Evaluation of Imminent Take-Over Requests With Real Automation on a Test Track

Philipp Wintersberger1,2, Clemens Schartmüller2,3, Shadan Shadeghian-Borojeni4, Anna-Katharina Frison2, and Andreas Riener2, 1TU Wien, Austria, 2CARISSMA, Technische Hochschule Ingolstadt (THI), Germany, 3Johannes Kepler University Linz (JKU), Austria, 4University of Siegen, Germany

Objective: Investigating take-over, driving, non-driving related task (NDRT) performance, and trust of conditionally automated vehicles (AVs) in critical transitions on a test track.

Background: Most experimental results addressing driver take-over were obtained in simulators. The presented experiment aimed at validating relevant findings while uncovering potential effects of motion cues and real risk.

Method: Twenty-two participants responded to four critical transitions on a test track. Non-driving related task modality (reading on a handheld device vs. auditory) and take-over timing (cognitive load) were varied on two levels. We evaluated take-over and NDRT performance as well as gaze behavior. Further, trust and workload were assessed with scales and interviews.

Results: Reaction times were significantly faster than in simulator studies. Further, reaction times were only barely affected by varying visual, physical, or cognitive load. Post-take-over control was significantly degraded with the handheld device. Experiencing the system reduced participants’ distrust, and distrusting participants monitored the system longer and more frequently. NDRTs on a handheld device resulted in more safety-critical situations.

Conclusion: The results confirm that take-over performance is mainly influenced by visual-cognitive load, while physical load did not significantly affect responses. Future take-over request (TOR) studies may investigate situation awareness and post-take-over control rather than reaction times only. Trust and distrust can be considered as different dimensions in AV research.

Application: Conditionally AVs should offer dedicated interfaces for NDRTs to provide an alternative to using nomadic devices. These interfaces should be designed in a way to maintain drivers’ situation awareness.

Précis: This paper presents a test track experiment addressing conditionally automated driving systems. Twenty-two participants responded to critical TORs, where we varied NDRT modality and take-over timing. In addition, we assessed trust and workload with standardized scales and interviews.

Keywords: automated driving, automated vehicles, driver take-over, take-over request, handover

INTRODUCTION

Automated vehicles (AVs) promise to relieve drivers from driving and enable them to perform non-driving related tasks (NDRTs). However, this requires an appropriate level of automation—given the taxonomy by the society of automotive engineers (SAE, 2018), at least automation level 3 (L3) or higher. In L3 driving, the automation performs the dynamic driving task, but the driver is expected to take-over in situations that the system is not able to handle. Until receiving a Take-Over Request (TOR), drivers can engage in NDRTs and do not need to monitor the AV. Relying on human fallback can be problematic, though (Parasuraman et al., 2000). Upon a TOR, drivers must stop the NDRT, shift their attention to the driving environment, perceive and understand it, make decisions, and act accordingly. Various studies have addressed this issue and investigated a wide range of factors potentially influencing TOR responses, including TOR notification modalities, driver characteristics, traffic volume, driving environments, or NDRTs with varying cognitive and physical demands (McDonald et al., 2019; Zhang et al., 2019). Many of these experiments have focused on “imminent” or “emergency” take-over scenarios, where drivers must provide a split-second reaction to prevent a safety-critical situation (Eriksson & Stanton, 2017). Due to the large number of
studies, de Winter et al. (2021) have even argued that TOR studies “have characteristics of a self-sustaining convenience.” We argue that one critical challenge in this line of research is to get out of the lab and conduct more realistic experiments with higher fidelity. This is important to (a) validate results from studies with lower fidelity and (b) uncover effects of the more realistic high-fidelity environment. Most experiments so far have been conducted in driving simulators. Comparisons have shown that simulators can reproduce relative to absolute validity in specific settings (Eriksson et al., 2017; Parduzi et al., 2019), but some factors can only be assessed with limitations (Radlmayr et al., 2014). A literature review by Zhang et al. (2019) shows that 84 of 129 investigated papers used simulators without motion cues. We claim that realistic perception of motion cues and risk is essential, especially in AV-related dual-task experiments. Previous work has shown that motion can influence the results of TOR experiments (Sadeghian Borojeni et al., 2018) and McDonald et al. (2019) have highlighted “a need to confirm simulator findings in naturalistic settings.” There already exist test track and real road experiments in SAE L2 (Eriksson et al., 2017; Victor et al., 2018), and some initial evaluations of noncritical TORs in realistic settings have been presented (Morales-Alvarez et al., 2020; Naujoks et al., 2019). However, to our best knowledge, no scientific experiment has yet investigated imminent/emergency TORs, as described in many driving simulator studies, in a real vehicle (see Table 1).

With this work, we contribute to the large body of TOR studies by evaluating imminent/emergency TORs under close to real-world conditions. We imitated an L3 vehicle with a driving robot on a test track and aimed at investigating factors that (a) have been frequently addressed in driving simulators, (b) might depend on motion cues and risk perception, and (c) are important for the safe implementation of L3 systems. We included as many TOR-related measurements as technically and organizationally possible given our setting (reaction times, driving performance, NDRT performance, gaze behavior, standardized subjective scales, and interviews) to answer a set of research questions related to three main themes: TOR responses in general, effects of the physical, visual, and cognitive demands of NDRTs (reading on a handheld device and an auditory display), and psychological factors (trust and workload).

### TORs

SAE (2018) defines a “Request to Intervene” (i.e., TOR) as a “notification by an automated driving system to a fallback-ready user indicating that s/he should promptly resume performing the dynamic driving task” to either continue driving or bringing the vehicle to a safe state (“minimal risk condition”). By including the term “fallback-ready” in the latest definition, SAE (2018) accounts for potential ambiguities arising that TOR studies “have characteristics of a self-sustaining convenience.” We argue that one critical challenge in this line of research is to get out of the lab and conduct more realistic experiments with higher fidelity. This is important to (a) validate results from studies with lower fidelity and (b) uncover effects of the more realistic high-fidelity environment. Most experiments so far have been conducted in driving simulators. Comparisons have shown that simulators can reproduce relative to absolute validity in specific settings (Eriksson et al., 2017; Parduzi et al., 2019), but some factors can only be assessed with limitations (Radlmayr et al., 2014). A literature review by Zhang et al. (2019) shows that 84 of 129 investigated papers used simulators without motion cues. We claim that realistic perception of motion cues and risk is essential, especially in AV-related dual-task experiments. Previous work has shown that motion can influence the results of TOR experiments (Sadeghian Borojeni et al., 2018) and McDonald et al. (2019) have highlighted “a need to confirm simulator findings in naturalistic settings.” There already exist test track and real road experiments in SAE L2 (Eriksson et al., 2017; Victor et al., 2018), and some initial evaluations of noncritical TORs in realistic settings have been presented (Morales-Alvarez et al., 2020; Naujoks et al., 2019). However, to our best knowledge, no scientific experiment has yet investigated imminent/emergency TORs, as described in many driving simulator studies, in a real vehicle (see Table 1).

With this work, we contribute to the large body of TOR studies by evaluating imminent/emergency TORs under close to real-world conditions. We imitated an L3 vehicle with a driving robot on a test track and aimed at investigating factors that (a) have been frequently addressed in driving simulators, (b) might depend on motion cues and risk perception, and (c) are important for the safe implementation of L3 systems. We included as many TOR-related measurements as technically and organizationally possible given our setting (reaction times, driving performance, NDRT performance, gaze behavior, standardized subjective scales, and interviews) to answer a set of research questions related to three main themes: TOR responses in general, effects of the physical, visual, and cognitive demands of NDRTs (reading on a handheld device and an auditory display), and psychological factors (trust and workload).

### Table 1. Existing TOR studies conducted on test tracks or in real environments, including the level of automation, vehicle setup, included NDRTs, TOR criticality, and obstacles. The table indicates that emergency TORs have not yet been investigated outside of driving simulators.

<table>
<thead>
<tr>
<th>Paper</th>
<th>LoA/Vehicle</th>
<th>NDRT</th>
<th>TOR</th>
<th>Obstacle</th>
</tr>
</thead>
<tbody>
<tr>
<td>Victor et al. (2018)</td>
<td>L2, Volvo XC90</td>
<td>—</td>
<td>—</td>
<td>AEB Target, Garbage bag</td>
</tr>
<tr>
<td>Eriksson and Stanton (2017)</td>
<td>L2/L3, Tesla model S</td>
<td>—</td>
<td>Noncritical</td>
<td>—</td>
</tr>
<tr>
<td>Morales-Alvarez et al. (2020)</td>
<td>L3, research test vehicle</td>
<td>Reading, texting (smartphone)</td>
<td>Noncritical</td>
<td>—</td>
</tr>
</tbody>
</table>
from NDRT engagement without specifying the circumstances—for example, would using smartphones satisfy these safety requirements? Given the large number of studies, we will base our related work analysis on existing literature reviews. TOR performance can be expressed in terms of reaction times and post-take-over control, that is, driving performance (McDonald et al., 2019). Zhang et al. (2019) report an average reaction time of 2.72 ± 1.45 s. Similar values were obtained by Eriksson and Stanton (2017), who report a mean of 2.96 ± 1.96 s. An important factor in this regard is the time budget (“lead time”), where larger budgets also increase reaction times—about 0.27 s per additional second (McDonald et al., 2019). The minimum lead time which will allow all drivers to respond safely is part of discussions (Petermann-Stock et al., 2013; Radlmayr & Bengler, 2015), but it was argued as the distribution is right-tailed, typical lead times of 5 or 7 s cannot guarantee a safe response (de Winter et al., 2021; Eriksson & Stanton, 2017). At least for noncritical TORs, Eriksson et al. (2017) showed reaction times being shorter in real environments. Considering post-take-over control, Merat et al. (2014) have shown that it takes drivers up to 40 s to regain full driving fitness. Many factors influence TOR performance, such as situation awareness (Lu et al., 2017), traffic volume (Gold et al., 2016), road curvature (Sadeghian Borojeni et al., 2018), TOR modality and presentation patterns (Politis et al., 2015), driver state (Wiedemann et al., 2018), and others; see McDonald et al. (2019) or Zhang et al. (2019) for a comprehensive discussion.

NDRTs

Non-driving related tasks are often highlighted as a significant advantage of AVs, and investigations have revealed that users desire internet browsing, texting, reading, or other activities involving digital media (Pfleging et al., 2016). Many drivers may use their private, handheld devices (i.e., smartphones)—although legislation in most countries (Linz, 2017; Njus, 2017) prevents that. Even in manual driving, handheld device usage is increasing (Kubitzki & Fastenmeier, 2016), and in a longitudinal L3 simulator study conducted by Large et al. (2019), the mobile phone was the item most commonly used by participants consistently throughout the study period. Theoretically, L3 would allow handheld NDRTs, but this could reduce traffic safety since in transitions, drivers must “find out where to put down the device, and [make] arm movements to move the device to a safe storing position” (McDonald et al., 2019). However, it is not completely clear how the varying physical, visual, and cognitive demands of NDRTs impact TOR responses. Zeeb et al. (2015) have argued that “cognitive and not motor processes determine the take-over time,” as establishing motor readiness “is mostly reflexive and not, or only marginally, influenced by the driver’s level of visual distractedness.” In contrast, recent literature reviews have highlighted the role of physical processes. A linear mixed-effects model by Zhang et al. (2019) showed that handheld NDRTs increase the take-over time by 1.3 s, while auditory NDRTs or cognitive demand did not “show significant associations.” Also, the literature review by McDonald et al. (2019) indicates that handheld devices affect both take-over time and post-take-over control. Jarosch et al. (2019) have explicitly addressed this issue in a literature review and concluded that “the effects of NDRTs on the take-over performance were less pronounced than one might expect.”

Driver Support Systems for Improved Take-Over

Still, NDRTs reduce drivers’ situation awareness (SA), that is, their capabilities to perceive, comprehend, and project the state of the environment (Endsley & Kiris, 1995). To counter, researchers have proposed different support systems which can be split into two categories: systems that (a) support a fast regain of SA upon TOR or (b) reduce a loss of SA while performing NDRTs. Category (a) includes hints for maneuvering, such as vibrating or morphing steering wheels (Mok et al., 2017; Sadeghian et al., 2017) and seats (Telpaz et al., 2015), visual cues (Sadeghian Borojeni et al., 2016), pre-alerts and priming (Sadeghian et al., 2018; van der Heiden et al., 2017), or better timing of
TORs (P. Wintersberger, A. Riener, et al., 2018). Category (b) aims at reducing the visual load and keeping drivers’ gaze near the road. Proposed solutions include the provision of NDRT content on Head-up-Displays (Gerber et al., 2020; X. Li et al., 2020), but also nonvisual modalities like auditory displays and speech interaction (Martelaro et al., 2019; Schartmüller et al., 2021). The auditory modality has shown to be less SA degrading than others already in manual driving (Maciej & Vollrath, 2009; Tsimhoni et al., 2004). Consequently, auditory interaction could be promising also for higher levels of automation.

**Trust in Automation as Important Psychological Factor**

Various driver characteristics like age, training, intoxication, or fatigue have been investigated in the context of TOR (McDonald et al., 2019; Wiedemann et al., 2018). One crucial psychological construct influencing drivers’ behavior in AVs is trust in automation (Gold et al., 2015; Wintersberger et al., 2018), that is, “the attitude that an agent will help achieve an individual’s goals in a situation characterized by uncertainty and vulnerability” (Lee & See, 2004). Trust plays an important role in the interplay of NDRT engagement, monitoring behavior, and TOR responses. Drivers with a high trust may monitor the system less frequently, potentially reducing SA. However, studies in this regard came to different conclusions. While a study by Hergeth et al. (2016) suggested a correlation between trust and eye movements, Strauch et al. (2019) could not confirm that (at least from the passenger perspective). Further, existing studies suggest that system experience fosters trust in AVs. For example, Gold et al. (2015) showed that trust increased after experiencing a ride in a driving simulator, and this was confirmed by Hergeth et al. (2017). As trust is strongly associated with risk perception (Lee & See, 2004; M. Li et al., 2019), which can hardly be simulated, we aimed at validating the findings discussed above.

**AIMS**

With this experiment, we wanted to determine how drivers engaged in NDRTs respond to imminent TORs in a realistic environment where motion cues and real risk are present. In particular, we aimed at answering the following research questions (RQs, grouped into three major themes as discussed above):

- **RQ1 TOR Response:** (a) How long do drivers need to take back control, (b) how do the results compare to existing simulator studies, and (c) can all drivers respond to imminent TOR safely?
- **RQ2 NDRTs:** (d) How do physical, visual, and cognitive load influence TOR responses, (e) which NDRT modality leads to higher NDRT performance, and (f) which NDRT modality is preferred by users?
- **RQ3 Psychological Factors:** (g) Is there an influence of the NDRT modality on drivers’ workload during take-over, (h) how does the presence of real risk influence drivers’ trust in automation, and (i) does trust correlate with drivers’ monitoring behavior?

**METHOD**

**Experimental Design**

Our study included two independent variables (IVs) varied on two levels (2 × 2 factorial design). We chose reading (text) comprehension as NDRT and varied the modality. Participants performed the task either on a handheld device (a visual NDRT that requires a physical response, condition visual-motoric), or using an auditory speech display (no visual and physical load, condition auditory). We utilized the task structure (i.e., the reading task as a series of subsequent trials) to vary the cognitive load. Research has shown that interruptions (such as TOR) are more cognitively demanding when interrupting an active sub-task, rather than when issued during breaks in between (Bailey & Konstan, 2006; Naujoks et al., 2017; Wintersberger et al., 2019). Thus, we varied the TOR timing as the second IV. In one condition, participants were interrupted by a TOR during a trial (high cognitive load), while in the other, we stopped the NDRT shortly.
Table 2. Independent variables: The NDRT modality was varied on two levels to evaluate visual and physical load, and the TOR timing was varied to either interrupt drivers or issue TOR after a trial.

<table>
<thead>
<tr>
<th></th>
<th>TOR Interrupting NDRT Trial</th>
<th>TOR After Completion of NDRT Trial</th>
</tr>
</thead>
<tbody>
<tr>
<td>Visual-motoric</td>
<td>High physical/visual and high cognitive load</td>
<td>High physical/visual and low cognitive load</td>
</tr>
<tr>
<td>Auditory</td>
<td>Low physical/visual and high cognitive load</td>
<td>Low physical/visual and low cognitive load</td>
</tr>
</tbody>
</table>

before the transition so that no trial was active when the TOR was issued (low cognitive load, condition after see Table 2).

**Apparatus**

To imitate L3 driving, we installed a driving robot by Stähle GmbH (2019) in an Audi A4, which actuated the steering wheel and the pedals (see Figure 1) based on a differential global positioning system (dGPS) and pre-recorded maneuvers. This setup led to minor restrictions; participants were not allowed to apply the pedals and had to grab the steering wheel only on two designated handlebars. To maintain safety, the maximum speed was set to 30 km/h, while two experimenters (one on the passenger seat, another outside) and a redundant sensor system (observing the driving robot) could stop the vehicle anytime. The TORs were issued by the USE software framework (Schartmüller et al., 2017), which also synchronized the involved technical components (measurement devices and NDRT prototypes) to provide a reproducible setting. More details on the setup can be found in A.-K. Frison, P. Wintersberger, C. Schartmüller, and A. Riener (2019).

Take-over scenario. The track was defined as a closed-loop containing two longer straight sections, one right- and two left-hand curves, and a hairpin curve. In one straight section, the vehicle accelerated to maximum speed (30 km/h/
h). In the next curve, the track was divided into two lanes, and at the end of the curve, an AEB target was placed as an obstacle, see Figure 1. While the vehicle was driving multiple laps, the obstacle was (in each lap) quasi-randomly placed on one of the two lanes using a remote-controlled platform. Thus, participants could not know in advance which lane was free to pass. We defined five different scenarios, where a TOR was issued in the first, second, third, fourth, or sixth lap (one lap took about 1 minute of driving), with the most common lead time of 7 s (McDonald et al., 2019). When a TOR was issued, the driving robot was disabled, and participants had to grab the steering wheel to maneuver into the free lane to avoid crashing into the obstacle. The TOR was issued multimodal, including a “beep” sound and the tablet display changing from “Autopilot” (green background) to “TAKE OVER” (red background; see Figure 1, bottom right). In condition visual-motoric, the TOR notification was mirrored on the smartphone.

NDRT. The NDRT was based on the reading-span task by Daneman and Carpenter (1980), where participants had to evaluate the semantic correctness of single sentences (correct, incorrect). The sentences were displayed one-by-one either on a smartphone (Android, condition visual-motoric) or via a text-to-speech engine (condition auditory). Participants rated the correctness of these sentences using a computer mouse in condition auditory, or two buttons directly on the smartphone in condition visual-motoric (marked with green: correct and red: incorrect). The mouse was mounted in the center stack (in front of the middle armrest) to simulate typical in-vehicle interaction (e.g., BMW iDrive, see Figure 1. Based on prior findings (Schartmüller et al., 2019), the voice output speed was set to 1.5 times the default speed. A sentence timed out after 10 s if no response was given. Otherwise, the correctness of the response (green checkmark/auditory “correct” or red cross/auditory “incorrect”) was indicated, and shortly after, the next sentence was issued. To realize the IV TOR Timing, the experimenter stopped the reading task remotely before an upcoming TOR so that the next sentence would not be interrupted by the TOR (condition: after) or not (condition: interrupted).

Measurements

TOR performance. We included three reaction times measured from the time of the TOR notification: time to first gaze on the road (RTeyesOn), time to hands on the steering wheel (RThandsOn), and the time to the first steering action (RTsteer). RTeyesOn and RThandsOn were assessed manually using the iMotions software suite. RTsteer was calculated based on the first steering action >2° (Wintersberger et al., 2017). To assess driving performance after a TOR, we calculated the standard deviation of lateral position (SDLP) as described by Knappe et al. (2007) and lane exceedances by assessing if the maximal lateral deviation exceeds 0.95 m (half of the German lane width of 3.75 m minus half of the vehicle width). For these parameters, the “ideal” maneuvers of the driving robot (dGPS coordinates for both escape trajectories) acted as the reference lines.

NDRT performance. To evaluate performance in the reading task, we calculated the reaction time (RT, time between a sentence appearing and the users’ response) and the F1 score (harmonic mean of precision and recall, \( F1Score = \frac{TruePositives}{TruePositives + 0.5 \ast (FalsePositives + FalseNegatives)} \) based on the ratings of sentences’ semantic correctness. Further, we assessed the time out rate (Time-OutRate), as each sentence was dismissed in case no rating was provided within 10 s.

Gaze behavior. For eye-tracking, we utilized the Tobii (2019) Pro Glasses 2 eye-tracker with 60 Hz and corrective lenses with the iMotions software for recording and Area of Interest (AoI) tracking/mapping. The gathered eye-movement data is indicative of visual attention (relative total gaze time on the windshield: EyesOnWSrel; average glance duration on the windshield: DurGlanceOnWS) and monitoring behavior between observing the driving situation and engaging in the NDRT (number of glances on the windshield: NrGlancesOnWS; number of glances on the windshield per second: NrGlancesPerSec). A glance hereby is defined
as “maintaining of visual gaze within an AoI, and may be comprised of more than one fixation and saccade to and from it,” see ISO 15007, (Automobiltechnik, 2015). To generate comparable measures between conditions of varying length, we calculated relative or rate measures, that is, relative to the drive duration or number of glances per time unit of a drive (cf. section Participants and Procedure). The average of each respective figure (overall drives of a condition) has been evaluated.

Self-ratings and qualitative assessment. Participants’ trust and workload were measured with standardized scales. For workload the (raw) NASA-TLX questionnaire was utilized (Hart & Staveland, 1988). Trust/Distrust was assessed using the Automation Trust Scale by Jian et al. (2000) pre-test and post-condition after both NDRT conditions. In addition, we conducted semi-structured interviews with participants pre-test (expectations) and post-test (reflection).

Participants and Procedure

Twenty-two participants (12 female, 10 male, M = 22.59, SD = 2.70 years old, all students and University staff holding a valid driving license) were recruited. After expressing consent and completing the pre-test questionnaires and interviews (expectations), participants received a safety briefing were equipped with the eye-tracker, and completed a training TOR scenario (2 laps) without NDRT, supported by instructions of the experimenter. Then, participants performed 4 drives with all manifestations of the two IVs NDRT modality (visual-motoric or auditory) and TOR timing (TOR interrupting drivers or after NDRT trials). The scenario (i.e., number of laps before TOR, lasting between 1.5 and 7 minutes per trial) and the IV TOR timing were fully randomized. In contrast, NDRT modality was always bundled to two subsequent drives in counterbalanced order. After such two drives, participants completed a postcondition survey addressing the just experienced NDRT modality. Every trial started with a re-calibration of the eye-tracker and ended after participants handled the TOR, where the car came to a full stop (automated stop with the driving robot at the obstacle’s height). After all four trials, we performed an additional interview with all participants (reflection). The experiment was conducted in a single session and lasted about 2 hours per participant.

RESULTS

All tests were conducted with IBM SPSS V24 and are reported as statistically significant at p < .05. Additionally, Bonferroni correction was applied in the case of multiple tests. To perform repeated measurement ANOVA, we applied log transformation (base 10) for non-normally distributed data. Descriptive statistics and test results for TOR/driving performance are depicted in Tables 3 and 4, NDRT performance and eye-tracking in Table 5, and subjective scales in Table 6.

Table 3. Descriptive statistics (mean M and standard deviation SD) of take-over responses (take-over reaction times and driving performance).

<table>
<thead>
<tr>
<th></th>
<th>Visual-motoric, M (SD)</th>
<th>Auditory, M (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>After</td>
<td>Interrupting</td>
</tr>
<tr>
<td>Take-over reaction times</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RTeyesOn (sec.)</td>
<td>.65 (.42)</td>
<td>.72 (.42)</td>
</tr>
<tr>
<td>RThandsOn (sec.)</td>
<td>.81 (.36)</td>
<td>.97 (.61)</td>
</tr>
<tr>
<td>RTsteer (sec.)</td>
<td>1.29 (.52)</td>
<td>1.17 (.30)</td>
</tr>
<tr>
<td>Driving performance</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SDLP (m)</td>
<td>.34 (.59)</td>
<td>.29 (.31)</td>
</tr>
</tbody>
</table>
Table 4. Results of the statistical evaluation using repeated measures ANOVA including p-values and effect sizes (in brackets). In case a preceding test for normality failed, we applied log transformation to be able to perform parametric analyses.

<table>
<thead>
<tr>
<th></th>
<th>Visual-motoric versus Auditory</th>
<th>After versus Interrupting</th>
<th>Interaction</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>F</td>
<td>p ((\eta^2_p))</td>
<td>F</td>
</tr>
<tr>
<td>Take-over reaction times (df = 1, 19), (\alpha = .016)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RTeyesOn (sec.)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RTHandsOn (sec.)</td>
<td>2.13</td>
<td>.161 (.101)</td>
<td>2.17</td>
</tr>
<tr>
<td>RTsteer (sec.), (\log^{10})</td>
<td>6.26</td>
<td>.022 (.248)</td>
<td>.314</td>
</tr>
<tr>
<td>Driving performance (df = 1,19), (\alpha = .05)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SDLP (m), (\log^{10})</td>
<td>9.79</td>
<td>.006 (.340)</td>
<td>.36</td>
</tr>
</tbody>
</table>

Table 5. Descriptive statistics (Median Md, and interquartile range IQR) and results of the statistical evaluation of drivers’ glance behavior during automated driving, as well as their performance in the reading comprehension task.

<table>
<thead>
<tr>
<th></th>
<th>Visual-motoric</th>
<th>Auditory</th>
<th>Wilcoxon</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Md (IQR)</td>
<td>Md (IQR)</td>
<td>Z</td>
</tr>
<tr>
<td>Glance behavior, (\alpha = .0125)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EyesOnWSrel</td>
<td>.020 (.07)</td>
<td>.807 (.16)</td>
<td>−3.92</td>
</tr>
<tr>
<td>NrGlancesOnWS</td>
<td>7.25 (37.50)</td>
<td>14.25 (15.25)</td>
<td>−.187</td>
</tr>
<tr>
<td>NrGlancesPerSec</td>
<td>.044 (.11)</td>
<td>.084 (.09)</td>
<td>−1.12</td>
</tr>
<tr>
<td>DurGlanceOnWS</td>
<td>5.269 (2.16)</td>
<td>9.30 (17.729)</td>
<td>−3.296</td>
</tr>
<tr>
<td>NDRT performance, (\alpha = .0125)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RT (sec)</td>
<td>4.605 (.88)</td>
<td>6.66 (.62)</td>
<td>−3.582</td>
</tr>
<tr>
<td>F1Score</td>
<td>.900 (.14)</td>
<td>.881 (.09)</td>
<td>−.563</td>
</tr>
<tr>
<td>MCC</td>
<td>.801 (.20)</td>
<td>.778 (.19)</td>
<td>−.805</td>
</tr>
<tr>
<td>TimeOutRate</td>
<td>.007 (.01)</td>
<td>.277 (.18)</td>
<td>−3.823</td>
</tr>
</tbody>
</table>

Table 6. Descriptive statistics (Median Md and interquartile range IQR) of the trust and the NASA-TLX scales.

<table>
<thead>
<tr>
<th>Scale</th>
<th>Pre-test Md (IQR)</th>
<th>Visual-motoric Md (IQR)</th>
<th>Auditory Md (IQR)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trust</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Trust</td>
<td>4.86 (1.18)</td>
<td>5.50 (1.75)</td>
<td>5.36 (1.77)</td>
</tr>
<tr>
<td>Distrust</td>
<td>2.80 (.60)</td>
<td>2.30 (.90)</td>
<td>2.10 (1.25)</td>
</tr>
<tr>
<td>NASA-TLX</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Overall score</td>
<td>—</td>
<td>31.00 (17.75)</td>
<td>19.50 (13.25)</td>
</tr>
</tbody>
</table>
General Results

Participants performed 88 TORs (excluding training) with an average reaction time ($RT_{steer}$) of $M = 1.12$, SD = .16 s. A single-sample t-test ($t(19) = -52.1, p < .001$) confirms that this was significantly shorter than the 3.04 s average reaction time for transitions with 7 s lead time in the literature review by Eriksson and Stanton (2017). Further, 2 of the 88 TORs—both in condition visual-motoric—yielded to an accident. One participant crashed into the AEB target, while another one failed to react timely before steering into the curve (see Figure 2). Further, 7 of the 22 participants exceeded their driving lane (lane exceedances) at least once in condition visual-motoric, compared to 2 participants in condition auditory.

Take-Over Reaction Times and Driving Performance

Because of technical issues (incomplete data tuples), we could only conduct ANOVA with 19
participants. As the parameter RT\(\text{eyesOn}\) showed no variance, we could not apply a statistical test. Obviously, in condition auditory participants’ gaze was already on the road, while in condition visual-motoric it took them around two-thirds of a second to visually attend the situation. There was no significant effect revealed for RT\(\text{handsOn}\); participants needed a similar time to grab the steering wheel in all conditions. Regarding (RT\(\text{steer}\)), participants’ reactions in condition auditory were slightly faster (.2 s on average), yet the statistical evaluation yielded no significant effect due to the lowered significance threshold of \(p < .016\) to account for multiple tests (\(F(1, 19) = 6.26, p = .022, \eta^2_p = .248\)), see Figure 4. Still, reactions in condition auditory showed less variance (see Figure 3). While only little differences were found for reaction times, performing the NDRT with the handheld device resulted in worse driving performance after TOR: the SDLP was significantly lower in condition auditory than in condition visual-motoric (\(F(1, 19) = 9.79, p = .006, \eta^2_p = .340\), see Figure 2 and Table 4). Neither interaction effects nor other main effects regarding the second IV TOR timing were visible.

**Glance and Monitoring Behavior**

Glance behavior was assessed during all phases of automated driving (independent of TORs) and thus can only consider the IV NDRT modality, see Table 5. Consequently, Wilcoxon signed rank tests (data not normally distributed) were conducted. When performing the NDRT in condition auditory, participants monitored the environment significantly more. This is indicated by a significant difference of the parameter EyesOnWSrel (\(Z = -3.92, p < .001, r = .877\)), which expresses the ratio of participants’ glancing through the windshield, and also by the average duration of glances on/through the windshield (DurGlancesOnWs) that was (on average) 10 s longer with the auditory display, compared to condition visual-motoric (\(Z = -3.296, p < .001, r = .88\)). Further, we ran Spearman correlation analyses to relate the eye-tracking parameters with participants’ subjective trust levels (see Table 6 for descriptive statistics of standardized scales). A significant correlation was visible only in condition visual-motoric. The parameters EyesOnWSrel (\(r = .612, p = .004\), NrGlancesOnWS (\(r = .465, p = .039\)), and NrGlancesPerSec (\(r = .530, p = .016\)) correlated positively with distrust. For the trust dimension, as well as condition auditory, no correlation was significant.

**NDRT Performance**

Wilcoxon tests show that participants were significantly faster in their ratings (RT, \(Z = -3.58, p < .001, r = .822\)) and had a lower TimeOutRate (\(Z = -3.82, p < .001, r = .877\)) in condition visual-motoric than in condition auditory (see Table 5). Although the success rate (F1Score) did not differ, this indicates a performance benefit in favor of the smartphone. In this condition, we also investigated where participants put the device before responding to the TOR. In 36 of the 44 TORs (as each participant performed two TORs in condition visual-motoric), the device was put on the lap. Sometimes, this meant that participants let the device “fall” on the seat rather than performing a controlled movement. In six trials, the device was put in the center console, and in two TOR situations (by two different participants), it was kept in hand, and the maneuver was performed with only one hand on the steering wheel.

**Subjective Ratings and Interviews**

The investigated scales showed acceptable reliability with a Cronbach’s \(\alpha > .6\). Descriptive statistics of the assessed scales are depicted in Table 6.

*Expectation versus reflection*. We used Friedman ANOVA to evaluate differences between participants’ trust before (expectations) and after (reflection) both conditions. Only distrust was affected (\(\chi^2(2) = 10.89; p = .004, w = .248\)), and post-hoc analyses (Wilcoxon) indicated a significant decrease for both NDRT modalities after experiencing the system (auditory \(z = 2.94, p = .010, r = .627\); visual-motoric \(z = 2.49, p = .039, r = .531\)). Half of the participants (n=11) articulated positive expectations towards L3 driving: “transport is more relaxed
if you do not drive yourself; if you do not have to be constantly alert and maybe you can do other things” (P2). Especially high expectations towards working, reading, texting, that is, doing “something useful” were expressed (n=9). Only (n=5) participants expressed skepticism: “it is still an immature system, and as long as I have not used it, I am not convinced that it works.” (P12).

Post-test (reflection), participants stated the experience to be interesting and exciting. Overall, most participants felt safe: “I trusted the technology pretty quickly. However, you are always nervous when you have to intervene. In real life, it can always happen, and then I have always been tense.” (P21).

NDRT Modality – Visual-Motoric versus Auditory

While no significant differences (Wilcoxon) were found for trust/distrust, we can report a significantly higher NASA-TLX overall score for condition visual-motoric (z = 3.22, p = .001, r = .687). Participants rated the auditory condition to be less demanding. This was also visible in interviews, where (n=18) participants preferred condition auditory. The main reason was an increase in situation awareness: “I liked the voice output better, as my gaze and focus were even more on the road, which allowed me to watch the drive,” (P8). Further, it was mentioned that having a device in the hands impairs TOR responses: “You first have to put it somewhere. It is a waste of time on something trivial, (P9); “while listening you can look out [of the window]. This is relaxing” (P18). Those preferring the smartphone argued with the possibility to self-pace the reading task: “If I was safe and everything was correct, then I could simply read the text” (P1). Further, handheld devices could allow more complex tasks, and for some, paying attention to the voice was perceived as stressful. P6 opted for changing modalities: “In a traffic jam the voice output would be better [...] on a highway I could imagine to look down and pay less attention.”

DISCUSSION

In the following, we discuss the results along the three main themes TOR response, NDRTs, and psychological factors.

TOR Response

Overall, the obtained reaction times were comparably short. On average, it took drivers slightly less than a second to grab the steering wheel and slightly above a second to execute a response (RQ1.a). The reaction time to the first steering action (RTsteer) was significantly shorter than in similar simulator studies, see Eriksson and Stanton (2017); McDonald et al. (2019). This confirms Eriksson et al. (2017), who obtained shorter reaction times in real vehicles for non-critical TORs, compared to simulator results, for imminent TORs too (RQ1.b). Two TOR responses resulted in a potential crash, both in condition visual-motoric. Further, 7 participants using the smartphone exceeded the lane, compared to only 2 in condition auditory (RQ1.c). This was not an issue on the test track but could become critical in real environments with curbs or soft shoulders. These results highlight the downsides of visual NDRTs on handheld devices. In real driving situations, imminent TORs will be even more surprising for distracted drivers. Consequently, we would not suggest using smartphones or other handheld devices when drivers act as fallback. Further, given that also in condition auditory 2 drivers exceeded the lane boundaries, relying on human fallback may be problematic in general. We argue that drivers should at least not use L3 systems without prior training and supervision.

NDRTs

Effects of physical or visual demand are barely visible when solely considering reaction times (see Figure 4). Neither the time to hands on the steering wheel (RThandsOn) nor the reaction time to the first steering action (RTsteer) were significantly affected in the comparison of visual-motoric and auditory conditions. We did not observe a significant increase in reaction time when handheld devices were involved, which contradicts existing findings obtained in driving simulators (McDonald et al., 2019; Zhang et al., 2019). Rather, our results (no difference in RThandsOn) indicate that imminent TOR responses are mostly reflexive, meaning that they
are not, or are only marginally influenced by the level of visual distractedness, as suggested by Zeeb et al. (2015). However, a potential effect regarding RTsteer (not significant in our experiment due to the lowered significance threshold) suggests that this issue should be investigated in less predictable scenarios in the future. Further, varying cognitive load did not affect reaction times. Drivers performed a TOR similarly when interrupted during the NDRT as when no reading task was active. However, visual distraction resulted in higher variance among reaction times and less stable maneuvering than auditory. In condition auditory, participants deviated significantly less from the lane center (SDLP), which indicates better post-take-over control (RQ2.d). Further, in condition auditory, participants, as expected, monitored the road environment longer and had their gaze already on the road when a TOR was issued. Thus, future L3 vehicles should provide NDRT modalities that allow drivers to maintain situation awareness and monitor the road environment. Previous works have already suggested alternatives such as augmented reality (Schartmüller et al., 2018).

Still, drivers should also use these dedicated NDRT modalities when available. In our study, participants performed better in the NDRT in condition visual-motoric (RQ2.e). However, most participants preferred the auditory modality (RQ2.f). Given that some drivers favor their own devices over in-vehicle systems (Oviedo-Trespalacios et al., 2019), convenience and performance benefits may foster private device use. Considering the downsides of visual-motoric NDRTs (see above), future driver monitoring systems could mediate drivers’ NDRT demands to warn and educate them. For example, when short lead times are expected (i.e., in dense traffic or bad weather), such systems could partly lock smartphone functionalities, similar to already existing systems such as Android Auto.

**Psychological Factors: Trust and Workload**

The results of the NASA-TLX questionnaire ratings show that participants perceived the TORs in combination with the visual-motoric NDRT as significantly more demanding than in the auditory condition (RQ3.g). The Trust Scale
indicated that participants’ subjective trust levels were high before the experiment and remained stable. In contrast, distrust significantly decreased for both NDRT conditions (RQ3.h). Further, participants’ monitoring behavior was correlated with distrust, that is, higher distrust was accompanied by more frequent and more prolonged gazes on the road environment (RQ3.i), which confirms the results obtained by (Hergeth et al. (2015) but only for visual-motoric NDRTs. The results also indicate that distrust seems to be a separate dimension rather than the opposite of trust, confirming (Frison et al., 2019). Further, as distrust vanishes quickly, we might have to expect overtrust issues to become relevant in L3 driving.

LIMITATIONS AND FUTURE WORK

There exist several limitations in our experimental design that need to be discussed. First, the TOR always happened in the same section of the test track. Although it was not clear in which lap a TOR would happen and participants had to maneuver the vehicle into the free lane, there was still some predictability involved. For example, all reactions required participants to steer into a right-hand curve, which could be one explanation for the lack of statistically significant differences in reaction times. Second, the TOR response can only be considered as partial, as participants could not apply the brakes and also could only grab the steering wheel at the designated handlebars. Previous studies have suggested that different TOR scenarios may affect drivers’ reactions in terms of braking and steering differently (Gold et al., 2018; Zeeb et al., 2017), which could not be covered by our experiment. Third, the safety restrictions only allowed a speed up to 30 km/h, and higher speeds could result in different behavior because of higher associated risks for the driver. Also, the presence of the driving robot must be considered as a limitation, as it may have influenced participants’ subjective perceptions (i.e., the system appeared as a prototype rather than a sophisticated system what may have influence participants’ trust levels) and immediately disengaged upon TOR. We also sometimes felt that test track studies may not be beneficial per se—for example, participants in driving simulators might be more relaxed, which might resemble everyday driving better than operating expensive experimental technology. Still, we believe our study to be an interesting first step towards more realistic L3 user research settings. Finally, the choice for our NDRT aimed at a holistic comparison of L3 NDRT interfaces (i.e., brought-in consumer device vs. vehicle-integrated auditory interaction). In natural NDRTs, drivers’ motivation may differ from our experiment, which used a performance-oriented task. Overall, we suggest developing dedicated user interfaces to allow drivers to perform NDRTs in a safe manner while maintaining SA, such as augmented reality displays or context-aware adaptation to the driving environment. However, not all tasks can be substituted with technology (i.e., eating and reading a book), which should be considered in-vehicle interior designs. Finally, our experiment included only a small sample with young participants from a technical university. Previous research has shown that dual-task performance differs between age groups (Brouwer et al., 1991). Consequently, future experiments should validate the presented results in more detail and include other participant groups, less predictable TOR scenarios, and other (potentially more natural) NDRTs.

CONCLUSION

As (to our best knowledge) no scientific experiment has yet addressed imminent (i.e., “emergency”) TORs outside of driving simulators, we investigated this issue by simulating an L3 vehicle with a driving robot on a test track. We evaluated factors that have been frequently addressed in classic driving simulators and are relevant for implementing L3 driving. In particular, we investigated (a) how take-over responses on a test track relate to driving simulators, (b) how the visual, physical, and cognitive demand of potentially realistic NDRTs (reading on a handheld device vs. an auditory display) influences automated driving with take-overs, and (c) how the presence of real risk affects driver workload, trust, and monitoring behavior. Our results indicate that
reactions in real vehicles are mostly reflexive and faster than in simulators. Still, the loss of situation awareness due to visual load negatively influences post-take-over control. Further, we found that monitoring behavior with visual NDRTs correlated with subjective (dis)trust, while distrust vanished during the experiment in both conditions. Finally, while using the handheld device, two drivers crashed and a high proportion of drivers exceeded the lane boundaries.

Besides validation of typical driving simulator results, our study shows that drivers (a) may have problems taking back control in emergencies, (b) should be supported by technology when engaging in demanding NDRTs, and (c) should be provided adequate warning systems that reduce unjustified engagement in NDRTs. Ultimately, the research community should define the term “fallback ready” more concretely and design appropriate interfaces for NDRTs that allow drivers to maintain situation awareness—our experiment indicates that “eyes-free” driving could become a safety risk even on SAE L3.

**KEY POINTS**

- Varying visual, physical, and cognitive load of NDRTs barely influence reaction times in imminent take-over situations. Reactions appear mostly reflexive, while physical and cognitive load may not have as much of a significant impact on reaction times as simulator studies suggest.
- Visual load and the associated loss of situation awareness significantly degraded post-take-over control. SAE L3 systems should support modalities for NDRTs that maintain drivers’ situation awareness.
- We recorded two potential crashes and more lane exceedances with a visual-motoric NDRT on a handheld device. Given that participants’ NDRT performance was higher when using the handheld device, future driver monitoring systems may mediate drivers’ NDRT demands and provide adequate warnings when detecting potentially safety-critical behavior.
- Participants’ trust was already high before the experiment, and the “increase in trust” due to system experience could solely be attributed to a reduction of distrust. Thus, treating trust and distrust as separate dimensions can be valuable in the AV context.
- Results show that distrust can be inferred from monitoring behavior, but only during engagement in visual NDRTs.

**ACKNOWLEDGMENTS**

We want to thank Dr. Missie Smith for proofreading and editing this work.

**DECLARATION OF CONFLICTING INTERESTS**

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

**FUNDING**

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work is supported under the FH-Impuls program of the German Federal Ministry of Education and Research (BMBF) under Grant Number 13FH71011A (SAFIR). Clemens Schartmüller was supported by the “Innovative Hochschule” program of the German Federal Ministry of Education and Research (BMBF) under Grant No. 03IHS109 A (MenschINBewegung).

**REFERENCES**


Telpaz, A., Rhindress, B., Zelman, I., & Tsimhoni, O. (2015). Haptic seat for automated driving: Preparing the driver to take control effectively. In Proceedings of the 7th International Conference on Automotive User Interfaces and Interactive Vehicular Applications, Nottingham United Kingdom, 1–3 September 2015 (pp. 23–30).


Philipp Wintersberger is a researcher at TU Wien (Vienna University of Technology, Austria). He received his Ph.D. in Technical Engineering Sciences from Johannes Kepler University Linz in 2020.

Clemens Schartmüller is a researcher in the Human-Computer Interaction Group at Technische Hochschule Ingolstadt, Ingolstadt, Germany. He is a PhD student in Computer Science at Johannes Kepler University, Linz, Austria.

Shadan Shadeghian-Borojeni is a postdoctoral researcher in the Experience and Interaction group at the University of Siegen, Germany. She received her Ph.D. in Human-Computer Interaction from the University of Oldenburg in 2018.

Anna-Katharina Frison is a Senior UX Researcher and founder of frisUX a consultancy for Research and Design based in Germany. She received her Ph.D. in Human-Computer Interaction from the Johannes Kepler University of Linz in Austria in 2020.

Andreas Riener is a professor at the Faculty of Computer Science and head of the Human-Computer Interaction Group at Technische Hochschule Ingolstadt, Germany. He received his PhD in applied computer science from Johannes Kepler University Linz in 2009.