelity in more dynamic scenarios, where the estimation could be recorded continuously and in real time.

By adopting the proposed TRASS method, where the understandability and effort required to understand the AV HMI were standardized and evaluated, the design and evaluation process for AV HMI could be more efficient since objective measures for understandability and workload demanded to understand the HMI design were incorporated, allowing future researchers to evaluate and modify the HMI designs based on a standardized objective metric. Demonstration of utilizing the proposed metric was also showcased by identifying critical information icons that allowed participants to understand the automated driving system with minimum effort, which could be further used to improve the HMI design.

Another potential contribution of the proposed method is that, using this standardized metric, the relationship between Functional Transparency (FT) and multiple driving safety-related constructs could be identified. The insight into how the AV HMI transparency would affect these factors could be investigated and used to acquire suitable directions and solutions to achieve automated driving safety. For instance, how would the FT of the AV HMI affect the lateral deviation of the vehicle during the control transition could be identified.

With a standardized and objective metric, models surrounding the proposed FT and TRASS could also be built and utilized.

In order to develop suitable automated driving system training programs or customized HMI design layout, proper FT of the HMI designs and procedures should be attained. To achieve this, factors such as user characteristics, level of automation, user experiences, or logic behind the automated driving system could be combined with the proposed TRASS method to develop and obtain a model that could identify different needs for each participant and scenario, and facilitate driving safety by providing suitable training program and HMI designs that could reach higher FT for the user.

This work took the first step in exploring the transparency of AV HMI designs, allowing human users to easily and correctly understand the automated driving system. Nevertheless, continuous developments are still required to achieve better ecological validity for the proposed TRASS method. Explorations of the sensitivity and validity of the proposed method in real-driving scenarios are required. Before fully autonomous vehicles are generally adopted, more researches have to be done to contribute to a safer environment for automated driving.

References

Acerra, E., Pazzini, M., Ghasemi, N., Vignali, V., Lantieri, C., Simone, A., ... others (2019). Eeg-based mental workload and perception-reaction time of the drivers while using adaptive cruise control. In *Human mental workload: Models and applications: Third international symposium, h-workload 2019, rome, italy, november 14–15, 2019, proceedings 3* (pp. 226–239).

Adnan, N., Nordin, S. M., bin Bahruddin, M. A., & Ali, M. (2018). How trust can drive forward the user acceptance to the technology? in-vehicle technology for autonomous vehicle. *Transportation research part A: policy and practice*, *118*, 819–836.

Albers, D., Radlmayr, J., Loew, A., Hergeth, S., Naujoks, F., Keinath, A., & Bengler, K. (2020). Usability evaluation—advances in experimental design in the context of automated driving human–machine interfaces. *Information*, *11*(5), 240.

Ayoub, J., Zhou, F., Bao, S., & Yang, X. J. (2019). From manual driving to automated driving: A review of 10 years of autoui. In *Proceedings of the 11th international conference on automotive user interfaces and interactive vehicular applications* (p. 70–90). New York, NY, USA: Association for Computing Machinery. Retrieved from `https://doi.org/10.1145/3342197.3344529` doi: 10.1145/3342197.3344529

Bansiya, J., & Davis, C. G. (2002). A hierarchical model for object-oriented design quality assessment. *IEEE Transactions on software engineering*, *28*(1), 4–17.

Beggiato, M., Hartwich, F., & Krems, J. (2019). Physiological correlates of discomfort in automated driving. *Transportation research part F: traffic psychology and behaviour*, *66*, 445–458.

Beggiato, M., Hartwich, F., Schleinitz, K., Krems, J., Othersen, I., & Petermann-Stock, I. (2015). What would drivers like to know during automated driving? information needs at different levels of automation. In *7. tagung fahrerassistenzsysteme.*

Bellem, H., Thiel, B., Schrauf, M., & Krems, J. F. (2018). Comfort in automated driving: An analysis of preferences for different automated driving styles and their dependence on personality traits. *Transportation research part F: traffic psychology and behaviour*, *55*, 90–100.

Bengler, K., Rettenmaier, M., Fritz, N., & Feierle, A. (2020). From hmi to hmis: Towards an hmi framework for automated driving. *Information*, *11*(2), 61.

Bhaskara, A., Duong, L., Brooks, J., Li, R., McInerney, R., Skinner, M., ... Loft, S. (2021). Effect of automation transparency in the management of multiple unmanned vehicles. *Applied Ergonomics*, *90*, 103243.

Borghini, G., Aricò, P., Di Flumeri, G., Salinari, S., Colosimo, A., Bonelli, S., ... Babiloni, F. (2015). Avionic technology testing by using a cognitive neurometric index: a study with professional helicopter pilots. In *2015 37th annual international conference of the ieee engineering in medicine and biology society (embc)* (pp. 6182–6185).

Borghini, G., Vecchiato, G., Toppi, J., Astolfi, L., Maglione, A., Isabella, R., … others (2012). Assessment of mental fatigue during car driving by using high resolution eeg activity and neurophysiologic indices. In *2012 annual international conference of the ieee engineering in medicine and biology society* (pp. 6442–6445).

Boumann, H., Hamann, A., Biella, M., Carstengerdes, N., & Sammito, S. (2023). Suitability of physiological, self-report and behavioral measures for assessing mental workload in pilots. In *International conference on human-computer interaction* (pp. 3–20).

Brooke, J. (1996). Sus: a "quick and dirty'usability. *Usability evaluation in industry*, *189*(3), 189–194.

Brookhuis, K. A., van Driel, C. J., Hof, T., van Arem, B., & Hoedemaeker, M. (2009). Driving with a congestion assistant; mental workload and acceptance. *Applied ergonomics*, *40*(6), 1019–1025.

Brown, A., Gonder, J., & Repac, B. (2014). An analysis of possible energy impacts of automated vehicles. *Road vehicle automation*, 137–153.

Bueno, M., Dogan, E., Selem, F. H., Monacelli, E., Boverie, S., & Guillaume, A. (2016). How different mental workload levels affect the take-over control after automated driving. In *2016 ieee 19th international conference on intelligent transportation systems (itsc)* (pp. 2040–2045).

Calvi, A., D'Amico, F., Ciampoli, L. B., & Ferrante, C. (2020). Evaluation of driving performance after a transition from automated to manual control: a driving simulator study. *Transportation research procedia*, *45*, 755–762.

Cantin, V., Lavallière, M., Simoneau, M., & Teasdale, N. (2009). Mental workload when driving in a simulator: Effects of age and driving complexity. *Accident Analysis & Prevention*, *41*(4), 763–771.

Capiluppi, A., Morisio, M., & Lago, P. (2004). Evolution of understandability in oss projects. In *Eighth european conference on software maintenance and reengineering, 2004. csmr 2004. proceedings.* (pp. 58–66).

Chen, J. Y., & Barnes, M. J. (2012). Supervisory control of multiple robots: Effects of imperfect automation and individual differences. *Human Factors*, *54*(2), 157–174.

Chen, J. Y., Lakhmani, S. G., Stowers, K., Selkowitz, A. R., Wright, J. L., & Barnes, M. (2018). Situation awareness-based agent transparency and human-autonomy teaming effectiveness. *Theoretical issues in ergonomics science*, *19*(3), 259–282.

Chen, J. Y., Procci, K., Boyce, M., Wright, J., Garcia, A., & Barnes, M. (2014). Situation awareness-based agent transparency. *US Army Research Laboratory*, 1–29.

Chen, W., Sawaragi, T., & Hiraoka, T. (2022). Comparing eye-tracking metrics of mental workload caused by ndrts in semi-autonomous driving. *Transportation research part F: traffic psychology and behaviour*, *89*, 109–128.

Chen, W., Sawaragi, T., & Horiguchi, Y. (2019). Measurement of driver's mental workload in partial autonomous driving. *IFAC-PapersOnLine*, *52*(19), 347–352.

Choi, J. K., & Ji, Y. G. (2015). Investigating the importance of trust on adopting an autonomous vehicle. *International Journal of Human-Computer Interaction*, *31*(10), 692–702.

Chowdhury, M. S. N., Dutta, A., Robison, M. K., Blais, C., Brewer, G. A., & Bliss, D. W. (2020). Deep neural network for visual stimulus-based reaction time estimation using the periodogram of single-trial eeg. *Sensors*, *20*(21), 6090.

Conti-Kufner, A. S. (2017). *Measuring cognitive task load: An evaluation of the detection response task and its implications for driver distraction assessment* (Unpublished doctoral dissertation). Technische Universität München.

Dargahi Nobari, K., Albers, F., Bartsch, K., Braun, J., & Bertram, T. (2022). Modeling driver-vehicle interaction in automated driving. *Forschung im Ingenieurwesen*, *86*(1), 65–79.

Dawson, M. E., Schell, A. M., & Filion, D. L. (2007). The electrodermal system. *Handbook of psychophysiology*, *2*, 200–223.

Debernard, S., Chauvin, C., Pokam, R., & Langlois, S. (2016). Designing human-machine interface for autonomous vehicles. *IFAC-PapersOnLine*, *49*(19), 609–614.

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

Dickmanns, E. D., Mysliwetz, B., & Christians, T. (1990). An integrated spatio-temporal approach to automatic visual guidance of autonomous vehicles. *IEEE Transactions on Systems, Man, and Cybernetics*, *20*(6), 1273–1284.

Di Flumeri, G., Borghini, G., Aricò, P., Sciaraffa, N., Lanzi, P., Pozzi, S., ... others (2018). Eeg-based mental workload neurometric to evaluate the impact of different traffic and road conditions in real driving settings. *Frontiers in human neuroscience*, *12*, 509.

DIN EN ISO 9241-11. (2018, 11). *DIN EN ISO 9241-11:2018-11, ergonomie der mensch-system-interaktion - teil 11: Gebrauchstauglichkeit: Begriffe und konzepte (ISO 9241-11:2018); deutsche fassung EN ISO 9241-11:2018* (Norm No. DIN EN 9241-11:2018). Beuth Verlag GmbH. Retrieved from `https://doi.org/10.31030%2F2757945` doi: 10.31030/2757945

Dixit, V. V., Chand, S., & Nair, D. J. (2016). Autonomous vehicles: disengagements, accidents and reaction times. *PLoS one*, *11*(12), e0168054.

Dozza, M. (2013). What factors influence drivers' response time for evasive maneuvers in real traffic? *Accident Analysis & Prevention*, *58*, 299–308.

Droździel, P., Tarkowski, S., Rybicka, I., & Wrona, R. (2020). Drivers' reaction time research in the conditions in the real traffic. *Open Engineering*, *10*(1), 35–47.

Dussault, C., Jouanin, J.-C., Philippe, M., & Guezennec, C.-Y. (2005). Eeg and ecg changes during simulator operation reflect mental workload and vigilance. *Aviation, space, and environmental medicine*, *76*(4), 344–351.

Ebnali, M., Lamb, R., Fathi, R., & Hulme, K. (2021). Virtual reality tour for first-time users of highly automated cars: Comparing the effects of virtual environments with different levels of interaction fidelity. *Applied Ergonomics*, *90*, 103226.

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

Figalová, N., Bieg, H.-J., Reiser, J. E., Liu, Y.-C., Baumann, M., Chuang, L., & Pollatos, O. (2024). From driver to supervisor: Comparing cognitive load and eeg-based attentional resource allocation across automation levels. *International Journal of Human-Computer Studies*, *182*, 103169.

Fitton, R. S., & Wadsworth, A. P. (1958). *The strutts and the arkwrights, 1758-1830: A study of the early factory system*. Manchester University Press.

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

Gold, C., Körber, M., Hohenberger, C., Lechner, D., & Bengler, K. (2015). Trust in automation–before and after the experience of take-over scenarios in a highly automated vehicle. *Procedia Manufacturing*, *3*, 3025–3032.

Gonçalves, J., & Bengler, K. (2015). Driver state monitoring systems–transferable knowledge manual driving to had. *Procedia Manufacturing*, *3*, 3011–3016.

Hancock, P. A., Kajaks, T., Caird, J. K., Chignell, M. H., Mizobuchi, S., Burns, P. C., ... others (2020). Challenges to human drivers in increasingly automated vehicles. *Human factors*, *62*(2), 310–328.

Hart, S. G. (2006). Nasa-task load index (nasa-tlx); 20 years later. In *Proceedings of the human factors and ergonomics society annual meeting* (Vol. 50, pp. 904–908).

Hecht, T., Kratzert, S., & Bengler, K. (2020). The effects of a predictive hmi and different transition frequencies on acceptance, workload, usability, and gaze behavior during urban automated driving. *Information*, *11*(2), 73.

Heikoop, D. D., de Winter, J. C., van Arem, B., & Stanton, N. A. (2019). Acclimatizing to automation: Driver workload and stress during partially automated car following in real traffic. *Transportation research part F: traffic psychology and behaviour*, *65*, 503–517.

Heine, T., Lenis, G., Reichensperger, P., Beran, T., Doessel, O., & Deml, B. (2017). Electrocardiographic features for the measurement of drivers' mental workload. *Applied ergonomics*, *61*, 31–43.

Helldin, T., Ohlander, U., Falkman, G., & Riveiro, M. (2014). Transparency of automated combat classification. In *Engineering psychology and cognitive ergonomics: 11th international conference, epce 2014, held as part of hci international 2014, heraklion, crete, greece, june 22-27, 2014. proceedings 11* (pp. 22–33).

Hoc, J.-M., Young, M. S., & Blosseville, J.-M. (2009). Cooperation between drivers and automation: implications for safety. *Theoretical Issues in Ergonomics Science*, *10*(2), 135–160.

Jasper, P., Sibley, C., & Coyne, J. (2016). Using heart rate variability to assess operator mental workload in a command and control simulation of multiple unmanned aerial vehicles. In *Proceedings of the human factors and ergonomics society annual meeting* (Vol. 60, pp. 1125–1129).

Kaduk, S. I., Roberts, A. P., & Stanton, N. A. (2021). Driving performance, sleepiness, fatigue, and mental workload throughout the time course of semi-automated driving—experimental data from the driving simulator. *Human Factors and Ergonomics in Manufacturing & Service Industries*, *31*(1), 143–154.

Kaleefathullah, A. A., Merat, N., Lee, Y. M., Eisma, Y. B., Madigan, R., Garcia, J., & Winter, J. d. (2022). External human-machine interfaces can be misleading: An examination of trust development and misuse in a cave-based pedestrian simulation environment. *Human factors*, *64*(6), 1070–1085.

Käthner, I., Wriessnegger, S. C., Müller-Putz, G. R., Kübler, A., & Halder, S. (2014). Effects of mental workload and fatigue on the p300, alpha and theta band power during operation of an erp (p300) brain–computer interface. *Biological psychology*, *102*, 118–129.

Kohn, S. C., de Visser, E. J., Wiese, E., Lee, Y.-C., & Shaw, T. H. (2021). Measurement of trust in automation: A narrative review and reference guide. *Frontiers in psychology*, *12*, 604977.

Kokil, U., & Scott, S. (2017). Usability testing of a school website using qualitative approach. In *International conference on human computer interaction theory and applications* (Vol. 3, pp. 55–64).

Körber, M., Baseler, E., & Bengler, K. (2018). Introduction matters: Manipulating trust in automation and reliance in automated driving. *Applied ergonomics*, *66*, 18–31.

Kous, K., Pušnik, M., Heričko, M., & Polančič, G. (2020). Usability evaluation of a library website with different end user groups. *Journal of Librarianship and Information Science*, *52*(1), 75–90.

Kraft, A.-K., Maag, C., Cruz, M. I., Baumann, M., & Neukum, A. (2020). Effects of explaining system failures during maneuver coordination while driving manual or automated. *Accident Analysis & Prevention*, *148*, 105839.

Kraft, A.-K., Naujoks, F., Wörle, J., & Neukum, A. (2018). The impact of an in-vehicle display on glance distribution in partially automated driving in an on-road experiment. *Transportation research part F: traffic psychology and behaviour*, *52*, 40–50.

Kraus, J., Scholz, D., Stiegemeier, D., & Baumann, M. (2020). The more you know: trust dynamics and calibration in highly automated driving and the effects of take-overs, system malfunction, and system transparency. *Human factors*, *62*(5), 718–736.

Kunze, A., Summerskill, S. J., Marshall, R., & Filtness, A. J. (2019). Automation transparency: implications of uncertainty communication for human-automation interaction and interfaces. *Ergonomics*, *62*(3), 345–360.

Lee, J., Abe, G., Sato, K., & Itoh, M. (2020). Impacts of system transparency and system failure on driver trust during partially automated driving. In *2020 ieee international conference on human-machine systems (ichms)* (pp. 1–3).

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human factors*, *46*(1), 50–80.

Lin, J.-C., & Wu, K.-C. (2006). A model for measuring software understandability. In *The sixth ieee international conference on computer and information technology (cit'06)* (pp. 192–192).

Liu, Y.-C., Figalova, N., Baumann, M., & Bengler, K. (2023). Human-machine interface evaluation using eeg in driving simulator. In *2023 ieee intelligent vehicles symposium (iv)* (pp. 1–6).

Liu, Y.-C., Figalová, N., & Bengler, K. (2022). Transparency assessment on level 2 automated vehicle hmis. *Information*, *13*(10), 489.

Liu, Y.-c., Figalová, N., Pichen, J., Hock, P., Baumann, M., & Bengler, K. (2023). Workload assessment of human-machine interface: A simulator study with psychophysiological measures. *Human Systems Engineering and Design (IHSED 2023): Future Trends and Applications*, *112*(112).

Loeches De La Fuente, H., Berthelon, C., Fort, A., Etienne, V., De Weser, M., Ambeck, J., & Jallais, C. (2019). Electrophysiological and performance variations following driving events involving an increase in mental workload. *European transport research review*, *11*, 1–9.

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

Longo, L., Wickens, C. D., Hancock, G., & Hancock, P. A. (2022). Human mental workload: A survey and a novel inclusive definition. *Frontiers in psychology*, *13*, 883321.

Lu, Z., Happee, R., Cabrall, C. D., Kyriakidis, M., & De Winter, J. C. (2016). Human factors of transitions in automated driving: A general framework and literature survey. *Transportation research part F: traffic psychology and behaviour*, *43*, 183–198.

Ma, R. H., Morris, A., Herriotts, P., & Birrell, S. (2021). Investigating what level of visual information inspires trust in a user of a highly automated vehicle. *Applied Ergonomics*, *90*, 103272.

Maarten Schraagen, J., Kerwien Lopez, S., Schneider, C., Schneider, V., Tönjes, S., & Wiechmann, E. (2021). The role of transparency and explainability in automated systems. In *Proceedings of the human factors and ergonomics society annual meeting* (Vol. 65, pp. 27–31).

Makishita, H., & Matsunaga, K. (2008). Differences of drivers' reaction times according to age and mental workload. *Accident Analysis & Prevention*, *40*(2), 567–575.

Matthews, G., Reinerman-Jones, L. E., Barber, D. J., & Abich IV, J. (2015). The psychometrics of mental workload: Multiple measures are sensitive but divergent. *Human factors*, *57*(1), 125–143.

Mazaheri, A., & Picton, T. W. (2005). Eeg spectral dynamics during discrimination of auditory and visual targets. *Cognitive Brain Research*, *24*(1), 81–96.

Melnicuk, V., Thompson, S., Jennings, P., & Birrell, S. (2021). Effect of cognitive load on drivers' state and task performance during automated driving: Introducing a novel method for determining stabilisation time following take-over of control. *Accident Analysis & Prevention*, *151*, 105967.

Miller, J. (2020). Futurama: An immersive experience of america's automotive future. *Journal of Motorsport Culture & History*, *1*(1), 6.

Misra, S., & Akman, I. (2008). Comparative study of cognitive complexity measures. In *2008 23rd international symposium on computer and information sciences* (pp. 1–4).

Morales-Alvarez, W., Marouf, M., Tadjine, H. H., & Olaverri-Monreal, C. (2021). Real-world evaluation of the impact of automated driving system technology on driver gaze behavior, reaction time and trust. In *2021 ieee intelligent vehicles symposium workshops (iv workshops)* (p. 57-64). doi: 10.1109/IVWorkshops54471.2021.9669230

Mueller, A. S., Cicchino, J. B., Singer, J., & Jenness, J. W. (2020). Effects of training and display content on level 2 driving automation interface usability. *Transportation Research Part F: Traffic Psychology and Behaviour*, *69*, 61-71. Retrieved from `https://www.sciencedirect.com/science/article/pii/S1369847819303997` doi: https://doi.org/10.1016/j.trf.2019.12.010

Mulder, L. B. J., de Waard, D., & Brookhuis, K. A. (2004). Estimating mental effort using heart rate and heart rate variability. In *Handbook of human factors and ergonomics methods* (pp. 227–236). CRC Press.

Müller, A. L., Fernandes-Estrela, N., Hetfleisch, R., Zecha, L., & Abendroth, B. (2021). Effects of non-driving related tasks on mental workload and take-over times during conditional automated driving. *European transport research review*, *13*(1), 1–15.

Naujoks, F., Wiedemann, K., Schömig, N., Hergeth, S., & Keinath, A. (2019). Towards guidelines and verification methods for automated vehicle hmis. *Transportation research part F: traffic psychology and behaviour*, *60*, 121–136.

Nygren, T. E. (1992). Cognitive and affective components of mental workload: Understanding the effects of each on human decision making behavior. *Hampton Univ., NASA (American Society for Engineering Education (ASEE) Summer Faculty Fellowship Program 1992 p 154-156 (SEE N93-16760 05-80).*

Oliveira, L., Burns, C., Luton, J., Iyer, S., & Birrell, S. (2020). The influence of system transparency on trust: Evaluating interfaces in a highly automated vehicle. *Transportation research part F: traffic psychology and behaviour*, *72*, 280–296.

Passchier, T. (2021). Do drivers take effort to learn about their car? the effect of knowledge on trust in level 2 automation vehicles.

Paula, D., Bauder, M., Pfeilschifter, C., Petermeier, F., Kubjatko, T., Böhm, K., ... Schweiger, H.-G. (2023). Impact of partially automated driving functions on forensic accident reconstruction: a simulator study on driver reaction behavior in the event of a malfunctioning system behavior. *Sensors*, *23*(24), 9785.

Pawar, N. M., & Velaga, N. R. (2020). Modelling the influence of time pressure on reaction time of drivers. *Transportation research part F: traffic psychology and behaviour*, *72*, 1–22.

Paxion, J., Galy, E., & Berthelon, C. (2014). Mental workload and driving. *Frontiers in psychology*, *5*, 1344.

Peng, C., Merat, N., Romano, R., Hajiseyedjavadi, F., Paschalidis, E., Wei, C., ... Boer, E. (2022). Drivers' evaluation of different automated driving styles: Is it both comfortable and natural? *Human factors*, 00187208221113448.

Pensabe-Rodriguez, A., Lopez-Dominguez, E., Hernandez-Velazquez, Y., Dominguez-Isidro, S., & De-la Calleja, J. (2020). Context-aware mobile learning system: Usability assessment based on a field study. *Telematics and Informatics*, *48*, 101346.

Perello-March, J. R., Burns, C. G., Woodman, R., Elliott, M. T., & Birrell, S. A. (2021). Driver state monitoring: Manipulating reliability expectations in simulated automated driving scenarios. *IEEE transactions on intelligent transportation systems*, *23*(6), 5187–5197.

Pouliou, A., Kehagia, F., Bekiaris, E., Pitsiava-Latinopoulou, M., & Poulios, G. (2022). Mental workload influence of drivers reaction time on unexpected events: A driving simulation study. In *Road safety and simulation conference.*

Pouliou, A., Kehagia, F., Poulios, G., Pitsiava-Latinopoulou, M., & Bekiaris, E. (2023). Drivers' reaction time and mental workload: A driving simulation study. *Transport and Telecommunication Journal*, *24*(4), 397–408.

Radhakrishnan, V., Louw, T., Gonçalves, R. C., Torrao, G., Lenné, M. G., & Merat, N. (2023). Using pupillometry and gaze-based metrics for understanding drivers' mental workload during automated driving. *Transportation research part F: traffic psychology and behaviour*, *94*, 254–267.

Radhakrishnan, V., Merat, N., Louw, T., Gonçalves, R. C., Torrao, G., Lyu, W., ... Lenné, M. G. (2022). Physiological indicators of driver workload during car-following scenarios and takeovers in highly automated driving. *Transportation research part F: traffic psychology and behaviour*, *87*, 149–163.

Recarte, M. A., & Nunes, L. M. (2003). Mental workload while driving: effects on visual search, discrimination, and decision making. *Journal of experimental psychology: Applied*, *9*(2), 119.

Reilhac, P., Hottelart, K., Diederichs, F., & Nowakowski, C. (2017). User experience with increasing levels of vehicle automation: Overview of the challenges and opportunities as vehicles progress from partial to high automation. *Automotive User Interfaces: Creating Interactive Experiences in the Car*, 457–482.

Reimer, B., Mehler, B., Pohlmeyer, A., Coughlin, J. F., & Dusek, J. (2006). The use of heart rate in a driving simulator as an indicator of age-related differences in driver workload. *Advances in Transportation Studies an International Journal, Special Issue*, 9–20.

Rezaei, A., & Caulfield, B. (2021). Safety of autonomous vehicles: what are the insights from experienced industry professionals? *Transportation research part F: traffic psychology and behaviour*, *81*, 472–489.

Richardson, N. T., Lehmer, C., Lienkamp, M., & Michel, B. (2018). Conceptual design and evaluation of a human machine interface for highly automated truck driving. In *2018 ieee intelligent vehicles symposium (iv)* (pp. 2072–2077).

Scalabrino, S., Bavota, G., Vendome, C., Linares-Vasquez, M., Poshyvanyk, D., & Oliveto, R. (2019). Automatically assessing code understandability. *IEEE Transactions on Software Engineering*, *47*(3), 595–613.

Selkowitz, A. R., Larios, C. A., Lakhmani, S. G., & Chen, J. Y. (2016). Displaying information to support transparency for autonomous platforms. In *Advances in human factors in robots and unmanned systems: Proceedings of the ahfe 2016 international conference on human factors in robots and unmanned systems, july 27-31, 2016, walt disney world®, florida, usa* (pp. 161–173).

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

Serter, B., Beul, C., Lang, M., & Schmidt, W. (2017). *Foreseeable misuse in automated driving vehicles-the human factor in fatal accidents of complex automation* (Tech. Rep.). SAE Technical Paper.

Shakouri, M., Ikuma, L. H., Aghazadeh, F., & Nahmens, I. (2018). Analysis of the sensitivity of heart rate variability and subjective workload measures in a driving simulator: the case of highway work zones. *International journal of industrial ergonomics*, *66*, 136–145.

Shi, E., & Bengler, K. (2022). Overall effects of non-driving related activities' characteristics on takeover performance in the context of sae level 3: A meta-analysis. *Human Factors in Transportation*, *60*, 69.

Smyth, J., Ulahannan, A., Florek, F., Shaw, E., & Mansfield, N. (2021). Understanding misuse of partially automated vehicles–a discussion of ntsb's findings of the 2018 mountain view tesla crash.

Srinivasulu, D., Sridhar, A., & Mohapatra, D. P. (2014). Evaluation of software understandability using rough sets. In *Intelligent computing, networking, and informatics: Proceedings of the international conference on advanced computing, networking, and informatics, india, june 2013* (pp. 939–946).

Stapel, J., Mullakkal-Babu, F. A., & Happee, R. (2019). Automated driving reduces perceived workload, but monitoring causes higher cognitive load than manual driving. *Transportation research part F: traffic psychology and behaviour*, *60*, 590–605.

Storey, M.-A. (2005). Theories, methods and tools in program comprehension: past, present and future. In *13th international workshop on program comprehension (iwpc'05)* (pp. 181–191).

Suh, W., Park, P. Y.-J., Park, C. H., & Chon, K. S. (2006). Relationship between speed, lateral placement, and drivers' eye movement at two-lane rural highways. *Journal of transportation engineering*, *132*(8), 649–653.

Thrun, S., Montemerlo, M., Dahlkamp, H., Stavens, D., Aron, A., Diebel, J., ... others (2006). Stanley: The robot that won the darpa grand challenge. *Journal of field Robotics*, *23*(9), 661–692.

Ting, P.-H., Hwang, J.-R., Doong, J.-L., & Jeng, M.-C. (2008). Driver fatigue and highway driving: A simulator study. *Physiology & behavior*, *94*(3), 448–453.

Tjolleng, A., Jung, K., Hong, W., Lee, W., Lee, B., You, H., ... Park, S. (2017). Classification of a driver's cognitive workload levels using artificial neural network on ecg signals. *Applied ergonomics*, *59*, 326–332.

transparency, n. (n.d.). Oxford University Press. Retrieved 2023-12-08, from `https://www.oed.com/dictionary/transparency_n?tab=factsheet#17971238`

Van Acker, B. B., Parmentier, D. D., Vlerick, P., & Saldien, J. (2018). Understanding mental workload: from a clarifying concept analysis toward an implementable framework. *Cognition, technology & work*, *20*, 351–365.

Voinescu, A., Morgan, P. L., Alford, C., & Caleb-Solly, P. (2020). The utility of psychological measures in evaluating perceived usability of automated vehicle interfaces–a study with older adults. *Transportation research part F: traffic psychology and behaviour*, *72*, 244–263.

Walch, M., Mühl, K., Baumann, M., & Weber, M. (2018). Click or hold: usability evaluation of maneuver approval techniques in highly automated driving. In *Extended abstracts of the 2018 chi conference on human factors in computing systems* (pp. 1–6).

Walker, F., Forster, Y., Hergeth, S., Kraus, J., Payre, W., Wintersberger, P., & Martens, M. (2023). Trust in automated vehicles: constructs, psychological processes, and assessment. *Frontiers in Psychology*, *14*.

Walker, F., Wang, J., Martens, M. H., & Verwey, W. B. (2019). Gaze behaviour and electrodermal activity: Objective measures of drivers' trust in automated vehicles. *Transportation research part F: traffic psychology and behaviour*, *64*, 401–412.

Weichbroth, P. (2020). Usability of mobile applications: a systematic literature study. *Ieee Access*, *8*, 55563–55577.

Weyuker, E. J. (1988). Evaluating software complexity measures. *IEEE transactions on Software Engineering*, *14*(9), 1357–1365.

Wickens, C. D. (2008). Multiple resources and mental workload. *Human factors*, *50*(3), 449–455.

Wilson, K. M., Yang, S., Roady, T., Kuo, J., & Lenné, M. G. (2020). Driver trust & mode confusion in an on-road study of level-2 automated vehicle technology. *Safety Science*, *130*, 104845.

Winkle, T. (2016). Safety benefits of automated vehicles: Extended findings from accident research for development, validation and testing. *Autonomous driving: Technical, legal and social aspects*, 335–364.

Wörle, J., & Metz, B. (2023). Misuse or abuse of automation? exploring drivers' intentions to nap during automated driving. *Transportation research part F: traffic psychology and behaviour*, *99*, 460–472.

Yang, C. D., & Fisher, D. L. (2021). *Safety impacts and benefits of connected and automated vehicles: How real are they?* (Vol. 25) (No. 2). Taylor & Francis.

Yang, S., Hosseiny, S. A. R., Susindar, S., & Ferris, T. K. (2016). Investigating driver sympathetic arousal under short-term loads and acute stress events. In *Proceedings of the human factors and ergonomics society annual meeting* (Vol. 60, pp. 1905–1905).

Yoon, S. H., & Ji, Y. G. (2019). Non-driving-related tasks, workload, and takeover performance in highly automated driving contexts. *Transportation research part F: traffic psychology and behaviour*, *60*, 620–631.

Zander, T. O., Andreessen, L. M., Berg, A., Bleuel, M., Pawlitzki, J., Zawallich, L., . . . Gramann, K. (2017). Evaluation of a dry eeg system for application of passive brain-computer interfaces in autonomous driving. *Frontiers in human neuroscience*, *11*, 78.

A Human-machine Interfaces

To validate the proposed method, Automated Vehicle (AV) Human-Machine Interface (HMI) designs with different levels of understandability should first be developed. Since there has been very little literature adopting a similar transparency concept to evaluate AV HMI, the HMI designs depicted in Fig. A.1 were developed. Differences in understandability were designed based on the previous study where both easy-to-understand and confusing elements were identified (Y.-C. Liu et al., 2022). Plus, whether the design principles, such as color contrasts, size of icon, or text, were obeyed was also considered as a factor to manipulate the effort and understandability when interacting with this HMI designs.

The term “trans“ represented transparent, meaning that this HMI could be easily understood. On the contrary, the “fog“ HMI stood for foggy and blurry, meaning that the HMI design might result in confusion and undesired understandability. Lastly, the term “trans-fog“ was the HMI design that lies between, where transparent and understandable interface design and information icons were provided, but the system transparency and the logic behind the automated assistance system activation process were not easily comprehensible. Participants reported confusion when interacting with a certain automated driving system where the “standby mode“ exists. During the standby stage, the automated assistance system was activated but not engaged since the operation conditions were not all met, and the automated driving system would wait and stand by and activate the full function whenever all operation requirements were satisfied. This was considered confusing and thus should contribute to a more “foggy“ and less understandable AV HMI design.

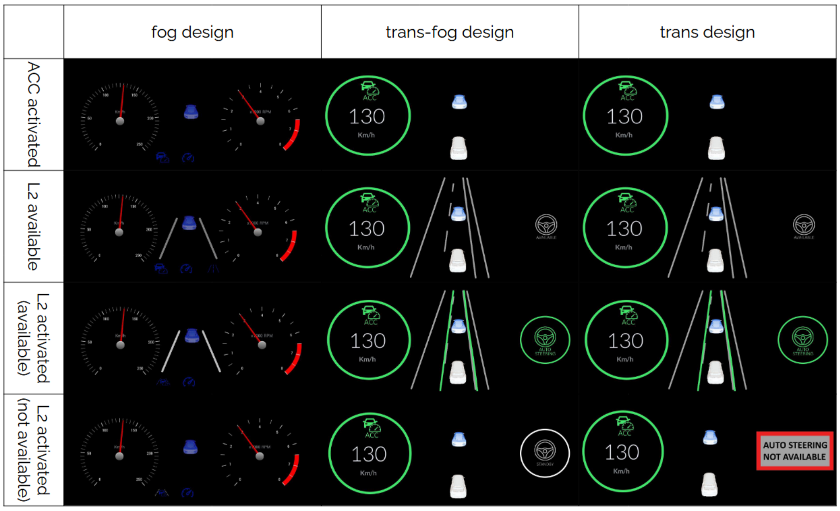


Figure A.1: HMI designs developed and adopted in Articles 2 and 3. Different system understandability was identified with labels fog, trans-fog, and trans designs.

B Safety analysis with TRASS

A potential application of the proposed Transparency Assessment (TRASS) method is that how the Human-Machine Interface (HMI) design could have effects on driving safety could be formulated. With the objective transparency assessment method, averaged Functional Transparency (FT) score for each HMI design could be identified and taken into consideration to evaluate how these FT scores are correlated to safety factors. For example, in Fig. B.1, the application of the TRASS method was demonstrated with the relationship between AV HMI with different averaged FT scores and the steering angles during automated driving. It was shown that the averaged steering angle raised rapidly at the end of the interaction when the automated assistance system was activated when using the Fog HMI, indicating that the corrections occurred too late and that huge correcting maneuvers were required. On the other hand, the averaged steering angle when using Trans HMI was raised earlier, showing that the participants noticed the deviation and made corrections more promptly, avoiding dangerous maneuvers (i.e., with high and rapidly changing steering angles) during the automated driving.

Further research could be done utilizing the proposed Transparency Assessment (TRASS) method and the resulting FT scores to construct models describing the relationship between the Functional Transparency (FT) scores and safety factors such as lateral deviation from the center line or steering angle, so that AV HMI designs could be modified to meet the needs for automated driving safety regulations. For instance, the relationship between FT scores and dangerous maneuvers could be determined. Then, the minimum FT score required to avoid these dangerous maneuvers could be identified, and this threshold could be part of the regulation in such a way that the AV HMI not reaching a certain level of FT scores should be prevented from using on public roads.

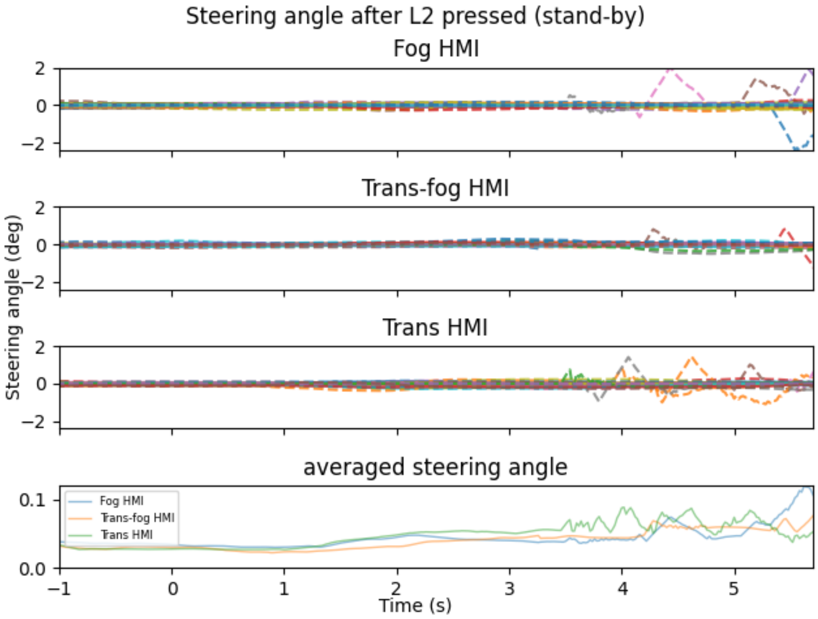


Figure B.1: Steering angle during automated driving after the Level 2 driving assistance system button was pressed. From top down were the steering angle profiles of severeal trials when using Fog HMI, Trans-fog HMI, and Trans HMI. The width of the road was set to be normal German lane width, which is 3.5 meters in the simulator.

C Article 1: “Transparency Assessment on Level 2 Automated Vehicle HMIs“

information

MDPI

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

check for updates

Citation: Liu, Y.-C.; Figalová, N.; Bengler, K. Transparency Assessment on Level 2 Automated Vehicle HMIs. *Information* **2022**, *13*, 489. https://doi.org/10.3390/info13100489

Academic Editor: Joaquim Ferreira

Received: 20 August 2022
Accepted: 30 September 2022
Published: 10 October 2022

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

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

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.

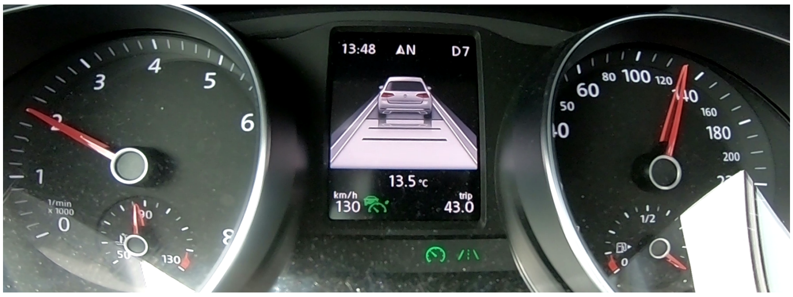


Figure 1. Example VW HMI design.

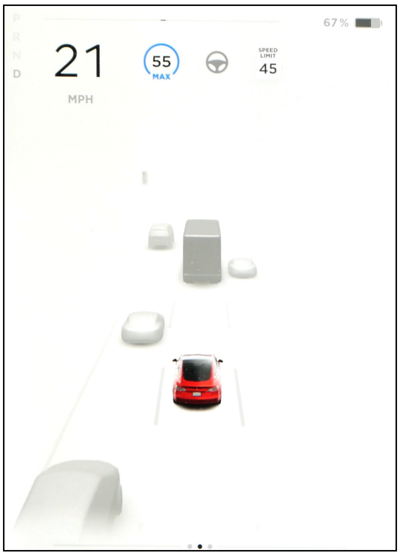


Figure 2. Example Tesla HMI design.

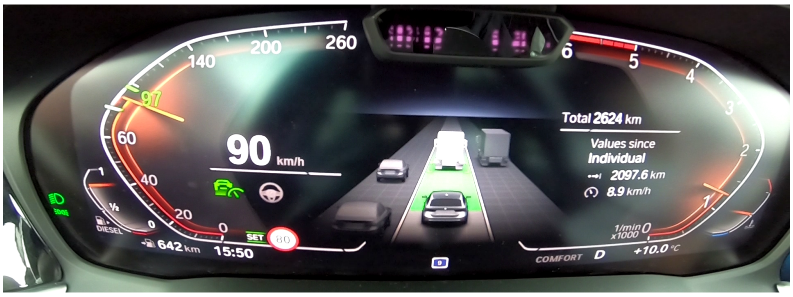


Figure 3. Example BMW HMI design.

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 subsystems (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:

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

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

*** $p < 0.01$, ** $p < 0.05$.

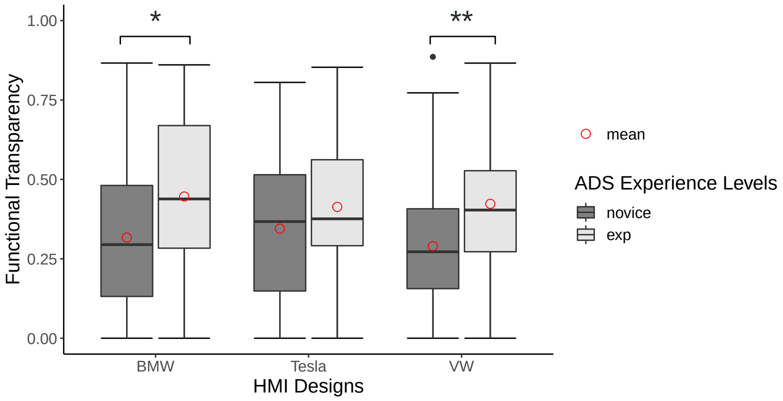


Figure 4. Comparing levels of ADS experience given HMI designs (exp stands for experienced in ADS experience levels). ** $p < 0.05$, * $p < 0.1$.

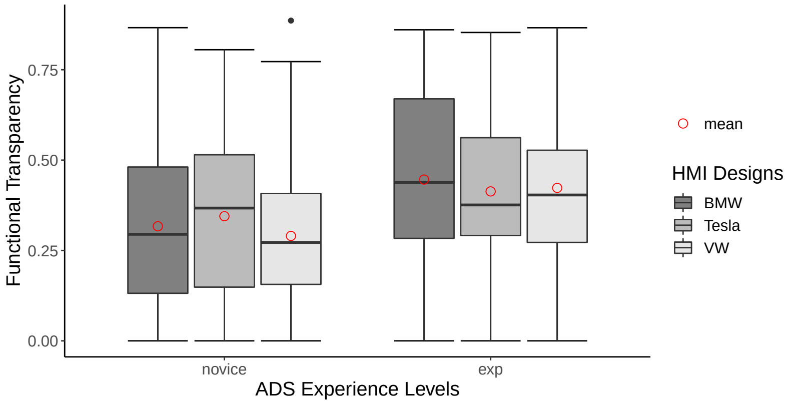


Figure 5. Comparing HMI designs given ADS experience levels (exp stands for experienced in ADS experience levels).

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).

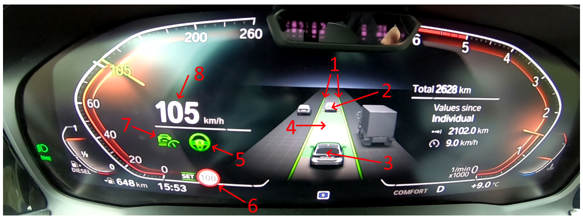


Figure 6. BMW HMI design and corresponding icon number.

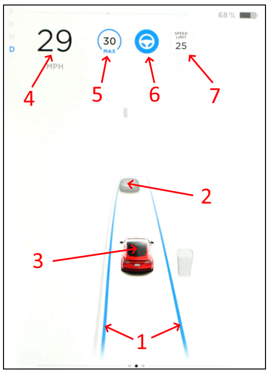


Figure 7. Tesla HMI design and corresponding icon number.

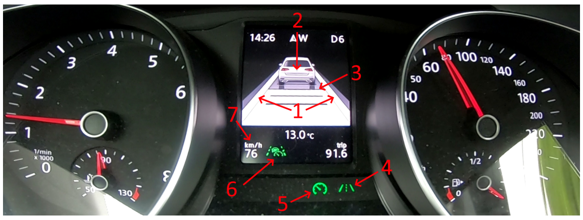


Figure 8. VW HMI design and corresponding icon number.

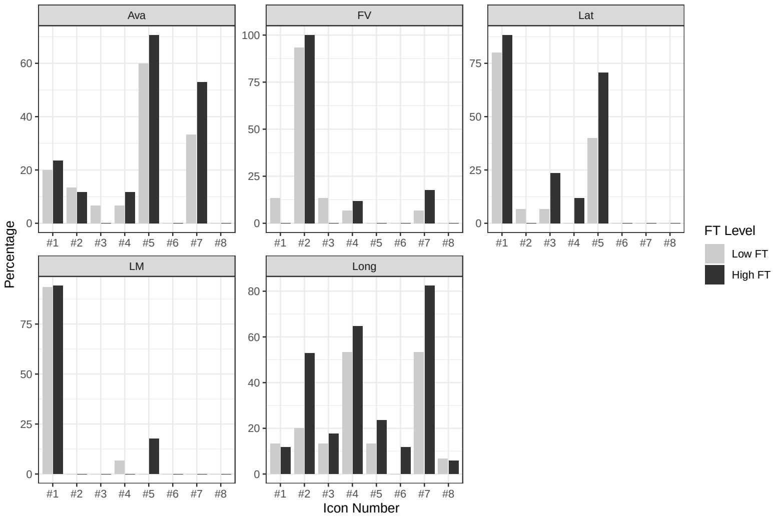


Figure 9. Percentages of BMW icons used to answer TRASS questions from participants with high and low FT.

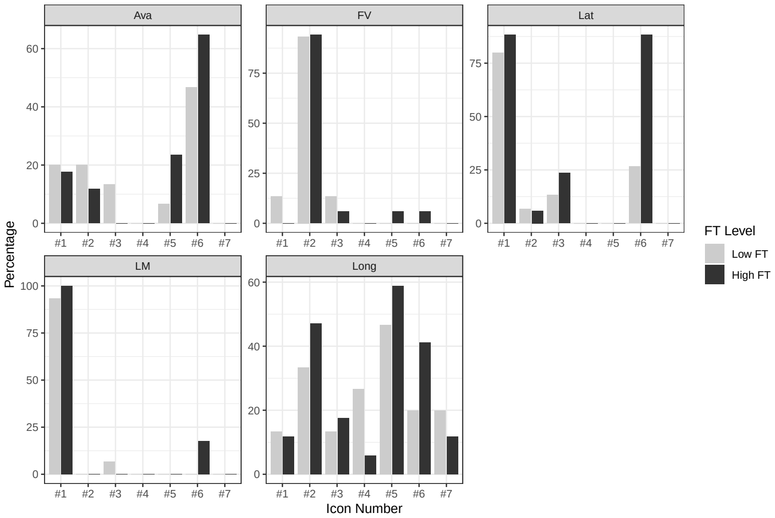


Figure 10. Percentages of Tesla HMI icons used to answer TRASS questions from participants with high and low FT.

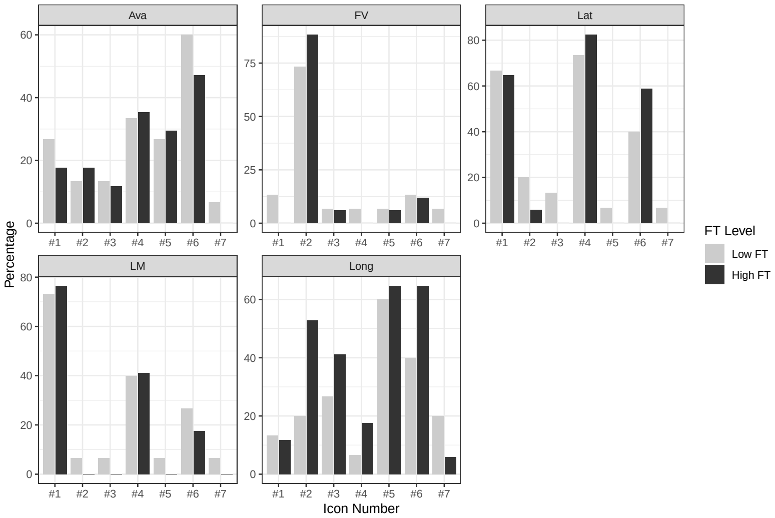


Figure 11. Percentages of VW HMI icons used to answer TRASS questions from participants with high and low FT.

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: Wasington, 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/5fcd24ef92851c00f8588cbf/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. *IFAC-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]

D Article 2: “Human-Machine Interface Evaluation Using EEG in Driving Simulator“

Human-Machine Interface Evaluation Using EEG in Driving Simulator

1st Yuan-Cheng Liu
Chair of Ergonomics
Technical University of Munich
Munich, Germany
yuancheng.liu@tum.de

2nd Nikol Figalova
Department of Clinical and health Psychology
Ulm University
Ulm, Germany
nikol.figalova@uni-ulm.de

3rd Martin Baumann
Department of Human Factors
Ulm University
Ulm, Germany
martin.baumann@uni-ulm.de

4th Klaus Bengler
Chair of Ergonomics
Technical University of Munich
Munich, Germany
bengler@tum.de

***Abstract*—Automated vehicles are pictured as the future of transportation, and facilitating safer driving is only one of the many benefits. However, due to the constantly changing role of the human driver, users are easily confused and have little knowledge about their responsibilities. Being the bridge between automation and human, the human-machine interface (HMI) is of great importance to driving safety. This study was conducted in a static driving simulator. Three HMI designs were developed, among which significant differences in mental workload using NASA-TLX and the subjective transparency test were found. An electroencephalogram was applied throughout the study to determine if differences in the mental workload could also be found using EEG's spectral power analysis. Results suggested that more studies are required to determine the effectiveness of the spectral power of EEG on mental workload, but the three interface designs developed in this study could serve as a solid basis for future research to evaluate the effectiveness of psychophysiological measures.**

I. Introduction

Automated vehicles (AV) have been considered a revolutionary technology for being economical and environmentally friendly, efficient in transportation, and able to increase driving safety [1]. To facilitate this, the interaction between users and AV has been found pivotal [2], making the human-machine interface (HMI) on AV crucial for users to operate it properly. Various studies regarding in-vehicle or external HMI designs have been conducted [3], [4], [10],

However, some researchers argued that current HMI designs are prone to error and confusing for users [5]. Whenever the transitions between different automation levels are made, the distribution of responsibilities between users and AV constantly changes. This could lead to confusion for users and even cause accidents while driving. To solve the problems, it would be essential for HMI designs to be transparent, i.e., transmit the information correctly, efficiently, and understandably.

There are currently numerous guidelines and evaluation methods aiming to help develop a transparent HMI. Still, most studies are either heavily based on experts' perspectives or gathered heuristically and subjectively. Some researchers developed a guideline to design and verify AV HMIs with a thorough itemized checklist to help evaluate whether the HMI design has fulfilled the recommendations [6]. Similarly, some researchers analyzed the HMI design based on usability heuristics and came up with suggestions to improve the HMI design [7]. These methods could provide valuable aspects and help improve HMI designs, but using only subjective and heuristic evaluations would not be enough. Hence, a standardized, objective, and efficient way for HMI design evaluation is urgently required to approach the optimal design for all sorts of different scenarios, levels of automation, and user characteristics.

In the previous study, we developed a standardized transparency assessment test to evaluate the HMI designs that were already available on the market. We evaluated how easy the HMI designs were for users to understand the information transmitted correctly [8]. The resulting TRASS (the proposed Transparency Assessment Method) was calculated by the user's answer accuracy of the given HMI state, and the estimated workload score based on the time to understanding (i.e., the time took to start answering questions). The results show that the proposed method effectively found the difference among the HMI designs, and was validated in an online study. However, the applicability is limited to static scenarios and needs exploration in a higher-fidelity environment. To do so, we need an alternative way to measure the workload in real time and still do it objectively. By doing so, the HMI design process could be handled in a more systematic way, and in return, be more efficient.

In this paper, we intend to evaluate the effectiveness of the electroencephalographic (EEG) in identifying differences in mental workload while interacting with different HMI designs during simulated driving. This evaluation would also be the first step in extending the proposed assessment method into a real driving scenario and developing a real-time HMI assessment method. We first developed three different SAE Level 2 (L2) HMI designs [9], as in Fig. 1, which had different transparencies according to the results from the previous study [8], and should result in different mental workloads during the interaction. To validate this idea, differences in workload among three HMI designs were first evaluated using subjective workload measures. Then, evaluations using the power spectral analysis of the EEG data were conducted, and comparisons between the EEG results and subjective workload measurements were also made.

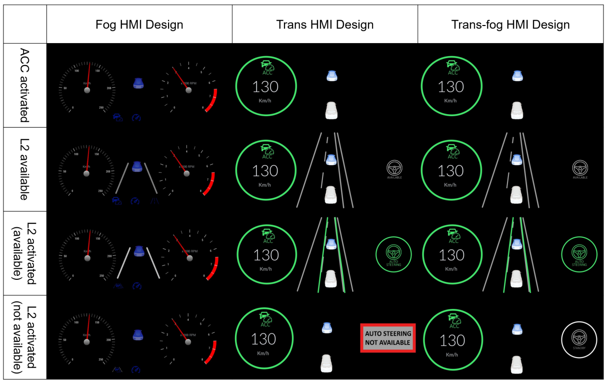


Fig. 1: HMI designs under given circumstances.

The research question is, whether any differences in mental workload among the HMI designs developed exist and if they could be found using analysis of the spectral power of EEG. Thus the corresponding hypotheses are set as follows:

- H1: Differences in mental workloads could be found among different HMI designs using NASA-TLX.
- H2: Differences in mental workloads could be found among different HMI designs using a subjective transparency questionnaire.
- H3: Differences in mental workloads could be found among different HMI designs using the spectral power of EEG.

II. Method

A. Human-machine interface designs

To evaluate whether EEG data is capable of telling the difference in mental workload among different HMI designs, we first need HMI designs that are distinct in understandability, which would result in different mental workloads. Owing to the fact that there hasn't been a standardized evaluation system to determine how understandable the HMI is, a heuristic approach is still needed at this stage. We followed the design principles for human-computer interface and also results from the previous study [8], [12], we developed the following three in-vehicle HMI that should be distinct in HMI understandability and correctness on information transmitted, i.e., HMI transparency:

1) Fog HMI design: The fog HMI design, as in the left column of Fig. 1, represents the HMI design that contains all the design elements which would mitigate the clarity of the information transmitted and hence diminish the HMI transparency for users. Looking at the HMI transparency side, those disadvantageous elements include small icons, which indicate levels of automation and whether the sensor is functioning properly, low contrast color, and redundant icons for the same function. These elements either violate the design principle or are pointed out by the participant in the last study stating that they were confusing and misleading. On the side of system transparency, there would be no indication of failing to activate L2 automation, and it would go straight into standby mode. That is if the L2 automation is activated by the user. Still, the activation failed due to the system limit being reached, there would be no indication that the activation failed and that the AV is not controlling the vehicle laterally.

2) Trans HMI design: The Trans HMI design, as in the middle column of Fig. 1, represents the HMI design that contains all the design elements which would intensify both the HMI and system transparencies. On the HMI transparency side, it fulfills all the design principles, including high contrast color, large icons, and no redundant icons, in contrast to fog HMI design. Regarding system transparency, there is no standby mode, which means if the L2 automation cannot be activated, the system would show a warning on the HMI to remind the users that the L2 automation failed to activate and that they should still control the steering (as shown in the rectangle at the bottom of the middle column of Fig. 1).

3) Trans-fog HMI design: The trans-fog HMI design, as in the right column of Fig. 1, represents the HMI design with the HMI transparency as Trans HMI design, but with minimum system transparency as fog HMI design. It satisfies all the design principles like Trans HMI design and has the same icons and colors. The only difference is when L2 automation is activated but turns out to be unavailable. In a scenario like this, the trans-fog HMI design would go into standby mode like the fog HMI design, resulting in minimum system transparency.

B. Psychophysiological and self-reported measures

In this study, multiple measures are used to evaluate users' cognitive state while interacting with the HMI designs, including psychophysiological and self-reported measures. When evaluating human cognitive states, using only the self-reported method has shown to be less accurate, as participants are usually not precise when it comes to judging their own cognitive state. Psychophysiological measures also have the advantage over other objective methods like behavioral measures. Since those physiological events are not under voluntary control, they could avoid being affected by unrelated constructs like drowsiness or stress during driving. It also has been suggested that psychophysiological measures could significantly increase the accuracy of mental workload measurements [13].

The EEG signal has been found to be a helpful psychophysiological measure when evaluating mental workload in driving scenarios [11], as it provides a non-intrusive way to measure neural activities while also allowing the high temporal resolution to acquire the data in real-time. In this study, we analyzed the oscillatory brain activity derived from EEG signals to determine its relation to specific neurocognitive functions. For instance, decreased alpha power activity (8-12 Hz) and increased theta power activity (4-7 Hz) are often associated with increased mental workload [14].

To the best of our knowledge, there has not been a study to investigate different workloads intrigued by HMI designs with different transparency, so we also included the self-reported measure (i.e., NASA-TLX) as a reference and compared it to results in the literature.

1) EEG signal recording and pre-processing: The EEG was recorded using 32 channels electrodes placed according to the international 10-20 system. ActiCAP set (Brain Products GmbH, Germany) was used, with active shielded electrodes and a LiveAmp amplifier. The data were recorded with a 1000 Hz sampling frequency and preprocessed in Matlab version R2022a.

We created an alternative dataset to perform adaptive mixture independent component analysis (AMICA). The alternative datasets were first downsampled to 500 Hz and bandpass-filtered between 0.1 Hz and 100 Hz. Line noise artifacts were removed using the ZapLine plugin [15], [16]. Channels that correlated with their own robust estimate less than r = .78 more than 50 percent of the time were interpolated, and the data were re-referenced to the common average. We then applied AMICA as the blind source separation using ten iterations [17]. The spatial filter produced by AMICA was then copied into the original, raw dataset, and the components were calculated using IClabel. We used the popularity classifier, which removes components that most likely do not originate in brain activity. Finally, we used a second-order Butterworth filter and bandpass filtered the data between 0.5 Hz and 30 Hz.

The pre-processed data were then epoched into fragments from -10 s to 10 s with regards to the activation of L2 automation, which was the time slot during which participants had to focus on the HMI design in order to do the transparency test that followed. Afterwards, fast Fourier transformations were applied to obtain the spectral power distribution. The relative mean spectral power of the alpha band (8-12 Hz) from bi-lateral parietal electrodes (Pz, P3, P4) and the relative mean spectral power of the theta (4-7 Hz) from bi-lateral frontal electrodes (Fz, F3, F4) were calculated with respect to the total power (0.5-30 Hz) [11].

During the spectral power analysis, all the 20-second epochs gathered with the same HMI design were combined for each participant. This resulted in one 80-second EEG pre-processed data for each HMI design, which was later used for the final statistic analysis.

2) Subjective evaluations: The first subjective evaluation was the mental workload evaluated by the National Aeronautics and Space Administration-Task Load Index (NASA-TLX) [18], which is a six-item questionnaire. In this study, the NASA-TLX scores were calculated without weighted parameters, so we averaged scores across six items. Besides, since each HMI design was measured four times for each participant, we also need to calculate the average again. The final scores calculated were then used as the subjective workload of each HMI design of the participant.

The second subjective evaluation consisted of three questions regarding HMI transparency and was derived from the TRASS test in the previous study [8]. First, participants were asked to evaluate whether they agreed that they could understand the HMI design. Then, they were asked if they agreed that they could obtain critical information from the HMI design. Lastly, they were asked whether they agreed that the HMI design was easy to understand. All three questions were scaled from 0 to 100.

C. Participants

Twelve participants were recruited for this study, where three were male, eight were female, and one was diverse. The ages ranged from 22 to 30, with mean age $= 26.92$, $SD = 3.87$. The data from two of the participants were excluded due to corrupted recordings. All of the participants came with valid driving licenses, and they had held them for at least three years $(M = 8.83, SD = 3.71)$

D. Procedures

The study was conducted in a static driving simulator with a field of view of 120°, as shown in Fig. 2. The front panel consists of three screens with the scenes projected from three projectors respectively. It also has rear, left, and right mirrors, which are small LCD displays. The software SILAB was used to create the 4-lanes highway scenario. A touch screen was fixed on the right-hand side of the driver's seat, where automated cruise control (ACC) and L2 automation activation buttons were shown.

Upon arrival, participants were welcomed and briefly introduced to the study. After finishing the demographic questionnaire and setting up the EEG electrodes, a pre-recorded video containing detailed instructions was played. Then, participants were brought to the driving simulator and started a familiarization test drive. During the test drive, the baseline EEG data were also collected. After the familiarization, the formal test would begin if no symptoms of simulator sickness were shown.

Each participant had to follow the procedure in each trial, as shown in Fig. 3. Five seconds after the activation of the L2 automation (regardless of whether it was successfully activated or not), the simulation ended with a fade-out animation to avoid simulator sickness, and the participant was asked to complete a survey consisting of NASA-TLX and subjective transparency test. Participants had to go through all three HMI designs in counterbalanced order, where each HMI design had four different traffic layouts, making a total of 12 trials for each participant. The EEG signal was recorded throughout the study. Finally, participants were compensated at the end of the study after the feedback section about the procedure, questionnaires, and HMI designs.

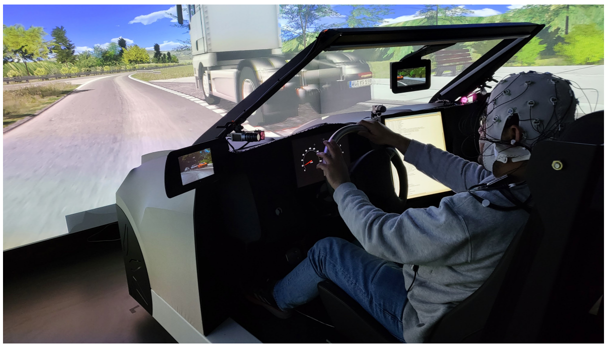


Fig. 2: Illustration of participant with EEG set up in the driving simulator.

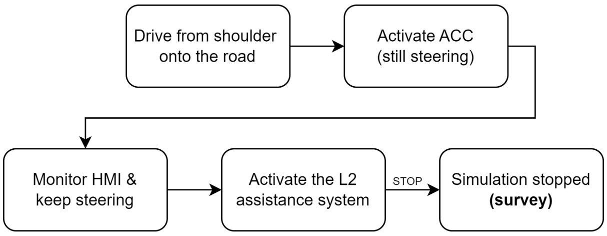

Drive from shoulder onto the road
Activate ACC (still steering)
Monitor HMI & keep steering
Activate the L2 assistance system
STOP
Simulation stopped **(survey)**

Fig. 3: Experimental procedure during each trial.

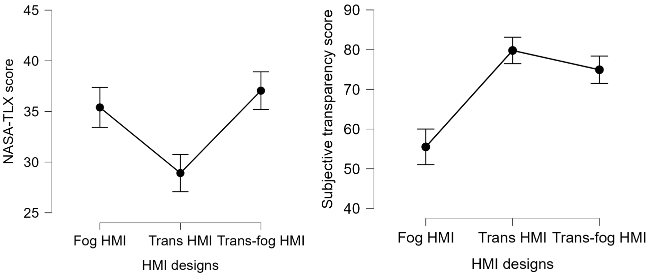
III. Results

To evaluate whether three different HMI designs had any effects on mental workload during driving, we conducted the repeated measure ANOVA to test the significance of the differences in the NASA-TLX and EEG data. The same approach was also used for subjective and objective transparency tests. Multiple comparisons were made with post hoc analysis (Holm's).

A. Subjective evaluations

Fig. 4 shows the resulting NASA-TLX and subjective transparency scores. The averaged NASA-TLX scores were found significantly different among those three HMI designs $F(2, 78) = 5.18, p = 0.008, \eta_p^2 = 0.12$, where Trans HMI had the lowest score among them. A similar outcome was found on subjective transparency scores, where the effect of HMI designs was found to be significant $F(2, 78) = 11.47, p < 0.001, \eta_p^2 = 0.56$.

To gain a better understanding of the relationships among the three HMI designs, post hoc comparisons were conducted for both NASA-TLX and subjective transparency scores. We can see from Table. I that, the Fog HMI possessed a significantly higher mental workload and significantly lower subjective transparency when comparing it to the Trans HMI. However, when compared to the Trans-fog HMI, the significance was only found in its lower subjective transparency. For the Trans and Trans-fog HMIs, no difference in subjective transparency was found, but in NASA-TLX scores, significantly lower mental workload for the Trans HMI was identified.



(a) NASA-TLX score. (b) Subjective transparency score.

Fig. 4: Subjective evaluations with mean values and standard errors of means.

TABLE I: Post Hoc Comparisons on HMI Designs for Subjective Evaluations.

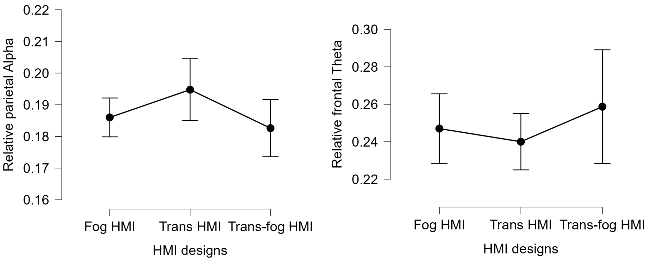
		NASA-TLX		**Subjective transparency**	
		Cohen's d	p_{holm}	Cohen's d	p_{holm}
Fog	Trans	0.33	**0.035**	-1.37	**<0.001**
	Trans-fog	-0.084	0.538	-1.097	**0.006**
Trans	Trans-fog	-0.42	**0.01**	0.27	1.00

B. Mental workload measurement with EEG

In Fig. 5 we can see the power spectral analyses of relative Alpha band power and Theta band power. We can see from Fig. 5 that the Trans HMI, which we expected to have the lowest mental workload, had the highest mean relative Alpha power. Although this result was in agreement with the literature, where decreased mental workload is associated in increased Alpha power, but the difference was not significant $F(2, 18) = 0.55, p = 0.58, \eta_p^2 = 0.058$. Similarly for relative Theta band power, despite the Trans HMI owned the lowest mean value, implying minimum effort was required to understand it, no significant difference was found $F(2, 18) = 0.18, p = 0.83, \eta_p^2 = 0.02$.

IV. Discussion

In this study, we were eager to determine whether different mental workloads could be found in those three developed



(a) Relative Alpha power. (b) Relative Theta power.

Fig. 5: EEG power spectral analysis with mean values and standard errors of means.

TABLE II: Descriptive Data for All Workload Measures and Subjective Transparency.

	EEG power spectral		**Subjective measurements**	
HMI designs	relative Alpha	relative Theta	NASA-TLX	Subjective transparency
Fog	0.186 (0.082)	0.247 (0.053)	35.4 (21.2)	55.5 (18.5)
Trans	0.195 (0.107)	0.240 (0.059)	28.9 (17.7)	79.8 (16.2)
Trans-fog	0.183 (0.074)	0.259 (0.146)	37.1 (19.7)	74.9 (18.3)

HMI designs and if the EEG spectral power analysis could be used to find the differences. This should be deemed as a starting point in developing a systematic and standardized HMI design assessment method, where we intend to incorporate psychophysiological measures into the proposed transparency assessment test, to make the HMI assessment process more objective and efficient.

Three HMI designs were developed and validated based on heuristics and results from the previous study, to have differences in the understandability and the easiness to understand them. The subjective mental workload measured using NASA-TLX confirmed Hypothesis 1. Significantly lower workloads were required by participants to understand the information transmitted by the L2 automated driving system (ADS) when using the Trans HMI. This HMI design was granted with both HMI and system transparency, which means that on the interface side, the information shown on the Trans HMI could be understood with minimum effort; while on the system side, the logic behind L2 automation activation on the Trans HMI is clearer: L2 automation could only be activated when it is available.

During the feedback section, a common reason for L2 ADS owners to turn the L2 function off was that they had no idea if it was on or off. Since users remain responsible throughout the whole time during L2 AV driving, most of the HMI designs expect users to monitor whether L2 is activated the whole time. Hence, even if the activation of L2 automation failed, instead of returning to the previous state and warning the users, the ADS is designed to enter the standby mode, where the lateral control will be automatically activated whenever the system is ready. And we consider this type of system logic intransparent, or as we put in the name of Fog HMI, foggy. The effect of the system transparency on mental workload was confirmed by the pairwise comparison between the Trans and Trans-fog HMIs, where it was the only difference between the two.

The results from the subjective transparency test confirmed Hypothesis 2, and it acted in accordance with the results from NASA-TLX, where Trans HMI remained to be the most transparent, i.e., easiest to understand, HMI design, given that it had the highest mean subjective transparency score. What we can note here is that the Fog and Trans-fog HMIs had around the same NASA-TLX score, but when it comes to subjective transparency score, the Trans-fog HMI had significantly higher values than that of Fog HMI. This could be explained by this additional construct included in the concept of subjective transparency, which is the capability of obtaining the information needed. When evaluating the subjective transparency, besides being asked if they agreed that the HMI design is easy to understand (which is a similar construct as workload), participants were asked if they agreed that they could obtain critical information from the HMI design. By design, the Trans-fog HMI had the advantage of HMI transparency, which allowed users to obtain correct information and thus resulted in higher subjective transparency.

After the confirmation of Hypothesis 1 and 2, we identified three HMI designs with differences in mental workload. Since workload plays an important role in estimating the transparency of HMI designs [8], it is urgently required for us to identify an objective workload measurement to make the proposed method efficient in real driving scenarios. Due to the high temporal resolution and ability to measure neural activities directly, EEG was chosen to be the psychophysiological measure in this study. However, Hypothesis 3 was not confirmed by the results. The possible reason is that the sample size was too small. Due to the loss of recording, the number of usable data dropped, which might explain why the statistical power is lower than expected. A possible solution besides recruiting more participants is to not combine all the epochs as one for each HMI design but keep them separated. When analyzing the spectral power of EEG, data with longer duration could average out the noises, which is why we combined all the epochs around the activations of L2 into one. Keeping them separated could, on the other hand, increase the sample size, and in return, increase the statistical power.

It might also be possible that it is difficult to identify differences in mental workload among HMI designs using the analysis of the spectral power of EEG. In this case, we should consider other psychophysiological measures to extend the transparency assessment method into real driving scenarios. But before that, more studies have to be done to come to that conclusion.

V. Conclusions and Future Work

In conclusion, the first part of the research question was successfully answered and confirmed by the results. This study proposed and validated three HMI designs with differences in mental workload and subjective transparency. We can also learn from the results that, in order to make the HMI design more understandable and easy to use, being only transparent on the interface might not be sufficient, as system transparency also plays a big part in it. More works and discussions on the logic behind the HMI and the ADS are necessary.

This study can also be regarded as the basis for the development of the transparency assessment method. In the future study, we would utilize the validated HMI designs to identify the psychophysiological measures that are effective in identifying differences in mental workload and transparency in real driving scenarios. By doing so, we could make the HMI more transparent, and the riding in automation safer.

References

[1] S. Smith, J. Bellone, S. Bransfield, A. Ingles, G. Noel, E. Reed, and M. Yanagisawa, “Benefits estimation framework for automated vehicle operations.,” Aug. 2015.

[2] M. Körber, A. Eichinger, K. Bengler and C. Olaverri-Monreal, ”User experience evaluation in an automotive context,” 2013 IEEE Intelligent Vehicles Symposium Workshops (IV Workshops), Gold Coast, QLD, Australia, 2013, pp. 13-18, doi: 10.1109/IVWorkshops.2013.6615219.

[3] M. Rettenmaier, M. Pietsch, J. Schmidtler and K. Bengler, ”Passing through the Bottleneck - The Potential of External Human-Machine Interfaces,” 2019 IEEE Intelligent Vehicles Symposium (IV), Paris, France, 2019, pp. 1687-1692, doi: 10.1109/IVS.2019.8814082.

[4] K. Bengler, M. Rettenmaier, N. Fritz, and A. Feierle, “From HMI to HMIs: Towards an HMI Framework for Automated Driving,” Information, vol. 11, no. 2, p. 61, Jan. 2020, doi: 10.3390/info11020061.

[5] O. Carsten and M. H. Martens, “How can humans understand their automated cars? HMI principles, problems and solutions”, Cognition, Technology Work, vol. 21, no. 1, pp. 3–20, 2019.

[6] F. Naujoks, K. Wiedemann, N. Schömig, S. Hergeth, and A. Keinath, ‘Towards guidelines and verification methods for automated vehicle HMIs’, Transportation research part F: traffic psychology and behaviour, vol. 60, pp. 121–136, 2019.

[7] E. L. Parkhurst, L. B. Conner, J. C. Ferraro, M. E. Navarro, and M. Mouloua, ‘Heuristic evaluation of A tesla model 3 interface’, in Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 2019, vol. 63, pp. 1515–1519.

[8] Y.-C. Liu, N. Figalová, and K. Bengler, “Transparency Assessment on Level 2 Automated Vehicle HMIs,” Information, vol. 13, no. 10, p. 489, Oct. 2022, doi: 10.3390/info13100489.

[9] Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles, SAE International, Warrendale, PA, 2021.

[10] N. T. Richardson, C. Lehmer, M. Lienkamp and B. Michel, ”Conceptual design and evaluation of a human machine interface for highly automated truck driving,” 2018 IEEE Intelligent Vehicles Symposium (IV), Changshu, China, 2018, pp. 2072-2077, doi: 10.1109/IVS.2018.8500520.

[11] K. Pollmann, O. Stefani, A. Bengsch, M. Peissner, and M. Vukelić, ‘How to work in the car of the future? A neuroergonomical study assessing concentration, performance and workload based on subjective, behavioral and neurophysiological insights’, in Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, 2019, pp. 1–14.

[12] D. A. Norman, ‘Design principles for human-computer interfaces’, in Proceedings of the SIGCHI conference on Human Factors in Computing Systems, 1983, pp. 1–10.

[13] S. Yang, S. A. R. Hosseiny, S. Susindar, and T. K. Ferris, ‘Investigating driver sympathetic arousal under short-term loads and acute stress events’, in Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 2016, vol. 60, pp. 1905–1905.

[14] S. Mun, M. Whang, S. Park, and M.-C. Park, ‘Effects of mental workload on involuntary attention: A somatosensory ERP study’, Neuropsychologia, vol. 106, pp. 7–20, 2017.

[15] A. de Cheveigné, ‘ZapLine: A simple and effective method to remove power line artifacts’, Neuroimage, vol. 207, no. 116356, p. 116356, Feb. 2020.

[16] M. Klug and N. A. Kloosterman, ‘Zapline-plus: A Zapline extension for automatic and adaptive removal of frequency-specific noise artifacts in M/EEG’, Hum. Brain Mapp., vol. 43, no. 9, pp. 2743–2758, Jun. 2022.

[17] J. A. Palmer, S. Makeig, K. Kreutz-Delgado and B. D. Rao, ”Newton method for the ICA mixture model,” 2008 IEEE International Conference on Acoustics, Speech and Signal Processing, Las Vegas, NV, USA, 2008, pp. 1805-1808, doi: 10.1109/ICASSP.2008.4517982.

[18] S. G. Hart and L. E. Staveland, ‘Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research’, in Advances in psychology, vol. 52, Elsevier, 1988, pp. 139–183.

E Article 3: “Workload Assessment of Human-Machine Interface: A Simulator Study with Psychophysiological Measures“

AHFE 2023: Vol. XX, 2023
doi: 10.54941/ahfeXXXX

Workload Assessment of Human-Machine Interface: A Simulator Study with Psychophysiological Measures

Yuan-Cheng Liu[1], Nikol Figalová[2], Jürgen Pichen[2], Philipp Hock[2], Martin Baumann[2], and Klaus Bengler[1]

[1] Chair of Ergonomics, School of Engineering and Design, Technical University of Munich, 85748 Garching, Germany

[2] Department of Human Factors, Ulm University, 89069 Ulm, Germany

ABSTRACT

Human-machine Interface (HMI) is critical for safety during automated driving, as it serves as the only media between the automated system and human users. To enable a transparent HMI, we first need to know how to evaluate it. However, most of the assessment methods used for HMI designs are subjective and thus not efficient. To bridge the gap, an objective and standardized HMI assessment method is needed, and the first step is to find an objective method for workload measurement for this context. In this study, two psychophysiological measures, electrocardiography (ECG) and electrodermal activity (EDA), were evaluated for their effectiveness in finding differences in mental workload among different HMI designs in a simulator study. Three HMI designs were developed and used. Results showed that both workload measures were able to identify significant differences in objective mental workload when interacting with in-vehicle HMIs. As a first step toward a standardized assessment method, the results could be used as a firm ground for future studies.

Keywords: Human-Computer Interaction, Automated Vehicles, Human-Machine Interface, Mental Workload, Psychphysiological Measure, Transparency

INTRODUCTION

Human-machine Interface (HMI) of Automated Driving Systems (ADS) is a critical component that aims to facilitate an intuitive way for humans to interact with automation (Bengler et al. 2020). However, depending on the level of automation, the roles and responsibilities of human users may be constantly changing. Thus, the design of the HMI plays a crucial part in enabling a safe and efficient transition between the roles by providing critical and understandable information and making the automated system transparent.

When evaluating HMIs, existing methods are mostly based on subjective questionnaires (Richardson et al. 2018; Voinescu et al. 2020), making them prone to biases and difficult to standardize. To resolve the problem, we proposed the Transparency Assessment Method (TRASS) in a previous study (Liu, Figalová, and Bengler 2022), where the transparency toward the HMI is estimated by evaluating the actual understanding and the workload of the user during the interaction. To further apply the TRASS in a dynamic environment (e.g., in simulator or test track studies), the workload estimation method must be adapted.

Multidisciplinary approaches have been used to measure workload in driving scenarios (Stapel, Mullakkal-Babu, and Happee 2019; Kim et al. 2018; Lim, Sourina, and Wang 2018; Matthews et al. 2019). Subjective mental workload measures like NASA-TLX are easy to assess and come directly from participants. However, these subjective mental workload assessment methods alone are usually inaccurate and unable to operate in real-time. On the other hand, psychophysiological measures enable a more objective and continuous assessment of the mental workload. With the non-intrusive sensors, the experimental process becomes more efficient, and the resulting measurements are also more reliable.

In this paper, we present a simulator study where electrocardiography (ECG) and electrodermal activity (EDA) are used to estimate the objective workload of participants when interacting with AV HMIs. The aim of this paper is to evaluate the effectiveness of these psychophysiological measures in finding differences in mental workload when interacting with different HMI designs in simulated driving. To the best of our knowledge, this is the first study that applies these two psychophysiological measures to estimate the mental workload of different AV HMIs. The main novel contribution of this paper is that the results of these real-time and objective workload measures could be further applied for researches in high-fidelity AV environments and HMI design and evaluation processes.

RELATED WORKS

Psychophysiological measures have been widely adopted in the driving context to assess many psychological constructs, such as workload and stress, owing to their sensitivity and accuracy in detecting mental workload (Meng, Zheng, and Huang 2022; Lohani, Payne, and Strayer 2019). Besides, they can also capture dynamic changes in workload that might be difficult to detect with subjective or behavior measures (Charles and Nixon 2019). In this study, we use ECG and EDA for their reliability in short task duration and sensitivity in continuous mental workload detection (Baek et al. 2015; Yoshida et al. 2014).

Electrocardiography

Electrocardiography (ECG) is a method commonly used to capture the electrical activity of the heart. From a series of heart beat waves, heart rate (HR) and heart rate variability (HRV) could be calculated, analyzed, and used to estimate mental workload (Heine et al. 2017; Shakouri et al. 2018). HR was found to increase with the increase in the mental workload, while the HRV decreased.

Heart period (R-R interval) could be derived from the time intervals between heart beats (R peaks). By converting the heart period (usually in milliseconds), we could obtain the heart rate (usually in beats per minute). On the other hand, HRV metrics are more versatile and could be categorized into frequency domain and time domain. In the frequency domain, methods such as low-frequency (LF) power, high-frequency (HF) power, or LF/HF ratio are often used (Alaimo et al. 2020). While in the time domain, the standard deviation of R-R intervals (SDRR) and root mean square of successive differences between normal heartbeats (RMSSD) are widely adopted. However, literature shows that the RMSSD is one of the most robust workload measurements, which is also reliable in ultra-short-term analysis (measurement duration less than 5 mins) (Shaffer and Ginsberg 2017; Baek et al. 2015).

Electrodermal Activity

Electrodermal activity (EDA) reflects changes in the electrical potential of the skin. It has been commonly used as an objective workload indicator in simulators and real driving studies (Yoshida et al. 2014; Daviaux et al. 2020). Two components are usually derived from the EDA signals, which are tonic and phasic. The tonic component is the slowly changing in the electrical conductivity level, also known as skin conductance level (SCL). In comparison, the phasic component describes the event-related change that increases the magnitude of electrical conductance, also called skin conductance response (SCR). Both SCL and SCR are found sensitive to changes in mental workload. SCL was found to increase with the higher workload during real-driving scenarios (Mehler and Reimer 2019), while higher SCR values were found with higher mental demands (Foy and Chapman 2018).

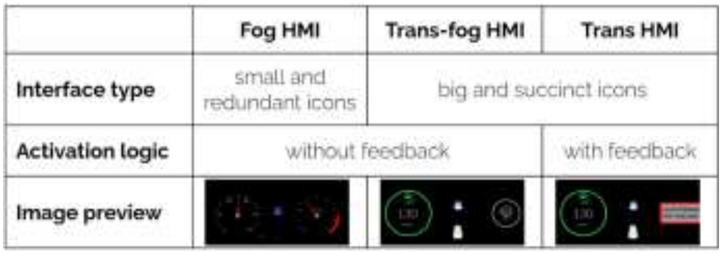
METHODOLOGY

Participants

Twenty-four participants were recruited for this study, where eight were male, 15 were female, and one was diverse. The ages ranged from 22 to 30, with mean age $= 27.92, SD = 4.34$. All of the participants came with valid driving licenses, and they had held them for at least three years ($M = 8.03, SD = 3.78$)

Human-machine Interface Designs

Three different SAE Level 2 HMI designs were developed and applied in the driving simulator, as shown in Figure 1. Following the design principles for human-computer interface and results from the previous study (Liu, Figalová, and Bengler 2022), we developed these three in-vehicle HMIs that are distinct in transparency and workload required to understand. The *Fog HMI design* should require the highest amount of mental workload during the interaction, owing to its small icons, low contrast color, and the fact that there would be no feedback when the system fails to activate. In contrast, the *Trans HMI* provide clear graphical designs, with big and high contrast icon and feedback information when necessary, hence should provide participants with the highest transparency and relatively low mental workload during the automated driving. The last HMI design is the *Trans-fog HMI*, which shares the same visual clarity as the *Trans HMI* but has no feedback information, similar to the *Fog HMI*.



	Fog HMI	Trans-fog HMI	Trans HMI
Interface type	small and redundant icons	big and succinct icons	
Activation logic	without feedback		with feedback
Image preview			

Figure 1: Illustration of HMI designs used in the driving simulator.

In the previous study, these three HMI designs show significant differences in subjective workload measured (NASA-TLX). The same approach would also be included in this study as a reference to the previous one and others in the literature.

Psychophysiological Measures

The ECG electrodes were placed in the common Lead II Configuration, while the EDA electrodes were mounted on the left foot to minimize the noise from the hand and right foot movement during driving. Usually, the EDA electrodes are placed on palmar sites or fingers, but foot placement is recommended when palmar sites or fingers are not available or suitable for placing EDA electrodes (Hossain et al. 2022). All electrodes are connected to a wireless trigger for LiveAmp, and the continuous signals were recorded and saved on a remote device throughout the experiment. The data were recorded with a 500 Hz sampling frequency and preprocessed in the Matlab version R2022a.

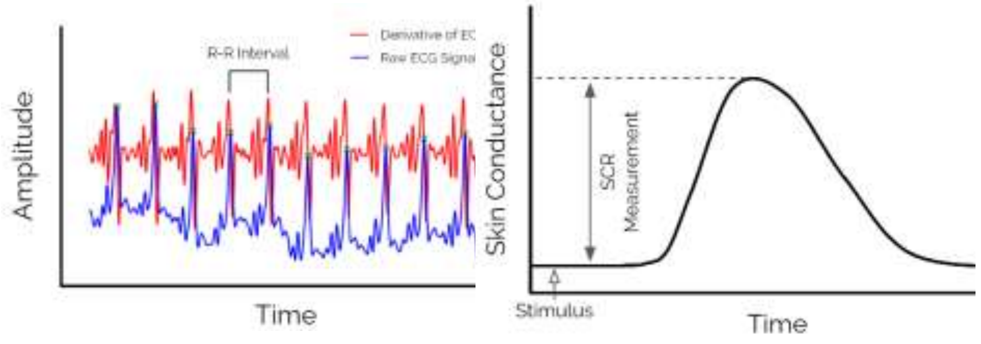


Figure 2: Illustration of the raw ECG signal and processed R-R interval (left) and skin conductance response (right).

Heart Rate Variability (HRV)

In this study, we use the root mean square of successive differences between normal heartbeats (RMSSD) as the metric among the time domain heart rate variability (HRV) for its robustness and sensitivity in ultra-short-term workload measurement. To obtain the RMSSD, we first need to calculate the time differences between consecutive heart beats (or between R peaks). Then, average the squared values of those time differences. Finally, we take the square root of the average obtained and have the RMSSD over the designated duration. Before the RMSSD calculation, the raw ECG signal was filtered and differentiated for clearer R-R intervals, as shown in Figure 2. Ten seconds before and after the activation of the Level 2 ADS were used as an epoch representing the interaction period between the participant and the HMI design. During this time, participants were required to monitor the HMI closely to answer the following question when the driving simulator was paused or stopped.

Skin Conductance Response (SCR)

In this study, we intend to capture the rise of skin conductance with the EDA signal. As the cognitive load rises, so does the skin conductance value, and this change in EDA is the skin conductance response (SCR). After the initial stimulus, there is usually a 2-5 seconds delay before the skin conductance begins to rise (see Figure

2). After the rising phase, the skin conductance value reaches the peak and begins to descend. This difference in skin conductance value between the initial stimulus and the peak after the delay is the SCR magnitude. We collected a 10 seconds epoch after the activation of the Level 2 ADS (stimulus) and calculated the corresponding SCR values.

Self-reported Workload Measures

Two subjective workload measurements were included in this study. The first one is the National Aeronautics and Space Administration-Task Load Index (NASA-TLX) (Hart and Staveland 1988), which is a six-item questionnaire. The average across the six workload related factors was calculated without the weighted parameter. The second subjective workload measurement was derived from the previous study (Liu, Figalová, and Bengler 2022), where participants were asked to evaluate whether they agreed they could understand and obtain critical information from the HMI design. They were asked if they agreed that the HMI design was easy to understand. All three questions were scaled from 0 to 100.

Procedure

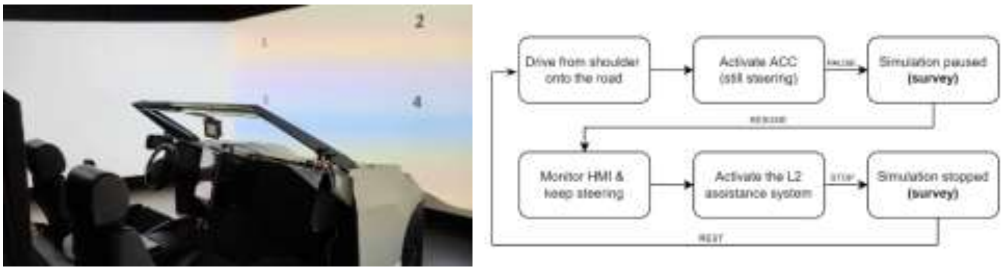

2
4
Drive from shoulder onto the road
Activate ACC (still steering)
PAUSE
Simulation paused (survey)
RESUME
Monitor HMI & keep steering
Activate the L2 assistance system
STOP
Simulation stopped (survey)
REST

Figure 3: Illustration of the driving simulator setup and experimental procedure.

The study was conducted in a static driving simulator with a field of view of 120°, as illustrated in Figure 3. The front panel consists of three screens with the scenes projected from three projectors respectively. It also has rear, left, and right mirrors, which are small LCD displays. The software SILAB was used to create the 4-lanes highway scenario. A touch screen was fixed on the right-hand side of the driver's seat, where automated cruise control (ACC) and L2 automation activation buttons were shown.

When participants arrived, we welcomed them and provided a brief study overview. A pre-recorded video containing detailed instructions was played to ensure all participants received the same information. Afterward, participants were asked to fill in a demographic questionnaire. In the meantime, experimenters set up the ECG and EDA electrodes and ensured signals from the amplifier were normal. Participants were then brought to the driving simulator and started a familiarization test drive. During the test drive, the baseline ECG and EDA data were also collected. After the familiarization, the formal test would begin if no symptoms of simulator sickness were shown.

The experimental procedure of this study is shown in Figure 3. In each trial, one of the three HMI designs was applied randomly, and two surveys were presented to the participants at different stages. Both surveys consisted of a NASA-TLX questionnaire and a subjective transparency test and would later be used to compare to the psychophysiological measures. The first and second surveys were used to estimate the subjective workload the HMI design required when the ACC was activated, and when the Level 2 (L2) ADS was activated. respectively. In each trial, participants were asked to start the vehicle and drive from the shoulder to the right lane on the highway. Meanwhile, they could activate the ACC function on the panel right to the steering wheel whenever they felt comfortable. At the moment the ACC button was pressed, the ADS began the longitudinal control, followed by a five seconds countdown before the simulator went into a pause. When the simulator was paused, all the screens dimmed and the sound effects volume lowered in a fade-away pattern to avoid simulator sick. Then, participants were asked to complete the first survey and inform the experimenter when they finished. After the first survey, the simulation resumed, and participants were asked to activate the L2 ADS whenever they felt comfortable. Five deconds after the L2 button was pressed, the simulation stopped (also in a fade-away pattern), and the second survey was presented and asked to complete carefully. Participants had to go through all three HMI designs in a counterbalanced order, where for each HMI design there would be four different traffic layouts, making a total of 12 trials for each participant. The ECG and EDA signal was recorded throughout the study. Finally, participants were compensated at the end of the study after the feedback section about the procedure, questionnaires, and HMI designs.

RESULTS

In this study, we attempted to determine if different HMI designs would significantly affect the mental workload assessed with psychophysiological measures. However, self-reported workload measures were also used as a reference to the previous work and also to future studies. Repeated measure ANOVA was used to test the significance of the differences in ECG, EDA, NASA-TLX, and subjective transparency data. Multiple comparisons with Holm's correction were made if necessary.

Psychophysiological Measures

We can see from Figure 4 that the same correlation in mental workload among HMI designs is shown, where the Trans HMI demanded the lowest workload during the interaction, followed by the Trans-fog HMI, and lastly, the Fog HMI. The time domain HRV, RMSSD, were found to be significantly different among those three HMI designs $F(2{,}522) = 7.25, p < 0.001, \eta_p^2 = 0.027$, where the Trans HMI had the highest RMSSD, representing that it demanded the lowest cognitive load during the interaction. A similar outcome was found on the SCR values, where the effect of HMI designs was also found to be significant $F(2{,}522) = 3.372, p = 0.035, \eta_p^2 = 0.013$, and that the Trans HMI had the lowest SCR, which again confirm the finding that the Trans HMI demand the lowest cognitive load when interacting with the ADS.

As detailed in the Table. 1., post hoc comparisons were conducted for both psychophysiological measures. The RMSSD of the Trans HMI was found to be significantly higher than both the Fog and Trans-fog HMI designs, representing a lower workload demand than the rest two. No significant difference in RMSSD was found between the Fog and Trans-fog HMI designs. A similar result was found that the SCR of the Trans HMI design was significantly lower than that of the Fog HMI design. However, no significant difference was found between the Fog and the Trans-fog HMI designs and the Trans and the Trans-fog HMI designs.

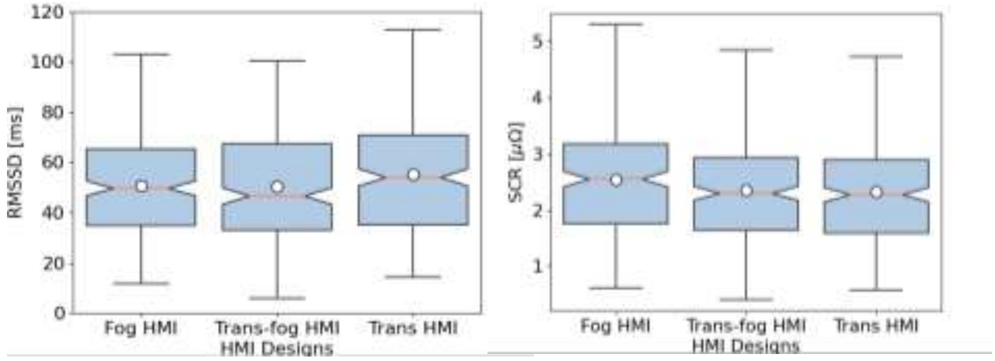


Figure 4: Relationships between psychophysiological measures for workload and HMI designs. (RMSSD: root mean square of successive differences between normal heartbeats. SCR: Skin conductance response)

Table 1. Post Hoc Comparisons on HMI Designs for Psychophysiological Measures.

		RMSSD		SCR	
		t	P_{holm}	t	P_{holm}
Fog	Trans	-3.31	**0.003**	2.43	**0.047**
	Trans-fog	-0.024	0.98	2.09	0.075
Trans	Trans-fog	3.29	**0.003**	-0.33	0.744

Self-reported Workload Measure

We see the same pattern in self-reported mental workload among the HMI designs from Figure 5, where the Trans HMI had the lowest subjective workload and highest self-reported transparency. The NASA-TLX scores were found to be significantly different among those three HMI designs $F(2{,}260) = 3.80, p = 0.028, \eta_p^2 = 0.027$, where the Trans HMI got the lowest averaged NASA-TLX score, representing that it required the lowest subjective mental workload. A similar outcome was found on the subjective transparency scores, where the effect of HMI designs was also found to be significant $F(2{,}260) = 49.88, p < 0.001, \eta_p^2 = 0.27$.

Post hoc comparisons were conducted for self-reported workload measures, as shown in Table. 2. The Trans HMI design had significantly lower NASA-TLX and higher subjective transparency scores than the Fog HMI design. But no significant difference was found between the Trans HMI and the Trans-fog HMI designs in NASA-TLX or subjective transparency scores. Between the Fog HMI and the Trans-fog HMI designs, the subjective transparency score of the Trans-fog HMI design was significantly higher than that of the Fog HMI design. But the effect was not found between these two in the NASA-TLX scores.

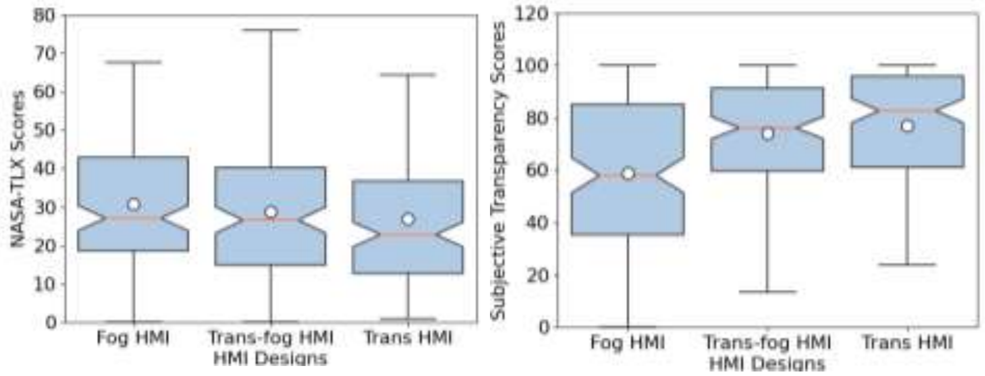


Figure 5: Relationships between self-reported measures for workload and HMI designs.

Table 2. Post Hoc Comparisons on HMI Designs for Psychophysiological Measures.

		NASA-TLX		Subjective Transparency	
		t	p_{holm}	*t*	p_{holm}
Fog	Trans	2.75	**0.019**	-8.71	< **0.001**
	Trans-fog	1.25	0.27	-7.15	< **0.001**
Trans	Trans-fog	-1,51	0.27	1.77	0.079

DISCUSSIONS

We attempted to evaluate the sensitivity and effectiveness of psychophysiological measures in identifying different mental workloads required when interacting with different in-vehicle HMI designs in this study. We used the three HMI designs developed in the previous study, which were found to result in different self-reported workloads as the workload variable. Results from RMSSD suggested that the ECG measurement and the HRV analysis can be used to find differences in mental workload effectively. The Trans HMI was found to have the highest RMSSD during the interaction with participants, which suggests that it required the lowest workload among the three HMI designs. This was in accordance with the HMI designs and the other psychophysiological measures. The SCR was found to be the lowest on the Trans HMI design, representing that the Trans HMI design demanded the lowest workload since the rise of the SCR corresponded to the increase of the workload. Hence, the EDA, together with the SCR, could also be applied to investigate the differences in workload among HMI designs.

The self-reported workload measures were also confirmed to have the same correlation in workload with the previous study and psychophysiological measures, where the Trans HMI design had higher NASA-TLX scores than the Trans-fog HMI design, and the Trans-fog HMI design than the Fog HMI design. It was also not surprising that the Trans HMI design had the highest score in the subjective transparency test since the lower the workload needed to understand the HMI design, the higher the transparency should be.

Traditionally, the HMI evaluation processes are usually subjective and heuristic, making it difficult to be efficient and standardized. Hence, to develop a standardized HMI evaluation method, the inclusion of objective measures is critical and necessary. In our previous work, we developed a transparency assessment method, which evaluates HMI transparency by combining the true understanding of the HMI and the workload required during the interaction. Now with the results from this study, the mental workload during the interaction between the users and the HMI designs could be estimated continuously and efficiently. Combining that with the proposed transparency assessment method, the assessment of the HMI designs could be adapted to environments with more dynamic interactions and be more efficient and reliable.

CONCLUSIONS AND FUTURE WORKS

In this study, we confirmed the effectiveness of two psychophysiological measures in evaluating the mental workload when interacting with in-vehicle HMI design. This finding could be a strong basis for the HMI evaluation process. The next step is to include the psychophysiological measure in the proposed transparency assessment method, to develop an objective and standardized HMI assessment method, and use it to increase the efficiency of the HMI design process.

ACKNOWLEDGMENT

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.

REFERENCES

Alaimo, Andrea, Antonio Esposito, Calogero Orlando, and Andre Simoncini. 2020. "Aircraft Pilots Workload Analysis: Heart Rate Variability Objective Measures and NASA-Task Load Index Subjective Evaluation." *Aerospace* 7 (9): 137.

Baek, Hyun Jae, Chul-Ho Cho, Jaegeol Cho, and Jong-Min Woo. 2015. "Reliability of Ultra-Short-Term Analysis as a Surrogate of Standard 5-Min Analysis of Heart Rate Variability." *Telemedicine and e-Health* 21 (5): 404–14.

Bengler, Klaus, Michael Rettenmaier, Nicole Fritz, and Alexander Feierle. 2020. "From HMI to HMIs: Towards an HMI Framework for Automated Driving." *Information* 11 (2): 61.

Charles, Rebecca L, and Jim Nixon. 2019. "Measuring Mental Workload Using Physiological Measures: A Systematic Review." *Applied Ergonomics* 74: 221–32.

Daviaux, Yannick, Emilien Bonhomme, Hans Ivers, Étienne de Sevin, Jean-Arthur Micoulaud-Franchi, Stéphanie Bioulac, Charles M Morin, Pierre Philip, and Ellemarije Altena. 2020. "Event-Related Electrodermal Response to Stress: Results from a Realistic Driving Simulator Scenario." *Human Factors* 62 (1): 138–51.

Foy, Hannah J, and Peter Chapman. 2018. "Mental Workload Is Reflected in Driver Behaviour, Physiology, Eye Movements and Prefrontal Cortex Activation." *Applied Ergonomics* 73: 90–99.

Hart, Sandra G, and Lowell E Staveland. 1988. "Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research." In *Advances in Psychology*, 52:139–83. Elsevier.

Heine, Tobias, Gustavo Lenis, Patrick Reichensperger, Tobias Beran, Olaf Doessel, and Barbara Deml. 2017. "Electrocardiographic Features for the Measurement of Drivers' Mental Workload." *Applied Ergonomics* 61: 31–43.

Hossain, Md-Billal, Youngsun Kong, Hugo F Posada-Quintero, and Ki H Chon. 2022. "Comparison of Electrodermal Activity from Multiple Body Locations Based on Standard EDA Indices' Quality and Robustness Against Motion Artifact." *Sensors* 22 (9): 3177.

Kim, Jungsook, Woojin Kim, Hyun-suk Kim, and Daesub Yoon. 2018. “Effectiveness of Subjective Measurement of Drivers’ Status in Automated Driving.” In *2018 IEEE 88th Vehicular Technology Conference (VTC-Fall)*, 1–2. IEEE.

Lim, WL, O Sourina, and Lipo P Wang. 2018. “STEW: Simultaneous Task EEG Workload Data Set.” *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 26 (11): 2106–14.

Liu, Yuan-Cheng, Nikol Figalová, and Klaus Bengler. 2022. “Transparency Assessment on Level 2 Automated Vehicle HMIs.” *Information* 13 (10): 489.

Lohani, Monika, Brennan R Payne, and David L Strayer. 2019. “A Review of Psychophysiological Measures to Assess Cognitive States in Real-World Driving.” *Frontiers in Human Neuroscience* 13: 57.

Matthews, Gerald, Ryan Wohleber, Jinchao Lin, Gregory Funke, and Catherine Neubauer. 2019. “Monitoring Task Fatigue in Contemporary and Future Vehicles: A Review.” In *Advances in Human Factors in Simulation and Modeling: Proceedings of the AHFE 2018 International Conferences on Human Factors and Simulation and Digital Human Modeling and Applied Optimization, Held on July 21–25, 2018, in Loews Sapphire Falls Resort at Universal Studios, Orlando, Florida, USA 9*, 101–12. Springer.

Mehler, Bruce, and Bryan Reimer. 2019. “How Demanding Is" Just Driving?" a Cognitive Workload-Psychophysiological Reference Evaluation.” In *Driving Assesment Conference*. Vol. 10. 2019. University of Iowa.

Meng, Xiaorong, Wei Zheng, and Kang Huang. 2022. “Cognitive Load Evaluation of Human-Computer Interface Based on EEG Multi-Dimensional Feature.” In *2022 IEEE 25th International Conference on Intelligent Transportation Systems (ITSC)*, 1536–41. https://doi.org/10.1109/ITSC55140.2022.9922203.

Richardson, Natalie Tara, C Lehmer, Markus Lienkamp, and Britta Michel. 2018. “Conceptual Design and Evaluation of a Human Machine Interface for Highly Automated Truck Driving.” In *2018 IEEE Intelligent Vehicles Symposium (IV)*, 2072–77. IEEE.

Shaffer, Fred, and Jay P Ginsberg. 2017. “An Overview of Heart Rate Variability Metrics and Norms.” *Frontiers in Public Health*, 258.

Shakouri, Mahmoud, Laura H Ikuma, Fereydoun Aghazadeh, and Isabelina Nahmens. 2018. “Analysis of the Sensitivity of Heart Rate Variability and Subjective Workload Measures in a Driving Simulator: The Case of Highway Work Zones.” *International Journal of Industrial Ergonomics* 66: 136–45.

Stapel, Jork, Freddy Antony Mullakkal-Babu, and Riender Happee. 2019. “Automated Driving Reduces Perceived Workload, but Monitoring Causes Higher Cognitive Load Than Manual Driving.” *Transportation Research Part F: Traffic Psychology and Behaviour* 60: 590–605.

Voinescu, Alexandra, Phillip L Morgan, Chris Alford, and Praminda Caleb-Solly. 2020. “The Utility of Psychological Measures in Evaluating Perceived Usability of Automated Vehicle Interfaces–a Study with Older Adults.” *Transportation Research Part F: Traffic Psychology and Behaviour* 72: 244–63.

Yoshida, Ryuichi, Tomohiro Nakayama, Takeki Ogitsu, Hiroshi Takemura, Hiroshi Mizoguchi, Etsuji Yamaguchi, Shigenori Inagaki, et al. 2014. “Feasibility Study on Estimating Visual Attention Using Electrodermal Activity.” *International Journal on Smart Sensing and Intelligent Systems* 7 (5): 1–4.