SN Computer Science (2023) 4:56
https://doi.org/10.1007/s42979-022-01455-9

ORIGINAL RESEARCH

How Older Drivers Perceive Warning Alerts? Insights for the Design of Driver–Car Interaction

Luka Rukonić[1] · Marie-Anne Pungu Mwange[2] · Suzanne Kieffer[1]

Received: 2 October 2021 / Accepted: 11 October 2022
© The Author(s), under exclusive licence to Springer Nature Singapore Pte Ltd 2022

Abstract

The automotive industry is working toward driving automation and driver-assistance technology is becoming a norm in modern cars. Warning alert systems support the driver–car interaction and inform drivers about automation system status, upcoming obstacles, or dangers ahead. However, older drivers' needs are not always addressed in research studies, although they make up a large segment of drivers. Therefore, we conducted a qualitative three-round formative evaluation of a warning alert system using video prototypes in lab and remote settings. The goal was to evaluate visual-, sound-, and speech-based alerts based on: (a) their efficiency in informing drivers about the road situation ahead, and (b) participants' subjective opinions. We evaluated the system's efficiency using self-reported data measuring participants' cognitive load, usability, UX, and ease of use. Also, we conducted interviews to collect subjective feedback about proposed prototypes. In this article, we describe the design of warning alerts and report on their evaluation results. Our results show that speech-based warnings, especially when coupled with visual warnings, are efficient and accepted well by the participants. This article illustrates older drivers' attitude toward the use of different warning modalities in the driving context.

Keywords Warning alerts · Modalities · Older drivers · Video prototypes · Formative evaluation

Introduction

This article is an extended version of our previous work on the UX design and evaluation of warning alerts for semi-autonomous cars [48]. The primary contribution of this article is the qualitative analysis of older drivers' subjective responses from three experiments to a warning alert system using speech, visuals, and sound as output modalities. Also, driver-car voice interaction regarding driving automation capabilities was investigated. Specifically, this article builds on top of the previous version by providing additional:

- in-depth qualitative analysis
- subjective statements from the participants
- analysis of usability and cognitive workload data
- background data about the participants
- and improved discussion of the implications related to the warning modalities.

The article is organized as follows: The section "Context and motivation" provides the background concerning the issues older drivers face in semi- autonomous driving. The section "Design of warning alerts" describes how we designed our prototypes. The section "Methodology" describes the methodology and the data collection process. Subsequent sections "Experiment 1" , "Experiment 2", and "Experiment 3" present the experiments we conducted as part of this research. For each of the three experiments, we present and discuss their respective results and findings. Afterward, the section "Positioning against the related work" reports on the position of our research against related work. Next, the section "Discussion" provides a summary of the most important findings and discusses some methodological implications of the study. Finally, the section "Conclusion" concludes this article.

This article is part of the topical collection "Computer Vision, Imaging and Computer Graphics Theory and Applications" guest edited by Jose Braz, A. Augusto Sousa, Alexis Paljic, Christophe Hurter, and Giovanni Maria Farinella.

✉ Luka Rukonić
luka.rukonic@uclouvain.be

[1] Institute for Language and Communication, Université catholique de Louvain, Louvain-la-Neuve, Belgium

[2] AISIN Europe, 1420 Braine-L'Alleud, Belgium

Published online: 16 November 2022

Context and Motivation

Drivers face many challenges when using new in-car technologies, as the automotive industry is making progress toward autonomous vehicles (AVs). Various advanced driver-assistance systems (ADAS) are already available in new cars and promise to increase the safety, comfort and pleasure while driving [21]. Previous research on driving automation discussed how to enhance the driver–car interaction and provide support to adapt to the different levels of autonomy. For example, dialogue-based interfaces were tested to increase the situation awareness of the driver while driving a semi-autonomous car [42]. Additionally, Kasuga et al. [24] designed a system to guide drivers safely to take back the control of the car using a multimodal human-machine interface (HMI) including speech, auditory alarms, and ambiance lightning. Research efforts were made to identify the information necessary to design a transparent automated driving system [10]. Also, Stromberg et al. [54] proposed a framework focusing on the driver–+car communication related to the vehicle's operation.

According to the human-centered design (HCD) specifications, including end-users in the design process is key to develop useful and usable design solutions [23]. However, in a recent literature review of the automotive HMI design guidelines applied in industry, Young et al. [59] found that the physical, mental, and sensory abilities of older drivers are seldom considered or do not provide any references on how to design for them. Also, the research on specific HMI design solutions for older drivers is scarce. We are aware of two publications discussing the concrete design solutions for older drivers with impairments. Caird et al. [6] and Fernandes [13] have presented some design guidelines and functional limitations to address HMI design for older drivers, considering their physical and mental impairments. Study by Fernandes et al. [13] presented design improvement opportunities for seats, mirrors, driver assistance, and on-board devices. Nevertheless, older drivers' opinion or attitude toward the HMI design addressing their needs are, to the best of our knowledge, rarely analyzed in the literature. This is also confirmed by Agudelo et al. [1] in their literature review on the use of value-sensitive design (VSD). They found that most studies applying VSD did not use it to improve the interface design of automotive infotainment interfaces. Also, there are no standard guidelines for the design of infotainment interfaces, which makes their safety of use questionable.

Yet, seniors represent a growing market for AVs. First, while people above 65 accounted for almost one-fifth of the European population in 2017 [12], they might account for one-sixth of the total population by 2050 [56]. Second, older adults are much more favorable toward future autonomous cars compared to younger drivers [47]. Third, older adults have a large interest for AVs despite some concerns about security issues, system failures or hacking attacks [51]. Therefore, it is vital to ensure that the design of automotive HMI caters to their specific needs and provides them with a sufficient amount of trust to adopt the technology. Additionally, older people have not grown up with digital technologies; thus, their process of adopting it significantly differs from the adoption process among younger people. Bolanos et al. [5] suggest that the cognitive status and physical condition of older adults are of utmost importance when developing technological solutions for them. The authors also suggest to use the TAMUX user acceptance model, because it offers high flexibility to adapt to the context, characteristics, and the user group for which the solution is being developed. Previous research showed that older drivers' acceptance of partial (i.e., facilitated by ADAS) or full driving automation depends on their perceived usefulness of automation, followed by perceived safety [36]. Older drivers' acceptance of full-AVs is conditional to demonstration of their reliability, and increased further if the full-AV follows the driver's driving style [17].

Lane-keeping assistant (LKA) and adaptive cruise control (ACC) are driver-assistance systems that support the driver by controlling the lateral (steering) and longitudinal (acceleration/braking) movement of the car, respectively. The Society of Automotive Engineering (SAE) defines six levels of driving automation. At level 0, the driving support features are limited to providing warnings. At level 1, the driving support features include LKA or ACC. At level 2, both steering and braking support are activated together with the LKA. While in use, those systems still require the driver to supervise the roadway and hold the steering wheel in case of a takeover request. At level 3, the car can operate under various conditions and make complex decisions, but requires the driver to always be aware of the road situation and be ready to drive when requested. Whenever the vehicle cannot cope with the road situation, it issues a takeover request and the driver needs to take back the control [49]. Levels 4 and 5 comprise the vehicles which do not have controls installed and can drive in restricted or unrestricted conditions with no driver's engagement. In sum, the first three levels (L0, L1, L2) are driver support systems also referred to as semi-AVs, while the last three levels (L3, L4, L5) are automated driving systems or full-AVs. In near future, semi-AVs will be dominant in the market. At present, mandatory fitting of driving safety features such as emergency braking and lake departure warning is already required by law. Also, from 2022, all new vehicles will need to be fitted with driver drowsiness warning, intelligent speed assistant, and driver distraction warning [11]. These warnings need to be attentively designed so

as not to increase driver's cognitive load while maintaining positive user experience (UX).

Warning alerts are intended not only to warn the driver about a danger (e.g., obstacle on the road), but also to communicate the automation system's decision (e.g., changing lane). Currently, warning alert systems rely mainly on visual cues, such as icons displayed in the instrument cluster on the dashboard. Such visual cues may not be appropriate for older users: they may be too small or they may be overlooked by them [13, 54]. To address this issue, it is advised that car interior designers adjust the typography, lightning, and size of information signs and control devices [13]. Also, older drivers pay more attention to the road than younger drivers when they are engaged in secondary tasks even in semi-autonomous driving mode [19]. Consequently, older drivers might miss the visual warnings. Similarly, haptic cues may not be suitable either, as older drivers have difficulties detecting tactile stimuli [19] and haptics should be combined with another modality such as speech [40].

Therefore, speech and auditory modalities generally seem to be a promising solution to address older drivers' decline in visual and cognitive capabilities caused by aging. For example, Porter et al. [43] found that auditory alerts resulted in faster braking response times for older drivers than for young drivers. Past research suggests that speech alerts result in better memory of the events ahead [37]. Also, as driving is a visually demanding task, using the auditory channels to convey danger-related warnings is more adequate [13]. Moreover, vocal message including spatial indications can help visual target spotting, provided that the visual search task is not too complex or too easy [9]. Finally, another argument to further investigate the use of speech-based alerts is that we are involved in a research project focusing on the use and design of voice-based systems for semi-autonomous vehicles.

This paper explores the efficiency of visual- and/or sound-based warning alerts to support older drivers' awareness of both the decisions made by the automation system and the road situation ahead in automation levels 2 and 3. Within a test-and-refine qualitative approach, we carried out a three-round experiment (XP1, XP2, and XP3) comparing several modalities of warning alerts (e.g., beep versus speech, speech versus visual cues, etc.). For each experiment, we created a low-fidelity video-based prototype corresponding to each combination of modalities, and tested each prototype with six participants in a lab or remote setting. We collected and analyzed data about UX, cognitive load, usability, and the participants' subjective responses to the proposed prototypes. The test-and-refine approach allowed us to set up user tests rapidly and to improve the warning alerts between each round, based on the findings of the previous round.

Zhou et al. [60] identified four topics pertinent to the takeover situations in autonomous driving:

1. Driver's awareness of whether the vehicle can continue operating safely in given conditions;
2. The system's capability of warning drivers for any dangers or conditions requiring the driver to take over;
3. Automation capability awareness;
4. Warning effectiveness.

This paper focuses on topics 3 and 4 and reports on how we designed and evaluated a warning alert system for semi-AVs. The goal was to investigate how older drivers of L2 or L3 cars can be informed of upcoming dangers while driving.

In the next section, we explain how we approached the design of warning alerts.

Design of Warning Alerts

The following is an explanation of how we designed the warning alerts and created the video simulations we used in the UX evaluation with participants.

First, we searched through open source databases to find real-world driving video recordings. We sourced the videos from the DR(eye)ve project repository [38]. We searched for videos showing a road situation involving an automatic lane change and vehicle avoidance maneuver. The videos were silent; there was no road or engine noise recorded.

Then, we designed the content of warnings to accommodate for the road situation in the videos. As we stated previously, literature suggests using either visual or audible warnings [16]. For this study, we compared variations and combinations of those alerts. Each warning alert involved three levels of urgency, similar to [41]: low-urgency (LU) at the beginning of the scenario, medium-urgency (MU) before reaching the obstacle, and high-urgency (HU) immediately before reaching the obstacle. For each event in the video prototype, a low–medium–high sequence of urgency was followed. The level of urgency was defined as a function of distance between the car and the obstacle or a dangerous event. We used the words *"Danger"*, *"Warning"*, and *"Notice"* to convey this notion in both voice and visual alerts, as [2] reported that the perceived urgency of the word *"Danger"* was higher than the words *"Warning"* and *"Notice"*. Regarding sound, previous research suggests that using more annoying sounds results in faster reaction times in handover and takeover situations [28]. These studies also recommend keeping the duration of the sounds short, not to delay the driver's reaction. Thus, for each urgency level of each event, the warning alert sound was played only once, without the possibility of repeating it.

The speech alert messages were digitally produced using a Text-To-Speech (TTS) system. The aim was to create voice alerts similar to those of an infotainment system, thus giving a familiar context to older drivers. We used a warm-toned,

mature male voice, speaking with a standard British accent. Voice synthesis experts provided us with advice to write the content of the messages using informal language style. The prototype also involved auditory alerts using an arbitrary chime sounds we selected and downloaded from http://freesound.org available under Creative Commons license, free for anyone to use, modify, and distribute. We used two different sounds in our experiments. The first selected sound was an alarm-like sound that had an average frequency of 1100 Hz, which is within the average middle-aged hearing frequency range of 20 Hz to 15 kHz [34]. The duration of the sound was 0.31 s, consisting of a double beep, each lasting around 0.12 s. The second selected sound was a soft, single-chime sound, lasting for 4.1 s, but the main chime part lasting for 1 s, with an average frequency of 226 Hz.

Among visual icons most commonly used in cars, we decided to use pictographs drivers are familiar with to design the visual warning components. The visual warnings consisted of textual messages along with the commonly used red triangle, which is known to be the most recognized sign by drivers [33]. For road event such as lane changes, a descriptive image of the upcoming car operation was added (Figs. 4d and 5c). Visual icons were coupled with a short text to make them more explicit. The visuals were designed to be simple, in order to be displayed in a large size and make them noticeable by older adults. In our warning design approach, we paid attention to the driver's situational awareness. Therefore, the sequence of warnings should prepare the driver well enough and in advance for the upcoming danger. Thus, assuming that drivers would already be aware of the road situation when HU warning messages are given, their role was to provide a last alert or notice about the car's following action or about the upcoming obstacle. In fact, that is why, the triangle was omitted in HU (see Fig. 4), because drivers would not have enough time to see it and process the visual warning.

In the last experiment (XP3), we only used LU and HU levels of urgency, which we clarify more in detail in "Experiment 3".

In addition, we wanted to identify which of the following seven types of alerts would provide users with the best possible warning: (C1) speech only (VB), (C2) sound only (S), (C3) visual-only (V), (C4) speech + sound (VB + S), (C5) speech + visual (VB + V), (C6) visual + sound (V + S), and (C7) speech + sound + visual (VB + S + V). We used C1; C3–C7 in XP1 and XP2, and C1–C6 in XP3. We eliminated C2 from XP 1 and XP2 because we assumed that the sole use of beeping sounds with no additional information regarding the upcoming danger would not be useful to drivers. We reintroduced it in XP3, because we assumed this might trigger the interaction between the car and the driver. We removed C7 from XP3 as this condition was judged too complex by participants from XP1 and XP2.

Methodology

We performed three experiments (designated in the following as XP1, XP2 and XP3) within a formative UX design approach to (1) investigate elderly drivers' subjective responses about the proposed danger alert system, (2) redesign the warning alerts based on the findings of the previous iteration, and (3) explore the application of voice interaction in the context of warning effectiveness in semi-autonomous cars. The formative approach allowed us not only to receive early feedback to be incorporated into the prototype of the future system, but also to identify target users' needs. Furthermore, we relied on low-budget prototyping methods, namely video prototyping and Wizard of Oz to quickly iterate on our design solutions in an industrial agile development setting. Adopting these UX methods gave us a high level of flexibility in adapting to changes of the experimental setting from in-lab to remote testing.

Data Collection Methods

In this section, we describe the data collection methods we used. We adopted a mixed-method approach in this qualitative study, using interviews and questionnaires.

Interviews

Semi-structured interviews allowed us to capture participants' point of view on the warning alerts and understand the questionnaire responses better. The experimenter asked questions, such as *"What happened in the video?"*, *"What is your understanding of the warning alerts?"*, *"What were you focused on during the drive?"* or similar follow-up questions to better understand the participant's reasoning. The goals of the interview was to check participants' understanding of the warning alerts and the road situation. This way we were able to check whether participants' understanding matched the intended meaning of warning alerts. Additionally, we asked them to explain their preferences for the design of warnings and describe their feelings related to the warning alerts and the road situation.

UEQ

We used the user experience questionnaire (UEQ), a standard instrument for evaluating UX constructed and validated by Laugwitz et al.[29], that measures the perceived UX across six scales: Attractiveness (AT), Perspicuity (PS), Dependability (DP), Efficiency (EF), Novelty (NV), and Stimulation (ST). PS, DP, and EF measure the pragmatic attributes of UX, ST, and NV the hedonic attributes of UX.

Attractiveness is considered separately. UEQ helps designers determine which experiential qualities need to be targeted to reach the highest impact on the product's UX [52]. A benchmark for UEQ was developed that allows designers to interpret whether a new product offers sufficient UX [53].

NASA-TLX (TLX)

The TLX is a post-task six-dimensional scale designed to assess the subjective workload of the participants while performing a task. It is widely used in the research community, due to its easy administration and a relatively wide range of application domains, such as aviation, military, automotive industry, and healthcare [18]. Most studies report its use in relation to user interface design and evaluation. This also makes it suitable for our study. NASA-TLX comprises two parts. In the first part, participants need to identify the sources of workload to obtain the weights for each of the six subscales. In the second part, they need to rate the workload of the task on each of the six scales by giving it a score between 0 and 100.

System Usability Scale (SUS)

The SUS is a widely used post-test questionnaire for subjective usability assessment of a product or a system. It yields a single SUS score as an output. However, Lewis and Sauro [31] have found that the score can be decomposed to extract measures for Usability and Learnability components. Therefore, we calculated the Learnability and Usability scores using their method. Usability is a software quality related to how easy it is to use a system. ISO-9241 [22] standard specifies effectiveness, efficiency, and satisfaction as usability measures. Effectiveness is about whether users can achieve their goals. Efficiency represents the resources used (e.g., time or effort) to complete the task and the quality of produced outputs. Satisfaction is the attitude of users toward the system. In our context, the participants' task was to, while watching the video, monitor the road situation and interpret the warning alerts that appeared. Although the task was broadly defined, the focus was solely on the evaluation of the warning alerts. Efficiency is represented by the cognitive load measured using NASA-TLX. Satisfaction is represented through the subjective feedback collected using interviews.

Single Ease Questionnaire (SEQ)

In XP3, we added an SEQ questionnaire to measure the perceived (subjective) ease of use of the warning system and the related voicebot. It is a one-question, post-task questionnaire that uses a 7-point Likert-type scale to evaluate how easy or difficult a task was for the participant [50].

Participants: User Profile

All participants were recruited through an external recruiting agency. We specified a user profile which was sent to the agency to recruit the participants for the study. The user profile included the following requirements: active senior drivers older than 50, male or female, who hold a driver's license, and have more than 20 years of driving experience. In addition, candidates had to be frequent users of ADAS (LKA, AAC) and GPS systems. Finally, they were all interested in AVs and driving assistance technologies. There were no conditions set regarding participants' vision, hearing, or physical ability.

Procedure

Figure 1 shows the experimental procedure. First, the researcher explained to participants the study procedure without revealing the underlying research questions. After signing a consent form, participants answered a few demographic questions and filled out a questionnaire regarding their attitude toward AVs [3]. XP1 took place in a lab room where the participants sat in front of a large-screen TV, next to which a smaller 14-in. laptop was placed, simulating the car's infotainment screen and displaying the visual warning messages. Speech messages and beep sounds were played through a set of stereo speakers placed behind the TV. In experiments 2 and 3, we conducted the tests remotely due to the coronavirus outbreak. In the remote evaluation setting, participants and researchers were at different physical

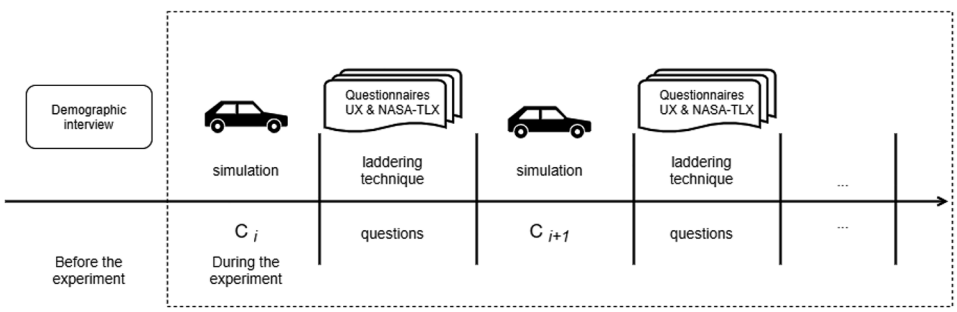


Fig. 1 The experimental procedure

locations, the latter moderating the session using an online tool for remote user research. We video-recorded each session.

We instructed participants to imagine that they were driving a semi-AV with a limited ability to deal with on-road situations and that the car might request the driver to take over. However, no driver response was ever required in our study. Thus, their task was to pay attention to and understand the road situation and warning alerts, similar to what they would have done in an SAE level 3 car. The experiment began after explaining the scenario to the participants. After each condition, the researcher would conduct a brief interview with the participants. Afterward, the researcher administered the NASA-TLX. In XP3, the SEQ rating sheet preceded the NASA-TLX. At the end of the session, participants completed the UEQ.

We modeled the first two experiments similarly to [14], in which authors used the laddering technique defined by [44] to probe participants to discover the underlying psychological needs while watching the videos showing an automated vehicle driving on the road. We used the laddering technique to ask questions about their understanding of warnings, their opinion about them, and the further clarification of ratings in the standard questionnaires administered to them. The quantitative data collected through UEQ, SEQ, and TLX helped us explain the qualitative findings from the interviews and explore the thoughts and attitudes of participants as a part of the iterative UX design process.

Experiment 1

The same video-based driving simulation was used in all experimental conditions. We reused a part of the video no. 41 from the public dataset containing 74 videos made available by the DR(eye)VE project [38]. The video lasted 80 s and it comprised a highway driving scenario, in which the car warned the driver about an obstacle ahead, overtook a long truck in the right lane by going to the left lane, and subsequently returned to the right lane. The scenario simulated a car equipped with ADAS functions, such as LKA, AAC, and an automatic lane change. Such simulation corresponded to SAE L3 automation. The goal of XP1 was to collect initial feedback from participants regarding their understanding of the warning alerts. Figure 2 shows the lab setup. Six conditions (C1, C3–C7) were presented to each participant, resulting in 36 trials overall.

Participants

Six participants (1 woman) aged between 64 and 75 ($M = 68.5$, SD = 3.8) participated in XP1. They were all in good health and did not have any type of physical disabilities. All participants held a driver's license.

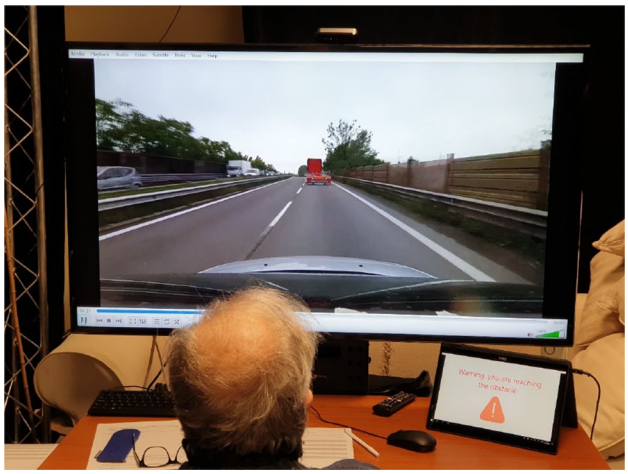


Fig. 2 The lab setup used in experiment 1

Warnings

The content of the voice-based (VB) warning alerts is shown in Table 1. Figure 4 shows the visual warnings. In XP1, the visual warnings (V) were displayed on a separate screen, on the right-hand side of the participant, as shown in Fig. 2.

Pre-study Questionnaire Attitude Toward AVs

Five out of six participants who took part in XP1 and XP2 filled out a pre-study questionnaire about their attitude toward AVs. The questionnaire contained four questions, which were rated on a 5-point scale from 1 (very negative) to 5 (very positive). Figure 3 shows their responses. Overall, the participants expressed a positive attitude toward AVs. Although they all stated that they would be very excited to drive an AV, there is still a moderately strong feeling of fear toward trying it for the first time.

Quantitative Results

Figure 6 shows the mean UEQ scores per dimension. According to the UEQ benchmark [53], obtained scale values can be categorized into five categories: excellent, good, above average, below average, and bad. The prototype in XP1 scored lowest on the pragmatic qualities DP (1.08—below average), and EF (1.88—excellent). These low scores can be attributed to a lack of direct interaction with the prototype. The prototype scored well on AT (2.08—excellent), PS (2.17—excellent), ST (2.17—excellent), and NV (1.92—excellent). It is important to note that although the values we obtained are high, our sample is probably too small to

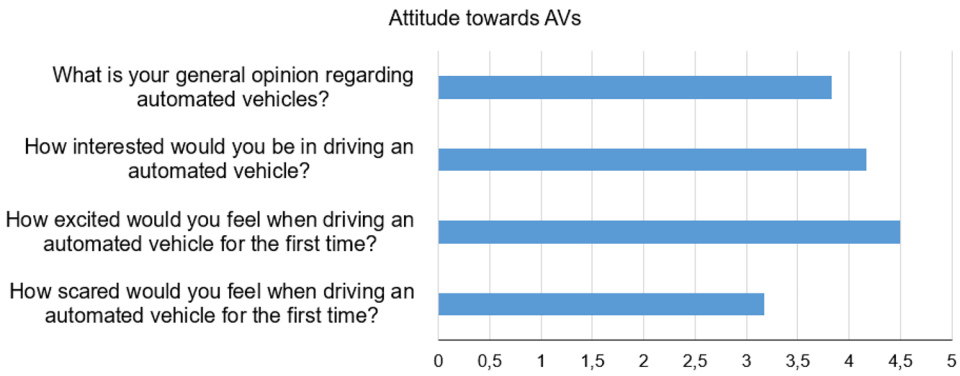
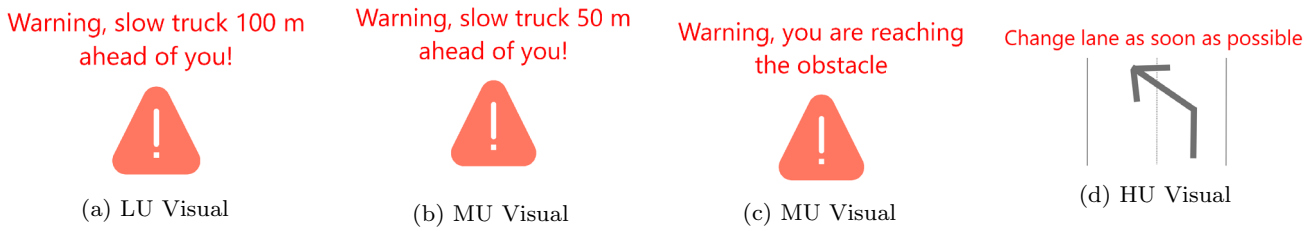


Fig. 3 Self-reported attitude toward AVs

Table 1 Design of warnings for experiment 1

Condition	Low urgency (LU)	Medium urgency (MU)	High urgency (HU)
C1 (VB)	Voice message: "Hey, there is a slow truck on the road in 100 m. Pay attention!"	Voice messages: "Be vigilant, a slow truck in the right lane in 50 m."; "Warning! You are approaching an obstacle!"	Voice message: "Slow down and change the lane as soon as possible."
C3 (V)	Fig. 4a	Fig. 4b, c	Fig. 4d
C4 (VB+S)	Voice message as in C1 + beep sound	Voice message as in C1 + beep sound	Voice message as in C1 + beep sound
C5 (VB+V)	Voice message as in C1 + visual warnings as in C3	Voice message as in C1 + visual warnings as in C3	Voice message as in C1 + visual warnings as in C3
C6 (V+S)	Visual warnings as in C3 + beep sound	Visual warnings as in C3 + beep sound	Visual warnings as in C3 + beep sound
C7 (VB+V+S)	Voice message as in C1 + Visual warnings as in C3 + beep sound	Voice message as in C1 + Visual warnings as in C3 + beep sound	Voice message as in C1 + Visual warnings as in C3 + beep sound



(a) LU Visual (b) MU Visual (c) MU Visual (d) HU Visual

Fig. 4 Visual warnings used in XP1

achieve meaningful comparisons with the benchmark. Thus, we treat these values as descriptors of UX rather than strict measurements. A high PS score hints that the warnings might be easy to get familiar with, while high indicates participants' positive impression of the warning system. Overall, the prototype scored 1.71 on pragmatic qualities, and 2.04 on hedonic qualities.

Qualitative Results

We performed an exploratory thematic analysis of the data collected during interviews (Table 2). Generally, participants reported that the voice messages were too friendly and too long. As a reference to other systems, participants P2 and P5 said that using polite words such as *please* or having a friendly tone is not something they liked. For the others, (overall) timely and coherent warning about upcoming dangers should suffice. This finding is in line with previous research on autonomous car voice alerts, where a higher level of assertiveness results in faster reaction times and conveys a higher level of perceived urgency [58]. Thus, we decided to shorten the voice alerts and make them more assertive.

Regarding the speech-based alerts, P2,P3, and P6 expressed their dissatisfaction with the length of the speech alerts. P2 said that *"... it is talking too much, imagine hearing this all the time... I can't listen to my music*

Table 2 Thematic analysis of the participants' feedback per condition in experiment 1

Condition	Theme	Codes	*N*	Participants
C1	Attention	Voice warnings allow me to focus on the road	1	P4
	Attractiveness	Voice is too friendly	1	P5
	Content	Truck should not be described as obstacle	3	P2
	Cognitive Load	Voice messages are too long	3	P6
	Cognitive Load	System talks too much	2	P2
	Usefulness	Voice warnings are useful	2	P4
C3	Attention	Beep draws attention	2	P5
	Attention	Visuals can easily be missed	2	P2, P4
	Attention	Visual warnings are distracting	2	P4, P6
	Usefulness	Visual warnings are not useful	3	P6
C4	Cognitive Load	Too much information	1	P5
	Cognitive Load	System talks too much	2	P2
	Stress	Beep is annoying	3	P2, P4
	Stress	Voice messages become annoying on long-term	2	P3
	Stress	Beep is stressing	1	P4
C5	Attention	Simple to follow with voice and visuals	1	P3
	Attractiveness	Voice is too friendly	2	P2, P5
	Cognitive Load	Reading and listening at the same time is difficult	4	P2, P4, P5
	Cognitive Load	Listening is easier than reading	1	P2
	Usefulness	Voice warnings are useful	2	P4
C6	Attention	Visuals can easily be missed	1	P3
	Attractiveness	Voice would interrupt music	1	P2
	Cognitive Load	Reading is difficult	1	P2
	Stress	Beep is annoying	1	P2
	Stress	Voice messages become annoying on long-term	2	P3
	Usefulness	Beep sound is confusing	1	P3
C7	Attention	Beep draws attention	2	P4, P6
	Content	Truck is not an obstacle	3	P2
	Cognitive Load	Too much information	3	P2, P3, P5
	Cognitive Load	Voice messages are too long	3	P6

anymore". P6 said *"I would just like to adjust the length of sentences, they should be shorter"*.

Participants often mentioned that visual warnings are distracting and that reading the textual warning messages was not easy. P1 said about C3: *"I am thinking what happens if I am talking to the passengers and I don't see the sign"*, while P2 commented about C6: *"I have to go and read to understand what is going on"*. When using voice and visual warnings (e.g., C5, C7) together, the text in visual warnings should correspond to the spoken content of the voice messages. That reduces the driver's workload while comparing what is being said and what is being shown on the screen. The beep sound is considered useful for preparatory purposes and drawing the driver's attention before hearing or seeing the actual warning. However, the visual or speech warning alert should follow the beep without a delay. Additionally, the beep sound was not deemed as very pleasant. P2 was not happy and recounted *"That beep sound, makes me nervous."*. Participants consistently reported that the warnings should give precise information about the upcoming event. In our case, they pointed out a confusing situation caused by using a non-specific vocabulary. Specifically, instead of using the word "truck", the warning alert mentioned an "obstacle" ahead, which made the drivers think about static objects on the road that impede the traffic flow. Participants reported that this increased confusion and stress. P2 had an interesting comment about that and said that *"The problem is, if you hear truck, and you read obstacle, it is not good for your brain. You start panicking that you have missed it."*. Finally, there should be a difference between warning messages and simple informative messages.

Experiment 2

XP2 aimed to fix some issues in the design of warnings from XP1, confirm the findings from XP1 and collect participants' feedback once again. We built a new prototype based on a

Table 3 Design of warnings for experiment 2

Condition	Low Urgency (LU)	Medium Urgency (MU)	High Urgency (HU)
C1 (VB)	Voice message: "Notice! There are roadworks in the left lane in 200m. Pay attention!"	Voice messages: "Warning! I am going to the middle lane now."; "Notice! Pay attention to the truck in the right lane."	Voice message: "Warning! You are approaching the roadworks in the left-most lane. Be careful!"
C3 (V)	Fig. 5a, b	Fig. 5c, d	Fig. 5e
C4 (VB+S)	Voice message as in C1 + beep sound	Voice message as in C1 + beep sound	Voice message as in C1 + beep sound
C5 (VB+V)	Voice message as in C1 + visual warnings as in C3	Voice message as in C1 + visual warnings as in C3	Voice message as in C1 + visual warnings as in C3
C6 (V+S)	Visual warnings as in C3 + beep sound	Visual warnings as in C3 + beep sound	Visual warnings as in C3 + beep sound
C7 (VB+V+S)	Voice message as in C1 + Visual warnings as in C3 + beep sound	Voice message as in C1 + Visual warnings as in C3 + beep sound	Voice message as in C1 + Visual warnings as in C3 + beep sound

(a) LU Visual (b) MU Visual (c) MU Visual (d) HU Visual (e) HU Visual

Fig. 5 Visual warnings used in XP2

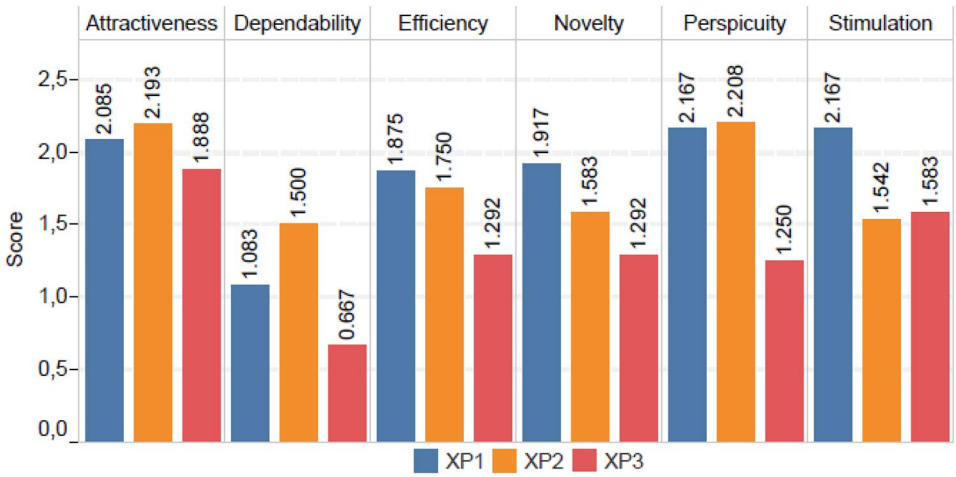


Fig. 6 Mean UEQ scores between XP1, XP2, and XP3

video that we recorded ourselves on the highway. Using a front-mounted camera in the car, we recorded several situations while driving on the highway. We selected the road situation displaying the roadworks in the left-most lane of a three-lane highway. Lane change was required. The video lasted 80 s. We used the same six conditions (C1, C3–C7), resulting in 36 trials overall. We added the TLX questionnaire: after the first condition was presented, participants identified the sources of workload; after each condition, participants filled out a TLX rating sheet. This is the strength of the formative evaluation, which allows UX practitioners to be flexible in terms of choice of evaluation methods they use. Since we had the same participants as in XP1, we could compensate for not measuring their cognitive load in the XP1. Finally, XP2 was conducted remotely using Lookback (https://www.lookback.com/), an online user research tool, instead of in a lab.

Participants

Six participants (2 women) aged between 52 and 75 ($M = 66.5$, SD = 7.2) participated in XP2. Five had participated to XP1; one had not. We recruited the same participants for two reasons. First, we worked in a formative approach where we frequently design and evaluate system prototypes with a small sample of users. Second, for

organizational purposes, we did not have to redo the recruitment process again, which included finding new participants and interviewing them.

Improved Warnings

We synchronized visual warnings with other modalities and integrated them into the videos, instead of showing them on a separate screen, thus simulating a Head-Up Display (HUD). The new placement of visual warnings increased their efficiency as the drivers could see them more easily, compared to when they were displayed aside (e.g., on a car's infotainment screen). Based on the findings from XP1, we reduced the amount of text included in visual warnings. In addition, we rewrote the content of the voice alerts to be more informative and direct, but appear less friendly. We also selected another beep sound, which had a duration of 2.797 s. Finally, the TTS voice remained the same as in XP1. Table 3 shows the design of warning alerts in XP2.

Quantitative Results

In this section we present and discuss the results obtained from analyzing the self-reported data collected using UEQ, SUS, and TLX questionnaires.

UEQ

Compared to XP1, the scores for AT (2.19—excellent), PS (2.20—excellent), and DP (1.50—good) increased. However, the scores for EF (1.75—good), ST (1.54—good), and NV (1.58—excellent) decreased, which is also shown in Fig. 6. The DP score was the lowest in both experiments, 1.08 and 1.50, respectively. The EF score is second lowest in both experiments. This could be attributed to the properties of the experimental design. Participants were only watching the video and observing the situation. Therefore, the results reflect the lack of control and interaction between the car and the driver, as well as imply that participants needed to make a lot of effort to understand the warnings.

Usability and Learnability

We calculated the overall SUS score and extracted the Learnability and Usability scores for each condition. Figure 7 shows that there was an increase of SUS scores for all three components, which might imply that the design improvements were successful, especially with respect to Learnability. In general, the Learnability scores in XP1 were significantly low across conditions ($M = 48.7$, SD = 6.6), while they increased dramatically in XP2 ($M = 71.9$, SD = 3.7). However, this might also be caused by the participants' previous experience with the prototype in XP1, thus making the familiarity act as a confounding variable, because the participants could adapt to the updated version faster. Moreover, the mean overall SUS scores for XP1 ($M = 64.8$, SD = 4.6) and XP2 ($M = 71.9$, SD = 2.7) across all conditions show a slight increase in usability in the XP2. According to the SUS benchmark [32], the reference SUS score for an acceptable level of usability is 68. This was achieved in XP2.

In XP1, conditions C7 and C4 scored highest on the overall SUS score, as well as on the Usability component, while scoring second and third on the Learnability component. However, participants often commented negatively about C7 as they found it too complex: P5 said *"I think it was too much this time. Either the screen or the voice, but not both. I think it should be possible to choose which modalities you*

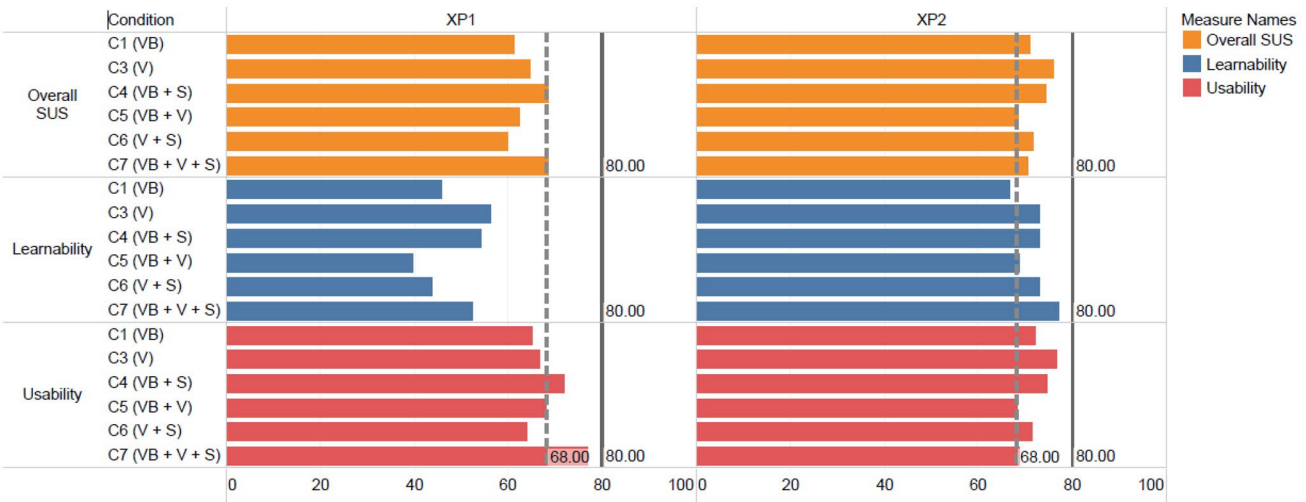


Fig. 7 Comparison of SUS scores between XP1 and XP2

want to use"; P3 commented *"We had the beep, visuals and the voice. That is too much. I prefer not to have the voice"*. On a positive note, P4 said *"These warnings are really efficient. It warns you quite a long time ahead. The sound tells you that there is something going on"*. The combination C6 (i.e., visual+auditory) scored lowest on usability.

In XP2, the unimodal visual-only warning (C3) scored highest on the SUS overall score in XP2. A plausible explanation could be that the visual warnings were easy to understand, did not require drivers to invest time and effort in learning how to use them, and many other systems, such as navigation systems, already use them. However, the condition C3 received negative feedback. Participants said that reading was too distracting, the visuals were not sufficiently descriptive, and they were poorly positioned on the screen, and easy to miss while driving. Nevertheless, visuals seem to be more preferred when combined with sound or speech modalities.

Comparison of Unimodal and Multimodal Warnings

We compared two groups of conditions: unimodal (C1,C3) and multimodal (C4–C7). Descriptive statistics showed no significant differences between these groups in XP2: neither regarding SUS score (73.33 vs 71.14), nor regarding TLX score (35.28 vs 39.16). The qualitative analysis provides a better comparison between unimodal and multimodal conditions. Attention, Attractiveness, Cognitive Load, and Preference were the most frequent themes related to unimodal conditions. On the other hand, the most frequent themes for multimodal conditions were Attention, Attractiveness, Cognitive Load, and Stress. The themes were similar, but Cognitive Load was assigned to multimodal conditions nine times, and only four times to unimodal conditions in XP2. In addition, Stress was always mentioned in relation to multimodal conditions, occurring only once in XP2 for a unimodal, speech-only condition C1. One possible implication of this finding is that the number of warning modalities must be carefully determined depending on the context and situational awareness of the driver. Inappropriate choice or number of modalities may increase driver's cognitive load and thus reduce their driving capabilities in potentially demanding situations.

TLX Cognitive Load

We calculated the means of the unweighted TLX scores for each subscale and the overall TLX score (Table 7). Mental demand and temporal demand were the dominant sources of workload in XP2. Conditions containing visual warnings had a higher mean TLX score (34.6), compared to the conditions without visual warnings (29.1). When grouped, unimodal combinations (C1,C3) and multimodal combinations (C4–C7) had a mean score of 31.3 and 33.5, respectively.

We further checked for the correlation between usability and cognitive load. Table 4 shows Pearson's correlation coefficients for XP2. When calculating the overall TLX score, we used the weighting scheme computation, whereas when comparing the individual TLX scales with the SUS components, we used the raw TLX ratings from the rating sheet. This is a common practice among researchers to simplify the procedure of data collection and analysis [18]. The results show a strong negative correlation between the two variables. For all participants, TLX and SUS across all six conditions were correlated, $r(34) = .59, p < .001$. Additionally, between conditions, significant correlations between usability and cognitive load were found only for the condition C1,

Table 4 Pearson's correlation coefficients per condition between SUS and TLX scores in XP2

SUS & TLX	.r	TLX Scale	.r C1	.r C3	.r C4	.r C5	.r C6	.r C7
C1	**−0.911**	MD & SUS	**−0.837**	−0.595	−0.446	−0.569	**−0.774**	−0.017
C3	−0.322	PD & SUS	**−0.717**	−0.497	**−0.769**	−0.308	**−0.555**	−0.106
C4	**−0.533**	TD & SUS	**−0.706**	**−0.619**	−0.401	−0.35	−0.444	0.434
C5	**−0.764**	PF & SUS	**−0.598**	0.413	−0.391	**−0.732**	−0.518	**−0.691**
C6	**−0.758**	EF & SUS	**−0.932**	0.105	**−0.845**	**−0.941**	**−0.834**	−0.61
C7	−0.263	FR & SUS	**−0.908**	**−0.919**	**−0.833**	**−0.685**	**−0.987**	**−0.734**
All conditions	**−0.591**	LEAR & MD	−0.327	−0.258	−0.337	−0.426	−0.366	−0.263
		LEAR & FR	**−0.695**	**−0.765**	−0.172	0.075	**−0.558**	0.009
		LEAR & EF	**−0.711**	0.292	−0.304	**−0.531**	**−0.531**	−0.094
		USAB & MD	**−0.921**	**−0.646**	−0.407	**−0.506**	**−0.795**	0.077
		USAB & FR	**−0.898**	**−0.85**	**−0.906**	**−0.829**	**−0.978**	**−0.801**
		USAB & EF	**−0.922**	0.014	**−0.877**	**−0.901**	**−0.804**	**−0.629**

Values in bold represent strong correlations (.r > ± 0.5)

(*MD* mental demand, *PD* physical demand, *TD* temporal demand, *EF* effort, *FR* frustration, *LEAR* learnability, *USAB* usability)

$r(4) = .91, p < .05$. This suggests that low usability and/or learnability of speech warning alerts might increase the perceived cognitive load. To further investigate the relationship between usability and cognitive load, we selected three TLX subscales (i.e. mental demand, frustration, and effort) to check for correlations with all three SUS scores. Figure 8 graphically represents the correlations between scores from selected subscales in SUS and TLX. With few exceptions for the visual-only condition (C3), the TLX and SUS scores seem to be overall negatively correlated. Interestingly, in the condition C3, Learnability and the overall SUS score were positively correlated to mental effort (EF) (Fig. 8). A likely explanation is that the unimodal visual warnings required more cognitive resources and effort to be processed and understood. The TLX EF scale measures the amount of perceived mental and physical work needed to accomplish the indicated level of performance. Indeed, the TLX P score for C3 in XP2 was highest across all conditions (39.2). Qualitative data reflect also this, as participants said that visuals were distracting and easy to miss.

Qualitative Findings

Table 5 summarizes the thematic analysis of participants' subjective feedback. Most of them found the beep warning useful and pleasant and agreed that it drew their attention and worked well in combination with other modalities. Similarly, participants agreed that voice warnings were useful and concise, although sometimes lacking dynamics or being bothersome. Still, there is a concern that voice warnings might be intrusive, disturbing for some drivers, which could lead to a loss of interest for them. One participant expressed concerns related to the integration of voice warnings with other voice-based systems already present in the car.

Four participants (P1, P2, P4, P6) found visual warnings distracting, which is also reflected in higher TLX scores when visual warnings were present (Fig. 11). As in XP1, some participants stated that listening is easier and some that

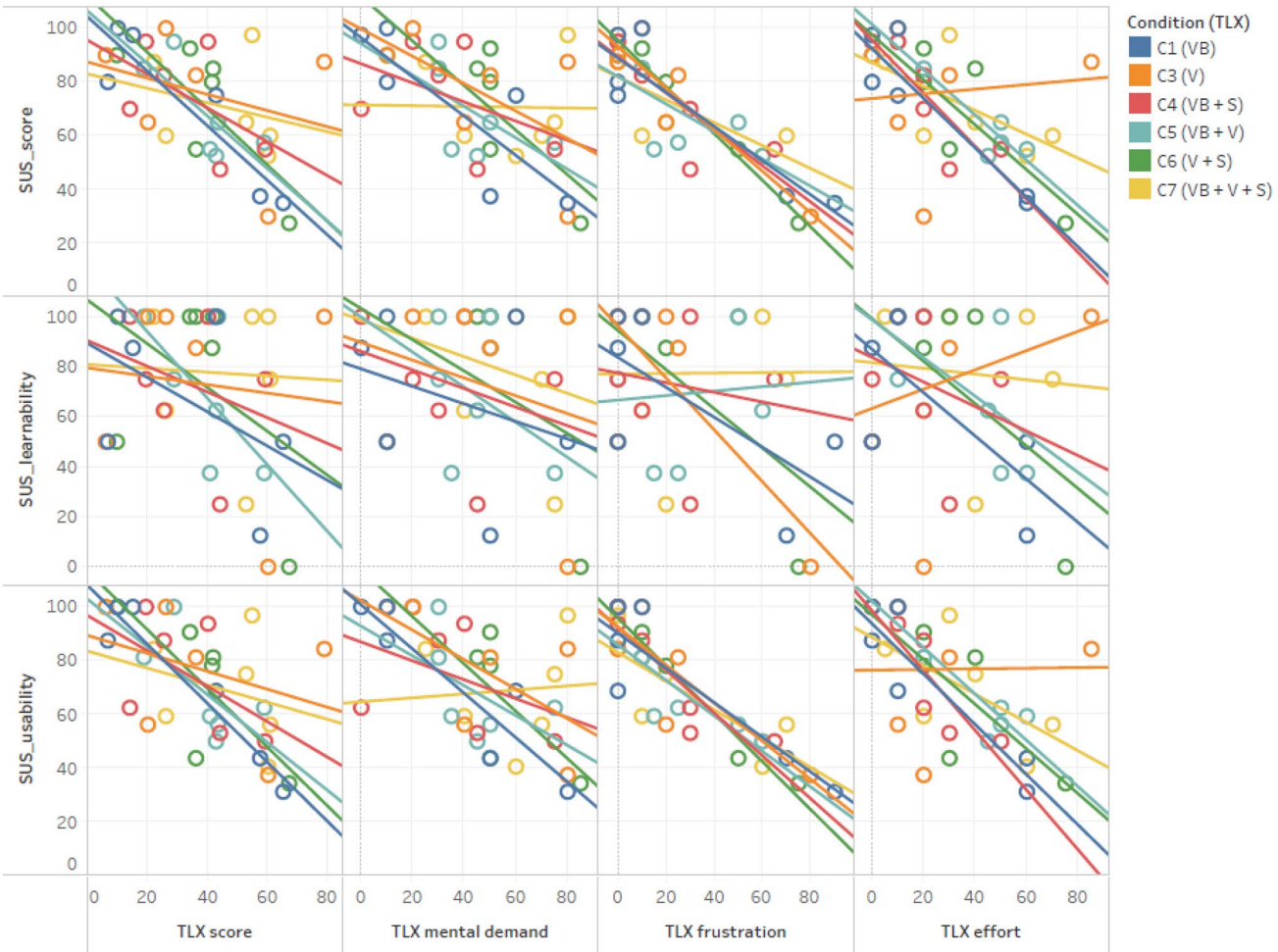


Fig. 8 Correlation coefficient charts per condition between TLX scores and SUS scores in XP2

Table 5 Thematic analysis per condition from experiment 2

Condition	Theme	Codes	*N*	Participants
C1	Adaptation	Attention to voice messages could fade over time	1	P4
	Attractiveness	Length of voice messages is appropriate	1	P2
	Attractiveness	The voice messages lack dynamics	1	P6
	Attractiveness	Voice messages are too long	2	P5
	Attention	Repeating the warnings would be distracting	1	P4
	Attention	Voice messages are disturbing	2	P4
	Customization	On-demand explanation is needed	1	P3
	Comfort	Voice messages are tiring	1	P4
	Cognitive load	Listening is easier	1	P2
	Cognitive load	Visual warnings are distracting	2	P2
	Stress	Missing the warning is stressful	1	P6
C3	Attention	Reading is distracting	2	P1
	Attention	Beep draws attention	1	P6
	Completeness	Visual warnings only are not sufficient	3	P6
	Comfort	Effective and timely warnings	1	P1
	Cognitive load	Listening is easier	1	P1
	Cognitive load	Visual warnings are distracting	1	P2
	Interruption	Only important alerts should interrupt music	1	P3
	Layout	Position of the visual warnings is not good	3	P2
	Perception	Visuals-only are too easy to miss	2	P6
	Reassurance	Feeling safe	1	P3
	Visual appearance	Visual warnings are clear	3	P1, P4
C4	Attractiveness	Voice messages are concise and friendly	1	P4
	Attractiveness	Beep is pleasant	2	P2, P4
	Attractiveness	Voice messages are too long	2	P5
	Attention	Beep draws attention	1	P1
	Attention	Beep signifies importance	1	P1
	Customization	Repeat option is needed	1	P6
	Comfort	Beep is discreet	1	P1
	Interruption	Only important alerts should interrupt music	1	P6
	Perception	Voice warnings are easy to miss	2	P4, P6
	Reassurance	Feeling safe	1	P3
C5	Attention	Beep draws attention	1	P1
	Cognitive load	Voice messages are lighter to process	1	P4
	Cognitive load	Voice warnings are too detailed	1	P1
	Cognitive load	Listening is easier	1	P2
	Cognitive load	Visual warnings are distracting	1	P2
	Integration	How to combine voice alerts with other voice-based systems	1	P4
	Integration	Voice messages and visual messages combined are heavy	1	P5
	Interruption	Voice messages interrupting the music would be annoying	1	P4
	Interruption	Beep interruption is better	1	P4
	Layout	Position of the visual warnings is not good	3	P2
	Reassurance	Visual warnings are a good backup	2	P3, P6
	Stress	Voice messages are stressful	1	P6
	Visual appearance	Visual warnings are too large	1	P6

Table 5 (continued)

Condition	Theme	Codes	*N*	Participants
C6	Attractiveness	Beep is pleasant	1	P5
	Attractiveness	Visual warning is not appealing	1	P1
	Attention	Beep draws attention	2	P3, P6
	Attention	The beep announces an event	1	P4
	Customization	Warning and voice personalization is needed	1	P3
	Cognitive load	Listening is easier	5	P1, P2
	Cognitive load	Visual warnings are distracting	4	P1, P4, P6
	Cognitive load	Visual messages are concise	1	P5
	Stress	Missing the warning is dangerous	1	P6
	Usefulness	Visual description of lane change is useful	1	P1
C7	Attractiveness	Voice is friendly	1	P6
	Attention	Voice messages are disturbing	2	P4
	Customization	Repeat option is needed	1	P1
	Cognitive load	Too much information	1	P5
	Cognitive load	Listening is easier	1	P1
	Reassurance	Feeling safe	1	P6
	Stress	Warnings are stressful and make you feel uncomfortable	2	P6
	Stress	Warnings are too close to the event	1	P3

visual warnings are easy to miss while driving. For example, P1 said *"by hearing the messages, it still leaves your attention free to check what's going around you"*. Even after we decreased the amount of text, the visuals still required a high level of attention. Participants stated visual warnings could serve well as a backup to check them manually when they miss the voice warning. P6 commented that *"if you didn't hear the message, you can always see the indications on the screen"*. This finding is in line with previous research reporting that text-only warnings are the longest to process and are the least preferred by participants [7].

For the voice-based warnings, two participants expressed the need for a repeat option for speech warning alerts. This additional feature might be useful in the case drivers would miss the alert, such as while listening to the radio or talking to other passengers in the car. A common remark was that voice-based warnings would not work well in certain contexts, as P4 illustrated by saying *"Now, again if you are talking with somebody in the car, the music is playing, the kids are playing in the back seat, the voice would not be good in that situation"*. P3 expressed concern by saying *"Cutting your music? That would be bad experience"*. Additionally, P2 stated that *"it's easier to focus on the road while you hear the message and don't have to read it"*. However, having the voice warnings always on might become tiresome for drivers and make them lose interest for it. P4 commented: *"After a while you won't be paying much attention to the voice. I think hearing the voice all the time I would get tired quickly."*.

In addition, it is important to say that the participants were pleased that there was a beep sound drawing their attention to the warnings or traffic situation. P1 said that *"In fact, every time you have to know something that's important, you get a little ding sound, so that it draws your attention to it [the warning]"*.

Experiment 3

XP3 aimed to investigate if and how drivers would interact with a warning alert system assuming they were driving an SAE level 3 car. The driving context comprised city driving with medium traffic density. The conditions C1–C6 were shown to the participants. In a between-subjects design, half of the participants were exposed to conditions without visual warnings (C1,C2,C4), and the other half to the conditions with visual warnings (C3,C5,C6). We also administered the SEQ after each condition to evaluate the users' task difficulty on a 7-point scale. The SUS questionnaire was not used in this experiment, because the focus of this iteration was to collect user behavior and explore their feedback when interacting with the voicebot [30]. Also, because we did not fundamentally change the warning alerts compared to the previous two iterations, we decided not to measure usability anymore and to exclude the SUS.

We simulated the voicebot system of the car using a WOz technique. WOz prototyping method involves a human operator, called a wizard, that simulates one or more parts of the system, while the user is interacting with it. WOz is mostly used in the early design stages of systems involving speech and gestures, as it allows the exploration of different design alternatives. WOz is often used to study the design of automotive user interfaces [39], commuter experience in autonomous cars [27], and for real-time observation and interaction prototyping in vehicles [35]. To support the remote WOz technique in this study, we developed a simple web application that contained an interface with buttons coupled to a TTS module to simulate the conversation with the voicebot (Fig. 9). The first author moderated the session, while the second author took the role of the wizard, invisible to the participants.

Participants

There were six participants (2 women) aged between 55 and 69 ($M = 61.17$, SD = 5.2) in XP3. None of them participated in the previous two iterations, as we decided to recruit a new sample. We divided the participants into two groups. One group only saw the conditions containing visual warnings (C3,C5,C6), while the second group only saw the conditions without visual warnings (C1,C2,C4), resulting in 18 trials overall.

WOz Interaction Design

We instructed participants that their task was to monitor the road and follow the warning alerts. Then, we informed them about the tasks the voicebot supported: alerting the driver about the obstacles ahead, repeating the warning, explaining the warning to the driver, and explaining why is the car making certain maneuvers. They had to start each new interaction sequence with the keyword "Tony", which was the name of the voice assistant. We told participants to use natural language when interacting and to try to stay within the limits of the supported tasks. However, the wizard still tried to address participants' requests that were out of scope whenever possible. This led us to even more discoveries. To better understand whether drivers would interact with their car about the warnings, we counted the number of times participants started the interaction. This measure was called "pull requests". We hypothesized that there would be more pull requests when participants are exposed to unimodal conditions, such as C1 and C3. We selected a video from DR(eye)VE project [38] that lasted 5 min.

Warnings

We hypothesized that the drivers would interact with the voicebot to extract more information about the upcoming danger or road situation. Hence, we decided to include the condition C2 and exclude the condition C7, because using three modalities proved to be too complex in the two previous experiments. Furthermore, the new scenario did not include the HU level, but only LU and MU. This decision was made, because HU situations would not give enough time to the participant to interact with the voicebot. Conditions C1, C4, and C5 contained voice warnings that were always pushed to the driver. LU voice warnings all started with the word "notice", and consisted of traffic jam alerts ("Notice! Queue of cars in 50 meters. Slowing down."), side hazards ("Notice! Stopping to give priority."), and lane change ("Notice! Slowing down to change the lane."). MU voice warnings started with the word "warning" and consisted of pedestrian alerts ("Warning! Pedestrian crossing the road!"), collision avoidance ("Warning! Slowing down

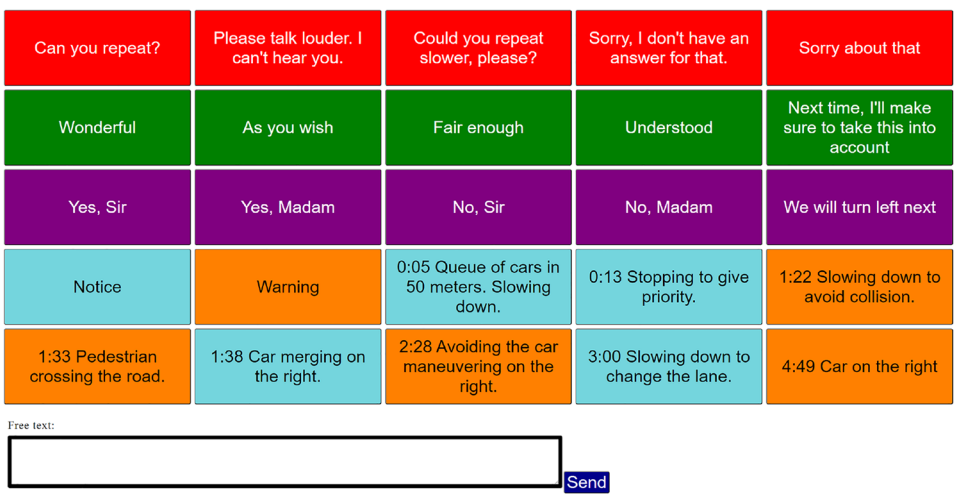


Fig. 9 TTS dashboard for the wizard used in XP3

to avoid collision."), and side hazards ("Warning! Car on the right!"). We used green (LU) (Fig. 10a) and orange (MU) bounding boxes (Fig. 10b) to highlight the obstacles visually and simulate the obstacle detection system.

Quantitative Results

Both experimental groups had a similar number of pull requests. Specifically, the group exposed to visual warnings and the group exposed to non-visual warnings had 5.33 and 5.57 pull requests on average, respectively. Regarding cognitive load, as in XP2, visual warnings were also positively correlated with higher cognitive load in XP3. Figure 11 shows that the mean TLX score in the group containing visual warnings was generally higher (57.5), compared to the mean TLX score in a group without visual warnings (31.4). Surprisingly, multimodal combinations of warning alerts in XP3 generated slightly lower cognitive load (42.7), compared to unimodal combinations (46.3). In Table 7, we can see the individual scores for each subscale and condition in both XP2 and XP3. Overall, we see that mental demand, temporal demand, and effort were the dominant sources of workload. The prototype used in XP3 scored higher on overall cognitive load compared to XP2. This could be attributed to the novelty with regard to the voice interaction.

The analysis of mean SEQ scores for each condition revealed that it was easier to monitor the road situation and follow the warnings when there were no visual warnings (SEQ 5.78), compared to when the visual warnings were used (SEQ 4.89). To investigate whether the ease of use and cognitive workload were correlated, we calculated the Pearson's correlation between unweighted TLX and SEQ scores. The two variables were negatively correlated, $r(7) = -.79, p < .05$ in case of unimodal conditions, and $r(7) = -.72, p < .05$ in case of multimodal conditions.

Regarding UEQ, AT score (1.89—excellent) confirmed that participants generally liked the prototype. However, DP (0.67—bad) decreased significantly, indicating that users did not feel in control. Also, compared to XP2, EF (1.29—above average) also decreased significantly, confirming our finding that the system's reaction time was a little slow. Low PS score (1.25—above average) tells us that participants had difficulties getting familiar with the system, which is also reflected in the qualitative findings and confirmed by the fact that participants often tried unsupported tasks. Therefore, improving the pragmatic aspects of the prototype and evaluating it in a more immersive context, would perhaps result in better UX. ST (1.58—excellent) remained almost unchanged, meaning that participants found the system relatively fun to use. Although NV (1.29—good) decreased, we concluded that the prototype scored well on hedonic qualities.

Qualitative Results

Table 6 presents the main findings from the thematic analysis. Generally, it was unclear for the participants how they should interact with the car. First, because they might not had been used to voice interaction. Second, because they did not know what they should ask. However, participants still tried some unsupported features and commented on the car's driving style, often requested the car to slow down, tried control the speed or change the route. Lack of control frustrated some participants, such as P7 who said *"I have a feeling that the system does not take my injections [input] into account. For example, slow down, it doesn't take this into account when I say it"*. The low score for DP scale of 0.67 in UEQ illustrates this lack of control. From the thematic analysis, we can see that "disapproving with driving style" was a common remark across conditions.

Participants also expected the car to react to their negative feedback and thought that the car would learn based on it. They often asked why is the car making certain decisions, such as taking turns or giving way to other cars from side streets. This was sometimes due to different traffic regulations between countries. Occasionally, they would not understand the warnings and would ask for an explanation from the car. In all conditions, except C4, there was one participant that did not record any interaction with the car. Additionally, some participants would just respond with simple "okay" or "thank you" when they heard the warning. This might indicate that at least a third of drivers are not willing to use their voice as a primary communication modality with their car. Regarding the content of the driver–car interaction, the majority of the utterances concentrated on the car's driving style. That is also present as the most common code in the thematic analysis (Table 6). Examples of participants' remarks include: *"Tony, why do you stay on the left lane?"*, *"Tony, did you see the car on the right?"*, *"Tony, slow down there are lots of cars"*, or *"Tony, it's raining. You should adapt your driving behavior accordingly"*.

Another issue was the response speed that was too high. Due to the architecture of the WOz system and the fact that it was done remotely, there was always a delay present between the participant's request and the answer from the wizard. P8 said that *"It takes a lot of time to answer. The possibility to ask is fine, but we're driving, I would expect it to respond faster. My personal feeling is that the system is asking itself "What am I going to answer?"*.

Regarding the beep sound, it was described as useful. For instance, P7 said *"The sound was not stressful, but a sound tells me that the car has seen the danger"*. Regarding the voice alerts, four participants (P8,P9,P11,P12) said that voice alerts are a way to get to know the system and build their trust. For example, P8 commented that *"the voice messages are fine when you have just started using the system.*

Table 6 Thematic analysis per condition in experiment 3

Condition	Theme	Codes	*N*	Participants
C1	Responsiveness	System not responsive to commands	1	P7
	Responsiveness	System not responding fast enough	1	P7
	Responsiveness	Response time is good	1	P9
	Cognitive Load	Lot of thinking required	1	P7
	Confusion	Not clear how to interact with VA	1	P8
	Confusion	Lack of understanding	1	P8
	Driving style	Disapproving with driving style	3	P7, P8, P9
	Confusion	Difference between yellow and green boxes	2	P9
	Usefulness	Useful visuals	1	P9
C2	Adaptability	Appropriate reaction to negative feedback is needed	3	P7, P8
	Preference	Beep sound is pleasant	1	P7
	Comfort	Useful beep sound	1	P7
	Cognitive Load	The beep sound lowers effort	2	P7
	Safety	Beep sounds reduces insecurity	1	P7
	Trust	Voice notifications are reassuring	1	P7
	Frequency of use	Familiarity makes it easier	2	P8, P9
	Driving style	Disapproving with driving style	3	P7, P8, P9
	Confusion	Mismatch between driving style and notification	1	P8
	Usefulness	Visuals indicate events well	1	P8
	Responsiveness	System not responding fast enough	1	P8
	Confusion	Difference between yellow and green boxes unclear	1	P9
	Interaction	Participant did not hear the beep	1	P9
C3	Trust	Voice notifications are reassuring	2	P7, P8
	Driving style	Disapproving with driving style	3	P7, P8
	Frequency of use	Familiarity makes it easier	1	P7
	Responsiveness	System not responding fast enough	1	P7
	Comfort	Voice notifications are interrupting	2	P7, P9
	Usefulness	Redundant warnings for traffic events	1	P8
	Learnability	Voice notifications are a way to get to know the system	3	P8, P9
	Adaptability	Appropriate reaction to negative feedback is needed	2	P8
	Usefulness	Voice notifications are informative	2	P9
	Usefulness	Voice notifications are helping in the city	1	P9
	Safety	Voice notifications increase safety	1	P9
C4	Responsiveness	System not responsive to commands	1	P10
	Trust	Voice notifications are reassuring	2	P10, P12
	Adaptability	Appropriate reaction to negative feedback is needed	1	P12
	Comfort	Voice notifications are interrupting	1	P12
	Comfort	Less voice notifications are expected on highway	1	P12
	Usefulness	Voice notifications are helping in the city	1	P12
C5	Interaction	Trigger is needed for interaction	1	P10
	Cognitive Load	Beep draws attention	1	P10
	Information	Navigation information missing	1	P10
	Responsiveness	System not responsive to commands	1	P11
	Adaptability	Reaction to negative feedback is needed	1	P11
	Driving style	Disapproving with driving style	1	P11
	Interaction	Lack of system's feedback	1	P11
	Confusion	Beep is unclear	1	P11
	Comfort	Asking the voice assistant is unnatural	1	P11
	Preference	Important messages can interrupt me	1	P11
	Accuracy	Voice assistant gives accurate answers	1	P12

Table 6 (continued)

Condition	Theme	Codes	*N*	Participants
C6	Adaptability	Appropriate reaction to negative feedback is needed	1	P10
	Responsiveness	System not responding fast enough	1	P10
	Cognitive Load	Assistant is reducing stress	1	P10
	Learnability	Voice notifications are a way to get to know the system	2	P11, P12
	Confusion	Beep is unclear	1	P11
	Information	Voice assistant is incomplete	1	P12
	Information	Navigation information missing	2	P12

Table 7 NASA-TLX subscale scores

CD	MD		PD		TD		P		E		F		M	
	XP2	XP3	XP2	XP3	XP2	XP3	XP2	XP3	XP2	XP3	XP2	XP3	XP2	XP3
C1	35	52.5	21.7	16.7	35.8	35	30	11.7	23.3	16.7	28.3	20	29	25.4
C2	–	56.7	–	15	–	41.7	–	15	–	48.3	–	38.3	–	35.8
C3	46.7	75	32.5	45	35	87.5	39.2	87.5	25.8	80	22.5	90	33.6	77.5
C4	35	51.7	20.8	15	40.8	50	34.2	26.7	21.7	30	22.5	25	29.2	33.1
C5	44.2	66.7	18.3	33.3	40	56.7	37.5	35	39.2	40	26.7	26.7	34.3	43.1
C6	48.3	65	30.8	33.3	45.8	63.3	22.5	46.7	30.8	60	27.5	43.3	34.3	51.9
C7	58.3	–	29.2	–	50	–	34.2	–	37.5	–	26.7	–	39.3	–
M	44.6	61.3	25.6	26.4	41.2	55.7	32.9	37.1	29.7	45.8	25.7	40.6	33.3	44.5

MD mental demand, *PD* physical demand, *TD* temporal demand, *P* performance, *E* effort, *F* frustration, *M* mean, *CD* condition

(a) Green box- low urgency

(b) Orange box - high urgency

Fig. 10 Bounding boxes as visual warnings used in XP3

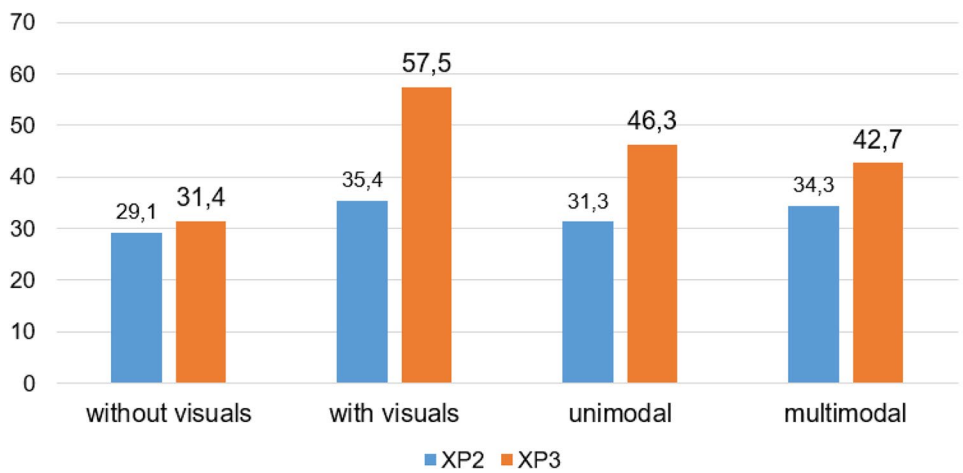


Fig. 11 Mean TLX scores per type of modality

After a while, the visuals are enough". P8 also said *"It gives you more confidence, but after a while I would put it off."* and P9 said *"I think that the voice is very useful in the beginning to build up the trust in the system."*, while P11 said *"In the beginning, I would use the explanation to know that the car can be trusted"*. However, they would switch off the voice alerts after they have accepted the system, illustrated by P12 *"After a while, you put quiet mode on"* and P8 *"It's like with the GPS, and then you turn it off after a while."*.

A possible explanation could be that the voice alerts reassured participants about what the driving automation system sees and does. Based on the content of the voice alerts, the participants could compare that to the situation they see themselves and judge the accuracy of the warning alert system. An open issue is how to integrate the voice alerts with other voice-based systems, but also not disturb the conversation between passengers. The correct timing to issue the voice alert needs to be determined, not to interrupt an ongoing conversation, but also to notify the driver about the danger on time. For P9, interruptions are not acceptable, although the voice messages were informative. This is illustrated by P9's comment; *"Voice messages are informative. But, I think that the priority is to talk to the other people in the car or to listen to the radio"*.

These findings might provide useful considerations for the design of future HMIs for driver-assistance and driving automation systems. Clearly, there is no one-size-fits-all solution; however, a warning system that would adapt to user's preferences (i.e., personalization), or at least allow the user to configure it manually, would represent a viable approach to the HMI design. This study demonstrates potential in using speech-based warning alerts, although its application should be considered carefully and contextually. Thus, further research is needed to examine how the voice pitch, tone, and language manipulations affect the perceived danger severity. Also, further research is required to determine the right moment to use the voice-based warning alerts based on driver's state.

Positioning Against the Related Work

Human–computer interaction (HCI) community is mainly contributed to the design of driving safety features for older drivers. However, the research remains insufficient with regard to infotainment and driving assistance systems design from HCI and human factors perspectives [45]. Several gaps were identified in the literature regarding the use of modalities for warning alert design, and HMI design for driver-car interaction more generally. First, subjective aspects of the interaction, such as users' likes, dislikes, perceived effectiveness, and preferences are rarely analyzed in the literature. Second, most research studies evaluated HMI design solutions experimentally on large sample sizes to enable statistical data analysis and make design comparisons based on predefined dependent variables. Third, most studies conduct research using static driving simulators, which, although suitable for this type of research, are not easily available or necessary for early prototyping phases. Therefore, we argue that more qualitative and formative research is needed to demonstrate its usefulness for HMI design and evaluations, and better understand how older driver perceive HMIs for warning alerts.

To fill this gap, this article provides users' perspective that helps derive additional requirements and understanding for HMI design. Table 8 shows the comparison of research studies that describe the evaluation of car HMIs for takeover and handover of control, effects of warning alerts on driver's state, and use of different modalities to design the alert systems.

Older drivers represent a user group that potentially has age-related impairments, such as decrease in visual, sensory, physical, or cognitive performance. Young et al. [59] conducted a systematic literature review of the HMI design guidelines for older drivers. They stated that further research is required to support the development of HMIs addressing age-related issues. This is problematic especially considering that older drivers are more prone to be involved in a car crash because of visual and cognitive impairments [20]. In Table 8, we see that only four studies included older drivers, while only two included drivers above the age of 70.

Most research studies in the domain of autonomous driving and human factors research focus on collecting and analyzing behavioral data (i.e., describing what users do) or investigating the effect of certain HMI designs on the driving style. Others focus only on objective measurements of driver's performance such as measuring reaction times to certain warning alerts, or collecting sensor and log data [7, 28, 40, 41]. In other words, much attention has been given to experimental research and less on investigating user's subjective reactions, attitude, and needs. Our methodology relies on the collection of early stage qualitative UX data that facilitates decision-making and ensures early user involvement. The advantage of our approach is that users can interact with an early prototype and collected data can be used to reduce the late design changes.

Conversely, we used a video prototyping method for early stage prototyping that is not resource-intensive. Most studies describe the use of advanced driving simulators to conduct experiments and rely on quantitative data collected from a larger sample of participants [15, 26, 28, 41]. In this work, we used a video-based driving simulation that allowed us to conduct tests remotely during the coronavirus outbreak. In addition, we decided to focus on a thorough qualitative analysis collected from a smaller sample of users. We made

Table 8 Review of the related work and comparison of methods and inclusion of older drivers

Year	Ref	Older drivers	Population segment	Evaluation methods	Platform
2019	[3]	x	28 participants (10 women, 18 men), (M = 38.64, SD 11.21)	questionnaire, log files analysis, take over suitability	video-based driving simulator
2009	[7]	x	10 drivers (2 women, 8 men), 25–45 years old	Secondary task, warning perception time, voting	Static driving simulator
2010	[8]	Δ	32 drivers (16 women, 16 men), 20–62 years old (M = 32.6, SD 10.8)	Questionnaires; reaction times	Driving simulator
2020	[19]	O	24 younger (M = 21.9, SD 1.4), 24 older (M = 71.7, SD 4.9) Drivers	Questionnaire, probes	Driving simulator
2020	[24]	x	41 young drivers (18–23 years old)	Questionnaire, interview, eye tracking	Driving simulator
2015	[26]	x	64 students (32 women, 32 men), 18–27 years old (M = 21.11, SD 1.42)	Questionnaires, log files analysis	Driving simulator STISIM
2019	[28]	x	24 students (6 women, 18 men) (M = 20, SD 1.1)	Time measurements, questionnaire	Static driving simulator
2016	[37]	x	85 participants (52 women, 33 men), (M = 19, SD 1.03)	Questionnaire, probe memory recall	3 video prototypes with ambient sound
2014	[40]	x	22 participants (9 women, 13 men), 18–44 years old (M = 25.04, SD 5.95)	Self-reported evaluation	n/a—exposure to designed cues
2015	[41]	x	21 participants (3 women, 18 men), 18–29 years old (M = 21.00, SD 2.84)	Self-reported evaluation	Static driving simulator
2018	[42]	x	49 participants (24 women, 25 men), 17–86 years old (M = 45.51, SD 17.36)	Self-reported evaluation	Static driving simulator
2008	[43]	Δ	16 younger men (M = 39.6, SD 7.1), 14 older men (M = 76.6, SD 4.3)	Log files analysis, measurements	Static driving simulator
2018	[46]	n/a	n/a	Expert evaluation, questionnaires, sketching	Paper prototype video prototype
2020	[57]	x	50 participants exp 1: (M = 34.3, SD 13.5) exp 2: (M = 39.9, SD 9.6)	Log files analysis	Static driving simulator

Inclusion of older drivers: x, does not include; O, includes; Δ, includes partially; *n/a* not applicable

this choice to be able to characterize the underlying user needs and attitudes toward different warning modalities. Consequently, our objective was to generate findings that could provide more insights for the future development of HMI prototypes for driver-car interaction.

Video prototypes, although not a novelty, were not used in a large number of previous studies on driving automation. For example, Nees et al. [37] used three pre-recorded videos of a car driving in a small town and highway context. They examined drivers' event recall performance, cognitive load, perceived usefulness and annoyance of speech, auditory, and visual alerts [37]. The study by Boelhouwer et al. [4] investigated whether reading about system information from owner's manuals helps drivers better understand car's automation capabilities. Participants watched videos of an autonomous car driving in various urban scenarios. Participants had to imagine they were driving the car in a real world and determine whether a takeover is necessary or not for each situation. The results showed that reading about system information does not improve driver's takeover decision-making. Furthermore, Pettersson et al. [39] used video prototypes as a design technique for human-vehicle interactions [39]. Other researchers [46] used them to conduct an expert evaluation of HMI designs for takeovers in highly automated truck driving.

Nees et al. [37] found that the visual alerts were less annoying compared to auditory icons and speech alerts. Additionally, their experiment showed a greater recall of events when speech alerts were used, compared to auditory and visual displays. For this reason, speech alerts could be effective in maintaining the situational awareness in automated driving. Similarly, in our experiments, we have discovered that visual alerts generated higher cognitive load with participants, which was not the case in [37]. A recent study [57] evaluated two interfaces for advisory traffic information using visual and auditory modalities. The interfaces

were compared in three different scenarios and driver's performance was measured. They found that visual and auditory modalities are complementary; however, the visual modality better conveyed the position of other road users, while the sound was better at grabbing driver's attention. Similarly, our findings also confirm that the sound is effective in drawing the attention to the warning alert.

Cao et al. [8] conducted a user study to evaluate the usability of speech and visual danger warnings with and without active driving suggestions. They studied driver's reactions to those warnings to avoid obstacles on the road. Their study suggests that combining speech and visual modalities results in highest usability of obstacle warning. However, in our study, that combination scored lowest on overall usability (C5 in Fig. 7). Nevertheless, our sample was significantly smaller and we treat our quantitative data as indicative and for design guidance. Further, they reported that speech-only resulted in the highest percentage of unsafe behavior and cognitive load, and lowest usability. However, combining visual and speech modalities resulted in very positive evaluation and good driver's performance. Additionally, the study reported that drivers were also in favor of the visual and speech combination. Conversely, our study revealed that the use of speech and visuals (i.e., C5) was considered too heavy to process. Several participants reported that if the speech message did not contain all the necessary information, they would have used the visual warning component as an additional source of information. This behavioral pattern enabled them to understand the warning alerts and the road situation better. However, this is highly context-dependent and specific to our warning alert design. Additional research needs to be done to determine the type of information each modality should convey, based on the driving context and driver's state.

Design Guidelines for Older Drivers

To the best of our knowledge, only three articles reviewed or proposed design guidelines for older drivers [13, 45, 59]. One survey research study on the use of ADAS involved drivers above the age of 80 [36]. Table 9 shows the articles that reviewed HMI design guidelines for ADAS and In-Vehicle Information Systems (IVIS). The results of those reviews show that capabilities and limitations of older drivers are rarely considered by both car manufacturers and design guidelines. Older drivers need to be accounted not just for their reduced physical, sensory, and cognitive capabilities, but also for the process of learning new skills necessary for using new technologies [59]. Two articles [13, 59] provide concrete suggestions and guidelines for how to design HMIs for older drivers and present their related benefits from safety and comfort perspective.

Discussion

Both XP1 and XP2 shared the same experimental design. Based on the feedback received in XP1, we improved the prototype in XP2 and recruited the same participants to evaluate the changes we made. This choice resulted in receiving similar user feedback in both experiments. A possible explanation is that these similarities were caused by the learning effect among participants between the two experiments, as the same people participated in both. Thus, recruiting the participants to match the user profile of the target users rather than having the same participants would be a better methodological choice, as it would compensate for the learning effects. However, this is not always easy to mitigate because of business or organizational constraints. The purpose of formative evaluations is to test-and-refine prototypes based on the qualitative feedback received from a small sample of participants. Users' subjective impressions of the prototype in question and their behavior are collected and analyzed to produce new requirements and redesign the prototype.

In contrast, summative evaluations rely on experiments involving a larger sample of participants, check for statistical differences, and produce generalizable results and conclusions. We recruited six participants for each experiment, which complied with both the formative approach [55] and our organizational constraints, namely time and budget. However, regardless of whether we conducted the

Table 9 List of review articles related to design guidelines for older drivers

Year	Reference	Methods used	Topic	Lists design guidelines
2017	[13]	Literature review	Identification of improvement opportunities for car interiors that improve safety, comfort and inclusion of older drivers	Yes
2021	[36]	Online survey	Investigation of older drivers' intention to use ADAS systems and full driving automation	No
2015	[45]	Literature review	Classification of research studies related to smart car technologies and elderly drivers' issues	No
2017	[59]	Literature review	Review of ADAS and HMI design guidelines for older drivers regarding their sensory, cognitive and physical capabilities	Yes

experiment remotely or in the lab, the results in XP1 and XP2 were comparably similar. The evidence from this study suggests that it is possible to obtain insightful and accurate results regardless of the test environment in which the study was conducted.

The advantage of the formative evaluation approach is that we could easily select the appropriate evaluation methods and techniques between iterations. For example, in XP1, we collected qualitative subjective feedback and usability ratings. Then, in XP2, we recognized the need to measure the subjective cognitive load to better understand the effects of three types of modalities on participants' cognitive demand. The aim was to first determine the usefulness and user satisfaction with different combinations of modalities and then evaluate them more in detail. Therefore, the number of variables measured may vary from one iteration to another depending on the received user feedback.

Using remote testing, we have identified several advantages and drawbacks. The advantages count for the easy setup of live remote testing as many tools are available nowadays. Also, it is relatively easy to create prototypes using video editing software once the warning alerts were designed. Open-source databases provide large amounts of video material to simulate various road situations. We also showed that recording footage on public roads is possible and rather easy to do. The drawbacks include lack of control on the user's side and lack of immersion compared to real or simulated driving. Furthermore, video prototypes have limited flexibility and can only provide predefined scenarios without the ability to respond user's inputs. This was especially visible in XP3 when some participants complained about the car not adapting its driving style after they requested it to do so. A reasonable approach would be to use video prototypes in early stage product development and testing. Video prototyping demonstrated its efficiency for collecting early feedback from target users regarding the warning modalities and the content of warning alerts.

In this study, we administered four questionnaires to collect quantitative data, coupled with semi-structured interviews to collect qualitative data. While it enriched data collection, this combination also substantially increased the duration of the sessions. For example, the TLX questionnaire requires a long administration procedure where participants must read the instructions, identify the weights of each dimension, understand the rating scales well, and consistently recall their definitions throughout the study. This disrupted a consistent flow of the experiment and required the researcher to remind the participants of the meaning of the rating scales. Relying on the TLX Raw would have been much more efficient as it would remove the need to use the weighted rankings of TLX dimensions, as proposed in the literature [18]. In addition, we decided to exclude the SUS from future evaluations and not to combine it with other questionnaires, such as UEQ, because we have not observed a substantial benefit between rounds. Using SUS would be more appropriate during usability testing with clearly specified user tasks. Although we could extract the Learnability score from the SUS, we realized that we should rather rely on interviews to check for participants' understanding and learnability of the warning alerts. UEQ provided more useful information regarding the potential areas of improvement and it focuses on experiential qualities rather than the usability of a product, which we believe is more important in exploratory phases of system development.

We have not included the measurement of ecological validity in this work, as it was out of the scope of our study. Ecological validity is the statistical correlation between a proximal cue and the distal variable to which it relates [25]. In our study, the proximal cues refer to the traits or characteristics of the setting perceived by participants during the controlled experiments (e.g., driving on the highway), while the distal variables refer to the actual traits of the environment (e.g., watching a video of highway driving on a screen). In future work, we intend to compare the ecological validity between the three following experimental settings. First, the remote testing of video prototypes. Second, the immersive simulator studies in the lab. Third, driving a real car on the road. In particular, we will assess ecological validity by comparing participants' feelings of immersion between the three settings and participants' behavior between the three settings. This will allow us to assess the extent to which participants' experimental behavior corresponds to the expected functional behavior toward which we wish to generalize [25].

Conclusion

This article presented a study on the UX design and evaluation of a senior-friendly warning alert system for semi-autonomous vehicles. To that end, we created low-fidelity video-based prototypes and investigated several combinations of output modalities to notify the driver about the road situation ahead. The analysis of qualitative and quantitative data shows that speech messages were effective in conveying the warning information to drivers. We also found that visual warnings are generally considered distracting and cause a higher workload. Still, participants considered the visual warnings as a good backup to voice warnings. Voice interaction with a car seems to be a novelty to older drivers. Within a formative approach, we recommend recruiting new participants for each experiment while maintaining the sample size between 6 and 8 individuals to control the learning effect with the task and discover new design opportunities. Regarding standardized questionnaires, we found the usage of TLX and UEQ to gather information on the potential areas of

improvement suitable for our domain. These findings should be useful for practitioners and researchers involved in the design and development of features for semi-autonomous vehicles, such as voice-based interfaces, chat-bots, or road sign assistance.

Acknowledgements The authors would like to thank all participants to the experiments and the anonymous reviewers for their constructive comments on earlier versions of this manuscript.

Funding This research was funded by the Service Public de Wallonie (SPW), Belgium (Grant number 7982).

Data Availability Study materials are available at: https://osf.io/xbpu7/?view_only=8c0ccd5839714443ad25194be3ab0ea9.

Code Availability Not applicable.

Declarations

Conflict of Interest The authors declare that they have no conflict of interest.

References

1. Agudelo AF, Bambague DF, Collazos CA, Luna-García H, Fardoun H. Design guide for interfaces of automotive infotainment systems based on value sensitive design: a systematic review of the literature. In: Proceedings of the VI Iberoamerican conference of computer human interaction (HCI 2020), Arequipa, Perú, 2020;16–18.
2. Baldwin CL, Lewis BA. Perceived urgency mapping across modalities within a driving context. Appl Ergon. 2014;45(5):1270–7. https://doi.org/10.1016/j.apergo.2013.05.002.
3. Boelhouwer A, van Dijk J, Martens MH. Turmoil behind the automated wheel: an embodied perspective on current HMI developments in partially automated vehicles. In: HCI in mobility, transport, and automotive systems, vol. 11596, p. 3–25. Springer, Cham; 2019. https://doi.org/10.1007/978-3-030-22666-4.
4. Boelhouwer A, van den Beukel A, van der Voort M, Martens M. Should I take over? Does system knowledge help drivers in making take-over decisions while driving a partially automated car? Transport Res Part F Traffic Psychol Behav. 2019;60:669–84. https://doi.org/10.1016/j.trf.2018.11.016.
5. Bolaños M, Collazos C, Gutiérrez F. Experiences in the application of some models of technology acceptance: adaptation for the elderly people. In: Proceedings of the XXI international conference on human computer interaction, Interacción '21. Association for Computing Machinery, New York; 2021. https://doi.org/10.1145/3471391.3471413.
6. Caird J, Chugh J, Wilcox S, Dewar R. A design guidelines and evaluation framework to determine the relative safety of in-vehicle intelligent transportation systems for older drivers. Ottawa: Transportation Association of Canada (TAC); 1998.
7. Cao Y, Castronovo S, Mahr A, Müller C. On timing and modality choice with local danger warnings for drivers. In: Proceedings of the 1st international conference on automotive user interfaces and interactive vehicular applications, AutomotiveUI'09, p. 75–78. Association for Computing Machinery, New York; 2009. https://doi.org/10.1145/1620509.1620524.
8. Cao Y, Mahr A, Castronovo S, Theune M, Stahl C, Müller CA. Local danger warnings for drivers: the effect of modality and level of assistance on driver reaction. In: Proceedings of the 15th international conference on intelligent user interfaces, IUI '10, p. 239-248. Association for Computing Machinery, New York; 2010. https://doi.org/10.1145/1719970.1720004.
9. Carbonell N, Kieffer S. Do oral messages help visual search? In: van Kuppevelt JCJ, Dybkjær L, Bernsen NO editors. Advances in natural multimodal dialogue systems, p. 131–157. Springer Netherlands, Dordrecht; 2005. https://doi.org/10.1007/1-4020-3933-6_7.
10. Debernard S, Chauvin C, Pokam R, Langlois S. Designing human–machine interface for autonomous vehicles. IFAC-PapersOnLine 2016;49(19), 609 – 614. https://doi.org/10.1016/j.ifacol.2016.10.629 (13th IFAC symposium on analysis, design, and evaluation of human–machine systems HMS 2016).
11. European Parliament, Council of the European Union: Regulation (EU) 2019/2144 2019. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32019R2144.
12. Eurostat: a look at the lives of the elderly in the EU today 2017. https://ec.europa.eu/eurostat/cache/infographs/elderly/index.html.
13. Fernandes SC, Esteves JL, Simoes R. Characteristics and human factors of older drivers: improvement opportunities in automotive interior design. Int J Veh Des. 2017;74(3):167–203. https://doi.org/10.1504/IJVD.2017.086418.
14. Frison AK, Wintersberger P, Liu T, Riener A. Why do you like to drive automated? A context-dependent analysis of highly automated driving to elaborate requirements for intelligent user interfaces. In: Proceedings of the 24th international conference on intelligent user interfaces, IUI '19, p. 528–537. Association for Computing Machinery, New York; 2019. https://doi.org/10.1145/3301275.3302331.
15. Gerber MA, Schroeter R, Vehns J. A video-based automated driving simulator for automotive UI prototyping, UX and behaviour research. In: Proceedings of the 11th international conference on automotive user interfaces and interactive vehicular applications, AutomotiveUI '19, p. 14–23. Association for Computing Machinery, New York; 2019. https://doi.org/10.1145/3342197.3344533.
16. Guo A, Brake J, Edwards S, Blythe P, Fairchild R. The application of in-vehicle systems for elderly drivers. Eur Transport Res Rev. 2010;2:165–74. https://doi.org/10.1007/s12544-010-0037-y.
17. Haghzare S, Campos JL, Bak K, Mihailidis A. Older adults' acceptance of fully automated vehicles: effects of exposure, driving style, age, and driving conditions. Accid Anal Prevent. 2021;150: 105919. https://doi.org/10.1016/j.aap.2020.105919.
18. Hart SG. Nasa-task load index (NASA-TLX); 20 years later. Proc Hum Factors Ergon Soc Annu Meet. 2006;50(9):904–8. https://doi.org/10.1177/154193120605000909.
19. Huang G, Pitts B. Age-related differences in takeover request modality preferences and attention allocation during semi-autonomous driving. In: Human aspects of IT for the aged population. Technologies, design and user experience: 6th international conference, ITAP 2020, held as part of the 22nd HCI international conference, HCII 2020, Copenhagen, Denmark, July 19–24, 2020, Proceedings, part I, p. 135–146. Springer, Berlin; 2020. https://doi.org/10.1007/978-3-030-50252-2_11.
20. Huisingh C, Levitan E, Irvin M, MacLennan P, Wadley V, Owsley C. Visual sensory and visual-cognitive function and rate of crash and near-crash involvement among older drivers using naturalistic driving data. Invest Ophthalmol Vis Sci. 2017;58:2959–67. https://doi.org/10.1167/iovs.17-21482.
21. Insurance Institute for Highway Safety (IIHS): Advanced Driver Assistance 2020. https://www.iihs.org/topics/advanced-driver-assistance.

22. ISO: ISO 9241: Ergonomic requirements for office work with visual display terminals (vdts)—Part 11: Guidance on usability. Technical report; 1998.
23. ISO 9241:210-2019: Ergonomics of human–system interaction—Part 210: Human-centred design for interactive systems; 2019.
24. Kasuga N, Tanaka A, Miyaoka K. Design of an HMI system promoting smooth and safe transition to manual from level 3 automated driving. Int J Intell Transport Syst Res. 2020;18(1):1–12. https://doi.org/10.1007/s13177-018-0166-6.
25. Kieffer S. ECOVAL: ecological validity of cues and representative design in user experience evaluations. AIS Trans Hum Comput Interact 2017;9(2):149–172. https://aisel.aisnet.org/thci/vol9/iss2/.
26. Koo J, Kwac J, Ju W, Steinert M, Leifer L, Nass C. Why did my car just do that? Explaining semi-autonomous driving actions to improve driver understanding, trust, and performance. Int J Interact Des Manuf. 2015;9(4):269–75. https://doi.org/10.1007/s12008-014-0227-2.
27. Krome S, Holopainen J, Greuter S. Autoplay: Unfolding motivational affordances of autonomous driving. In: Meixner G, Müller C editors. Automotive user interfaces: creating interactive experiences in the car, p. 483–510. Springer, Cham; 2017. https://doi.org/10.1007/978-3-319-49448-7_18.
28. Kutchek K, Jeon M. Takeover and handover requests using non-speech auditory displays in semi-automated vehicles. In: Extended abstracts of the 2019 CHI conference on human factors in computing systems, CHI EA '19. Association for Computing Machinery, New York; 2019. https://doi.org/10.1145/3290607.3313078.
29. Laugwitz B, Held T, Schrepp M. Construction and evaluation of a user experience questionnaire. HCI Usability Educ Work. 2008;5298:63–76. https://doi.org/10.1007/978-3-540-89350-9.
30. Lee SS, Lee J, Lee KP. Designing intelligent assistant through user participations. In: Proceedings of the 2017 conference on designing interactive systems, DIS '17, p. 173–177. Association for Computing Machinery, New York; 2017. https://doi.org/10.1145/3064663.3064733.
31. Lewis JR, Sauro J. The factor structure of the system usability scale. In: Kurosu M editor. Human centered design, pp. 94–103. Springer, Berlin; 2009. https://doi.org/10.1007/978-3-642-02806-9_12.
32. Lewis JR, Sauro J. Item benchmarks for the system usability scale. J Usability Stud. 2018;13(3):158–67. https://doi.org/10.5555/3294033.3294037.
33. Luoma J, Rämä P. Comprehension of pictograms for variable message signs. Traffic Eng Control. 2001;42(2):53–8.
34. Markandeya M, Abeyratne U. 0438 snore sound analysis: within and beyond human hearing range. Sleep. 2017;40:A163–A163. https://doi.org/10.1093/sleepj/zsx050.437.
35. Martelaro N, Ju W. Woz way: Enabling real-time remote interaction prototyping & observation in on-road vehicles. In: Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing, CSCW '17, p. 169–182. Association for Computing Machinery, New York; 2017. https://doi.org/10.1145/2998181.2998293.
36. Motamedi S, Masrahi A, Bopp T, Wang JH. Different level automation technology acceptance: older adult driver opinion. Transport Res Part F Traffic Psychol Behav. 2021;80:1–13. https://doi.org/10.1016/j.trf.2021.03.010.
37. Nees MA, Helbein B, Porter A. Speech auditory alerts promote memory for alerted events in a video-simulated self-driving car ride. Hum Factors. 2016;58(3):416–26. https://doi.org/10.1177/0018720816629279.
38. Palazzi A, Abati D, Calderara S, Solera F, Cucchiara R. Predicting the driver's focus of attention: the DR(eye)VE Project. IEEE Trans Pattern Anal Mach Intell. 2018;41(7):1720–33.
39. Pettersson I, Ju W. Design techniques for exploring automotive interaction in the drive towards automation. DIS '17, p. 147–160. Association for Computing Machinery, New York; 2017. https://doi.org/10.1145/3064663.3064666.
40. Politis I, Brewster S, Pollick F. Speech tactons improve speech warnings for drivers. In: Proceedings of the 6th international conference on automotive user interfaces and interactive vehicular applications, AutomotiveUI '14, p. 1–8. Association for Computing Machinery, New York; 2014. https://doi.org/10.1145/2667317.2667318.
41. Politis I, Brewster S, Pollick F. 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, AutomotiveUI '15, p. 3–10. Association for Computing Machinery, New York; 2015. https://doi.org/10.1145/2799250.2799262.
42. Politis I, Langdon P, Adebayo D, Bradley M, Clarkson PJ, Skrypchuk L, Mouzakitis A, Eriksson, A, Brown JW, Revell K, Stanton N. An evaluation of inclusive dialogue-based interfaces for the takeover of control in autonomous cars. In: International conference on intelligent user interfaces, proceedings IUI (March), p. 601–606; 2018. https://doi.org/10.1145/3172944.3172990.
43. Porter MM, Irani P, Mondor TA. Effect of auditory road safety alerts on brake response times of younger and older male drivers: a simulator study. Transport Res Rec. 2008;2069(1):41–7.
44. Reynolds TJ, Gutman J. Laddering theory, method, analysis, and interpretation. J Advertis Res. 1988;28(1):11–31.
45. Rhiu I, Kwon S, Bahn S, Yun MH, Yu W. Research issues in smart vehicles and elderly drivers: a literature review. Int J Hum-Comput Interact. 2015;31(10):635–66. https://doi.org/10.1080/10447318.2015.1070540.
46. Richardson N, Lehmer C, Lienkamp M, Michel B. Conceptual design and evaluation of a human machine interface for highly automated truck driving. In: 2018 IEEE intelligent vehicles symposium (IV), 2018;2072–2077. https://doi.org/10.1109/IVS.2018.8500520.
47. Rödel C, Stadler S, Meschtscherjakov A, Tscheligi M. Towards autonomous cars: the effect of autonomy levels on acceptance and user experience. In: Proceedings of the 6th international conference on automotive user interfaces and interactive vehicular applications, AutomotiveUI '14, p. 1–8. Association for Computing Machinery, New York; 2014. https://doi.org/10.1145/2667317.2667330.
48. Rukonic L, Pungu Mwange MA, Kieffer S. UX design and evaluation of warning alerts for semi-autonomous cars with elderly drivers. In: HUCAPP 2021, 5th international conference on human computer interaction theory and applications, p. 25–36. Scitepress; 2021. https://doi.org/10.5220/0010237000250036.
49. SAE International: Taxonomy and definitions for terms related to driving automation systems for on-road motor vehicles. SAE Standard J3016 2021. https://doi.org/10.4271/J3016_202104.
50. Sauro J, Dumas JS. Comparison of three one-question, post-task usability questionnaires. In: Proceedings of the SIGCHI conference on human factors in computing systems, 2009;1599–1608.
51. Schmargendorf M, Schuller HM, Böhm P, Isemann D, Wolff C. Autonomous driving and the elderly: perceived risks and benefits. In: Dachselt R, Weber G (eds) Mensch und computer 2018—workshopband. Gesellschaft für Informatik e.V., Bonn; 2018. https://doi.org/10.18420/muc2018-ws11-0524.
52. Schrepp M, Hinderks A, Thomaschewski J. Applying the user experience questionnaire (UEQ) in different evaluation scenarios. In: International conference of design, user experience, and usability, p. 383–392. Springer, Cham; 2014. https://doi.org/10.1007/978-3-319-07668-3_37.
53. Schrepp M, Thomaschewski J, Hinderks A. Construction of a benchmark for the user experience questionnaire (UEQ). Int J

Interact Multimed Artif Intell. 2017;4(4):40–44. https://doi.org/10.9781/ijimai.2017.445.
54. Strömberg H, Bligård LO, Karlsson M. HMI of autonomous vehicles—more than meets the eye. In Bagnara S, Tartaglia R, Albolino S, Alexander T, Fujita Y editors. Proceedings of the 20th congress of the international ergonomics association (IEA 2018), p. 359–368. Springer, Cham; 2019. https://doi.org/10.1007/978-3-319-96074-6_39.
55. Tullis T, Albert B. Measuring the user experience: collecting, analysing, and presenting usability metrics. Burlington: Morgan Kaufmann; 2013. https://doi.org/10.1016/B978-0-12-415781-1.00007-8.
56. United Nations, Department of Economic and Social Affairs, Population Division: World Population Ageing 2019: Highlights (ST/ESA/SER.A/430); 2019.
57. Wang M, Liao Y, Lyckvi SL, Chen F. How drivers respond to visual vs. auditory information in advisory traffic information systems. Behav Inf Technol. 2020;39(12):1308–19. https://doi.org/10.1080/0144929X.2019.1667439.
58. Wong PNY, Brumby DP, Babu HVR, Kobayashi K. Voices in self-driving cars should be assertive to more quickly grab a distracted driver's attention. In: Proceedings of the 11th international conference on automotive user interfaces and interactive vehicular applications, AutomotiveUI '19, p. 165–176. Association for Computing Machinery, New York; 2019. https://doi.org/10.1145/3342197.3344535.
59. Young KL, Koppe S, Charlton JL. Toward best practice in human machine interface design for older drivers: a review of current design guidelines. Accident Anal Prevent. 2017;106:460–7. https://doi.org/10.1016/j.aap.2016.06.010.
60. Zhou F, Yang XJ, Zhang X. Takeover transition in autonomous vehicles: a YouTube study. Int J Hum-Comput Interact. 2019;36(3):295–306. https://doi.org/10.1080/10447318.2019.1634317.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.