Investigating the Influence of Gaze- and Context-Adaptive Head-up Displays on Take-Over Requests

Henrik Detjen
University of Duisburg-Essen
Essen, Germany
University of Applied Sciences Ruhr West
henrik.detjen@uni-due.de

Sarah Faltaous
University of Duisburg-Essen
Essen, Germany
sarah.faltaous@uni-due.de

Jonas Keppel
University of Duisburg-Essen
Essen, Germany
jonas.keppel@uni-due.de

Marvin Prochazka
University of Duisburg-Essen
Essen, Germany
marvin.prochazka@stud.uni-due.de

Uwe Gruenefeld
University of Duisburg-Essen
Essen, Germany
uwe.gruenefeld@uni-due.de

Shadan Sadeghian
University of Siegen
Siegen, Germany
shadan.sadeghian@winfo.uni-siegen.de

Stefan Schneegass
University of Duisburg-Essen
Essen, Germany
stefan.schneegass@uni-due.de

ABSTRACT

In Level 3 automated vehicles, preparing drivers for take-over requests (TORs) on the head-up display (HUD) requires their repeated attention. Visually salient HUD elements can distract attention from potentially critical parts in a driving scene during a TOR. Further, attention is (a) meanwhile needed for non-driving-related activities and can (b) be over-requested. In this paper, we conduct a driving simulator study (N=12), varying required attention by HUD warning presence (absent vs. constant vs. TOR-only) across gaze-adaptivity (with vs. without) to fit warnings to the situation. We found that (1) drivers value visual support during TORs, (2) gaze-adaptive scene complexity reduction works but creates a benefit-neutralizing distraction for some, and (3) drivers perceive constant HUD warnings as annoying and distracting over time. Our findings highlight the need for (a) HUD adaptation based on user activities and potential TORs and (b) sparse use of warning cues in future HUD designs.

CCS CONCEPTS

• **Human-centered computing** → *Empirical studies in ubiquitous and mobile computing*; *User studies*; **Mixed / augmented reality**.

AutomotiveUI '22, September 17–20, 2022, Seoul, Republic of Korea

ACM ISBN 978-1-4503-9415-4/22/09…$15.00
https://doi.org/10.1145/3543174.3546089

KEYWORDS

Automated Vehicles, Head-up Displays, Take-Over Requests, SAE Level 3, Gaze-Interaction, Warning Cue Design, Warning Continuity

ACM Reference Format:
Henrik Detjen, Sarah Faltaous, Jonas Keppel, Marvin Prochazka, Uwe Gruenefeld, Shadan Sadeghian, and Stefan Schneegass. 2022. Investigating the Influence of Gaze- and Context-Adaptive Head-up Displays on Take-Over Requests. In *14th International Conference on Automotive User Interfaces and Interactive Vehicular Applications (AutomotiveUI '22), September 17–20, 2022, Seoul, Republic of Korea.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3543174.3546089

1 INTRODUCTION

Errors in human perception can cause hazardous situations. This is especially true for safety-critical tasks that heavily rely on visual perception, such as driving. For example, in 2019, 47% of all traffic accidents in Great Britain occurred because the drivers failed to look properly, and in 8% of the cases, drivers *overlooked* pedestrians [52]. Globally, more than half of all lethal traffic accidents are among vulnerable road users (e.g., pedestrians, cyclists, etc.) [56]. Consequently, it is necessary to design driving assistance systems that protect vulnerable road users by reducing errors in the driver's visual perception.

One approach to overcome this problem is using *head-up displays* (HUDs) or *windshield displays* (WSDs). These displays present the information to the drivers via infotainment systems without requiring them to look away. A recent expert workshop of Riegler et al. [44] emphasized the extension of human (visual) perception for *safety purposes*, such as *highlighting of potentially critical objects*, as one of the most important goals for the future design of HUDs and WSDs. Furthermore, these displays increase system transparency by communicating the vehicle's status or perception of the driving

environment, which can lead to enhanced user experience (UX) and acceptance, especially for highly- and fully-automated vehicles (SAE levels 4 and 5 [46]) [6, 10, 57].

However, communicating the status or perception on the HUD adds visual complexity for the driver, which in turn may reduce situational awareness [8]. Visually salient elements bind attention and make it harder for the human to parse all elements of the driving scene. This is especially challenging in situations that are safety-critical or time-limited. A very good example within *level 3 automated vehicles* is takeover requests. In these situations, after receiving a takeover request (TOR), the driver has to shift his/her attention from non-driving-related tasks (NDRTs) to the driving task and quickly regain situational awareness of the driving context (out-of-the-loop problem). Visually salient HUD elements may impede that process. Consequently, communicating potentially critical objects on the HUD as a safety and UX feature in level 3 cars involves at least three paradoxes:

P1 **Transparency Paradox**: *During NDRTs*, an intended increasing transparency feature, such as highlighting critical objects on the HUD, may *distract* the user from the NDRT, thus worsening UX.
P2 **Scene Parse Paradox**: *During a TOR*, an intended safety feature, such as highlighting critical objects on the HUD, may add *complexity and distraction* to the driving scenery and decrease the driver's situational awareness, reducing safety.
P3 **Exposure Paradox**: *During a TOR*, an intended safety feature, such as highlighting critical objects on the HUD, could be *ignored* over time due to repeated "false" alarm exposure in non-critical situations (stimulus overexposure: cf. Banner Blindness [2] or cry-wolf-effect [17]).

The described paradoxes raise the question is necessary to display visual warning on the HUD throughout the whole ride and what potential impacts for safety and UX are. And to address this, how we can *reduce visual complexity* and *if* and *when* displaying critical objects is beneficial in terms of UX and safety in level 3 automated vehicles (L3-AVs). Therefore, in this paper, we conduct a virtual reality (VR) simulator study with an L3-AV that uses a HUD to address all three paradoxes. First, to address the Transparency Paradox, our system communicates warnings, either constantly on the HUD or only during a takeover. Second, to address the Scene Parse Paradox, we test a gaze-interaction mechanism that removes visually salient warnings from the already seen objects to increase visual saliency of the remaining objects (remove object salience on gaze, ROSOG). Third, to address the Exposure Paradox, users perform multiple TORs during a workload-inducing NDRT. We measured participants' driving behaviors and user experiences and compared them to a baseline without any HUD elements. Our paper contributes to a better understanding of HUD design for safety and user experience in L3-AVs. Specifically, we provide an experimental investigation of (1) the benefits of constantly displaying critical objects on the HUD, (2) the effectiveness of the ROSOG-mechanism, and (3) the "banner blindness" problem in an automated driving context. Our findings can improve future HUD designs in level 3 automated cars.

2 BACKGROUND & RELATED WORK

We relate our work to visualizing automotive head-up displays, perception, attention, and control transitions during automated driving.

2.1 Driving in Level 3 Automation

One of the most challenging problems for Level 3 automated cars is the control transition between the vehicle and the human driver.

Demons of Situational Awareness. While driving, it is crucial to capture the driving environment correctly. Otherwise, critical objects can be missed, (e.g., vehicles in the blind spot during a lane change). Endsley [13] defines three levels of situational awareness (SA): (1) perception of the environment, (2) understanding of the scene objects, and (3) projection of their position into the future. Endsley further describes so-called demons of situational awareness, most relevant in the case of automated driving: *SA Demon 8 - Out-of-the-loop syndrome*. After being not fully engaged in the driving loop, the sudden re-engagement requires a fast shift of attention and assessment of the scene. Scene parsing becomes even more challenging when the driver is still mentally engaged in another task (cf. [41]; Endsley: *SA Demon 2 - Requisite Memory Trap*). Overall, the possibility of making a perceptual error (SA level 1) during a TOR grows.

Take-Over Performance. To avoid perceptual errors, it is necessary to know how long it takes drivers to evaluate the situation correctly. A meta-study of Eriksson and Stanton [14] found variances between 1s and 23s. Gold et al. [19] showed that even after 7 seconds, automation effects lead to worse driving quality (e.g., drivers missed mirror checks). Lu et al. [32] demonstrated that drivers detect surrounding cars after 7 seconds, but it takes more time for them them to correctly perceive speed (up to 20s). Another study of Gold et al. [20] showed that increased traffic complexity reduces TOR performance. In sum, driving performance is strongly influenced by both the environmental context of a TOR and the user's state and capabilities.

Take-Over Interfaces. TOR interfaces must catch the driver's attention quickly and shift it back to the driving scene. For that purpose, one can use auditory signals; however, a combination with visual cues, such as warning signs [35] or ambient light bands [3], can help catch attention and increase TOR performance. Pre-ride familiarization [25] with TORs, priming drivers via mobile phones [4], or auditory messages [53] can all help to further increase performance.

2.2 Visual Attention

One forms attention in two ways: Either through 1) *goal-directed capture* or 2) *stimulus-driven capture* [58]. Either the driver purposefully directs their attention to certain objects of interest in the environment (top-down processing), e.g., looking for street signs for navigation, or objects in the environment are automatically and unintentionally brought into the focus of attention (bottom-up processing), e.g., a traffic sign that suddenly changes colors. The color change will often catch our attention, even if we are not focused on the traffic light. There is evidence that under certain conditions, stimulus-driven capture outperforms goal-directed

capture (cf. [59]), e.g., when new stimuli appear or move in the environment (cf. [36, 59]). Thus, warnings about potential hazards make goal-directed capture more difficult. Especially in L3-AVs, constant visual warnings, e.g., about pedestrians and bicycles, interfere with goal-directed behaviors such as a) potential NDRTs or b) scene parsing during a TOR (cf. Endsley: *SA Demon 5 - Misplaced Salience*). In response, users may start to ignore the warnings (cf. Endsley: *SA Demon 1 - Attention Tunneling*).

2.3 Interacting with Head-Up Displays

While preventing off-road glances, HUDs provide drivers with relevant information (e.g., current speed or navigation cues [34]). Other HUD concepts aim to increase human perception by augmenting the driving scene, e.g., warning about nearby cyclists [39] or pedestrians at night [18]. Currently, two significant developments will likely change the way we interact with automotive HUDs: (1) Through the ongoing progress of display technologies, such as transparent OLEDs, future HUDs will be a part of virtual windshield displays [22] (WSDs). (2) Through ongoing vehicle automation, future vehicles may also assist NDRTs, such as watching a movie or working on the laptop while driving (cf. [9]).

While engaged in an NDRT, even if not necessary from a technical perspective, transparency about the system's driving behavior leads to a better user experience (UX) during an autonomous ride, increasing trust and acceptance [10, 31, 49]. Examples of this include highlighting detected objects in the scenery [6, 7] or warning about potentially critical objects [57]. Thus, augmenting the traffic scenery during autonomous driving phases can be advantageous.

Eisma et al. [12] found that augmented visual feedback leads to subjectively easier tasks, yet it also creates misunderstandings and leads to visual attention tunneling. In line with that, future HUDs may help to improve scene perception by communicating potential hazards. However, a study of Currano et al. [8] found that visual communication about the driving scenery, including potential hazards, increases complexity and negatively influences SA. Another study of Kim and Gabbard [29] found that HUDs can informative or distractive depending on the perceptual forms of graphical elements. Thus, there seems to be a trade-off between enhancing TOR performance by displaying SA-relevant cues and communicating too much (salient) information (cf. Endsley: *SA Demon 6 - Complexity Creep*).

A method to interact efficient, comfortably, and without paying much attention with a HUD while being engaged in a NDRT is *gaze interaction*. Eyetracting can be used as an input device [55] for HCI and has been used in the automotive context, e.g., for selection tasks [28, 43].

2.4 Conclusive Summary

In Level 3 automated cars, preparing drivers for TORs on the HUD requires their attention which is a sparse resource [37]. Visually salient HUD elements can distract attention from potentially critical parts in a driving scene during a TOR (cf. Scene Parse Paradox). Further, attention is a) meanwhile needed for NDRTs (cf. Transparency Paradox) and can b) be over-requested (cf. Exposure Paradox). The idea of this paper is to investigate if it is necessary to display visual warnings on the HUD throughout the whole ride and what potential impacts for safety and UX are. We, therefore, test 1) the time aspect of visual warnings and 2) the low-effort deactivation (ROSOG interaction) of already seen warnings through gaze in order to systematically reduce the required attention.

3 USER STUDY

We conducted a user study to examine the impact of HUD warning presence on safety and user experience in a VR setup. In the study, the experienced system's HUD varies by *presence* (no HUD warnings, during TOR only, or constant) and *gaze-adaptivity* (no gaze response, deactivation on gaze-focus), leading to five conditions (Figure 1.c) that were tested in a within-subjects study.

First, addressing the Exposure Paradox, users perform multiple TORs during a workload-inducing NDRT. Second, addressing the Transparency Paradox, our system communicates warnings either constantly on the HUD or only during a takeover. Third, addressing Scene Parse Paradox, we test a gaze-interaction mechanism that removes visually salient warnings from the already seen objects to increase visual saliency of the remaining objects (remove object salience on gaze, ROSOG).

3.1 Hypotheses

Based on the evidence of previous work, we pose the following hypotheses (cf. Section 2):

H1.1 The constant presence of visual warnings on the HUD make the system more transparent and will improve overall user experience during NDRTs (cf. [6, 10, 57]).

H1.2 The constant presence of visual warnings on the HUD will help prepare for takeover by increasing situational awareness and improving takeover performance (cf. [3]).

H1.3 The gaze-adaptivity of HUD elements reduces complexity and distraction; thus, it helps with scene parsing by removing visual salience from already seen objects, leading to better takeover performance (cf. paradox 3).

In addition:

H2.1 The constant presence of visual warnings in the peripheral field of view distracts the user during NDRT performance and decreases user experience (contrasting H1.1, cf. paradox 1).

H2.2 The constant presence of visual warnings becomes annoying over time and participants will start to ignore them, impeding the takeover performance (contrasting H1.2, cf. paradox 2).

3.2 Driving Scenario

To test our hypotheses, we created a virtual reality (VR) driving scenario in a suburban area that might occur after a phase of autonomous driving on the highway. The participants experienced sitting in a level 3 automated vehicle that drives on the right-hand side of a two-lane road, where the maximum permitted speed is *30km/h*. The road consisted of a long, straight street bordered by sidewalks and home gardens. The car's autonomous driving mode was activated via a push of a button on the driving wheel. Occasionally, a hazard would appear, giving the driver *5secs* to react. Hazards included a person crossing the road from behind a bus parked on the left side, a loose tire rolling into the road from the right side, and a ball rolling in from the left side.

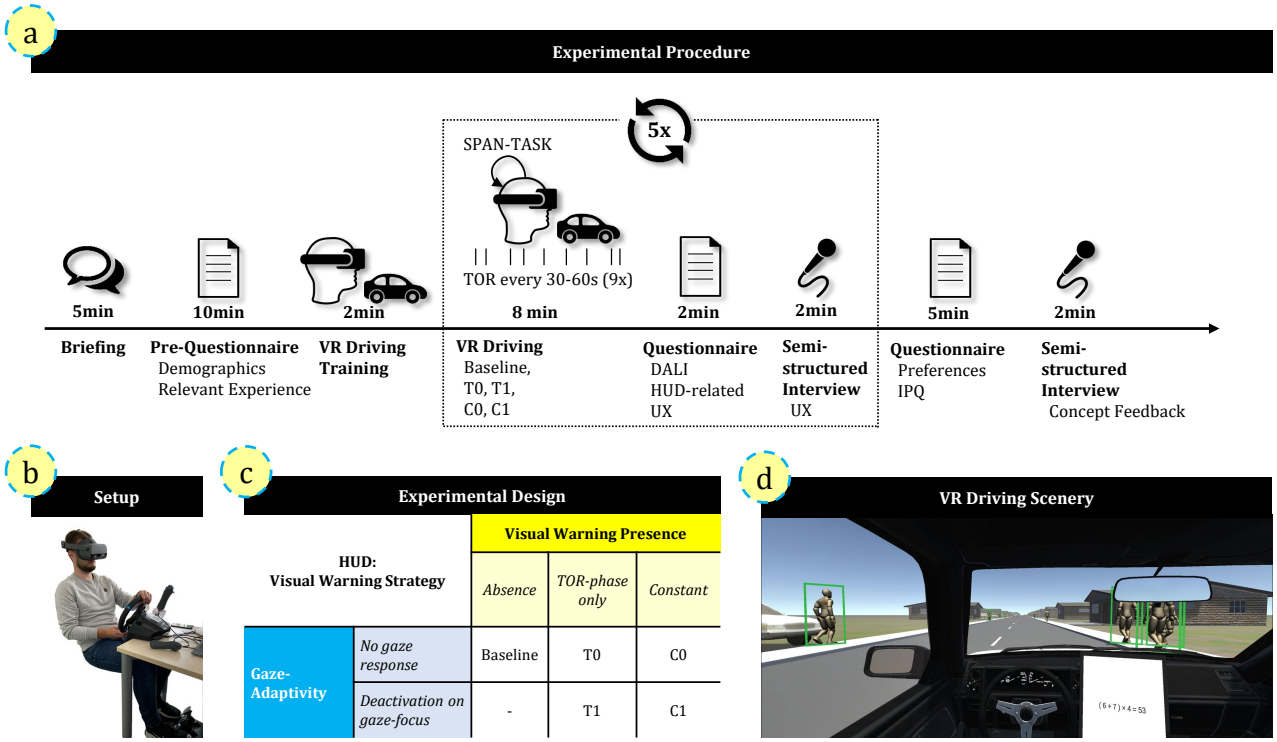

Figure 1: (a) Procedure of the experiment; (b) VR setup with a participant; (c) Experimental conditions and variation of the HUD warnings; (d) The driver's view, in which the NDRT is on the right side and potential hazards in the driver's field of view are highlighted

Simulation setup. The scenario was simulated in VR and we used a tile generator that allowed hazard events to be configured and intermixed on the track. Our VR setup consisted of Logitech G29 driving force (i.e., a driving wheel and standard pedals) and a pico neo G2 VR headset with native eye-tracking [1] (cf. Figure 1.b). One of the main features of the headset is its ability to track eye gaze, which is a major factor in our study design. The scenario was implemented using the Unity3D game engine, was displayed on the headset, and received the driving wheel and pedal inputs.

HUD Design. We designed the HUD so that its only elements were warnings. The *warnings* highlight any potential hazard in the scene by surrounding it with a green box, as in Figure 1.d. The *gaze mechanism* led to the behavior that, upon checking a hazard, the warning around the hazard would disappear. To avoid the problem of looking but not seeing, we based our determination of whether a hazard is checked by the duration of the fixations, which should be equal or greater than 300*msec* (comparable to dwell time used in previous work [43]). So, if there are, for example, five hazards in the scene and the participant looked at three of them, only two would be left with the warnings around them. Therefore, based on the 5*sec* time to collision (TTC), we had a sliding window that moves with the AV, highlighting all potential hazards within a time frame of 5*sec*.

[1]https://vr.tobii.com/integrations/pico-neo-2-eye/

Takeover Task. Upon detecting a hazardous situation, the AV communicates a takeover request in the form of a beeping sound that lasted for *1 sec*. We based the audio frequency on the previous work of Gray [21]. The participants then have to intervene and take over the system by steering the driving wheel and stepping on either the gas or brake pedals. Once the participants intervene, the autonomous driving stops, and the driver needs to execute the right decision by braking and waiting in the case of the crossing person, steering to the left in the case of the rolling tire, or steering to the right in the case of the rolling ball.

Non-Driving Related Task. To ensure a high level of cognitive engagement in a non-driving activity, we asked our participants to perform a working memory span task [54]. Specifically, the participants had to verbally verify whether a mathematical operation was true or false. Afterward, a consonant would appear for 1 second. The participants were asked to remember these consonants until they had to repeat them. After 5 operations (and thus 5 consonants), the participants were asked to recall and loudly state the previously displayed consonants. The task was displayed on a virtual tablet in the scene. The tablet appeared to the right-hand side of the driver, as in previous work [15] (cf. Figure 1.d).

3.3 Measurement

We applied the following measures to qualify and quantify our hypotheses within/after the driving scenario.

Presence in the Simulation. Since we used a virtual environment, we asked participants about their presence in order to determine the extent to which their virtual experience might be compared to a real-world experience. Therefore, we used the established IGROUP Presence Questionnaire (IPQ) [42, 50]. The questionnaire measures, on a 7-point scale with varying answer dimensions, the general "sense of being there." It also measures three related subscales: 1) Spatial Presence - the sense of being physically present in the virtual environment, 2) Involvement - the attention devoted to the virtual world and the involvement experienced, and 3) Experienced Realism, the subjective experience of realism in the virtual environment.

Workload and NDRT Performance. In level 3 automated cars, the driver is usually not busy handling takeovers, but is performing NDRTs, which may be influenced by the HUD warning design. We use a NDRT as described in Section 3.2 and measure the participants' performance in terms of success (calculations, memory) and speed. Additionally, we apply a subjective questionnaire. A standard measure used to assess the subjective task load is the NASA-TLX [24]. We used an index based on the TLX, the driving activity load index [38] (abbr.: DALI). In contrast to the NASA-TLX, the questionnaire removes the performance and physical demand dimensions because performance can be observed through other measures, and because modern cars are not designed to be physically demanding. To better distinguish the mental demand dimension of the TLX, the DALI is separated by perceptual and cognitive load, which are visual/auditory demand and effort of attention, respectively. Further, it adds the dimensions of interference to evaluate dual-task performance and situational stress to evaluate stress level during the driving task. These driving-task-specific adjustments make it easier to identify the origins of users' impressions, thus improving the interpretability of the results. We adjusted the interference dimension of the DALI to our takeover scenario (asking for the takeover interference through the NDRT, rather than the dual-task interference). Further, we compute a global score for the DALI – comparable to the RAW-TLX [23] score (unweighted average) for the NASA-TLX – to assess the overall workload per condition. We use a 100-point scale anchored from very low to very high for the DALI.

TOR Performance. The TOR in level 3 automated cars is a safety challenge and must be performed as efficiently and safely as possible. HUD concepts may influence situational awareness and thus TOR performance. In our case, we measure the takeover task (cf. Section 3.2) performance with the time from the TOR warning until participants start to react by braking/steering (TTR). To assess participants' subsequent driving quality, we log the drivers' applied braking/steering force and how close their path is to the hazard. Further, utilizing the gaze data, we check if they looked at the hazard or other potentially critical objects.

HUD Perception. We applied multiple measures to assess the participants' experiences and impressions of the different HUD warning strategies. For a quick assessment of the driving experience in the conditions, we used two scales (a positive and a negative) with a 7-point Likert scale agreement score (very low - very high) and asked for the reasoning in short interviews. The same Likert scale type measures the HUD's perception in terms of distraction, helpfulness for task switching, situational awareness, transparency, trust, safety, and acceptance. Further, we estimated the participants' overall preference for a particular HUD warning strategy on a 7-point Likert agreement scale.

3.4 Experimental Procedure

The experimental protocol was as follows (cf. Figure 1.a). First, the experimenter welcomed the participants and verbally informed them about the experimental procedure, which was followed by a written description and informed consent of the participants to the procedure and the use of data. Then, they answered a questionnaire about their sociodemographics and relevant system experiences, such as driving experience and familiarity with 3D technology. Then the experiment began.

The participants took their seats and familiarized themselves with the system. We informed the participants that the system could fail due to insecurities. In a two-minute drive without HUDs or side tasks and a total of 2 takeover prompts, they were able to get used to the takeover procedure. They experienced each of the five counterbalanced conditions in a short phase of 30 seconds. No side task or takeover was necessary to become acquainted with the (non)visualization. After the training phase, data recording of the driving and gaze behavior started.

Each recorded run contained repeated potential hazard events (HUD alerts). At intervals of 45±15 seconds, one of them requires a takeover. Thus, there are at least 30 seconds between two takeover requests. After each run, participants went to a PC station next to the VR setup and filled out questionnaires regarding their subjective workload (DALI [38] questionnaire) and system experience (custom questionnaire with questions addressing UX, SA, distraction, trust, transparency, safety perception, utility, and acceptance). They were encouraged to add positive/negative/other comments regarding the ride (interview). We provided all questionnaires via an online platform (www.soscisurvey.de). Completing a run (ride, questionnaires, interview) took about 15 minutes.

After all five runs, we did a conclusive qualitative interview with the participants about their general impressions regarding the HUD warning design and the takeover situations. In addition, we asked them about the strengths and weaknesses of the particular conditions. Finally, the participants were debriefed. Overall, the experiment took about 90 minutes.

Pilot Study. We did a pilot study with three participants ($N = 3$) to test our study procedure and VR driving perception. Generally, the participants found the simulation convincing. However, regarding the fixed time from warning to potential hazard impact of 3s, participants got used to the interval and started to react automatically. Thus, we added a variation of this interval for the final study procedure and excluded these participants from the analysis. For the main study, based on the results obtained from participants in the pilot study, we added a random variation to the time-to-react (duration: $3sec - 7sec$).

3.5 Analysis

To answer our research questions, we applied a mixed-linear effects model (LMEM) [1, 16] to our data – utilizing in *R*-script [40]

with the package *lme4* [11]. LMEMs are arguably robust to use on Likert-data [5, 30, 48]. We controlled for the participants' variation (random effect) while separating the effect of the conditions (fixed effect) on the responses, which results in the model: $response \sim condition + (1|participant)$. In a next step, using the *emmeans* [45] package, we calculated the estimated marginal means for the model before we conducted planned contrasts on the estimated marginal means of the conditions (i.e., we reformulated the regression coefficients in line with our research questions, for grouping the conditions into combined effects and for baseline comparisons). In other words, we applied *orthogonal sum contrasts* for our independent variables/factors (warning presence, gaze adaptivity, and interaction between them). For baseline comparison, we applied factor-wise *treatment contrasts*. LMEMs with planned contrasts provide a viable alternative to omnibus tests such as ANOVAs (cf. Schad et al. [47]). To account for multiple comparisons of the contrasts and the underlying t-tests, we used Šidák corrections [51]. We estimated the degrees of freedom with the Kenward-Roger procedure [27].

3.6 Sample

For the final experiment, we invited twelve persons ($N = 12$), 10 of whom identified as male and two as female, to the University of Duisburg-Essen. Participants were young ($M = 25.83$years, $SD = 3.56, MIN = 22, MAX = 32$) and affiliated with the University (9 students, 2 researchers, 1 university staff member). They had a strong affinity for technology (ATI; 6-point Likert scale: $M = 4.77, SD = 0.63$) and were very experienced (7-point Likert scale) with VR goggles ($M = 5.75, SD = 1.35$) and 3D apps ($M = 6, SD = 1.48$). Regarding their driving experience, participants held a valid driving license for around eight years ($M = 8.42, SD = 3.32$). In addition, they reported to be rather unfamiliar (7-point Likert scale) with current ACC and lane-keeping systems ($M = 3, SD = 2.09$) and used their cars relatively fewer times per year than the average driver (≤5k: 6, >5k-10k: 2, >10k-15k: 2).

4 RESULTS

4.1 Presence in the Simulation

The IPQ results (7-point scales) show that participants recognized the artificial driving scenery, as they reported a rather low experienced realism ($M = 3.48, SD = 1.15$) and medium involvement ($M = 4.06, SD = 1.08$). Nevertheless, they felt spatially ($M = 5.42, SD = 0.75$) and generally ($M = 5.25, SD = 1.29$) present in the simulation.

4.2 Workload and NDRT Performance

Figure 2 shows the distribution of participants' estimated workload (DALI [38]) by condition: Participants generally felt a high attention demand, a substantial interference between tasks, and rather stressed. They perceived the temporal and visual demand as mediocre and the auditory demand as relatively low. Participants' overall workload is right above the middle of the scale. For the auditory dimension, TOR-only warnings significantly reduce the subjective demand compared to the baseline (t(44)=-2.64, p=0.02, effect=-28.7, CI95[-50.5,-6.8]).

Participants solved approximately 60 math tasks in each condition, which translates to a speed of roughly one task every 10 seconds. We found no significant differences in participants' performance. The calculations and the remembered letters were mainly correct and are also comparable across conditions (cf. Table 1).

4.3 Takeover Performance

Since the driving performance was measured for 9 subsequent TOR situations, we added the time component (hazard) as an additional random intercept in our LMEM ($response \sim condition + (1|participant) + (1|hazard)$).

Table 1 shows the results for NDRT and TOR performance. Gaze-adaptive warnings seem to minimize the safety distance to a critical level during a TOR, i.e., reduce the distance to the hazard significantly compared to non-adaptive warnings (t(514)=-3.25, p<0.01, effect=0.35, CI95[-0.56,-0.14]). Constantly present warnings lead to a significantly higher percentage of applied brake force compared to TOR-only warnings (t(514)=2.42, p=0.05, effect=.021, CI95[.004,.037]). The eye-gaze data show that there is a cross-over interaction between factors for the detection of the critical objects (t(516)=2.85, p=0.01, effect=.01, CI95[.003,.172]). With gaze-adaptivity, constant warnings lead to better detection, while without gaze-adaptivity, TOR-only warnings perform better, and the detection efficiency of constantly present warnings decreases. The effects are not distinguishable from the baseline. The baseline and the TOR-only condition have the lowest TOR fail rate with $N = 1$. In the other conditions, the TOR fail rate ranged from 4 (Constant, TOR-only with gaze) to 5 (Constant with gaze).

4.4 HUD Perception

To assess HUD perception, we used a questionnaire and conducted interviews after each condition and after all were done.

HUD-related Questionnaire. For this set of questions, we leave out the baseline comparison because each question targets the perception of the HUD warning concept. Figure 3 shows the participants' responses to our HUD-related questions. The participants reported a low level of distraction (Q1) throughout the trials. In a TOR situation (Q2), users perceive the utility of the HUD concepts as significantly better when warnings are tor-only rather than constant (t(44)=4.09, p<0.001, effect=3.08, CI95[1.57,4.60]). The Situational Awareness (Q3–Q5) support, as well as transparency (Q6) and trust (Q7), are perceived as medium to rather high. Regarding safety (Q8), constant warnings were rated at approximately the middle of the scale, whereas TOR-only warnings led to a significant shift of that perception to a high level (t(44)=2.78, p=0.01, effect=2.41, CI95[0.78,4.01]). Similarly, TOR-only warnings significantly increase the participants' intention to use (Q9) the HUD (t(44)=2.65, p=0.03, effect=2.33, CI95[0.56,4.11]) to a high level compared to constant warnings.

Driving Experience. After each condition, we assessed the driving experience on two scales: a positive and a negative 7-point Likert scale. Figure 4.a shows the driving experience differential (positive ratings - negative ratings). Neither positive nor negative ratings differ between conditions, but they tend slightly towards a positive driving experience.

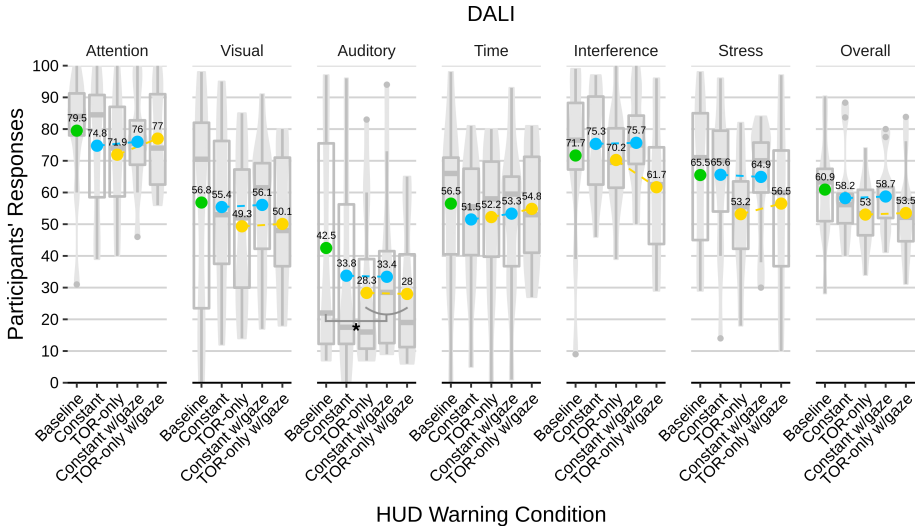

Figure 2: Results of the DALI questionnaire by dimension and HUD condition – Points and numbers show the mean response values, Whisker and Violin plots show the distribution of responses for conditions. For easier comparison, we color-coded the baseline green and the factor *Warning Presence* blue (constant) and yellow (TOR-only). Significant differences marked with * $p \leq 0.05$, ** $p \leq 0.01$, *** $p \leq 0.001$.

Table 1: NDRT performance and driving performance measures during TOR-situations by HUD condition – The "Significant Findings"-column contains the results of the LMEM orthogonal sum contrasts labeled as "factor effects" and of the treatment contrasts labeled as "vs baseline". Significant differences marked with * $p \leq 0.05$, ** $p \leq 0.01$, *** $p \leq 0.001$.

	HUD warning condition					Significant Findings	
	Baseline (B)	Constant (C0)	TOR-only (T0)	Constant w/gaze (C1)	TOR-only w/gaze (T1)		
	M(*SD*)	*M*(*SD*)	*M*(*SD*)	*M*(*SD*)	*M*(*SD*)	factor effects	vs baseline
NDRT Performance							
Completed tasks *n*	60.00 (12.06)	60.83 (13.46)	62.50 (16.58)	62.50 (16.03)	61.25 (13.34)	-	-
Correct calculations %	.97 (.02)	.97 (.04)	.97 (.03)	.98 (.02)	.97 (.06)	-	-
Correct memorized words %	.85 (.12)	.86 (.10)	.84 (.14)	.83 (.13)	.84 (.11)	-	-
TOR Performance							
Time-to-react *ms*	1825.53 (606.48)	1872.77 (797.29)	1922.48 (797.62)	1964.43 (779.64)	1941.54 (882.62)	-	-
Min. distance to hazard *m*	3.12 (0.59)	3.20 (0.66)	3.37 (0.57)	3.1 (0.63)	3.12 (0.58)	Gaze-adaptivity***	-
Mean braking %	.06 (.05)	.05 (.05)	.07 (.06)	.06 (.05)	.07 (.06)	Warning Presence*	-
Mean steering %	.04 (.03)	.03 (.03)	.03 (.02)	.03 (.03)	.03 (.03)	-	-
Looked at critical object %	.96 (.19)	.98 (.14)	.92 (.28)	.95 (.21)	.99 (.10)	Gaze-adaptivity x Warning Presence**	-
Looked at potentially crit. obj. %	.62 (.45)	.62 (.38)	.62 (.36)	.70 (.40)	.65 (.38)	-	-
Unresponded TORs *n*	1	4	1	5	4		

Preference. Participants expressed their overall preference on a 7-point Likert scale after experiencing all conditions. Regarding *Gaze-adaptivity* preference, without gaze-adaptivity, the ratings significantly increase towards the middle of the scale (t(22)=2.55, p=0.03, effect=1.67, CI95[0.23,3.11]). Regarding *Warning Presence* preference, ratings are rather low for the baseline, medium for constant warnings, and rather high for TOR-only (see Figure 4.c). TOR-only warnings score significantly better than the baseline (t(22)=2.74, p=0.04, effect=2.25, CI95[0.55,3.95]).

Qualitative Interviews. We conducted a qualitative content analysis for the interviews (cf. [33]) and quantified the number of codes by the number of mentions (i.e., $N = 12$). Regarding the driving simulation, participants got used to the virtual environment after a while ($n = 8$). They found the TOR scenario design repetitive and could foresee TOR situations over time ($n = 5$). However, the driving data could not support any learning effects. Regarding the NDRT, users generally perceived the task as hard ($n = 6$), stressful ($n = 7$), and exhausting ($n = 1$), as they became fatigued over time ($n = 7$). On the other hand, four participants also felt stimulated by the task or positively perceived it as exciting ($n = 3$) or challenging ($n = 1$). The task also made them feel competent, and they felt that it was fun ($n = 3$) to drive and that the TOR-assistive warning system was not really necessary ($n = 3$, P2: *"I don't really need help, I am a real driver!"*)

Many participants commented that beeping was enough to switch their attention ($n = 9$) from the NDRT to the TOR task. However, one person also found the beep sound annoying. Regarding the TOR task, participants found the visual warnings helpful, e.g., as they increased awareness of the hazards (P12: *"[...] the HUD really*

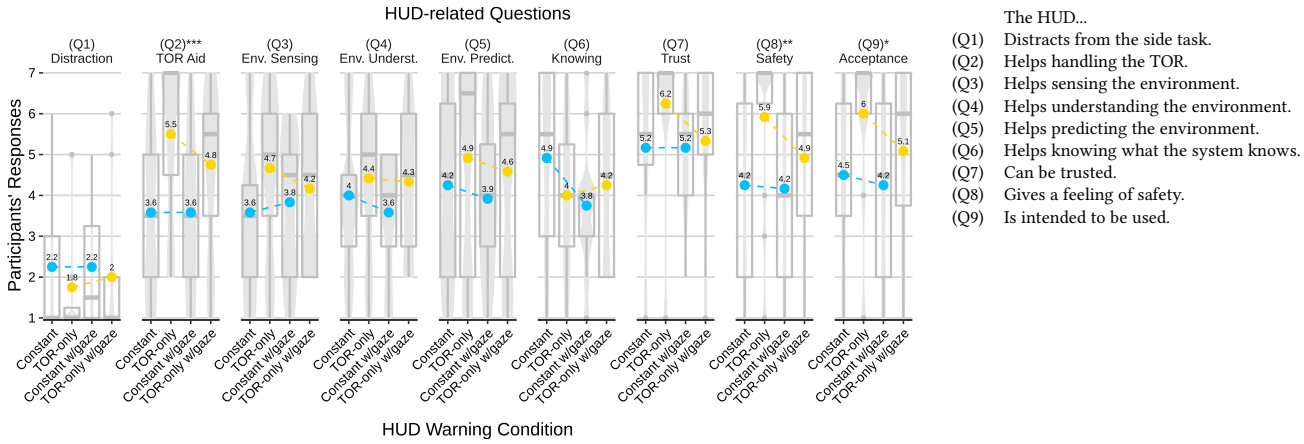

Figure 3: Results of the HUD-related questions by condition – Points and numbers show the mean response values, Whisker and Violin plots show the distribution of responses for conditions. For easier comparison, we color-coded the baseline green and the factor *Warning Presence* blue (constant) and yellow (TOR-only). Significant differences (next to question label) marked with * $p \leq 0.05$, ** $p \leq 0.01$, *** $p \leq 0.001$.

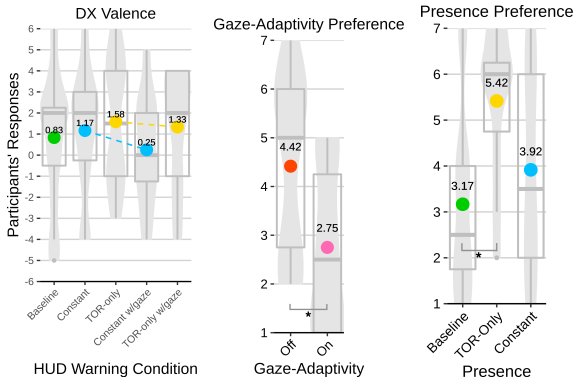

Figure 4: Results of the Driving Experience and Preference Ratings. Shape and color-coding as in previous figures, plus factor *Gaze-adaptivity* in red (off) and pink (on). Significant differences marked with * $p \leq 0.05$, ** $p \leq 0.01$, *** $p \leq 0.001$.

helps to estimate the situation when the TOR arises.") However, most participants also found the TOR situation challenging ($n = 7$), especially without a visual aid. Some participants even commented that they could not distinguish between the conditions with constant and TOR-only presence of the visual warnings because of the challenging NDRT ($n = 4$). However, nearly all participants said that the TOR-only warnings were less distracting and less annoying over time ($n = 11$, P5: *"While doing the tasks, I looked further down on purpose not to see the distracting boxes in my peripheral vision."*) The constant warning presence, on the other hand, increased trust (P7: *"You know what the car knows."*)

Regarding the gaze-adaptivity of the warnings, one participant appreciated the mechanism as a compromise to information complexity. However, half of the participants ($n = 6$) first found the disappearing warnings confusing (e.g., P5: *"Did the car identify the person behind the bus as a hazard and now the box is deactivated or is there another person coming?"*) or even distracting (e.g., P11: *"A disappearing box often triggers an attention switch to another box. That is the opposite of what was intended."*)

To improve the HUD warning mechanism, participants suggested highlighting occluded objects ($n = 8$), e.g., warning about a person behind the bus and adjusting the visualization ($n = 4$). Adjustments could be color-coding the different objects (to avoid becoming blind to important ones) or providing more information (hazard trajectory, velocity, etc.)

5 DISCUSSION

In the following, we discuss our data and relate it to the HUD paradoxes, indicate directions for future research, and point out the limitations of our study.

5.1 Warning Presence

Transparency Paradox (P1): While performing an NDRT, the HUD can distract and impede UX. We expected the constant presence of visual warnings on the HUD to make the system more transparent and increase the overall user experience during NDRTs (H1.1), as well as help prepare for takeover by increasing situational awareness. Overall, we cannot confirm these assumptions in our experiment. Instead, the opposite was the case. We found evidence for H2.1 in the interviews: The constant presence of visual warnings in the peripheral field of view distracts the user during NDRT performance and decreases user experience. Instead of having no warning mechanism, the participants' preference is significantly towards presenting warnings during TOR only, whereas constant presence is rated ambiguously. Users perceive the TOR-only presence as significantly safer and more acceptable than the constant presence in both the survey and the interviews: Nearly all reported that TOR-only warnings were less distracting and less annoying over time. So is it a bad idea to display warnings constantly in level 3 automation?

We think not. Instead, our findings indicate that the NDRT workload could moderate the trade-off in the Transparency Paradox: We chose a challenging NDRT for the experiment to make the TOR task not too simple, because a non-demanding task would have allowed participants to more easily parse the driving scene and handle TOR situations. A practical approach to the Transparency Paradox would be to adapt HUD warnings' presence to NDRT workload. This could be done by hiding the warnings or at least making them less salient when a potentially complex activity (e.g., smartphone use) is detected; conversely, it could also be done by increasing salience (e.g., increasing hue) when a less complex activity (e.g., looking out of the window) is detected. Future work should compare different NDRT load levels across warning presence to further investigate the Transparency Paradox.

Exposure Paradox (P3): Repeated exposure to salient stimuli during non-TOR situations leads to blindness for stimuli in TOR situations. One hypothesis was that the constant presence of visual warnings on the HUD would help the user to prepare for the takeover by increasing their situational awareness, thereby improving takeover performance (H1.2). The observed data does not support this, as the presence was not significantly different from the baseline in the observed TOR performance. Participants also rated the utility of warnings only during a TOR as significantly better than constantly presented warnings, though it was also not significantly different from the baseline. The reason for that might be, as discussed previously: the demanding NDRT. The annoying constant warnings may have contributed to a higher starting stress level during a TOR than highlighting the critical objects just as the TOR scene appears. As a result, participants may have had a lower capacity for scene parsing. The higher initial stress would also explain the increased braking in the constant warning conditions. The counter-hypothesis (H2.2) was that the constant presence of visual warnings becomes annoying over time, and participants will thus ignore them, impeding the takeover performance. We can partly confirm this hypothesis through the interview data: Most participants perceive constant warnings as annoying, but TOR performance is not measurably affected, and the baseline leads for the most reliable responses with only 1 missed TOR in total. In contrast, participants also said they found the TOR more challenging in the absence of warnings. Overall, beneficial effects, as in H2.1, could not be observed due to task demand. The adverse effects of H2.2 are partly supported by subjective perception, but are not manifested in participants' behavior. In our case, the adverse effect may have been induced just before the condition ended. It seems as if the anticipated trade-off of the Exposure Paradox is there. However, future work should test the warning presence with varying task demands to induce the beneficial effects of constant warnings during less challenging tasks and capture data over a longer time frame to induce more substantial banner blindness/alarm fatigue.

5.2 Gaze-Adaptivity and Scene Complexity

Scene Parse Paradox (P2): Using visually salient warning elements during TORs adds scene parsing complexity and distraction. We used a gaze mechanism to reduce visual complexity. We expected that the gaze-adaptivity of HUD elements reduces complexity and distraction and thus helps with scene parsing by removing visual salience from already seen objects, leading to better takeover performance (H1.3). Here, we observed the opposite: Participants' gaze-interaction preference was significantly lower for warnings with a reaction on gaze. The driving quality decreases because the minimal distance to the hazard is reduced via gaze-adaptivity. A reason for this could be that participants are not used to the gaze deactivation of the visual warnings: While parsing the scenery, the interaction might be unexpected or even undesired. This problem of unintended gaze-interaction is often referred to as the Midas touch problem [26]. The interviews support this assumption: Half of the participants found the gaze-interaction confusing or distracting. This seems plausible. Removing a visually salient HUD element (color) from the scene triggers another visually salient movement (disappearing), thus keeping the complexity level of the scene parse. Furthermore, and more importantly, after a HUD element disappears, focusing on the critical object and ignoring the others that become more salient is more demanding (stimulus-driven capture prevails over attention-driven capture). Future work could investigate other techniques to reduce complexity or test gaze-adaptivity across different levels of scene complexity in order to better understand the Scene Parse Paradox.

5.3 Limitations

We conducted the study in a VR driving environment. While participants reported relatively high general and spatial presence in the simulation, they rated the experienced realism as relatively low and the felt involvement as medium. The VR setup may have influenced their perceived safety, trust, and user experience. We expect these parameters to change in a setting with higher ecological validity, such as a test track. Further, our sample size was comparably small, and the found effects of the HUD conditions had wide confidence intervals. Therefore, some of the tendencies in our data may reach statistical significance with an increased number of participants (e.g., the general trend in visual data inspection that TOR-only warnings systematically performed better than constant warnings in terms of workload). In addition, our sample consisted mainly of male, educated, technologically skilled persons with a European background. Therefore, the results may vary for samples with different properties. Despite these limitations, we see our experiment as a first step towards understanding HUD paradoxes, which certainly require more investigation.

6 CONCLUSION & FUTURE WORK

Designing warnings on HUDs in level 3 automated cars is a double-edged sword. They can be helpful and supportive, increasing situational awareness and leading to better takeovers and system transparency. However, they can also be annoying and distracting, leading to the opposite of their design intention. This paper investigated if it is necessary to display visual warning on the HUD throughout the whole ride and what potential impacts for safety and UX are. We therefore varied the warnings' presence to systematically reduce and added a low-effort gaze-interaction mechanism to further economize the required attention. We found that (1) it is helpful for drivers to have visual support during the TOR phase, (2) reducing scene complexity is necessary, but adaptive scene complexity reduction through gaze bears the risk of distraction, and (3)

drivers perceive constantly presented HUD warnings as annoying and distracting after a while. These findings highlight the need for (a) HUD adaptation based on passenger activity and potential TORs and (b) sparse use of warning cues in future HUD designs. We encourage others to address HUD warning presence in terms of timing and complexity of level 3 HUDs in future work in order to better understand the design trade-offs before implementing these technologies.

REFERENCES

[1] R. H. Baayen, D. J. Davidson, and D. M. Bates. 2008. Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language* 59, 4 (2008), 390–412. https://doi.org/10.1016/j.jml.2007.12.005

[2] Jan Panero Benway. 1998. Banner Blindness: The Irony of Attention Grabbing on the World Wide Web. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 42, 5 (1998), 463–467. https://doi.org/10.1177/154193129804200504

[3] 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 (ACM Digital Library)*, Shadan Sadeghian Borojeni (Ed.). ACM, New York, NY, 237–244. https://doi.org/10.1145/3003715.3005409

[4] Shadan Sadeghian Borojeni, Lars Weber, Wilko Heuten, and Susanne Boll. 2018. From reading to driving. In *MobileHCI 2018 (ACM Digital Library)*, Lynne Balie and Nuria Oliver (Eds.). ACM, New York, 1–12. https://doi.org/10.1145/3229434.3229464

[5] F. Bross. 2019. Using mixed effect models to analyze acceptability rating data. www.fabianbross.de/mixedmodels.pdf

[6] Mark Colley, Benjamin Eder, Jan Ole Rixen, and Enrico Rukzio. 05062021. Effects of Semantic Segmentation Visualization on Trust, Situation Awareness, and Cognitive Load in Highly Automated Vehicles. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, Yoshifumi Kitamura, Aaron Quigley, Katherine Isbister, Takeo Igarashi, Pernille Bjørn, and Steven Drucker (Eds.). ACM, New York, NY, USA, 1–11. https://doi.org/10.1145/3411764.3445351

[7] Mark Colley, Svenja Krauss, Mirjam Lanzer, and Enrico Rukzio. 2021. How Should Automated Vehicles Communicate Critical Situations? *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 3 (2021), 1–23. https://doi.org/10.1145/3478111

[8] Rebecca Currano, So Yeon Park, Dylan James Moore, Kent Lyons, and David Sirkin. 05062021. Little Road Driving HUD: Heads-Up Display Complexity Influences Drivers' Perceptions of Automated Vehicles. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, Yoshifumi Kitamura, Aaron Quigley, Katherine Isbister, Takeo Igarashi, Pernille Bjørn, and Steven Drucker (Eds.). ACM, New York, NY, USA, 1–15. https://doi.org/10.1145/3411764.3445575

[9] Henrik Detjen, Bastian Pfleging, and Stefan Schneegass. 2020. A Wizard of Oz Field Study to Understand Non-Driving-Related Activities, Trust, and Acceptance of Automated Vehicles. In *12th International Conference on Automotive User Interfaces and Interactive Vehicular Applications (ACM Digital Library)*. Association for Computing Machinery, New York,NY,United States, 19–29. https://doi.org/10.1145/3409120.3410662

[10] Henrik Detjen, Maurizio Salini, Jan Kronenberger, Stefan Geisler, and Stefan Schneegass. 2021. Towards Transparent Behavior of Automated Vehicles. In *Proceedings of the 23rd International Conference on Mobile Human-Computer Interaction (ACM Digital Library)*. Association for Computing Machinery, New York,NY,United States, 1–12. https://doi.org/10.1145/3447526.3472041

[11] Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2015. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software* 67, 1 (2015), 1–48. https://doi.org/10.18637/jss.v067.i01

[12] Yke Bauke Eisma, Clark Borst, René van Paassen, and Joost de Winter. 2021. Augmented Visual Feedback: Cure or Distraction? *Human factors* 63, 7 (2021), 1156–1168. https://doi.org/10.1177/0018720820924602

[13] Mica R. Endsley. 2016. *Designing for Situation Awareness.* CRC Press, Boca Raton, Florida, USA. https://doi.org/10.1201/b11371

[14] Alexander Eriksson and Neville A. Stanton. 2017. Takeover Time in Highly Automated Vehicles: Noncritical Transitions to and From Manual Control. *Human factors* 59, 4 (2017), 689–705. https://doi.org/10.1177/0018720816685832

[15] Sarah Faltaous, Martin Baumann, Stefan Schneegass, and Lewis L. Chuang. 2018. Design Guidelines for Reliability Communication in Autonomous Vehicles. In *Proceedings of the 10th International Conference on Automotive User Interfaces and Interactive Vehicular Applications (ACM Conferences)*. ACM, New York, NY, 258–267. https://doi.org/10.1145/3239060.3239072

[16] W. Holmes Finch, Jocelyn E. Bolin, and Ken Kelley. 2014. *Multilevel modeling using R.* CRC Press, Boca Raton, FL. http://site.ebrary.com/lib/alltitles/docDetail.action?docID=10891557

[17] Ernestine Fu, Srinath Sibi, David Miller, Mishel Johns, Brian Mok, Martin Fischer, and David Sirkin. 2019. The Car That Cried Wolf: Driver Responses to Missing, Perfectly Performing, and Oversensitive Collision Avoidance Systems. In *IV19*. IEEE, Piscataway, New Jersey, 1830–1836. https://doi.org/10.1109/IVS.2019.8814190

[18] Klaus Fuchs, Bettina Abendroth, and Ralph Bruder. 2009. Night Vision - Reduced Driver Distraction, Improved Safety and Satisfaction. In *Engineering Psychology and Cognitive Ergonomics*, Don Harris (Ed.). Lecture Notes in Computer Science, Vol. 5639. Springer Berlin Heidelberg, Berlin, Heidelberg, 367–375. https://doi.org/10.1007/978-3-642-02728-4{_}39

[19] 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 (2013), 1938–1942. https://doi.org/10.1177/1541931213571433

[20] Christian Gold, Moritz Körber, David Lechner, and Klaus Bengler. 2016. Taking Over Control From Highly Automated Vehicles in Complex Traffic Situations: The Role of Traffic Density. *Human factors* 58, 4 (2016), 642–652. https://doi.org/10.1177/0018720816634226

[21] Rob Gray. 2011. Looming auditory collision warnings for driving. *Human factors* 53, 1 (2011), 63–74. https://doi.org/10.1177/0018720810397833

[22] Renate Haeuslschmid, Bastian Pfleging, and Florian Alt. 05072016. A Design Space to Support the Development of Windshield Applications for the Car. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, Jofish Kaye, Allison Druin, Cliff Lampe, Dan Morris, and Juan Pablo Hourcade (Eds.). ACM, New York, NY, USA, 5076–5091. https://doi.org/10.1145/2858036.2858336

[23] Sandra G. Hart. 2006. Nasa-Task Load Index (NASA-TLX); 20 Years Later. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 50, 9 (2006), 904–908. https://doi.org/10.1177/154193120605000909

[24] Sandra G. Hart and Lowell E. Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. In *Human Mental Workload*, P. A. Hancock and N. Meshkati (Eds.). Elsevier textbooks, s.l., 139–183. https://doi.org/10.1016/s0166-4115(08)62386-9

[25] Sebastian Hergeth, Lutz Lorenz, and Josef F. Krems. 2017. Prior Familiarization With Takeover Requests Affects Drivers' Takeover Performance and Automation Trust. *Human factors* 59, 3 (2017), 457–470. https://doi.org/10.1177/0018720816678714

[26] Robert J. K. Jacob. 1991. The use of eye movements in human-computer interaction techniques. *ACM Transactions on Information Systems* 9, 2 (1991), 152–169. https://doi.org/10.1145/123078.128728

[27] Michael G. Kenward and James H. Roger. 1997. Small Sample Inference for Fixed Effects from Restricted Maximum Likelihood. *Biometrics* 53, 3 (1997), 983. https://doi.org/10.2307/2533558

[28] 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*, Anind K. Dey (Ed.). ACM, New York, NY, 110–116. https://doi.org/10.1145/1969773.1969792

[29] Hyungil Kim and Joseph L. Gabbard. 2019. Assessing Distraction Potential of Augmented Reality Head-Up Displays for Vehicle Drivers. , 18720819844845 pages. https://doi.org/10.1177/0018720819844845

[30] Johannes Kizach. 2014. Analyzing Likert-scale data with mixed-effects linear models: a simulation study. https://pure.au.dk/portal/files/70360382/simulationposterjk.pdf

[31] Jeamin Koo, Jungsuk Kwac, Wendy Ju, Martin Steinert, Larry Leifer, and Clifford Nass. 2015. 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 (2015), 269–275. https://doi.org/10.1007/s12008-014-0227-2

[32] Zhenji Lu, Xander Coster, and Joost de Winter. 2017. How much time do drivers need to obtain situation awareness? A laboratory-based study of automated driving. *Applied ergonomics* 60 (2017), 293–304. https://doi.org/10.1016/j.apergo.2016.12.003

[33] Philipp Mayring. 2010. Qualitative content analysis. In *A companion to qualitative research*, Uwe Flick, Ernst von Kardorff, and Ines Steinke (Eds.). SAGE, London, 159–176.

[34] Zeljko Medenica, Andrew L. Kun, Tim Paek, and Oskar Palinko. 2011. Augmented reality vs. street views. In *Proceedings of the 13th International Conference on Human Computer Interaction with Mobile Devices and Services - MobileHCI '11*, Markus Bylund, Oskar Juhlin, and Ylva Fernaeus (Eds.). ACM Press, New York, New York, USA, 265. https://doi.org/10.1145/2037373.2037414

[35] Coleman Merenda, Hyungil Kim, Joseph L. Gabbard, Samantha Leong, David R. Large, and Gary Burnett. 2017. Did You See Me?. In *Proceedings of the 9th International Conference on Automotive User Interfaces and Interactive Vehicular Applications (ACM Digital Library)*, Ioannis Politis (Ed.). ACM, New York, NY, 40–49. https://doi.org/10.1145/3122986.3123013

[36] Coleman Merenda, Hyungil Kim, Kyle Tanous, Joseph L. Gabbard, Blake Feichtl, Teruhisa Misu, and Chihiro Suga. 2018. Augmented Reality Interface

Design Approaches for Goal-directed and Stimulus-driven Driving Tasks. *IEEE transactions on visualization and computer graphics* 24, 11 (2018), 2875–2885. https://doi.org/10.1109/TVCG.2018.2868531

[37] Donald A. Norman and Daniel G. Bobrow. 1975. On data-limited and resource-limited processes. *Cognitive Psychology* 7, 1 (1975), 44–64. https://doi.org/10.1016/0010-0285(75)90004-3

[38] A. Pauzié. 2008. A method to assess the driver mental workload: The driving activity load index (DALI). *IET Intelligent Transport Systems* 2, 4 (2008), 315. https://doi.org/10.1049/iet-its:20080023

[39] Jurgen Pichen, Fei Yan, and Martin Baumann. 10/19/2020 - 11/13/2020. Towards a Cooperative Driver-Vehicle Interface: Enhancing Drivers' Perception of Cyclists through Augmented Reality. In *2020 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, New York, NY, USA, 1827–1832. https://doi.org/10.1109/IV47402.2020.9304621

[40] R Core Team. 2021. R: A Language and Environment for Statistical Computing. https://www.R-project.org/

[41] Miguel A. Recarte and Luis M. Nunes. 2003. Mental workload while driving: effects on visual search, discrimination, and decision making. *Journal of experimental psychology. Applied* 9, 2 (2003), 119–137. https://doi.org/10.1037/1076-898x.9.2.119

[42] Holger Regenbrecht and Thomas Schubert. 2002. Real and Illusory Interactions Enhance Presence in Virtual Environments. *Presence: Teleoperators and Virtual Environments* 11, 4 (2002), 425–434. https://doi.org/10.1162/105474602760204318

[43] Andreas Riegler, Bilal Aksoy, Andreas Riener, and Clemens Holzmann. 2020. Gaze-based Interaction with Windshield Displays for Automated Driving: Impact of Dwell Time and Feedback Design on Task Performance and Subjective Workload. In *12th International Conference on Automotive User Interfaces and Interactive Vehicular Applications (ACM Digital Library)*. Association for Computing Machinery, New York,NY,United States, 151–160. https://doi.org/10.1145/3409120.3410654

[44] Andreas Riegler, Andreas Riener, and Clemens Holzmann. 11222020. A Research Agenda for Mixed Reality in Automated Vehicles. In *19th International Conference on Mobile and Ubiquitous Multimedia*, Jessica Cauchard and Markus Löchtefeld (Eds.). ACM, New York, NY, USA, 119–131. https://doi.org/10.1145/3428361.3428390

[45] Russell V. Lenth. 2022. emmeans: Estimated Marginal Means, aka Least-Squares Means. https://CRAN.R-project.org/package=emmeans

[46] SAE. 2018. SAE J3016B Standard: Taxonomy and Definitions for Terms Related to on-Road Motor Vehicle Automated Driving Systems. https://doi.org/10.4271/J3016

[47] Daniel J. Schad, Shravan Vasishth, Sven Hohenstein, and Reinhold Kliegl. 2020. How to capitalize on a priori contrasts in linear (mixed) models: A tutorial. *Journal of Memory and Language* 110 (2020), 104038. https://doi.org/10.1016/j.jml.2019.104038

[48] Holger Schielzeth, Niels J. Dingemanse, Shinichi Nakagawa, David F. Westneat, Hassen Allegue, Céline Teplitsky, Denis Réale, Ned A. Dochtermann, László Zsolt Garamszegi, and Yimen G. Araya-Ajoy. 2020. Robustness of linear mixed–effects models to violations of distributional assumptions. *Methods in Ecology and Evolution* 11, 9 (2020), 1141–1152. https://doi.org/10.1111/2041-210X.13434

[49] Tobias Schneider, Joana Hois, Alischa Rosenstein, Sabiha Ghellal, Dimitra Theofanou-Fülbier, and Ansgar R.S. Gerlicher. 05062021. ExplAIn Yourself! Transparency for Positive UX in Autonomous Driving. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, Yoshifumi Kitamura, Aaron Quigley, Katherine Isbister, Takeo Igarashi, Pernille Bjørn, and Steven Drucker (Eds.). ACM, New York, NY, USA, 1–12. https://doi.org/10.1145/3411764.3446647

[50] Thomas Schubert, Frank Friedmann, and Holger Regenbrecht. 2001. The Experience of Presence: Factor Analytic Insights. *Presence: Teleoperators and Virtual Environments* 10, 3 (2001), 266–281. https://doi.org/10.1162/105474601300343603

[51] Zbyněk Šidák. 1967. Rectangular Confidence Regions for the Means of Multivariate Normal Distributions. *J. Amer. Statist. Assoc.* 62, 318 (1967), 626–633. https://doi.org/10.1080/01621459.1967.10482935

[52] Statista. 2021. Factors leading to road accidents in Great Britain 2019. https://www.statista.com/statistics/323079/contributing-factors-leading-to-road-accidents-in-great-britain-uk/

[53] 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*. ACM, New York, NY, USA, 392–404. https://doi.org/10.1145/3025453.3025507

[54] Titus von der Malsburg. 2015. Py-Span-Task - A Software For Testing Working Memory Span. https://doi.org/10.5281/zenodo.18238

[55] Colin Ware and Harutune H. Mikaelian. 1987. An evaluation of an eye tracker as a device for computer input2. In *Proceedings of the SIGCHI/GI conference on Human factors in computing systems and graphics interface - CHI '87*, John M. Carroll and Peter P. Tanner (Eds.). ACM Press, New York, New York, USA, 183–188. https://doi.org/10.1145/29933.275627

[56] WHO. 2018. Global status report on road safety 2018: summary. http://apps.who.int/iris/bitstream/handle/10665/277370/WHO-NMH-NVI-18.20-eng.pdf?ua=1

[57] Philipp Wintersberger, Anna-Katharina Frison, Andreas Riener, and Tamara von Sawitzky. 2019. Fostering User Acceptance and Trust in Fully Automated Vehicles: Evaluating the Potential of Augmented Reality. *PRESENCE: Virtual and Augmented Reality* 27, 1 (2019), 46–62. https://doi.org/10.1162/pres{_}a{_}00320

[58] Steven Yantis. 1993. Stimulus-Driven Attentional Capture. *Current Directions in Psychological Science* 2, 5 (1993), 156–161. https://doi.org/10.1111/1467-8721.ep10768973

[59] Steven Yantis and John Jonides. 1996. Attentional capture by abrupt onsets: New perceptual objects or visual masking? *Journal of Experimental Psychology: Human Perception and Performance* 22, 6 (1996), 1505–1513. https://doi.org/10.1037/0096-1523.22.6.1505