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

Philipp Wintersberger[1,2], Clemens Schartmüller[2,3], Shadan Shadeghian-Borojeni[4], Anna-Katharina Frison[2], and Andreas Riener[2], [1]TU Wien, Austria, [2]CARISSMA, Technische Hochschule Ingolstadt (THI), Germany, [3]Johannes Kepler University Linz (JKU), Austria, [4]University 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

Address correspondence to Philipp Wintersberger, Institute of Visual Computing and Human-Centered Technology, TU Wien, Favoritenstraße 9, 1040 Wien, Austria.
Email: philipp.wintersberger@tuwien.ac.at

HUMAN FACTORS
Vol. 0, No. 0, ■■ ■, pp. 1-17
DOI:10.1177/00187208211051435
Article reuse guidelines: sagepub.com/journals-rmissions

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 control 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

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.

Paper	LoA/Vehicle	NDRT	TOR	Obstacle
Victor et al. (2018)	L2, Volvo XC90	—	—	AEB Target, Garbage bag
Naujoks et al. (2019)	L3, BMW 520d, Wizard-of-Oz	Reference task, audio book, search task, reading, playing Tetris	Noncritical	—
Eriksson and Stanton (2017)	L2/L3, Tesla model S	—	Noncritical	—
Morales-Alvarez et al. (2020)	L3, research test vehicle	Reading, texting (smartphone)	Noncritical	—

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

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) and 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.

	TOR Interrupting NDRT Trial	TOR After Completion of NDRT Trial
Visual-motoric	High physical/visual and high cognitive load	High physical/visual and low cognitive load
Auditory	Low physical/visual and high cognitive load	Low physical/visual and low cognitive load

Figure 1. Study setup: Driving robot operating the vehicle. Participants were emphasized to control the steering wheel with both hands, but only at the designated handlebars (top left); TOR scenario with multi-lane right-hand curve and movable automatic emergency braking (AEB) target (top right); reading comprehension task performed on a smartphone (bottom left); TOR scenario from the perspective of a participant (bottom right).

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). 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 stoppped the reading task remotely before an upcoming TOR so that the next sentence would notbe 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, $F1Score = \text{TruePositives}/\text{TruePositives} + 0.5 * (\text{FalsePositives} + \text{FalseNegatives}))$ based on the ratings of sentences' semantic correctness. Further, we assessed the time out rate (**TimeOutRate**), 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).

	Visual-motoric, M (SD)		Auditory, M (SD)	
	After	Interrupting	After	Interrupting
Take-over reaction times				
RTeyesOn (sec.)	.65 (.42)	.72 (.42)	0 (0)	0 (0)
RThandsOn (sec.)	.81 (.36)	.97 (.61)	.73 (.12)	.80 (.18)
RTsteer (sec.)	1.29 (.52)	1.17 (.30)	1.01 (.26)	1.02 (.23)
Driving performance				
SDLP (m)	.34 (.59)	.29 (.31)	.15 (.06)	.19 (.20)

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.

	Visual-motoric versus Auditory		After versus Interrupting		Interaction	
	F	p (η_p^2)	F	p (η_p^2)	F	p (η_p^2)
Take-over reaction times (df = 1, 19), $\alpha = .01\overline{6}$						
RTeyesOn (sec.)	—	—	—	—	—	—
RThandsOn (sec.)	2.13	.161 (.101)	2.17	.157 (.103)	.37	.551 (.019)
RTsteer (sec.), $\log^{10}$	6.26	.022 (.248)	.314	.582 (.016)	.98	.335 (.049)
Driving performance (df = 1,19), α = .05						
SDLP (m), $\log^{10}$	9.79	.006 (.340)	.36	.556 (.019)	.03	.871 (.001)

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.

	Visual-motoric	Auditory	Wilcoxon	
	Md (IQR)	Md (IQR)	Z	p ($r = \lvert Z/\sqrt{n}\rvert$)
Glance behavior, α = .0125				
EyesOnWSrel	.020 (.07)	.807 (.16)	−3.92	**<.001 (.877)**
NrGlancesOnWS	7.25 (37.50)	14.25 (15.25)	−.187	.852 (.041)
NrGlancesPerSec	.044 (.11)	.084 (.09)	−1.12	.263 (.25)
DurGlanceOnWS	5.269 (2.16)	9.30 (17.729)	−3.296	**.001 (.88)**
NDRT performance, α = .0125				
RT (sec)	4.605 (.88)	6.66 (.62)	−3.582	**<.001 (.822)**
F1Score	.900 (.14)	.881 (.09)	−.563	.573 (.129)
MCC	.801 (.20)	.778 (.19)	−.805	.421 (.185)
TimeOutRate	.007 (.01)	.277 (.18)	−3.823	**<.001 (.877)**

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

Scale	Pre-test Md (IQR)	Visual-motoric Md (IQR)	Auditory Md (IQR)
Trust			
Trust	4.86 (1.18)	5.50 (1.75)	5.36 (1.77)
Distrust	2.80 (.60)	2.30 (.90)	2.10 (1.25)
NASA-TLX			
Overall score	—	31.00 (17.75)	19.50 (13.25)

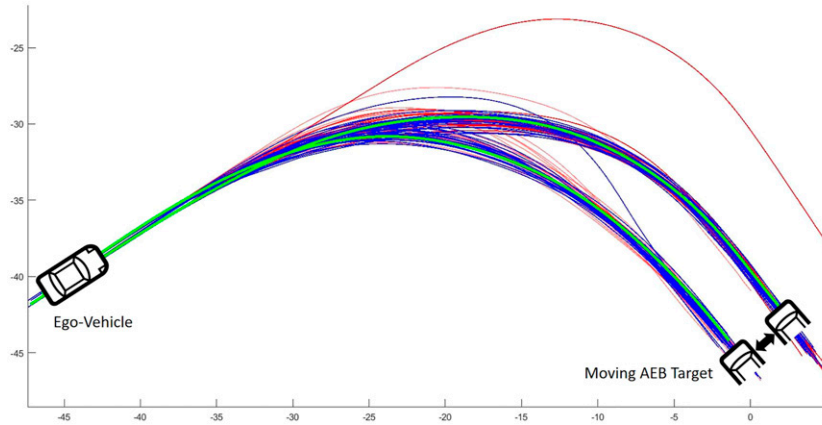

Figure 2. Plotted driving trajectories of all TOR responses. The thicker green lines represent the reference lines by the driving robot, TORs in condition *auditory* are colored in blue, and TORs in condition *visual-motoric* in red. The red outlier depicts one of the two potential accidents.

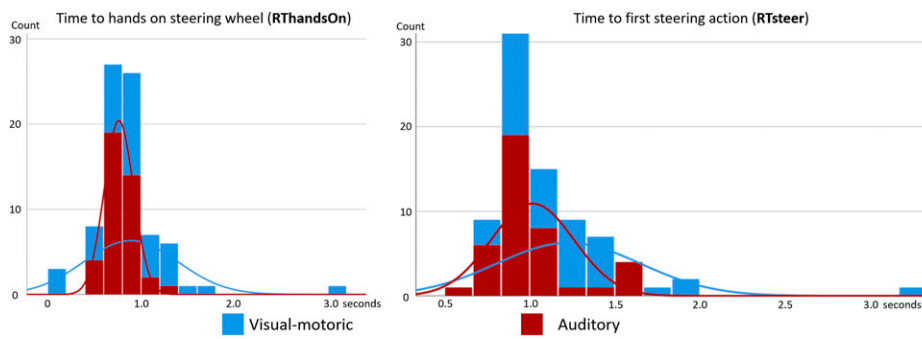

Figure 3. Histograms with normal lines showing the time to hands on steering wheel (RThandsOn, left) and the reaction time to the first steering action (RTsteer, right) after a take-over request. Reactions in condition *visual-motoric* are only slightly delayed compared to condition *auditory*, but show higher variance.

General Results

Participants performed 88 TORs (excluding training) with an average reaction time (**RTsteer**) 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 **RTeyesOn** 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 **RThandsOn**; participants needed a similar time to grab the steering wheel in all conditions. Regarding (**RTsteer**), 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 $p < .01\overline{6}$ to account for multiple tests ($F(1, 19) = 6.26, p = .022, \eta_p^2 = .248$), 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* ($F(1, 19) = 9.79$, $p = .006, \eta_p^2 = .340$, 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

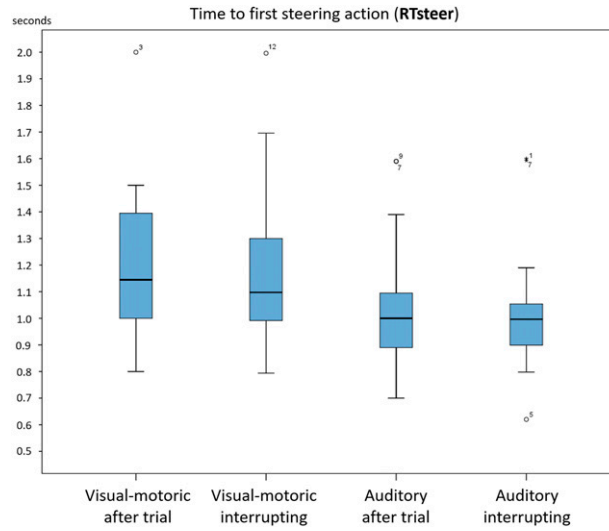

Figure 4. The comparison of the reaction times to the first steering action (**RTsteer**) indicates only marginal differences between the conditions. One outlier (see Figure 2) in condition *visual-motoric/after* is beyond the scale of the *Y*-axis.

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 adaption 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 13FH7I01IA (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

Automobiltechnik, D.-N. (2015). *Road vehicles – measurement of driver visual behaviour with respect to transport information and control systems – Part 1: Definitions and parameters (ISO15007-1:2014)*. (Tech. Rep.). ISO.

Bailey, B. P., & Konstan, J. A. (2006). On the need for attention-aware systems: Measuring effects of interruption on task performance, error rate, and affective state. *Computers in Human Behavior*, *22*(4), 685–708. (Attention aware systems). https://doi.org/10.1016/j.chb.2005.12.009.

Brouwer, W. H., Waterink, W., Van Wolffelaar, P. C., & Rothengatter, T. (1991). Divided attention in experienced young and older drivers: Lane tracking and visual analysis in a dynamic driving simulator. *Human Factors*, *33*(5), 573–582.

Daneman, M., & Carpenter, P. A. (1980, August). Individual differences in working memory and reading. *Journal of Verbal Learning and Verbal Behavior*, *19*(4), 450–466. https://doi.org/10.1016/S0022-5371(80)90312-6. https://www.sciencedirect.com/science/article/pii/S0022537180903126

de Winter, J., Stanton, N., & Eisma, Y. B. (2021). Is the take-over paradigm a mere convenience?

Endsley, M. R., & Kiris, E. O. (1995). The out-of-the-loop performance problem and level of control in automation. *Human Factors*, *37*(2), 381–394. https://doi.org/10.1518/001872095779064555.

Eriksson, A., Banks, V. A., & Stanton, N. A. (2017). Transition to manual: Comparing simulator with on-road control transitions. *Accident Analysis & Prevention*, *102*, 227-234.

Eriksson, A., & Stanton, N. A. (2017). Takeover time in highly automated vehicles: Noncritical transitions to and from manual control. *Human Factors*, *59*(4), 689–705. https://doi.org/10.1177/0018720816685832.

Frison, A.-K., Wintersberger, P., Schartmüller, C., & Riener, A. (2019). The real t (h) or: Evaluation of emergency take-over on a test track. In Proceedings of the 11th International Conference on Automotive User Interfaces and Interactive Vehicular Applications: Adjunct Proceedings, Utrecht Netherlands, 21–25 September 2019 (pp. 478–482).

Frison, A.-K., Wintersberger, P., Schartmüller, C., Riener, A., Boyle, L. N., Millner, E., & Weigl, K. (2019). In ux we trust - the impact of hmi design and system performance on the perception/experience of automated driving. In Proceedings of the Sigchi Conference on Human Factors in Computing Systems, Glasgow Scotland Uk, 4–9 May 2019 (pp. 159–166).

Gerber, M. A., Schroeter, R., Xiaomeng, L., & Elhenawy, M. (2020). Self-interruptions of non-driving related tasks in automated vehicles: Mobile self-interruptions of non-driving related tasks in automated vehicles: Mobile vs head-up display. In Proceedings of the Conference on Human Factors in Computing Systems CHI'20, Honolulu, HI, USA, 25–30 April 2020. ACM. (Issue: April). https://doi.org/10.1145/3313831.3376751.

Gold, C., Happee, R., & Bengler, K. (2018). jul). Modeling take-over performance in level 3 conditionally automated vehicles. *Accident Analysis & Prevention*, *116*, 3-13. https://doi.org/10.1016/J.AAP.2017.11.009. https://www.sciencedirect.com/science/article/pii/S0001457517303962.

Gold, C., Körber, M., Hohenberger, C., Lechner, D., & Bengler, K. (2015). Trust in automation–before and after the experience of take-over scenarios in a highly automated vehicle. *Procedia Manufacturing*, *3*, 3025-3032. https://doi.org/10.1016/j.promfg.2015.07.847.

Gold, C., Körber, M., Lechner, D., & Bengler, K. (2016). Taking over control from highly automated vehicles in complex traffic situations: The role of traffic density. *Human Factors*, *58*(4), 642–652. https://doi.org/10.1177/0018720816634226.

Hart, S. G., & Staveland, L. E. (1988). Development of nasa-tlx (task load index): Results of empirical and theoretical research. In P. A. Hancock & N. Meshkati (Eds.), *Human mental workload* (Vol. 52, pp. 139–183). North-Holland. https://doi.org/10.1016/S0166-4115(08)62386-9. http://www.sciencedirect.com/science/article/pii/S0166411508623869.

Hergeth, S., Lorenz, L., & Krems, J. F. (2017). Prior familiarization with takeover requests affects drivers' takeover performance and automation trust. *Human Factors*, *59*(3), 457–470. https://doi.org/10.1177/0018720816678714.

Hergeth, S., Lorenz, L., Krems, J. F., & Toenert, L. (2015). Effects of take-over requests and cultural background on automation trust in highly automated driving.

Hergeth, S., Lorenz, L., Vilimek, R., & Krems, J. F. (2016). Keep your scanners peeled: Gaze behavior as a measure of automation trust during highly automated driving. *Human Factors*, *58*(3), 509–519. https://doi.org/10.1177/0018720815625744.

Jarosch, O., Gold, C., Naujoks, F., Wandtner, B., Marberger, C., Weidl, G., & Schrauf, M. (2019). *The impact of non-driving related tasks on take-over performance in conditionally automated driving–a review of the empirical evidence*. 9. tagung automatisiertes fahren.

Jian, J.-Y., Bisantz, A. M., & Drury, C. G. (2000). Foundations for an empirically determined scale of trust in automated systems. *International Journal of Cognitive Ergonomics*, *4*(1), 53–71. https://doi.org/10.1207/S15327566IJCE0401_04.

Knappe, G., Keinath, A., Bengler, K., & Meinecke, C. (2007). Driving simulator as an evaluation tool–assessment of the influence of field of view and secondary tasks on lane keeping and steering performance. In 20th International Technical Conference on the Enhanced Safety of Vehicles (esv) National Highway Traffic Safety Administration, Lyon , France, 18–21 June 2007.

Kubitzki, J., & Fastenmeier, W. (2016). *Ablenkung durch moderne informations- und kommunikationstechniken und soziale interaktion bei autofahrern*. Allianz AG.

Large, D., Burnett, G., Salanitri, D., Lawson, A., & Box, E. (2019). A longitudinal simulator study to explore drivers' behaviour in level 3 automated vehicles.

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors*, *46*(1), 50–80. https://doi.org/10.1518/hfes.46.1.50_30392.

Li, M., Holthausen, B. E., Stuck, R. E., & Walker, B. N. (2019). No risk no trust: Investigating perceived risk in highly automated driving. In Proceedings of the 11th International Conference on Automotive User Interfaces and Interactive Vehicular Applications, Utrecht Netherlands, 21–25 September 2019 (pp. 177–185).

Linz, C. (2017, September 21). *Handy am steuer wird teurer. Stuttgarter Nachrichten*. https://goo.gl/FopUCo)

Li, X., Schroeter, R., Rakotonirainy, A., Kuo, J., & Lenné, M. G. (2020, April). Effects of different non-driving-related-task display modes on drivers' eye-movement patterns during take-over in an automated vehicle. *Transportation Research Part F: Traffic Psychology and Behaviour*, *70*(4), 135–148. (Publisher: Elsevier Ltd) https://doi.org/10.1016/j.trf.2020.03.001

Lu, Z., Coster, X., & De Winter, J. (2017). How much time do drivers need to obtain situation awareness? a laboratory-based study of automated driving. *Applied Ergonomics*, *60*, 293-304. https://doi.org/10.1016/j.apergo.2016.12.003.

Maciej, J., & Vollrath, M. (2009). Comparison of manual vs. speech-based interaction with in-vehicle information systems. *Accident Analysis & Prevention*, *41*(5), 924–930. https://doi.org/10.1016/j.aap.2009.05.007.

Martelaro, N., Teevan, J., & Iqbal, S. T. (2019). An exploration of speech-based productivity support in the car. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems - CHI '19, Glasgow, Scotland, UK, 4–9 May 2019 (pp. 1–12).

McDonald, A. D., Alambeigi, H., Engström, J., Markkula, G., Vogelpohl, T., Dunne, J., & Yuma, N. (2019). Toward computational simulations of behavior during automated driving takeovers: A review of the empirical and modeling literatures. *Human Factors*, *61*(4), 642–688. https://doi.org/10.1177/0018720819829572.

Merat, N., Jamson, A. H., Lai, F. C. H., Daly, M., & Carsten, O. M. J. (2014). Transition to manual: Driver behaviour when resuming control from a highly automated vehicle. *Transportation Research Part F: Traffic Psychology and Behaviour*, *27*(Part B), 274–282. https://doi.org/10.1016/j.trf.2014.09.005.

Mok, B., Johns, M., Yang, S., & Ju, W. (2017). Reinventing the wheel: Transforming steering wheel systems for autonomous vehicles. In Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology - UIST, *17*, 229-241, New York, NY: Association for Computing Machinery. https://doi.org/10.1145/3126594.3126655.

Morales-Alvarez, W., Smirnov, N., Matthes, E., & Olaverri-Monreal, C. (2020). Vehicle automation field test: Impact on driver behavior and trust. In 2020 ieee 23rd International Conference on Intelligent Transportation Systems (itsc), Rhodes, Greece, 20–23 September 2020 (pp. 1–6).

Naujoks, F., Purucker, C., Wiedemann, K., & Marberger, C. (2019). Noncritical state transitions during conditionally automated driving on german freeways: Effects of non–driving related tasks on takeover time and takeover quality. *Human Factors*, *61*(4), 596–613. https://doi.org/10.1177/0018720818824002.

Naujoks, F., Wiedemann, K., & Schömig, N. (2017). The importance of interruption management for usefulness and acceptance of automated driving. In Proceedings of the 9th International Conference on Automotive User Interfaces and Interactive Vehicular Applications, Oldenburg Germany, 24–27 September 2017 (pp. 254–263). ACM. https://doi.org/10.1145/3122986.3123000.

Njus, E. (2017, September 20). Oregon's new distracted driving law takes effect oct. 1: What you need to know. *The Oregonian*. https://goo.gl/W9bCKK)

Oviedo-Trespalacios, O., Nandavar, S., & Haworth, N. (2019). How do perceptions of risk and other psychological factors influence the use of in-vehicle information systems (ivis)? *Transportation Research Part F: Traffic Psychology and Behaviour*, *67*, 113-122. https://doi.org/10.1016/j.trf.2019.10.011.

Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2000). A model for types and levels of human interaction with automation. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, *30*(3), 286–297. https://doi.org/10.1109/3468.844354.

Parduzi, A., Venrooij, J., & Marker, S. (2019). A method to assess the effect of vertical dynamics on driving performance in driving simulators: A behavioural validation study. In International Conference on Human-Computer Interaction (pp. 176–189).

Petermann-Stock, I., Hackenberg, L., Muhr, T., & Mergl, C. (2013). *Wie lange braucht der fahrer-eine analyse zu übernahmezeiten aus verschiedenen nebentätigkeiten während einer hochautomatisierten staufahrt. 6*. Tagung Fahrerassistenzsysteme. Der Weg zum automatischen Fahren.

Pfleging, B., Rang, M., & Broy, N. (2016). Investigating user needs for non-driving-related activities during automated driving. In Proceedings of the 15th International Conference on Mobile and Ubiquitous Multimedia, Rovaniemi Finland, 12–15 December 2016 (pp. 91–99).

Politis, I., Brewster, S., & Pollick, F. (2015). Language-based multimodal displays for the handover of control in autonomous cars. In Proceedings of the 7th International Conference on Automotive User Interfaces and Interactive Vehicular Applications, Nottingham United Kingdom, 1–3 September 2015 (pp. 3–10). ACM. https://doi.org/10.1145/2799250.2799262.

Radlmayr, J., & Bengler, K. (2015). *Literature survey and description of methods for the development of highly automated driving*. FAT Schriftenreihe.

Radlmayr, J., Gold, C., Lorenz, L., Farid, M., & Bengler, K. (2014). How traffic situations and non-driving related tasks affect the take-over quality in highly automated driving. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *58*(1), 2063–2067. https://doi.org/10.1177/1541931214581434.

Sadeghian Borojeni, S., Boll, S. C. J., Heuten, W., Bülthoff, H. H., & Chuang, L. (2018). Feel the movement: Real motion influences responses to take-over requests in highly automated vehicles. In Proceedings of the 2018 chi Conference on Human Factors in Computing Systems, Montreal QC Canada, 21–26 April 2018 (pp. 246:1–246:13). ACM. https://doi.org/10.1145/3173574.3173820.

Sadeghian Borojeni, S., Chuang, L., Heuten, W., & Boll, S. (2016). Assisting drivers with ambient take-over requests in highly automated driving. In Proceedings of the 8th International Conference on Automotive User Interfaces and Interactive Vehicular Applications, Ann Arbor MI USA, 24–26 October 2016 (pp. 237–244).

Sadeghian, S., Wallbaum, T., Heuten, W., & Boll, S. (2017). Comparing shape-changing and vibro-tactile steering wheels for take-over requests in highly automated driving. In Proceedings of the 9th International Conference on Automotive User Interfaces and Interactive Vehicular Applications, Oldenburg Germany, 24–27 September 2017 (pp. 221–225).

Sadeghian, S., Weber, L., Heuten, W., & Boll, S. (2018). *From reading to driving: Priming mobile users for take-over situations in highly automated driving*. Association for Computing Machinery. https://doi.org/10.1145/3229434.3229464.

SAE. (2018). *Sae j3016 taxonomy and definitions for terms related to on-road motor vehicle automated driving systems*. SAE International.

Schartmüller, C., Riener, A., Wintersberger, P., & Frison, A.-K. (2018). Workaholistic: On balancing typing-and handover-performance in automated driving. In Proceedings of the 20th International Conference on Human-Computer Interaction With Mobile Devices and Services, Barcelona Spain, 3–6 September 2018 (p. 16).

Schartmüller, C., Weigl, K., Löcken, A., Wintersberger, P., & Riener, A. (2021) Displays for productive non-driving related tasks: Visual behavior and its impact in conditionally automated driving. 1–20.

Schartmüller, C., Weigl, K., Wintersberger, P., Riener, A., & Steinhauser, M. (2019). Text comprehension: Heads-up vs. auditory displays: Implications for a productive work environment in sae level 3 automated vehicles. In Proceedings of the 11th International Conference on Automotive User Interfaces and Interactive Vehicular Applications, Utrecht Netherlands, 21–25 September 2019 (pp. 342–354).

Schartmüller, C., Wintersberger, P., & Riener, A. (2017). Interactive demo: Rapid, live data supported prototyping with U.S.E. In Adjunct Proceedings of the 9th International Conference on Automotive User Interfaces and Interactive Vehicular Applications, Oldenburg Germany, 24–27 September 2017 (pp. 239–241). ACM. http://doi.org/10.1145/3131726.3132046.

Stähle GmbH. (2019). *Automated driving system sfphybrid for cars*. https://www.staehle-robots.com/english-1/products/sfphybrid-eng/.

Strauch, C., Mühl, K., Patro, K., Grabmaier, C., Reithinger, S., Baumann, M., & Huckauf, A. (2019). Real autonomous driving from a passenger's perspective: Two experimental investigations using gaze behaviour and trust ratings in field and simulator. *Transportation Research Part F: Traffic Psychology and Behaviour*, *66*, 15-28. https://doi.org/10.1016/j.trf.2019.08.013.

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).

Tobii, A. (2019). *Tobii Pro eye-tracking-Brille 2*. https://www.tobiipro.com/de/produkte/tobii-pro-glasses-2/.

Tsimhoni, O., Smith, D., & Green, P. (2004). Address entry while driving: Speech recognition versus a touch-screen keyboard.

Human Factors, *46*(4), 600–610. https://doi.org/10.1518/hfes.46.4.600.56813.

van der Heiden, R. M., Iqbal, S. T., & Janssen, C. P. (2017). Priming drivers before handover in semi-autonomous cars. In Proceedings of the 2017 chi Conference on Human Factors in Computing Systems, Denver Colorado USA, 6–11 May 2017 (pp. 392–404).

Victor, T. W., Tivesten, E., Gustavsson, P., Johansson, J., Sangberg, F., & Ljung Aust, M. (2018). Automation expectation mismatch: Incorrect prediction despite eyes on threat and hands on wheel. *Human Factors*, *60*(8), 1095–1116. https://doi.org/10.1177/0018720818788164.

Wiedemann, K., Naujoks, F., Wörle, J., Kenntner-Mabiala, R., Kaussner, Y., & Neukum, A. (2018). Effect of different alcohol levels on take-over performance in conditionally automated driving. *Accident Analysis & Prevention*, *115*, 89-97. https://doi.org/10.1016/j.aap.2018.03.001.

Wintersberger, P., Frison, A.-K., Riener, A., & von Sawitzky, T. (2018). Fostering user acceptance and trust in fully automated vehicles: Evaluating the potential of augmented reality. *Presence: Teleoperators and Virtual Environments*, *27*(1), 46–62. https://doi.org/10.1162/pres_a_00320.

Wintersberger, P., Green, P., & Riener, A. (2017). Am i driving or are you or are we both? A taxonomy for handover and handback in automated driving. In *9th International Driving Symposium on Human Factors in Driver Assessment, Training, and Vehicle Design, Equinox Manchester Village, Vermont, 26–29 June 2017*. University of Iowa.

Wintersberger, P., Riener, A., Schartmüller, C., Frison, A.-K., & Weigl, K. (2018). Let me finish before i take over: Towards attention-aware device integration in highly automated vehicles. In Proceedings of the 10th International Conference on Automotive User Interfaces and Interactive Vehicular Applications, Toronto ON Canada, 23–25 September 2018.

Wintersberger, P., Schartmüller, C., & Riener, A. (2019). Attentive user interfaces to improve multitasking and take-over performance in automated driving: The auto-net of things. *International Journal of Mobile Human Computer Interaction (IJMHCI)*, *11*(3), 40-58. https://doi.org/10.4018/IJMHCI.2019070103.

Zeeb, K., Buchner, A., & Schrauf, M. (2015). What determines the take-over time? An integrated model approach of driver take-over after automated driving. *Accident Analysis & Prevention*, *78*, 212-221. https://doi.org/10.1016/j.aap.2015.02.023.

Zeeb, K., Härtel, M., Buchner, A., & Schrauf, M. (2017). Why is steering not the same as braking? The impact of non-driving related tasks on lateral and longitudinal driver interventions during conditionally automated driving. *Transportation Research Part F: Traffic Psychology and Behaviour*, *50*, 65-79. https://doi.org/10.1016/j.trf.2017.07.008.

Zhang, B., de Winter, J., Varotto, S., Happee, R., & Martens, M. (2019). Determinants of take-over time from automated driving: A meta-analysis of 129 studies. *Transportation Research Part F: Traffic Psychology and Behaviour*, *64*, 285-307. https://doi.org/10.1016/j.trf.2019.04.020.

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.