# Why Do I Have to Take Over Control? Evaluating Safe Handovers with Advance Notice and Explanations in HAD

Frederik Wiehr
Anke Hirsch
Lukas Schmitz
Nina Knieriemen
Antonio Krüger
{firstname}.{lastname}@dfki.de
German Research Center for Artificial Intelligence (DFKI)
Saarland Informatics Campus, Germany

Alisa Kovtunova
Stefan Borgwardt
{firstname}.{lastname}@tu-dresden.de
Institute of Theoretical Computer Science
TU Dresden, Germany

Ernie Chang
Vera Demberg
{cychang,vera}@coli.uni-saarland.de
Saarland University
Saarland Informatics Campus, Germany

Marcel Steinmetz
Jörg Hoffmann
{steinmetz,hoffmann}@cs.uni-saarland.de
Saarland University
Saarland Informatics Campus, Germany

## ABSTRACT

In highly automated driving (HAD), it is still an open question how machines can safely hand over control to humans, and if an advance notice with additional explanations can be beneficial in critical situations. Conceptually, use of formal methods from AI – description logic (DL) and automated planning – in order to more reliably predict when a handover is necessary, and to increase the advance notice for handovers by planning ahead at runtime, can provide a technological support for explanations using natural language generation. However, in this work we address only the user's perspective with two contributions: First, we evaluate our concept in a driving simulator study (N=23) and find that an advance notice and spoken explanations were preferred over classical handover methods. Second, we propose a framework and an example test scenario specific to handovers that is based on the results of our study.

## CCS CONCEPTS

• **Human-centered computing** → **Empirical studies in HCI**; • **Applied computing** → *Transportation*; • **Computing methodologies** → *Planning for deterministic actions*; • **Theory of computation** → *Description logics.*

## KEYWORDS

highly automated driving; criticality; safety; cyber-physical system; car driving; multimodal interaction; human-computer interaction; natural language generation; description logic; automated planning; critical interaction; interdisciplinary

**ACM Reference Format:**
Frederik Wiehr, Anke Hirsch, Lukas Schmitz, Nina Knieriemen, Antonio Krüger, Alisa Kovtunova, Stefan Borgwardt, Ernie Chang, Vera Demberg, Marcel Steinmetz, and Jörg Hoffmann. 2021. Why Do I Have to Take Over Control? Evaluating Safe Handovers with Advance Notice and Explanations in HAD. In *Proceedings of the 2021 International Conference on Multimodal Interaction (ICMI '21), October 18–22, 2021, Montréal, QC, Canada.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3462244.3479884


*ICMI '21, October 18–22, 2021, Montréal, QC, Canada*

ACM ISBN 978-1-4503-8481-0/21/10…$15.00
https://doi.org/10.1145/3462244.3479884

## 1 INTRODUCTION

Highly automated driving (HAD, [4]) is about to hit public roads[1] and with that, computers will participate in actions and decisions that affect humans; this will soon also be the case for other cyber-physical systems (CPS) like drones, e.g. for delivery services. A recent study found that SAE level 2 [50] (Standard SAE International) Tesla drivers have significant driving experience and high self-rated computer expertise, and they care about how automation works. Very common automation failures like incorrectly detected lanes and hard braking were not perceived as risky by this user group, as they stressed the importance of being alert and having the technical limitations in mind [14]. If the car encounters a high-risk situation, it will still need to efficiently support the user in regaining situational awareness [6, 17, 27, 64]; otherwise, human life is inherently at stake.

The German Federal Office for Motor Vehicles (KBA) reported that the Tesla autopilot poses a significant traffic risk after testing it for several thousand kilometers. Among others, one criticism was that it did not provide an explanation or notification to the driver when it ran into a highly critical situation. The root cause of the problem is that most CPS do not have any built-in concepts to explain their behavior. These systems rely on their own static design instead of effectively communicating their state, current

[1]https://innovationatwork.ieee.org/new-level-3-autonomous-vehicles-hitting-the-road-in-2020/

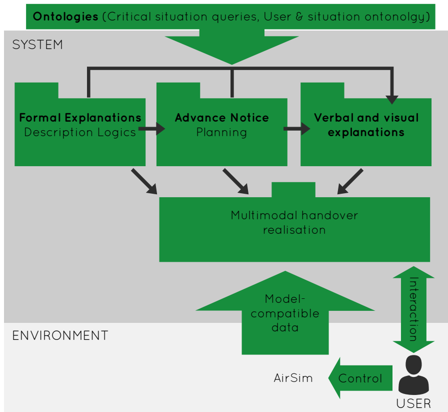


**Figure 1: The proposed system architecture**

problems, and possible solutions to the user. For HAD, a summary report containing only relevant information before the handover should be provided. It is crucial to explain the situation to the human in a concise and understandable manner. The nature of the explanation directly depends on the advance notice, i.e., the time available before a critical handover occurs.

To evaluate the approach of advance notice and explanations, we conducted a driving simulator study (N=23) in Microsoft AirSim [52] to compare advance notice and explanations to the classical handover without explanations (notification only). In Section 5 we discuss the lessons learned from our study, and derive suggestions for an evaluation framework for handover methods in HAD as an extended discussion.

## 2 RELATED WORK

The related work in HCI can be separated into user modeling, situation awareness/vigilance, and multimodal interaction. User modeling enables a system to adapt and make assumptions about the current user's goals, plans, knowledge and (possibly false) beliefs, i.e., maintain a conceptual understanding of the user (user model [32]), in which user differences need to be modeled explicitly – first in dialogue systems, and nowadays for a broad range of personalized applications ranging from museum guides [55] to recommender systems [48]. Clearly, handovers also must be tailored to individual users [43].

Situation awareness describes the human's awareness of the environment [17], including awareness of critical information for a task at hand, mainly known from aviation but also applied to HAD [56]. Such information depends greatly on the situation (e.g. a pilot approaching an airport vs. a driver navigating in dense traffic). The definition and recognition of situations in relation to machine operations have been studied. Adams et al. [3] identified human factors that can be measured during operation to assess the situation awareness of human operators.

Two approaches can be distinguished depending on sensor placements: body-worn sensors measure human activity directly [39], while control elements (e.g. the steering wheel of a car, the flight stick of a plane) allow indirect assessment of human activities [29]. Cognitive effects for assistive systems can be measured, e.g., using self-assessments and questionnaires [35] or the Index of Cognitive Activity (ICA; [40]), a pupillometric measure, in dual task studies involving simultaneous language processing [12, 13, 18].

In highly automated systems, the risk of humans being out of the loop increases and thus a potential handover is more difficult to achieve [17]. The degree of *vigilance* plays an important role; it describes the ability of a human to attend to the environment although the current situation might not require much attention. Prior research has developed vigilance measures based on questionnaires, and sensors assessing heart rate, eye movement and skin conductance [16].

Multimodal interaction has the potential to increase usability and thus the safety of operation [9]. It has been used for mobile applications and environments, including gesture and speech [62], eye tracking and face detection [45], and gaze-based [51, 61] and tangible interaction [28] to adapt to the user's needs. The style of visualization can improve trust in autonomous driving [24]. The authors found that using a spatial representation of the environment to visualize the situation and the actions of the autonomous vehicle improved performance significantly.

Prior work has also identified feedback factors that help to improve the understandability of and trust in system decisions made in autonomous driving, i.e., an explanation why ("obstacle ahead") is preferred over information about how something happens, e.g., "the car is braking" [33]. Visually, a world-in-miniature visualization or a human-like chauffeur avatar can increase trust in automation [24].

A closely related study [60] shows user preference in car-driver handovers prompted by multimodal (auditory and visual) warnings; however, in some cases, the visual component of the warning may obstruct the view of the driver. Other studies showed that an auditory alert ahead of time caused participants to look more at the road before the handover and disengage from the secondary task earlier [58]. Also, giving early additional audio-visual explanations about an appropriate upcoming maneuver leads to faster responses and longer time to collision with obstacles [7]. The latter two support the main motivation of generating an advance notice with explanations. Explanations that build trust in autonomous vehicles have been researched to be adapted in terms of context, cognitive skills, alertness, contextual knowledge, and time available. The authors describe such adaptations in the form of different roles, such as developer, assurance, end-user, and external explanation needs [22]. Shen et al. investigated in which situations explanations become necessary, as well as how they need to be changed for different scenarios or driver types and can be predicted with a learning-based model [53].

Other framework descriptions take the perspective of attention management [27] or provide an attention-aware architecture for the integration of handheld consumer devices to increase the lead time for transitions that increase safety and comfort [64].

The time for transfer of control (ToC) in critical situations has been examined in several studies [11, 23, 44]. A review of 25 studies on urgent take-over scenarios by Eriksson and Stanton [19] revealed that the mean allowed time for ToC was 6.37 ± 5.36 seconds and the mean reaction time was 2.96 ± 1.96 seconds. Mok et al. [44] examined driver behavior when drivers had to take over control before they encountered a road hazard with takeover times of 2, 5, and 8 seconds. Gold et al. [23] examined when a driver must engage with the driving task again so as to have safe handling in critical situations by comparing two take-over times, namely 5 and 7 seconds, and also compared this to manual driving. The results showed that for a shorter takeover time, the reaction of the driver was faster and the decision process was quicker. However, the quality of take-over was worse than it was given the longer ToC time. Following a similar approach, Dambök et al. [11] conducted a study which compared three different ToC times (4, 6 and 8 seconds). Their aim was to find the best case boundary where the driver would be able to handle even the most difficult case. In contrast to the studies on ToC for critical take-over situations, Eriksson and Stanton [19] investigated ToC time in a non-critical scenario. They conducted a within-subject experiment with three different driving conditions (manual driving, highly automated driving and highly automated driving with a secondary task). The results showed an increase in take-over time if the transfer of control happened in non-critical scenarios. Merat et al. [41] found that drivers were able to regain stable control over the vehicle after around 40 seconds.

A meta-analysis of 129 studies found that the mean takeover time in SAE level 2 or higher was lower when the urgency of the take-over situation was higher, and performing a secondary task task with a tablet or similar yielded increased mean takeover times [65]. In order to better prime drivers for the ToC, van der Heiden et al. [57] investigated auditory pre-alerts triggered well before the actual ToC request.

# 3 TECHNOLOGICAL SUPPORT FOR EXPLANATIONS WITH ADVANCE NOTICE

We outline the architecture that conceptualizes the necessary technological support to realize safe handovers in mixed-initiative control (see Figure 1). Our concept addresses the specific requirements during highly critical handover situations, through advance notice and explanations. Our contribution consists in preparing the ground for a long-term research agenda.

Our architecture concept leverages formal methods from AI – *description logic* and *automated planning* – in order to more reliably predict when a handover is necessary, to increase the advance notice for handovers, and to generate explanations of the reasons for handovers. It further uses methods from *human-computer interaction* and *natural language generation* to develop solutions for safe and smooth handovers.

The component "Advance Notice" enables the system, based on automated planning [21, 49], to decide about handovers ahead of time, at run-time, based on the current circumstances (e.g., traffic/weather conditions in HAD) and the likelihood of entering a critical situation. To predict such situations, automated planning uses an abstract description of the world. To assess the criticality of a situation and to generate explanations of critical situations, description logic (DL) [5] methods are used. This allows the system to specify constraints, to generate higher-level information, to detect inconsistencies, and to generate formal, logical explanations. The third component mainly uses techniques from natural language generation (NLG) [47] to generate verbal and visual explanations suitable for human consumption. It transforms the information generated via DLs and planning to inform the user about the system state. Finally, methods from HCI, such as user modeling [26], support a multimodal handover realization based on information.

Environment information (such as sensor data or any other knowledge) needs to be provided in a model-compatible way, i.e., translated into the ontologies underlying the system. For a prototypical realization, AirSim as indicated in Figure 1 is one option, but our concept is generic.

# 4 USER STUDY: USER HANDOVER PREFERENCES FOR ADVANCE NOTICE

In the user study, we decided to focus on users' handover preferences for advance notice when combined with additional verbal explanations for four fixed critical example handover scenarios, by creating mock-up scenarios. We thereby explicitly compare different behaviors of a simulated intelligent system (an approach similar to Wizard-of-Oz studies).

## 4.1 Methods

In our experiment, participants drove in a driving simulator in four different driving scenarios while being engaged in a secondary task. After each scenario, they answered two questionnaires. At the end, they took part in a semi-structured interview.

*4.1.1 Participants.* Overall, 23 participants, aged 21–29 ($M$ = 25.13, $SD$ = 2.42, 8 males), were recruited using social media, mailing lists and posters on campus. They were paid 10 euros. The screening criteria for the participants were that they were German native speakers and held a valid driver's license for cars, since (a) the questionnaire and the spoken instructions were in German and (b) the driver's license ruled out significant visual impairments. Overall, participants self-rated their prior driving experience rather high ($M$ = 4.04, $SD$ = 0.77, on a 5-point Likert-like rating scale) and reported a medium general trust in autonomous vehicles ($M$ = 2.78, $SD$ = 0.95, also on a 5-point Likert-like rating scale). Only two participants had already driven with driver-assistance systems (including lane assist and adaptive cruise control), but 16 participants would drive autonomous vehicles if they were available. Their driving experience, measured in kilometers driven in the last year, ranged from the categories 0–5,000km to 20,000–25,000km.

*4.1.2 Conditions.* Most previous studies about HAD analyzed the timing for the handover requests [10, 23], but they did not focus on an explanation of why the handover was occurring. Depending on the length of the explanation, more time to address the handover may be needed. In the current study, both are taken into account, which leads to the following timeline conception. Here, the handover timeline consists of three essential points. The first two, Point A and Point B, are in temporal relation to the last point, the critical situation onset (see Figure 3). The figure represents an example scenario with high criticality. However, in order to avoid learning effects, different but similarly critical scenarios were implemented for the user study as shown in the supplemental video. The critical situation onset was the beginning of a construction zone. Point A occurred 30 seconds before that and depending on the condition, different information was presented to the user.

At Point B, which was 15 seconds before the critical situation onset, an acoustic signal was given (a beep). The fifteen seconds between point B and the critical situation onset was the time window for the handover, where the participants were to take over control

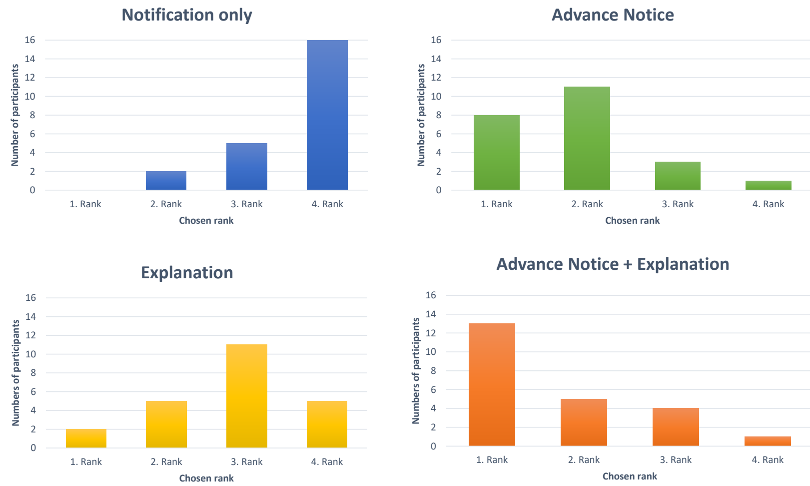


**Figure 2: The participants' preferred ranking of conditions with 1 = best and 4 = worst**

of the car. During that window, another car appeared in front of the participant in the driving simulator and showed unusual driving behavior. We decided on the 15 second time frame after reviewing literature with the criticality of the situation and the secondary task as determinants for takeover time [65] and piloting our specific implemented scenarios to allow for a safe but also quick response. The 12–15 second timeframe was found to allow for sufficient time to look at the road for 3–4 seconds, put one's hands on the wheel and feet on the pedals, push the button for confirmation (7–8 seconds) and take a first glance at the mirror and instruments (12–15 seconds) to indicate situational awareness [36].

In the four conditions of the study, we simulate highly autonomous cars that feature a planning component, an explanation component, both, or none. Planning enables the car to initiate the handover 15 seconds earlier (at point A) whereas the explanation component alone can only provide an explanation of the current situation (at point B, 15 seconds before the critical situation).

**Notification Only:** The driver is notified with a beep at point B without further explanation.

**Advance Notice:** At point A, the user gets a generic advance notice: 'Please pay attention to the road. In 15 seconds, you will get a handover request.' At point B, a beep sounds.

**Explanation:** At point B, the user gets an explanation directly after the beep, e.g.: 'Be alert to the erratic red car ahead with three lane changes in the last 10 seconds. The driver may be ill or drunk. You are required to take over before reaching the construction zone!'

**Advance Notice + Explanation:** At point A, the user gets an advance notice and the same explanation as in the previous condition. At point B, a beep sounds.

The participant had to take over control by the critical situation onset. We accompany this paper with a video that shows footage of all conditions and the study apparatus.

*Driving scenarios.* To avoid learning effects, we randomly combined the following driving scenarios with the previously defined four conditions and slightly adapted the spoken notifications. All scenarios used the same critical case, which in general was erratic behavior of another car leading to a takeover request and an additional explanation depending on the condition. The actual takeover request was always indicated with a beep in all conditions, with a break after the speech output. A generic speech output was also generated in the conditions without explanations.

All scenarios featured an approaching construction zone in an urban environment. Additional criticality was introduced by the following road users: **S1:** drunk/reckless driver (erratic steering), **S2:** car remaining in the left lane, **S3:** very slow driver with sudden braking, **S4:** truck with unsecured load.

When choosing the explanatory sentences, we followed two orthogonal goals: First, the explanations should provide sufficient information for the purpose of informing the subject of the surroundings. Secondly, it is also imperative not to overload the text with too many details on the situation, so as to reduce cognitive overhead.

*4.1.3 Tasks & Design.* Participants played the game 2048 [2] on a tablet as a secondary task while the driving simulator was in automation mode. Similar to prior studies [7, 58], the participants did not have to immediately give up their secondary task if an advance notice was given. This was in all conditions then clearly indicated with a beep as the handover request. At this point, they were required to pause the game, to take over control and confirm this with a button press on the steering paddle shifters. We chose the game 2048 [2] as it can easily be paused to regain situational awareness during automation phases.

The study was designed as a within-subjects experiment, i.e., after completing the training phase, each participant ran through four driving scenarios. The order of conditions was counterbalanced by a Williams design Latin square (LS) [63]. The conditions were randomly combined with the four aforementioned different driving scenarios, such that every scenario appeared only once per participant and was combined with all conditions over all participants at least once.

*4.1.4 Procedure.* First, participants were briefed about the experimental procedure and their primary and secondary task. For their

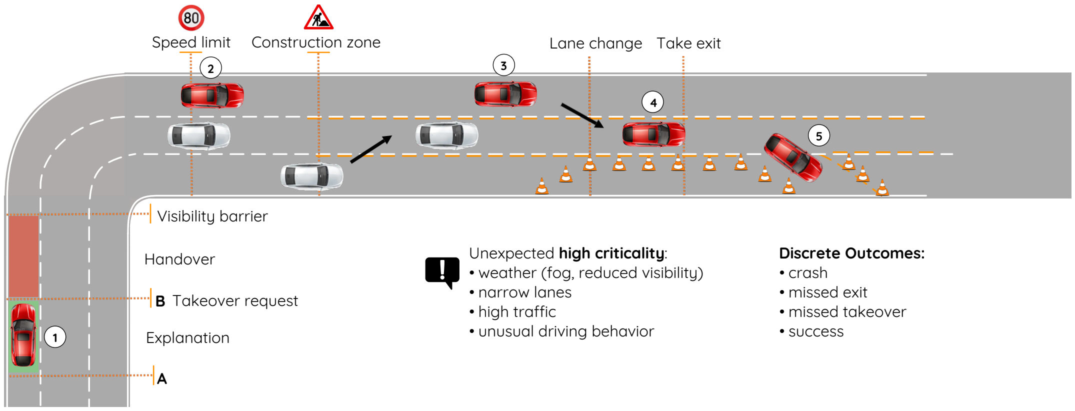


**Figure 3: Example handover task with high criticality: the numbers in circles indicate discrete time steps of the same car in a single critical scenario. The red car has to perform a lane change before entering a construction zone, where it has to take the exit amid other traffic (indicated by the white car). The criticality of the situation can be further increased/decreased by the factors listed and the discrete outcome can be used as a performance measure.**

primary task, the experimenter explained HAD and the concept of a handover to them. It was clearly stated that a beep always signals a handover request, after which the participant will have enough time to confirm the handover with a button press on the steering wheel. The participants were intentionally not informed about the technical background of the conditions, e.g., the role of the planner and description logic. For the secondary task, the experimenter explained the game 2048 [2] on a tablet and the participants could try it out briefly. In the following training phase of the driving simulator, participants could familiarize themselves with the handover, the manual control of the car, the specific beep sound, and the spoken instructions.

Afterwards, they were reminded of their task again, and the study phase started. They experienced four handover scenarios and after each trial, the participant filled in the Trust in Automation Questionnaire (TiA) [34] and NASA Task Load Index (NASA TLX) [25] on the tablet.

At the end, participants filled in a post-study questionnaire in which some demographic data and their driving experience during the experiment were assessed. Furthermore, they were asked to rank the four handover conditions according to their personal preference, and a short semi-structured interview about their reasons for the ranking was conducted. Overall, the experiment lasted approximately one hour per participant. During the experiment, the participants' handover behavior and the secondary task were recorded from behind.

*4.1.5 Apparatus.* A three-monitor integrated driving seat for gaming (InsideRace Sport Triple) with a Logitech G27 steering wheel and pedals was used. One of the two paddle shifter buttons had to be pressed to confirm the handover. The system was set to automatic shifting so only the throttle and brake pedals had to be controlled. The AirSim Simulator [52] ran on a PC with a 3.40GHz i7-6700 CPU, 24GB RAM, NVIDIA GeForce GTX 1080, and a modified version of the Windridge City environment [1]. A tripod with a DSLR camera was used for filming the secondary task interaction and driving simulator screen only.

*4.1.6 Variables & Hypotheses.* The four conditions were used as the four factor levels of the independent variable. The dependent variables, trust in automation and workload, were measured with the TiA [34] and the NASA TLX [25]. The TiA [34] consists of 19 items which are combined into the following six scales: Reliability/Competence, Understanding, Familiarity, Intention of Developers, Propensity to Trust and Trust. The items are answered on a 5-point rating scale from 1 (= strongly disagree) to 5 (= strongly agree). The unweighted NASA TLX [25] consists of the following six subscales with a scale of 20 points and a total workload over all six subscales: Mental Demand, Physical Demand, Temporal Demand, Performance, Effort, and Frustration. Additionally, the participants' ranking order of the conditions was assessed, from 1 (= best) to 4 (= worst). Although we assessed the two questionnaires, in this study we focus mainly on the qualitative results, since the aim was to get the individuals' opinions about the system to improve the experimental paradigm for further studies. Thus, it was assessed how the participants described their experiences in the semi-structured interview.

Since the focus is on the qualitative results, the hypotheses for the questionnaires are condensed. The workload as well as its subscales are expected to be higher for the condition *Notification Only* than for the other conditions. The trust in automation as well as its subscales are expected to be higher in the conditions *Advance Notice + Explanation* and *Advance Notice* than in the other conditions. We expect the condition *Advance Notice + Explanation* to be ranked highest by most participants and the condition *Notification Only* lowest.

## 4.2 Results

*4.2.1 Quantitative Results.* Participants rated their overall driving simulation experience during the study with a mean of 3.22 ($SD$ = 0.90) on a 5-point Likert-like rating scale from 1 (= very difficult) to 5 (= very easy). A multivariate analysis of variance (MANOVA) revealed no significant overall difference between the conditions; Pillai's Trace = .41, $F(36, 237) = 1.04$, $p = .422$. Correlations between dependent variables were low ($r < .90$), indicating that multicollinearity was not a confounding factor in the analysis [59]. Since we had directed hypotheses, the between-subject effects were still considered. For this, we chose Bonferroni adjustment, which resulted in a $p$ value of 0.004.

Levene's test showed homoscedasticity for all subscales. Only the subscale temporal demand of the NASA TLX turned out to differ significantly between the four conditions, $F(3, 88) = 5.08$, Bonferroni-adjusted $p < .004$. Participants rated the temporal demand higher in the *Notification Only* condition ($M = 6.57$, $SD = 4.34$) compared to the other three conditions (*Advance Notice*: $M = 2.48$, $SD = 2.43$; *Explanation*: $M = 5.53$, $SD = 3.74$; *Advance Notice + Explanation*: $M = 5.35$, $SD = 4.11$). Thus, participants felt more time-pressured in the *Notification Only* condition.

The other scales from the NASA TLX and the TiA did not differ significantly between the four conditions.

For the ranking of the four conditions (1 = best, 4 = worst), we used a Friedman test which revealed a significant difference in the ranking of the conditions, $\chi^2(3) = 33.13$, $p < .001$ with a moderate effect size (Kendall's $W = .48$); refer to Fig. 2. *Notification Only*, with a median ranking of 4, was never chosen as the first. 16 participants ranked it fourth. The ranks for the other three conditions ranged from 1 to 4 and also showed a clear tendency of participants' choice towards their calculated median. *Explanation* showed a median ranking of 3. The median ranking for *Advance Notice* was 2 and for *Advance Notice + Explanation* it was 1. 13 participants ranked that condition first and only one ranked it last.

For the post-hoc pairwise comparisons Bonferroni-Holm adjustment was used for correcting the alpha levels, resulting in four significant comparisons (see Table 1). The condition *Advance Notice + Explanation* was preferred over *Notification Only* ($z = 1.91$, $p < .001$). Additionally, *Advance Notice* was preferred over *Notification Only* ($z = 1.74$, $p < .001$). Furthermore, *Advance Notice + Explanation* was ranked higher than *Explanation*, ($z = 1.13$, $p = .003$), and *Advance Notice* was ranked higher than *Explanation* ($z = 0.96$, $p = .012$). The first two comparisons showed medium effect sizes, $r = 0.40$ and $r = 0.36$ respectively. The latter have small effect sizes, $r = 0.24$ and $r = 0.20$ respectively.

*4.2.2 Qualitative Results.* For the qualitative evaluation we transcribed the audio recordings into text and arranged the texts according to the condition. Then, we compared the participants' statements on each condition based on similarities and differences according to the hypotheses, grouped them content-wise, and finally translated them into English. *Notification Only* was referred to as "too fast", "too sudden" or "too pressuring" by nine participants. They described a feeling of not being able to adapt to the situation because there was not enough time. Eight participants said that only the beep is "not enough", "too little" or there is "not enough information." One participant rated this condition to be "terrible". On the other hand, two participants rated this condition as "simple enough".

The condition *Explanation*, where at point B the explanation appeared directly after the beep, was referred to as "the most complicated" condition by two participants, as well as "distracting". One participant summarized: "At first, there was the beep and it was clear I should take over, but then I forgot because the explanation distracted me." Another two were startled by this condition. Four participants felt pressured or unsure in this situation. On the other hand, one participant defined this condition in the ranking at the end as "okay" and two others said it helped them focus.

The condition *Advance Notice* came out as the best in the semi-structured interview at the end. One participant declared "it was the clearest for me and also the easiest to understand" which sums up six participants' perceptions of this condition as "clear" and "easy to understand". Another three participants described it as "efficient" and stated that they had "more trust" in this system. Five participants felt more "prepared". One participant preferred it to the condition *Advance Notice + Explanation* because "this way I could get an idea of the situation by myself."

The combined condition *Advance Notice + Explanation* was mostly rated in the context of preparation and attention. 13 participants felt "prepared", said their "attention was more focused" or described the system as "affirmative." One participant summed it up this way: "It was kind of reassuring, having the system telling you 'Take care about the car in front; it crossed the lanes multiple times.'" However, eight participants, partially overlapping with the aforementioned 13, gave the remark that this condition takes "too long" or delivers "too much information."

## 4.3 Discussion

The ranking and the qualitative data strongly support our hypothesis that the condition *Advance Notice + Explanation* are expected to be ranked highest by most participants and the condition *Notification Only* lowest. Unfortunately, only the significant difference in the temporal demand subscale of the NASA TLX supported the other hypotheses. Participants' rating of their perceived workload and most of its subconstructs, as well as their trust in the system, did not differ significantly between the conditions. It is possible that they really perceived no difference in their workload (and its subconstructs) and trust in the system. One reason could be the short timespan of the handover situation. Another possibility for the absence of significance in the subscales of NASA TLX and TiA (except for temporal demand) could be that due to the within-subjects design, we only had one data point per participant per condition. For future research, on the one hand, we suggest altering the design: either using more handover situations in each condition when using a within-subjects design, or using a between-subjects design. On the other hand, one should also consider using different constructs and questionnaires which point toward the direction of safety rather than trust.

While the participants consciously ranked some conditions better than others, this was only rarely shown in their self-assessed workload and trust in automation. For *Advance Notice + Explanation* a clear choice preference was visible since 13 participants ranked it as the best condition. On the other hand, *Notification Only* was never chosen as the best, but rather as the worst condition by 16 participants. Additionally, *Advance Notice* was placed first and second by eight and eleven participants, respectively, and was only ranked fourth by one participant. Thus, *Advance Notice + Explanation* and *Advance Notice* seem to be the best conditions for signaling a handover. They led to the feeling of having had a good, safe handover for our participants. In contrast, *Notification Only* should not be considered for handover situations.

| | | $z$ | $Z$ | $p$-value | alpha | $r$ |
|---|---|---|---|---|---|---|
| Advance Notice + Explanation | vs. Notification Only | 1.91 | 5.03 | <.001 | .008 | 0.40 |
| Advance Notice | vs. Notification Only | 1.74 | 4.57 | <.001 | .01 | 0.36 |
| Advance Notice + Explanation | vs. Explanation | 1.13 | 2.97 | .01 | .013 | 0.24 |
| Advance Notice | vs. Explanation | 0.96 | 2.51 | .003 | .017 | 0.20 |
| Explanation | vs. Notification Only | 0.78 | 0.78 | .040 | .025 | |
| Advance Notice + Explanation | vs. Advance Notice | 0.17 | 0.17 | .648 | .05 | |

**Table 1: Results for the pairwise comparisons using Bonferroni-Holm adjustment from lowest to highest $p$-value with the adjusted alpha levels. The conditions were ranked from 1 = best to 4 = worst. $\chi^2(3) = 33.13$, $p < .001$, Kendall's $W = .48$. $z$ (test statistic) = difference between the mean ranks from the Friedman test for the two groups, $Z$ = standardized $z$-value, $r$ = effect size. Standard error = 0.38 for each comparison. Mean ranks of the conditions: $M$(*Advance Notice + Explanation*) = 1.70, $M$(*Advance Notice*) = 1.87, $M$(*Explanation*) = 2.83, $M$(*Notification Only*) = 3.61.**

One possibility for why *Explanation* and *Notification Only* were seldom or never ranked first could be the time aspect. When the driver suddenly hears the beep and has to react immediately, they cannot plan their action beforehand; the same applies when, after the beep, the explanation is presented during the actual handover. Then, it could even be distracting from the handover. This was supported by the participants' statements in the semi-structured interviews. The advance notice, on the other hand, gives a timely notification, which is preferable for planning one's actions. From the participants' statements, the impression was that they rated the condition with only the advance notice better than the combined condition, since the latter took "too long." Some participants were overwhelmed by the amount of information the system delivered.

The difference in the two conditions is that the *Advance Notice* only tells the driver to focus on the road and warns them about the upcoming handover ("*Please pay attention to the road. In 15 seconds you will get a handover request.*"). The *Advance Notice + Explanation* additionally gives an explanation of why the car in front shows unusual driving behavior, which leads to a longer speech sequence ("*Be alert to the erratic red car ahead with 3 lane changes in the last 10 seconds. The driver may be ill or drunk. You are required to take over before reaching the construction zone!*") This additional information might shift the focus of the driver away from the actual upcoming handover because they might try to validate the explanation given by the system. To address this issue, one participant suggested that the driver should only get more information if they want to, which would depend on their trust in the system. Another participant suggested two beeps as notification: "the first one for attention, then the advance notice and then the other beep when one has to take control." These could be important improvements for further research on the paradigm.

Overall, the results of this first study suggest a clear research path for the explanation components of our system. The components generating explanations should do so as concisely as possible, extracting only the relevant facts about the current situation. The handover realization component should emphasize shortening this explanation even further, and using different modalities that are easier to grasp. The combination of verbalization, visual cues, and simple auditory signals should be chosen depending on a user model that takes into account the user's trust and awareness.

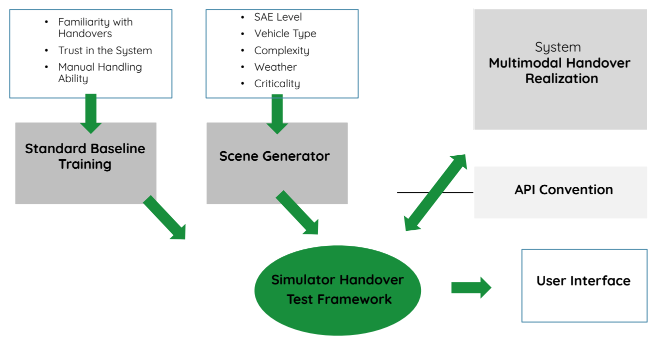


**Figure 4: Evaluation Framework for Handover Situations of Cyber-Physical Systems**

### 4.4 Problems and Limitations

Although we planned the study according to our power analyses (with G*Power) that yielded 48 participants, only 23 were actually tested because of the university shut-down due to the coronavirus pandemic. The limited number of participants is also the reason for the within-subject design. Given that all the participants were rather young drivers, the age range limits the results of the study as it might also affect the overall driving experience.

One participant told us in the semi-structured interview that they were "getting more attentive" with each driving scenario, leading to more focus on the driving simulator than on the secondary task. To improve the paradigm, further studies that use the same design should repeat each condition multiple times, which would lead to more data points and habituation to the task.

## 5 FRAMEWORK DESIGN SUGGESTIONS FOR HANDOVERS WITH EXPLANATIONS

Our study shows that there is a need for an evaluation framework for handover situations with explanations, especially in HAD, because of the dimensionality of possible scenarios, output modalities, system architectures and individual differences of users. In the automotive industry there exist international standards to evaluate the manual handling performance of a vehicle with the double lane change test (ISO 3888-2). However, with the introduction of HAD, we think there is a need to assess the interaction with the autopilot in a standard simulator setup, in which a controlled baseline can be established. As a result of our qualitative evaluation, we found that there is a challenge in establishing the baseline with regard to trust in the system and initial familiarity, i.e. the participants do not have the knowledge of the developers, nor do they have a way to assess the reliability of the system in the first place.

In the following, we outline a study framework and procedure as a result of our evaluation of how handover situations for CPS could be assessed in a more standardized way, to improve comparability and validity of future studies (see Figure 4). The design of the framework is driven by the question of what the system can or cannot handle from the perspective of the user. Secondly, we need to be able to quickly generate a variety of testing scenarios by simply modifying parameters. And lastly, to be able to compare different multimodal handover realizations and system implementations, a standard API to exchange the core of the system's behavior is necessary.

There are different dependent variables here that can be assessed. We used the TiA [34] and the NASA TLX [25] to assess trust in the automated system and the perceived workload (including temporal, physical, and mental demand as well as frustration), respectively. Derived from the NASA TLX, another tool to assess cognitive load while exercising a secondary task could be the Driver Activity Load Index (DALI; [30, 46, 54]) because it is more tailored to the automotive context. One possibility is to compare each condition with the baseline, e.g. only driving. Another is to compare questionnaires in the different conditions. Additionally, the usability of the system can be assessed with the System Usability Scale (SUS; [8, 54]). Llaneras, Salinger, and Green also evaluated trust in the system, but additionally they assessed the "comfort with the system, perceived vigilance and willingness to engage in secondary tasks" (p. 94, [38, 42]). They also assessed what kind of secondary task the participants would be willing to perform. Thus, a possible independent variable could be the secondary tasks or the (mental) demand they take [15, 31]. As another perspective, objective measures can be incorporated into the analysis as well, e.g. steering accuracy or takeover time.

If different groups are assessed, then the best statistical analysis would be comparing the means of the dependent variables between the groups. Thus, depending on the number of dependent variables, either a $t$-test would suffice, or a MANOVA should be taken into account. For a more *standardized baseline training*, trust in the HAD system is of the essence. The user has to know what the system is capable and incapable of to develop a trusting attitude [20, 33, 37]. Therefore, a longer training phase is necessary. This way, the user learns how the HAD works in critical situations, in which situations it interferes and in which ways it shows support. Thus, they are not taken by surprise in the study phase when it comes to using the system. For our study framework, we suggest a standard baseline training phase that consists of many common situations that would normally appear during the main phase of the study. This has the purpose of giving the user sufficient opportunity to get used to not only the normal manual handling, e.g. steering the car, but also what the autopilot can handle and what it cannot. We suggest including situations with higher criticality which the autopilot is still able to handle, to establish a baseline of trust in the system.

If the conducted study uses a within-subjects experimental design like we did, there should be more repetitions of each condition. But from what we learned, we suggest a between-subjects design, when the conditions vary as they do in our study. If the conditions do not vary that much, e.g. only the length of the explanation of the handover is altered, a within-subjects design is preferable.

As a first example for *scene generation*, we describe an example scenario for HAD in this framework: we suggest approaching a construction zone with the vehicle or driving simulator to induce high criticality (see Figure 3). The explanation is given between A and B, after which the actual handover occurs. The level of criticality in this scenario can be varied, e.g., by the number of cars (traffic), lane width, and reduced visibility (weather). The layout and location of the construction zone should be varied slightly to avoid learning effects.

In terms of the practicality of the system, the time between notification and the critical situation depends on how far ahead information can be collected. 30 seconds is around 300–800 meters ahead at normal inner city speeds, which is already quite far. There is known information from the map and prior 3D imaging but current AI systems have only a partial real-time model of other road users (e.g. Tesla Full Self Driving v9). If the static information can be collected *a priori* from the traffic network, e.g. on the road situation (traffic jam, etc.) and position of construction sites, the dynamic information about the behavior of a car on the road ahead could be received with the help of car-to-car communication. Certainly, it is unrealistic to foresee 30s in advance that someone in close proximity will run a red traffic light or drastically change their behavior. So in contrast to "long-term" critical situations where we will have time to explain more, there are more "short-term" ones where a different notification is necessary. However, if a "short-term" critical situation happens, the knowledge about new erratic drivers must be transmitted to all cars at a certain distance, such that an advanced notice can be generated for them instead.

The reliability of the systems for detecting critical situations mostly builds on situation modeling, a car's processor power to run the methods formalizing the model, and availability of the input data for these methods. And if, in the real world, the former two are no longer a big issue, the latter point heavily depends on how many users on a road have agreed to transmit information collected by their cars. Making an automated decision in a setting with partially available data about a road situation will be a major research question in future work.

## 6 CONCLUSION

With the advance of highly automated driving and other cyber-physical systems affecting humans, the design of suitable techniques for handover from machine to human becomes more and more pressing. Many such handover problems inherently require advance notice (to the extent possible), as well as an explanation of what is going on, to give the user context. The authors believe that these two features inherently require world behavior models and reasoning; certainly, it is natural to support the features in this manner. Our user study shows promise and yields important insights for the design of future studies.

## ACKNOWLEDGMENTS

This work was funded by DFG grant 389792660 as part of TRR 248.

# REFERENCES


[1] 2019. Windridge City. https://naturemanufacture.com/windridge-city/

[2] 2020. 2048 (video game). https://2048game.com

[3] Marilyn Jager Adams, Yvette J. Tenney, and Richard W. Pew. 1995. Situation Awareness and the Cognitive Management of Complex Systems. *Hum Factors* 37, 1 (1995), 85–104.

[4] Jackie Ayoub, Feng Zhou, Shan Bao, and X. Jessie Yang. 2019. From Manual Driving to Automated Driving. In *Proceedings of the 11th International Conference on Automotive User Interfaces and Interactive Vehicular Applications - AutomotiveUI '19*. ACM Press.

[5] Franz Baader, Ian Horrocks, Carsten Lutz, and Uli Sattler. 2017. *An Introduction to Description Logic*. Cambridge University Press.

[6] Shadan Sadeghian Borojeni, Lewis Chuang, Wilko Heuten, and Susanne Boll. 2016. Assisting Drivers with Ambient Take-over Requests in Highly Automated Driving. In *Proceedings of the 8th International Conference on Automotive User Interfaces and Interactive Vehicular Applications* (Ann Arbor, MI, USA) *(Automotive'UI 16)*. Association for Computing Machinery, New York, NY, USA, 237–244.

[7] Shadan Sadeghian Borojeni, Lars Weber, Wilko Heuten, and Susanne Boll. 2018. From Reading to Driving: Priming Mobile Users for Take-over Situations in Highly Automated Driving. In *Proceedings of the 20th International Conference on Human-Computer Interaction with Mobile Devices and Services* (Barcelona, Spain) *(MobileHCI '18)*. Association for Computing Machinery, New York, NY, USA, Article 14, 12 pages.

[8] John Brooke et al. 1996. SUS - A Quick and Dirty Usability Scale. *Usability Evaluation in Industry* 189, 194 (1996), 4–7.

[9] Philip R. Cohen and David R. McGee. 2004. Tangible Multimodal Interfaces for Safety-critical Applications. *Commun. ACM* 47, 1 (2004), 41–46.

[10] Daniel Damböck and Klaus Bengler. 2012. Übernahmezeiten beim hochautomatisierten Fahren. In *5. Tagung Fahrerassistenz*.

[11] D. Damböck, M. Farid, L. Tönert, and K. Bengler. 2012. Übernahmezeiten beim hochautomatisierten Fahren. In *Autofahren. 5. Tagung Fahrerassistenz*. München, Germany.

[12] Vera Demberg and Asad Sayeed. 2016. The Frequency of Rapid Pupil Dilations As a Measure of Linguistic Processing Difficulty. *PLOS One* 11, 1 (2016), e0146194.

[13] Vera Demberg, Asad Sayeed, Angela Mahr, and Christian Müller. 2013. Measuring Linguistically-induced Cognitive Load during Driving Using the ConTRe Task. In *ACM International Conference on Automotive User Interfaces and Interactive Vehicular Applications*. 176–183.

[14] Murat Dikmen and Catherine M. Burns. 2016. Autonomous Driving in the Real World: Experiences with Tesla Autopilot and Summon. In *Proceedings of the 8th International Conference on Automotive User Interfaces and Interactive Vehicular Applications* (Ann Arbor, MI, USA) *(Automotive'UI 16)*. Association for Computing Machinery, New York, NY, USA, 225–228.

[15] Thomas A. Dingus, Sheila G. Klauer, Vicki Lewis Neale, Andy Petersen, Suzanne E. Lee, Jeremy Sudweeks, Miguel A. Perez, Jonathan Hankey, David Ramsey, Santosh Gupta, et al. 2006. *The 100-car Naturalistic Driving Study, Phase II - Results of the 100-car Field Experiment*. Technical Report.

[16] Mica R. Endsley. 1995. Measurement of Situation Awareness in Dynamic Systems. *Hum Factors* 37, 1 (1995), 65–84.

[17] Mica R. Endsley. 1996. Automation and Situation Awareness. *Automation and Human Performance: Theory and Applications* (1996), 163–181.

[18] Nikolaos Engonopoulos, Asad B. Sayeed, and Vera Demberg. 2013. Language and Cognitive Load in a Dual Task Environment. In *Annual Meeting of the Cognitive Science Society (CogSci)*. 2249–2254.

[19] A. Eriksson and N. A. Stanton. 2017. Take-over Time in Highly Automated Vehicles: Non-critical Transitions to and from Manual Control. In *Human Factors*.

[20] David Garcia, Christine Kreutzer, Karla Badillo-Urquiola, and Mustapha Mouloua. 2015. Measuring Trust of Autonomous Vehicles: A Development and Validation Study. In *Communications in Computer and Information Science*. Springer International Publishing, 610–615.

[21] Malik Ghallab, Dana S. Nau, and Paolo Traverso. 2004. *Automated Planning - Theory and Practice*. Elsevier.

[22] Jon Arne Glomsrud, André Ødegårdstuen, Asun Lera St Clair, and Øyvind Smogeli. 2019. Trustworthy Versus Explainable AI in Autonomous Vessels. In *Proceedings of the International Seminar on Safety and Security of Autonomous Vessels (ISSAV) and European STAMP Workshop and Conference (ESWC)*. 37–47.

[23] Christian Gold, Daniel Damböck, Lutz Lorenz, and Klaus Bengler. 2013. "Take Over!" How Long Does It Take to Get the Driver Back Into the Loop? *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 57, 1 (Sept. 2013), 1938–1942.

[24] Renate Haeuslschmid, Max von Buelow, Bastian Pfleging, and Andreas Butz. 2017. Supporting Trust in Autonomous Driving. In *International Conference on Intelligent User Interfaces (IUI)*. 319–329.

[25] Sandra G. Hart and Lowell E. Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. In *Advances in Psychology*. Vol. 52. Elsevier, 139–183.

[26] Dominik Heckmann and Antonio Krüger. 2003. A User Modeling Markup Language (UserML) for Ubiquitous Computing. In *International Conference on User Modeling (UM)*. 393–397.

[27] Christian P. Janssen, Shamsi T. Iqbal, Andrew L. Kun, and Stella F. Donker. 2019. Interrupted by My Car? Implications of Interruption and Interleaving Research for Automated Vehicles. *Int. J. Hum. Comput. Stud.* 130 (2019), 221–233.

[28] Vaiva Kalnikaite, Yvonne Rogers, Jon Bird, Nicolas Villar, Khaled Bachour, Stephen Payne, Peter M. Todd, Johannes Schöning, Antonio Krüger, and Stefan Kreitmayer. 2011. How to Nudge in Situ: Designing Lambent Devices to Deliver Salient Information in Supermarkets. In *International Conference on Ubiquitous Computing (UbiComp)*. 11–20.

[29] Steven J. Kass, Kerstan S. Cole, and Claudia J. Stanny. 2007. Effects of Distraction and Experience on Situation Awareness and Simulated Driving. *Transportation Research Part F: Traffic Psychology and Behavior* 10, 4 (2007), 321–329.

[30] Dagmar Kern, Angela Mahr, Sandro Castronovo, Albrecht Schmidt, and Christian Müller. 2010. Making Use of Drivers' Glances onto the Screen for Explicit Gaze-based Interaction. In *Proceedings of the 2nd International Conference on Automotive User Interfaces and Interactive Vehicular Applications - AutomotiveUI '10*. ACM Press.

[31] Sheila G. Klauer, Feng Guo, Jeremy Sudweeks, and Thomas A. Dingus. 2010. *An Analysis of Driver Inattention Using a Case-crossover Approach on 100-car Data*. Technical Report.

[32] Alfred Kobsa and Wolfgang Wahlster. 1989. *User Models in Dialog Systems*. Springer.

[33] Jeamin Koo, Jungsuk Kwac, Wendy Ju, Martin Steinert, Larry Leifer, and Clifford Nass. 2014. Why Did My Car Just Do That? Explaining Semi-autonomous Driving Actions to Improve Driver Understanding, Trust, and Performance. *International Journal on Interactive Design and Manufacturing (IJIDeM)* 9, 4 (April 2014), 269–275.

[34] Moritz Körber. 2018. Theoretical Considerations and Development of a Questionnaire to Measure Trust in Automation. In *Congress of the International Ergonomics Association*. Springer, 13–30.

[35] Antonio Krüger, Ilhan Aslan, and Hubert Zimmer. 2004. The Effects of Mobile Pedestrian Navigation Systems on the Concurrent Acquisition of Route and Survey Knowledge. In *International Conference on Human-Computer Interaction with Mobile Devices and Services (MobileHCI)*. 446–450.

[36] M. Kuehn, Tobias Vogelpohl, and M. Vollrath. 2017. Takeover Times in Highly Automated Driving (level 3).

[37] J. D. Lee and K. A. See. 2004. Trust in Automation: Designing for Appropriate Reliance. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 46, 1 (Jan. 2004), 50–80.

[38] Robert E. Llaneras, Jeremy Salinger, and Charles A. Green. 2013. Human Factors Issues Associated with Limited Ability Autonomous Driving Systems: Drivers' Allocation of Visual Attention to the Forward Roadway. In *Proceedings of the 7th International Driving Symposium on Human Factors in Driver Assessment, Training, and Vehicle Design: Driving Assessment 2013*. University of Iowa.

[39] Paul Lukowicz, Holger Junker, and Gerhard Tröster. 2004. Automatic Calibration of Body Worn Acceleration Sensors. In *International Conference on Pervasive Computing*. 176–181.

[40] Sandra P. Marshall. 2002. The Index of Cognitive Activity: Measuring Cognitive Workload. In *IEEE Conference on Human Factors and Power Plants*. 7–5–7–9.

[41] Natasha Merat, A. Hamish Jamson, Frank C. H. Lai, Michael Daly, and Oliver M. J. Carsten. 2014. Transition to Manual: Driver Behaviour When Resuming Control from a Highly Automated Vehicle. *Transportation Research Part F: Traffic Psychology and Behaviour* 27 (2014), 274–282.

[42] Abhijai Miglani, Cyriel Diels, and Jacques Terken. 2016. Compatibility Between Trust and Non-driving Related Tasks in UI Design for Highly and Fully Automated Driving. In *Adjunct Proceedings of the 8th International Conference on Automotive User Interfaces and Interactive Vehicular Applications* (Ann Arbor, MI, USA) *(AutomotiveUI '16 Adjunct)*. Association for Computing Machinery, New York, NY, USA, 75–80.

[43] Tina Mioch, Liselotte Kroon, and Mark A. Neerincx. 2017. Driver Readiness Model for Regulating the Transfer from Automation to Human Control. In *Proceedings of the 22nd International Conference on Intelligent User Interfaces* (Limassol, Cyprus) *(IUI '17)*. Association for Computing Machinery, New York, NY, USA, 205–213.

[44] B. Mok, M. Johns, K. J. Lee, D. Miller, D. Sirkin, P. Ive, and W. Ju. 2015. Emergency, Automation Off: Unstructured Transition Timing for Distracted Drivers of Automated Vehicles. In *2015 IEEE 18th International Conference on Intelligent Transportation Systems*. 2458–2464.

[45] Jörg Müller, Juliane Exeler, Markus Buzeck, and Antonio Krüger. 2009. ReflectiveSigns: Digital Signs That Adapt to Audience Attention. In *IEEE International Conference on Pervasive Computing and Communications (PerCom)*. 17–24.

[46] A. Pauzié and G. Pachiaudi. 1997. Traffic and Transport Psychology: Theory and Application. Chapter 18. Subjective Evaluation of the Mental Workload in the Driving Context. *Traffic and Transport Psychology: Theory and Application* (1997).

[47] Ehud Reiter and Robert Dale. 1997. Building Applied Natural Language Generation Systems. *Nat Lang Eng* 3, 1 (1997), 57–87.

[48] Francesco Ricci, Lior Rokach, and Bracha Shapira. 2011. *Recommender Systems Handbook*. Springer.
[49] Silvia Richter and Matthias Westphal. 2010. The LAMA Planner: Guiding Cost-based Anytime Planning with Landmarks. *J. Artif. Intell. Res.* 39 (2010), 127–177.
[50] SAE International. 2016. Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-road Motor Vehicles. *SAE International, (J3016)* (2016).
[51] Holger Schmidt, Gottfried Zimmermann, and Albrecht Schmidt. 2019. Using Gaze-based Interactions in Automated Vehicles for Increased Road Safety. In *Proceedings of the 11th International Conference on Automotive User Interfaces and Interactive Vehicular Applications: Adjunct Proceedings* (Utrecht, Netherlands) *(AutomotiveUI '19)*. Association for Computing Machinery, New York, NY, USA, 321–326.
[52] Shital Shah, Debadeepta Dey, Chris Lovett, and Ashish Kapoor. 2018. AirSim: High-fidelity Visual and Physical Simulation for Autonomous Vehicles. In *Field and Service Robotics*. Springer, 621–635.
[53] Yuan Shen, Shanduojiao Jiang, Yanlin Chen, Eileen Yang, Xilun Jin, Yuliang Fan, and Katie Driggs Campbell. 2020. To Explain or Not to Explain: A Study on the Necessity of Explanations for Autonomous Vehicles. arXiv:2006.11684 [cs.AI]
[54] Annika Silvervarg, Sofia Lindvall, Jonatan Andersson, Ida Esberg, Christian Jernberg, Filip Frumerie, and Arne Jönsson. 2016. Perceived Usability and Cognitive Demand of Secondary Tasks in Spoken Versus Visual-Manual Automotive Interaction. In *Interspeech 2016*. ISCA.
[55] Oliviero Stock, Massimo Zancanaro, Paolo Busetta, Charles Callaway, Antonio Krüger, Michael Kruppa, Tsvi Kuflik, Elena Not, and Cesare Rocchi. 2007. Adaptive, Intelligent Presentation of Information for the Museum Visitor in PEACH. *User Modeling and User-Adapted Interaction* 17, 3 (2007), 257–304.
[56] Sandra Trösterer, Alexander Meschtscherjakov, Alexander G. Mirnig, Artur Lupp, Magdalena Gärtner, Fintan McGee, Rod McCall, Manfred Tscheligi, and Thomas Engel. 2017. What We Can Learn from Pilots for Handovers and (De)skilling in Semi-Autonomous Driving: An Interview Study. In *Proceedings of the 9th International Conference on Automotive User Interfaces and Interactive Vehicular Applications* (Oldenburg, Germany) *(AutomotiveUI '17)*. Association for Computing Machinery, New York, NY, USA, 173–182.
[57] Remo M.A. van der Heiden, Shamsi T. Iqbal, and Christian P. Janssen. 2017. Priming Drivers before Handover in Semi-autonomous Cars. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17)*. ACM, New York, NY, USA, 392–404.
[58] Remo M. A. van der Heiden, Shamsi T. Iqbal, and Christian P. Janssen. 2017. Priming Drivers before Handover in Semi-autonomous Cars. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17)*. Association for Computing Machinery, New York, NY, USA, 392–404.
[59] J. P. Verma. 2015. *Repeated Measures Design for Empirical Researchers*. John Wiley & Sons.
[60] Marcel Walch, Kristin Lange, Martin Baumann, and Michael Weber. 2015. Autonomous Driving: Investigating the Feasibility of Car-driver Handover Assistance. In *ACM International Conference on Automotive User Interfaces and Interactive Vehicular Applications*. 11–18.
[61] Marcel Walch, David Lehr, Mark Colley, and Michael Weber. 2019. Don't You See Them? Towards Gaze-based Interaction Adaptation for Driver-vehicle Cooperation. In *Proceedings of the 11th International Conference on Automotive User Interfaces and Interactive Vehicular Applications: Adjunct Proceedings* (Utrecht, Netherlands) *(AutomotiveUI '19)*. Association for Computing Machinery, New York, NY, USA, 232–237.
[62] Rainer Wasinger, Antonio Krüger, and Oliver Jacobs. 2005. Integrating Intra and Extra Gestures into a Mobile and Multimodal Shopping Assistant. In *International Conference on Pervasive Computing*. 297–314.
[63] E. J. Williams. 1949. Experimental Designs Balanced for the Estimation of Residual Effects of Treatments. *Aust. J. Chem.* 2, 2 (1949), 149–168.
[64] Philipp Wintersberger, Clemens Schartmüller, and Andreas Riener. 2019. Attentive User Interfaces to Improve Multitasking and Take-over Performance in Automated Driving: The Auto-net of Things. *International Journal of Mobile Human Computer Interaction (IJMHCI)* 11, 3 (2019), 40–58.
[65] Bo Zhang, Joost de Winter, Silvia Varotto, Riender Happee, and Marieke Martens. 2019. Determinants of Take-over Time from Automated Driving: A Meta-analysis of 129 Studies. *Transportation Research Part F: Traffic Psychology and Behaviour* 64 (2019), 285–307.