Proceedings of the 10th International ACM Conference on Automotive User Interfaces and Interactive Vehicular Applications (AutomotiveUI '18), September 23–25, 2018, Toronto, Canada .

Design Guidelines for Reliability Communication in Autonomous Vehicles

Sarah Faltaous[1,2], **Martin Baumann**[3], **Stefan Schneegass**[1], **Lewis L. Chuang**[2,4]

[1]University of Duisburg-Essen – HCI Group – {firstname.lastname}@uni-due.de
[2]Max Planck Institute for Biological Cybernetics – {firstname.lastname}@tuebingen.mpg.de
[3]Ulm University – {firstname.lastname}@tuebingen.mpg.de
[4]LMU Munich – {firstname.lastname}@um.ifi.lmu.de

ABSTRACT

Currently offered autonomous vehicles still require the human intervention. For instance, when the system fails to perform as expected or adapts to unanticipated situations. Given that reliability of autonomous systems can fluctuate across conditions, this work is a first step towards understanding how this information ought to be communicated to users. We conducted a user study to investigate the effect of communicating the system's reliability through a feedback bar. Subjective feedback was solicited from participants with questionnaires and semi-structured interviews. Based on the qualitative results, we derived guidelines that serve as a foundation for the design of how autonomous systems could provide continuous feedback on their reliability.

CCS Concepts

• **Human-centered computing** → **Human computer interaction (HCI); HCI design and evaluation methods; User studies; Laboratory experiments;**

Author Keywords

situation awareness; interface design; autonomous vehicles; reliability display; human-machine partnership.

INTRODUCTION

Recent advances in automotive technologies allow tasks that were previously performed by drivers to be carried out by a constellation of sensors and artificial intelligence [33] (e.g., autopilot). Most of these currently available technologies and most of the near future technologies require the user's intervention when the automation fails [6, 16, 38]. At the same time with these technologies the driver is granted features that can perform complex perception and decision-making tasks [3] (e.g. sign recognition). These features allow the driver to be engaged in different tasks other than driving [12]. Consequently, the driver becomes a supervisory operator [3, 28, 35], who is barely involved in the driving task [16] and mostly identified as being out-of-loop [3].

To date, much research has been performed on how to bring the driver back into the loop when vehicle automation encounters a situation that it cannot handle safely, that is if it reaches a system limit situation [8, 17]. Such research implicit assume that autonomous vehicle systems are entirely reliable partners; they can either handle a situation or are able to recognize when they are not, subsequently issuing a notification to introduce the driver back into the loop to handle the situation instead. In reality life, vehicle automation is more likely to operate on the basis of probabilistic certainty. In other words, their reliability will vary in accordance with their changing levels of certainty in coping with driving conditions that can be expected to vary continuously. Variability in an automated vehicle's reliability could result from at least two factors. First, it could be related to technological limitations (e.g., sensors' imprecise input registration, imperfections of the signal transmission in electronic circuits, analogue to digital conversions, etc) [20]. Second, it could result from the unpredictable behaviour of fellow drivers (e.g., sudden lane changing) [9]. These factors will undoubtedly influence the reliability (or certainty) of vehicle automation in safe vehicle handling. To some extent, automated vehicles will have access to understanding their current levels of reliability. For example, inputs such as the accident statistics of a given road or the confidence bounds of sensor readings. Therefore, users could benefit from being informed of an autonomous vehicle's current level of reliability. This could allow users to adapt their supervision strategy and level of secondary task engagement, so as to be able to intervene should vehicle automation fail. To the best of our knowledge, the topic of reliability communication in autonomous vehicles has yet to be investigated in depth.

In this paper, we explore how autonomous vehicles can communicate it's current level of system reliability to the

AutomotiveUI '18, September 23–25, 2018, Toronto, ON, Canada

ISBN 978-1-4503-5946-7/18/09…$15.00

DOI: https://doi.org/10.1145/3239060.3239072

user. We investigate how users might perceive the usefulness of such a communication system, using a simulation scenario that focuses on an automated collision avoidance system. The level of automation simulated could be loosely described as SAE L2/L3 given that our participants were expected to perform a primary non-driving task but were expected to intervene when automation failed. We developed a continuous visual feedback system that communicated the levels of automation reliability in order to understand how users responded to such communication. Through implementing it in a realistic driving simulation scenario, we elicited user feedback via questionnaires, and semi-structured interviews, which serve as a basis for general guidelines and recommendations for communicating a system's reliability to users of autonomous vehicle systems.

RELATED WORK

In the real world, no system is perfect. It means that any system can and will fail at some point. As it is not feasible for the engineering of automated systems to exhaustively consider the infinite possibilities that could transpire across different use scenarios, designing for effective human intervention will continue to be a crucial component of automated systems [37]. To understand how humans interact with automated systems (i.e., autonomous vehicles), it is necessary to first consider how user trust engenders reliance on automation.

Trust and Reliance in Autonomous Systems

Reliance is categorized as a behavior, while trust is an attitude [5] carried out by the human towards automation [36]. Both are different sides of the same coin, which can determine how users interact with a given system [32]. User trust in a Human-Machine Interaction (HMI) can be considered by comparisons to a Human-Human Interaction (HHI) [14, 25, 30]. To realize a certain task [36, 5, 41], for example, driving safely [16], both the human and the autonomous vehicle should form a team [14].

Muir stated that trust in HHI and HMI is comparable [31], in terms of predictability (i.e., action anticipation), dependability (i.e., action execution consistency) and faith (i.e., system's reliability) [25, 36]. For instance, whenever a driver turns an automatic parking feature on, he or she knows that the final target is entering the selected parking spot. The driver also knows that the car is going to be performing a specific sequence of maneuvers to achieve that. By using this feature, drivers entrust the system to do this task on their behalf.

Both the autonomous system and the user possess abilities and knowledge that the other may not. Therefore, each of them receives the changes occurring in the driving environment with different variance [39]. Drivers are more likely to rely on automated aid if they trust it to perform more reliably than their own performance (e.g. back-parking sensors in estimating the distance to collide). However, the converse could also be true. If they believe that their performance is likely to be superior, given their experience with this automated feature, they are unlikely to rely on automation [14, 30]. In both cases, they cautiously choose whom to rely on and trust (e.g., their performance vs. the system's) [14].

Reliability Communication

In a changeable precarious world, alerting systems and decision aids can be expected to fail in providing correct feedback. This leaves room for uncertainty and risk [22]. Flawed alerting systems may result in two different types of errors [27, 40, 42]. First of all, errors that result from an over-reliance on automation, which results in a lack of driver attention to the road and how vehicle automation is functioning. Consequently, this could prevent the drive from intervening appropriately [40]. A commonplace and relatively benign example would be when the driver misses to exit the highway as a result of over-reliance on a navigation system that does not update reliably [42]. Another type of error occurs from over-compliance, that is when the operator accepts the false alarms and recommendations of an automated system. For instance, entering a one-way road from the wrong direction based on the navigation system recommendations [42]. In both cases, the consequence of system error is problematic. However, it has been noticed that the latter is more grievous, as it eventually leads to total ignorance of the alert [40].

In general, people tend to invest more trust in autonomous systems that show more collaboration in a helpful and understandable manner, regardless of their actual level of reliability [5]. Dzindolet and colleagues showed that people did not refrain from using an unreliable system even though they were aware of its unreliability [13]. Another study reported similar findings when user interventions during system failures, varied with presenting alarm rates of changing reliabilities (e.g., 20, 25 and 70% true alarms). As the researchers showed that most of the participants (90%) response rates matched that of the expected probability of true alarms (e.g., probability matching) [7].

Communicating the changing reliability of an automated system can be expected to improve user interactions with automated vehicles, if users are able to accept and effectively interpret this information. In order for users to rely on autonomous vehicles, it is important for them to continuously moderate their levels of reliance and allocate their available resources according to the changing circumstances and capability of vehicle automation [5, 39]. The work of Parasuraman and colleagues [31] suggests that users are sensitive to variable feedback by evaluating the reaction times of system failure detections. Specifically, users were faster to respond to system failures when confronted with variable feedback as opposed to constant feedback. These results agree with those reported by Helldin and colleagues who showed that participants were able to shift to manual control faster when presented with variable feedback compared to when no feedback was issued [21].

Figure 1: Experiment setup with the feedback bar mounted on top of the tablet showing the non-driving task.

It has been suggested that good automated system designs ought to consider communicating possible error values [29]. Communicating the varying reliability of an automated system would engender more trust even if the system fails to perform as expected on occasion [14, 22]. Thus, it is observed that the driver's trust and the system's reliance is greatly influenced by the amount of data communicated through the system's feedback (e.g., accuracy) and the clarity and ease of understanding the presented cues (e.g., visual, auditory, or tactile) [30, 36].

USER STUDY

We designed a user study to examine how the changing reliability of autonomous vehicles could be continuously communicated to users engaged with a non-driving task. Specifically, we simulated use of an SAE level 3 autonomous vehicle whereby a user is still expected to intervene when the autonomous system does not operate as designed. Visual feedback was continuously provided to communicate the reliability levels of the system. To date, the design of reliability communication has not benefited from much investigation, especially in the context of autonomous vehicles given their commercial scarcity. Thus, the goal was to understand the subjective concerns of potential users and the conditions whereby feedback on system reliability would be appreciated. Pragmatically speaking, this contributes towards recommended guidelines for the subsequent designs of communicating system reliability.

Apparatus

The simulation consisted a driver seat with wheel and standard pedals, three adjacent displays, and a peripheral touch computing device (iPad2, iOS 9.3.5.) that was positioned to the right of the user (Figure 1). A customized driving scenario was programmed using Unity3D (5.0.0f3) game engine, which was presented on the display and received pedal inputs. The peripheral device independently presented the user with a standardized non-driving working memory task, which was the primary occupation of our participants. Most importantly, visual feedback on the system's reliability was communicated to the participants using a 30 LED strip that was horizontally positioned above the peripheral device. This LED strip was controlled with an Arduino UNO.

Driving Scenario

In the presented scenario, participants experienced sitting in an autonomous vehicle that travelled on the right lane of a straight two-lane road at a constant speed of $108km/hr$. A lead vehicle would appear occasionally, approximately every $35secs$, which travelled at a slower speed of $32km/hr$, prompting the ego-vehicle to automatically execute an overtaking manoeuvre on the left lane. This manoeuvre was delayed if the ego-vehicle detected another oncoming vehicle in the left lane, travelling at $108km/hr$ in the opposing direction. The ego-vehicle could experience two possible failures. It could either fail to detect an oncoming vehicle and perform an overtaking manoeuvre or it could falsely detect an oncoming vehicle when there was none and wait for $7secs$. During these failures, participants could intervene and override these erroneous actions by stepping on the brake or accelerator pedals respectively.

Participants experienced the same scenario twice, under two different levels of external visibility. The first level represented clear weather and permitted a visibility range of $995m$ ahead (Figure 2a). The second level represented foggy weather and permitted a visibility range of $200m$ ahead of the autonomous vehicle (Figure 2b).

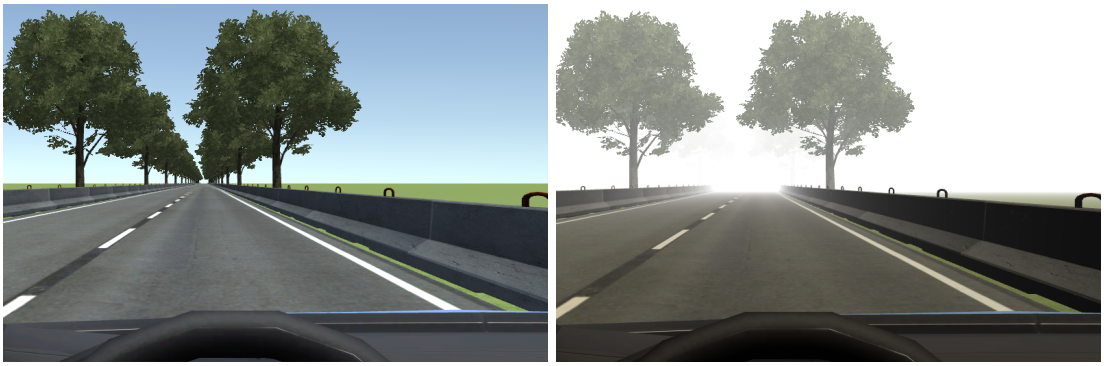
Non-driving Task

To simulate a cognitively engaging task that was similar to the potential activities that users of autonomous vehicles might perform, our participants were required to perform a working memory span task on the peripheral computing device [11]. Specifically, each participant was presented mathematical operations that were either correct (e.g., 2+16=18) or wrong (e.g., 7+11=12), which they had to respond by tapping a "Yes" or "No" button respectively. A letter of the Latin alphabet was presented for approximately $1sec$ above the presented operation. After three operations were presented, participants were required to recall the three letters that were presented.

Reliability Communication

For the reliability communication, we chose a half meter 30 LED RGB strip to be connected to Arduino UNO micro-controller programmed using Arduino. The strip was connected to an external 5V power source and was placed horizontally just above the tablet used in the non-driving task. This allowed the user to see both the tablet's display and the feedback all in one frame (Figure:1).

The LED strip that we used to communicate the reliability values, with one color at a time, a color gradient from red to green. The hue values for these colors ranged from 0 to 100. Each different color shade presented a reliability level, consequently, an 80% reliability would represent a yellowish green color (i.e., hue value of 80). As we wanted to know if a certain displayed color range would affect the perceived reliability, we separated the



(a) Non-foggy condition with a visibility range of $995m$. (b) Foggy condition with a visibility range of $200m$.

Figure 2: The two weather conditions, which permit different visibility ranges ahead of the autonomous vehicles

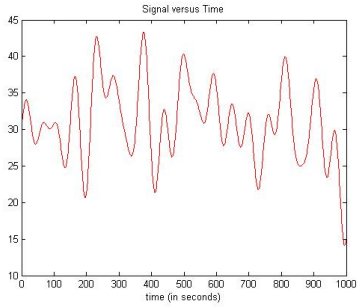


Figure 3: Sample of the waves generated from using the mean of seven sinusoidal signals across $1000s$. The values range from 10 to 50.

communicated reliability values into two levels. The first level is with low values ranging between 10% to 50% (i.e., red to yellowish orange). The second level is with high values ranging between 50% to 90% (i.e., yellowish orange to green). The displayed value (i.e., color) was updated each $1s$. This was implemented to match the real-world continuous uncertainty values change.

The presented values were generated from an algorithm, in which we calculate the mean value of seven sinusoidal waves having different phases and frequencies. We used prime numbers for phase shifts and frequencies to avoid any harmonic wave generation (Figure 3). We implemented the signal approach to avoid any sudden jump if a random representation was picked and to give the users a sense of the system's reliability in a consistent smooth way. It is important to mention that the displayed values were not linked to the real system's accuracy. The goal was just to detect whether or not the drivers would be affected by the reliability communication, as well as examining a new design approach to do so.

Study Design

We designed two experimental simulation blocks, in which we displayed the two weather conditions separately. We wanted to further examine the effect of communicating different color ranges in the reliability feedback. We divided each block into five intervals, where we manipulated the number of presented encounters and the displayed reliability values. Two of which had the same schema, in which we displayed 10 encounters each, with 5 encounters having an oncoming vehicle on the opposite lane. Starting from the beginning till the end of each of these intervals, the reliability feedback bar displayed either high range or low range values. We chose the system's accuracy to be 70% in these intervals, consequently, out of the 10 encounters 3 would be of failing. To avoid any fixed structure that the participant might notice, we added separating intervals, in which random number of encounters between 1 and 5 is presented with randomly selected scenarios. The LED strip in these intervals showed the corresponding accuracy value. Meaning, if the upcoming scenario reflects a system failure, the accuracy bar displayed low range values ranging from 10% to 50% of the predetermined hue gradients. Similarly, if the upcoming scenario was of a succession overtake, the high range values ranging from 50% to 90% were displayed as discussed in the feedback visualization design (Figure 4).

Participants and Procedure

Twenty participants (6 female and 14 male) participated in this study and were remunerated $8euros/hr$. They had different nationalities (i.e., 2 Americans, 7 Germans, 5 Italians, 3 Egyptians, 2 Dutch, and 1 Brazilian). Their ages ranged from 19 to 57 years old ($M = 29.65$, $SD = 8.86$) and all participants had a valid driving license for at least 2 months.

The study took approximately $90mins$ to complete and consisted of three parts: familiarization, use scenario, interview.. Before this, all the participants read a study description and explained the study to the experiment in their own words. They were corrected for any misunderstanding prior to providing signed consent.

In this user study, the participants were mostly required to perform a non-driving task. In addition, the participants were also required to monitor the performance of the highly autonomous driving system and to intervene when the system makes a wrong decision. The

Figure 4: Block design that starts with representing random scenarios while communicating the right reliability values. In between, the reliability communication is once set to be constant high or constant low with reliability level of 70%.

first part was training, where they started with the span task. Each participant was given a 3*mins* training, just to get familiarized with the task. Later, two training blocks of the non-foggy condition, five trials each, were presented on the simulator. The blocks were designed to give the participants the sense of how the accuracy bar works. In that context, one training block with a constant high accuracy value 80% was presented, displaying a scenario with one case of system failure out of the five trials. The other training block was with a low accuracy value 20%, displaying four failures of different scenarios out of the five trials. Both training blocks were with non-foggy condition. After finishing the training part, we presented the two different experimental blocks, which depict the two weather conditions. All the presented blocks were counter-balanced, to avoid any effect that might be caused by presenting a certain sequence. After being done with the two experimental blocks, we interviewed the participants and asked them to fill in a questionnaire.

RESULTS

Our study focused on the participants' assessment of the communicated reliability. To do so, we gathered their feedback using post-experiment questionnaires and conducted semi-structured interviews, to give us an insight into what users would be expecting from a system's feedback.

In the questionnaire, we first asked the participants to indicate on a Likert item from 1 (not much) to 5 (a lot) to which extent they trusted the system and to which extent they were pleased with the system's performance. Almost all the participants chose the same value for the two attributes (i.e., median=3). After that, we asked them if the system's performance improved block-wise. The main aim behind this question was to detect the effect of communicating reliability values differently (i.e., different color ranges). Although most of them (90%) indicated that no difference was noticed in the car performance, they mentioned that they trusted the car more when the feedback bar was displaying the *"reddish range"* as they called it.

In the interview, the participants were first asked whether or not they think that reliability values should be communicated to the users in autonomous systems. All of them (100%) said yes, with relating the reasons to one of these two aspects. The first aspect is safety, as some of them thought that it is safer to be given the choice to take over if the system is not quite sure about its performance. Therefore, preparing the driver beforehand to a potential Take Over Request (TOR) through this call for attention is more assuring. The second aspect relates more to the trust issue, as highlighted by some participants, that in order for them to trust the system, there has to be a clear communication that reflects the system's state.

When we asked the participants about in which circumstances they would be expecting the system to communicate the reliability values, most of them (70%) indicated that it should be communicated in risky situations. These are the situations where the system recognizes a failure or as described by one of the participants *" when the situation is on the verge of being dangerous"*. These answers were further supported by their definitions of the reliability communication. As they all indicated that the reliability communication is used to indicate risky situations in case of *"doubtful"* performance as referred to by one of the participants.

Regarding the design aspect, they all (100%) suggested different designs using various modalities. However, two main points were the focus of the majority. The first point is the importance of using a multi-modal feedback design. They reflected that it is important for them to be able to differentiate between a TOR and a continuous feedback about the system's state. The second point is the importance of using any modality other than visual communication. They justified their answer by saying that in real life, as well as in the current experiment, they would be visually *"over-whelmed"* by other tasks.

Afterwards, we focused on the result of the current design. Concerning the displayed colors, most the participants (70%) said that it was hard for them to notice the difference between the various shades. For them, it was either *"red or yellow or green"*. They even translated this three colours to three states. The red reflected an urgent take-over request, the yellow as a warning about a possible failure and consequently an attention request and the green as all good. They cited their inability of

perceiving the various colors to the high workload of the non-driving task that required most of their attention. They also mentioned that the high update rate of the displayed colours confused them about the communicated value.

In an attempt to know what would be a suitable update rate for them. They pointed that it's a situation dependent value. All of them said that it definitely needs to be in a real-time continuous form in risky situations, that is predefined by a specific reliability threshold value. In general, as they stated before it is good to have this clear communication. However, *"Not much, otherwise I will not trust the car at all"* as one of the participants said.

Second, positioning the LED strip over the tablet didn't serve the purpose of easy-reach communication as we expected. One participant said: *"It was easier to shift my eyes directly to the road if I'm to shift my eyes anyway"*. Another comment was: *"the bar position was on a different level not where I am focusing. It was an unreliable non-noticeable thing. It could be embedded in the wind-shield for example"*.

The last point that we were interested in knowing, was their opinion about granting the driver the option of turning the automation off. Some of them answered as per their own personal preference, it is a direct yes. As they would first need time to trust the system before they totally hand it the ultimate control. Second, they would like to feel superior to the system, in other words, controlling it not being controlled by it. For the general good, their answer was no, the automation feature should never be turned off. From the legal aspect, who would be the person to blame in fatal failing scenarios (i.e., an accident in case of off collision avoidance feature in a fully autonomous vehicle.). Another aspect is the fairness of the driving performance under different circumstances (i.e., reaction time in manual driving versus autonomous system).

IMPLICATIONS

User experience and acceptance.

When asked about the liberty of automation use and disuse, most participants indicated that they want to be given the chance to either turn it on or off., as they don't like to be forced to use a new technology, especially if they don't have any previous experience with it. They further elaborated that they lacked experience with system behaviour and the provided feedback. Therefore, they required more time before they were ready to grant the automated system "ultimate control of their lives". This agrees with what Dzindolet et al. discussed, namely that distrust in automation can arise from a lack of experience and understanding of how a given system behaves and operates. This can lead to system disuse, particularly when system failures occur that the operator cannot explain [14]. Based on that, we recommend applying one of the following approaches: (a) Providing tutorial programmes that include multiple scenarios that users are already familiar with in order to increase the intuitiveness of the communication display. This could familiarise targeted users with new technologies, (b) Monitoring a training phase whenever a new technology is offered. In this regard, users will have to use the technology in an incremental scale (e.g., using it for 20% of the time in the first week, 40% in the second week .... etc.). In brief, applying any training approach is highly favoured as it would grant the users (i.e., old generations) more familiarity with an automated system's performance and interface. As for new generations, part of passing a driving test could be examining their awareness with the autonomous system feedback.

Although the presented research showed the importance of communicating the autonomous' system reliability value, the question whether the car manufacturers would want to communicate the system's reliability at all still needs to be investigated. Communicating the reliability values would highlight the system's limitations, which could also affect the users' trust in the system and in a specific manufacturer. As a result, the use (i.e., buying) or disuse of such systems might highly depend on the user acceptance of such a new system. This puts the car manufacturers at risk, in-case of being the first to offer. Furthermore, it induces high pressure, in-case of competing with the other manufacturers, to achieve the best performance credibility.

Categorical perception of reliability levels.

Participants' feedback from the interview revealed that they perceived far fewer discrete levels of system reliability than the full spectrum that we strove to communicate, given the availability of the color levels of our visual display. In other words, they perceived color ranges categorically instead of continuously. In fact, research in color perception has demonstrated that our visual system separates colors into discrete categories rather as a continuous function of luminance [18]. Nonetheless, system reliability is a continuous parameter that automated systems compute and can provide as output. Therefore, more research is necessary to bridge this continuous measure of system reliability that is available to the categorical perception of what humans can perceive in the first place. Two options are available that are not mutually exclusive: (a) the use of a display parameter that humans can perceive continuously, (b) assessing the resolution of system reliability that users are able to understand and appreciate being informed about. To sum up, we find that just because system reliability is available as a continuous and high-resolution measure does not mean that humans need to or can perceive this information.

Feedback modalities.

On the one hand, several observations of the users behaviour towards the study indicated that our participants preferred to have multi-modal communication in critical situations that required their immediate attention. On the other hand, our participants also mentioned that they preferred a single modality for reliability communication and opined that visual communication was the

least "annoying" display modality, particularly for long journeys. A separate early study arrived at a similar conclusion [26], which investigated driver's workload variability when presented with warning information across different modalities (visual vs. auditory vs. multi-modal). Their result, based on the SWAT workload assessment, indicated that the perceived workload was least in the multi-modal, with a higher value in the auditory and the highest value was scored in the visual (i.e., 4.96 vs. 5.00 vs. 5.73 respectively.). Burke et al. [10], showed similar conclusion when they examined the driver's performance presented visual-tactile (VT) vs. visual only (V) and visual-auditory (VA) vs. visual only (V). They demonstrated that the performance was improved in the two examined multi-modalities compared to visual only (For effect size g in VT-V = 0.38 and g in VA-V = 0.42). These findings are further supported with an international survey carried out by Bazinlinsky et al. [4], as they showed that the participants' preference in the used TOR (Take-Over Request) modality was related to the situation urgency. They highlighted that multi-modal TORs (i.e., auditory, visual and vibrotactile) were most favoured in the *high-urgency situations*. Hence, we suggest a multimodal design for TORs and visual communication for continuous system's reliability.

Adaptive to the non-driving task.

When asked about the different display modalities that could be employed to communicate system reliability, participants pointed out that the feedback modality should differ from that of the non-driving task. In a previous study [34], no significant effect was found relating the reaction times to the modality of the communicated feedback and that of the non-driving task. Nonetheless, they recommended further research on understanding potential interactions between situation urgency and the nature of the non-driving task (e.g., hands-free task). This goes in line with the findings of Gold and colleagues who showed that the take-over performance in an autonomous vehicle was affected by the non-driving task nature (e.g., high demanding cognitive task) and the situation severity [15]. More specifically, the cognitive demanding tasks hindered the take-over performance in the case of cognitive demanding take-over situations, more than the situations where a take-over was *easy well practiced*. Based on previous findings and the current feedback, we conclude that adapting the communicated feedback to the nature of the non-driving task—for example, in terms of its cognitive demands—could result in a better overall performance. However, we do not recommend adapting feedback modalities to situation severity, as this could lead to even more confusion from the users' perspective.

Reliability update rate.

Our user feedback highlighted that there should be a clearly communicated threshold that defines the criticality of the situation. Previous studies on autonomous systems have shown that reliability values depend on the probabilistic error values that can be computed from different approaches (e.g., stochastic reachability approach for separate time zones based on the traffic participants' behaviour [2]). A simple example is an autonomous vehicle that drives in the right lane and turning left in a two opposite-direction road [2]. On the one hand, if the stochastic reachable set indicated a close intersecting range to the planned path of the autonomous vehicle, the situation should be marked as unclear and, therefore, the system's actual reliability should have a high-reliability update rate that changes with the proximity range to this computed set. On the other hand, if the planned path is marked as clear and safe, a slower update rate should be communicated, which would be granted less attention. Therefore, we recommend investigating the minimum threshold, for different systems (e.g., active cruise control), after which the participants should have a continuous update rate of the system's reliability values.

LIMITATIONS

We acknowledge the following limitations to our work. To begin, we generated specific reliability values using an algorithm. Although we aimed at creating a realistic values, we based these values on the results of an algorithm and not on real world data. Thus, the reliability values communicated to the participants were not linked to the real system certainty. Nevertheless, we opted for using an algorithm to be able to show a wider range of reliability values without creating a simulation environment that was too complex, which could lead to further distraction in the user.

Next, we provided a single display of system reliability that is presumably aggregated automatically from multiple sensor inputs (e.g., LIDAR, vehicle-vehicle communications, stereo cameras, traffic statistics, etc). It could be possible that users might prefer multiple reliability displays, each related to a given sensor. Recent research has shown that users prefer automatic aggregation of multiple data sources with related uncertainty, but that this preference varies with the perceived importance of the information that is presented [19]. It is worth noting that single-sensor-single-indicator displays are associated with a higher mental workload that can be further compromised in situations that induce high state anxiety [1]. Therefore, attempts to introduce more reliability indicators for different sensors should be justified for their expected utility.

Finally, we used a highly controlled driving simulation with a specific driving situation. This was deliberately designed to minimize influences from the driving scene. However, due to the relatively fixed timing used in the study, participants could predict the next occurrence of an overtaking situation. This could lead to a reduced attention towards the reliability communication during the times in-between the overtaking situations. Also, the limited number and type of possible encountered scenarios (i.e., four) could have affected the amount of attention paid to and the judgement of importance of the reliability communication.

GUIDELINES

The current work has revealed some important user findings with regards to how human-users of automated vehicles might respond to the communication of system reliability. We summarise the results in terms of five guidelines as follows:

- **Guideline 1:** Providing training sessions to increase the user experience with the system behaviour and interface. As providing the chance for the old generation to gain experience in a controlled condition, would overcome the fear of modern technology use.
- **Guideline 2:** Minimalistic color design in visualising specific system states. As using distinctively different colors to represent various values would be noticeably perceived.
- **Guideline 3:** Employ multi-modal system feedback, whereas reliability is preferred as visual and TORs as haptic or auditory.
- **Guideline 4:** Adapting different feedback modalities according to the nature of the non-driving task. This is inline with the work of Large and colleagues, who showed that the drivers' posture in an autonomous vehicle is relevant to the nature of the non-driving task [24].
- **Guideline 5:** Reflecting the situation criticality through the communicated reliability values update-rate. In that sense, slow update rate would relate to a stable and certain driving performance. While, fast update rate would indicate an uncertain and risky performance.

CONCLUSION

The introduction of autonomous vehicles into our daily life routines will give rise to more design challenges. How can we design communication systems that allow human users that are no longer "in-the-loop" to be aware not only of the operational environment but also of the changing reliability of the systems that they rely on? In this paper, we conducted a user study to investigate the effect of using visual feedback in communicating the system's reliability values to the user. Our qualitative results (e.g., based on questionnaires and semi-structured interviews) imply some basic guidelines that serve as bases for reliability communication design for autonomous vehicles.

More empirical work should be conducted around these guidelines to evaluate their generalizability and robustness to the complex and ever-changing landscape of autonomous driving [23]. These findings provide new evidence and a base for future inspection concerning the reliability communication design in autonomous systems.

ACKNOWLEDGEMENTS

This work is financially supported by the DFG within project SFB-TRR161 C03.

REFERENCES

1. Jonathan Allsop, Rob Gray, Heinrich Bülthoff, and Lewis Chuang. 2017. Eye movement planning on Single-Sensor-Single-Indicator displays is vulnerable to user anxiety and cognitive load. *Journal of Eye Movement Research* 10, 5 (2017).
2. Matthias Althoff. 2010. *Reachability analysis and its application to the safety assessment of autonomous cars*. Ph.D. Dissertation. Technische Universität München.
3. J Elin Bahner, Anke-Dorothea Hüper, and Dietrich Manzey. 2008. Misuse of automated decision aids: Complacency, automation bias and the impact of training experience. *International Journal of Human-Computer Studies* 66, 9 (2008), 688–699.
4. P Bazilinskyy, SM Petermeijer, V Petrovych, D Dodou, and JCF De Winter. 2018. Take-over requests in highly automated driving: A crowdsourcing survey on auditory, vibrotactile, and visual displays. *Transportation Research Part F: Traffic Psychology and Behaviour* 56 (2018), 82–98.
5. Johannes Beller, Matthias Heesen, and Mark Vollrath. 2013. Improving the driver–automation interaction an approach using automation uncertainty. *Human Factors: The Journal of the Human Factors and Ergonomics Society* (2013), 0018720813482327.
6. Myra Blanco, Jon Atwood, Holland M. Vasquez, Tammy E. Trimble, Vikki L. Fitchett, Joshua Radlbeck, Gregory M. Fitch, Sheldon M. Russell, Charles A. Green, Brian Cullinane, and Justin F. Morgan. 2015. Human Factors Evaluation of Level 2 and Level 3 Automated Driving Concepts. *(Report No. DOT HS 812 182)* August (2015), 300. DOI: http://dx.doi.org/10.13140/RG.2.1.1874.7361
7. James P Bliss, Richard D Gilson, and John E Deaton. 1995. Human probability matching behaviour in response to alarms of varying reliability. *Ergonomics* 38, 11 (1995), 2300–2312.
8. Shadan Sadeghian Borojeni, Lewis Chuang, Wilko Heuten, and Susanne Boll. 2016. Assisting drivers with ambient take-over requests in highly automated driving. In *Proceedings of the 8th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*. ACM, 237–244.
9. Sebastian Brechtel, Tobias Gindele, and Rüdiger Dillmann. 2014. Probabilistic decision-making under uncertainty for autonomous driving using continuous POMDPs. In *Intelligent Transportation Systems (ITSC), 2014 IEEE 17th International Conference on*. IEEE, 392–399.
10. Jennifer L Burke, Matthew S Prewett, Ashley A Gray, Liuquin Yang, Frederick RB Stilson, Michael D Coovert, Linda R Elliot, and Elizabeth Redden. 2006. Comparing the effects of visual-auditory and visual-tactile feedback on user

performance: a meta-analysis. In *Proceedings of the 8th international conference on Multimodal interfaces*. ACM, 108–117.

11. Andrew R Conway, Michael J Kane, Michael F Bunting, D. Zach Hambrick, Oliver Wilhelm, and Randall W. Engle. 2005. Working memory span tasks : A methodological review and user ' s guide. *Psychonomic Bulletin & Review* 12, 5 (2005), 769–786. DOI:http://dx.doi.org/10.3758/BF03196772
12. Stephen R Dixon and Christopher D Wickens. 2006. Automation reliability in unmanned aerial vehicle control: A reliance-compliance model of automation dependence in high workload. *Human Factors* 48, 3 (2006), 474–486.
13. Mary T Dzindolet, Hall P Beck, and Linda G Pierce. 2000. *Encouraging human operators to appropriately rely on automated decision aids*. Technical Report. ARMY RESEARCH LAB FORT SILL OK HUMAN RESEARCH AND ENGINEERING DIR.
14. Mary T Dzindolet, Scott A Peterson, Regina A Pomranky, Linda G Pierce, and Hall P Beck. 2003. The role of trust in automation reliance. *International Journal of Human-Computer Studies* 58, 6 (2003), 697–718.
15. Christian Gold, Ilirjan Berisha, and Klaus Bengler. 2015. Utilization of drivetime–performing non-driving related tasks while driving highly automated. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 59. SAGE Publications Sage CA: Los Angeles, CA, 1666–1670.
16. Christian Gold, Daniel Damböck, Lutz Lorenz, and Klaus Bengler. 2013. "Take over!" How long does it take to get the driver back into the loop?. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 57. SAGE Publications Sage CA: Los Angeles, CA, 1938–1942.
17. Christian Gold, Moritz Körber, David Lechner, and Klaus Bengler. 2016. Taking over control from highly automated vehicles in complex traffic situations: the role of traffic density. *Human factors* 58, 4 (2016), 642–652.
18. Robert L Goldstone and Andrew T Hendrickson. 2010. Categorical perception. *Wiley Interdisciplinary Reviews: Cognitive Science* 1, 1 (2010), 69–78.
19. Miriam Greis, Emre Avci, Albrecht Schmidt, and Tonja Machulla. 2017. Increasing Users' Confidence in Uncertain Data by Aggregating Data from Multiple Sources. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM, 828–840.
20. Miriam Greis, Aditi Joshi, Ken Singer, Albrecht Schmidt, and Tonja Machulla. 2018. Uncertainty Visualization Influences how Humans Aggregate Discrepant Information. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 505.
21. Tove Helldin, Göran Falkman, Maria Riveiro, and Staffan Davidsson. 2013. Presenting system uncertainty in automotive UIs for supporting trust calibration in autonomous driving. In *Proceedings of the 5th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*. ACM, 210–217.
22. Kevin Anthony Hoff and Masooda Bashir. 2015. Trust in automation integrating empirical evidence on factors that influence trust. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 57, 3 (2015), 407–434.
23. Andrew L Kun, Susanne Boll, and Albrecht Schmidt. 2016. Shifting gears: User interfaces in the age of autonomous driving. *IEEE Pervasive Computing* 15, 1 (2016), 32–38.
24. David R Large, Gary Burnett, Andrew Morris, Arun Muthumani, and Rebecca Matthias. 2017. A longitudinal simulator study to explore drivers' behaviour during highly-automated driving. In *International Conference on Applied Human Factors and Ergonomics*. Springer, 583–594.
25. John D Lee and Katrina A See. 2004. Trust in automation: Designing for appropriate reliance. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 46, 1 (2004), 50–80.
26. Yung-Ching Liu. 2001. Comparative study of the effects of auditory, visual and multimodality displays on drivers' performance in advanced traveller information systems. *Ergonomics* 44, 4 (2001), 425–442.
27. Masha Maltz and David Shinar. 2007. Imperfect in-vehicle collision avoidance warning systems can aid distracted drivers. *Transportation research part F: traffic psychology and behaviour* 10, 4 (2007), 345–357.
28. Neville Moray and T Inagaki. 1999. Laboratory studies of trust between humans and machines in automated systems. *Transactions of the Institute of Measurement and Control* 21, 4-5 (1999), 203–211.
29. Donald A Norman. 1990. The'problem'with automation: inappropriate feedback and interaction, not'over-automation'. *Philosophical Transactions of the Royal Society of London B: Biological Sciences* 327, 1241 (1990), 585–593.
30. Raja Parasuraman and Christopher A Miller. 2004. Trust and etiquette in high-criticality automated systems. *Commun. ACM* 47, 4 (2004), 51–55.
31. Raja Parasuraman, Robert Molloy, and Indramani L Singh. 1993. Performance consequences of automation-induced'complacency'. *The International Journal of Aviation Psychology* 3, 1 (1993), 1–23.

32. Raja Parasuraman, Thomas B Sheridan, and Christopher D Wickens. 2008. Situation awareness, mental workload, and trust in automation: Viable, empirically supported cognitive engineering constructs. *Journal of Cognitive Engineering and Decision Making* 2, 2 (2008), 140–160.
33. Raja Parasuraman and Christopher D Wickens. 2008. Humans: Still vital after all these years of automation. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 50, 3 (2008), 511–520.
34. Sebastiaan Petermeijer, Fabian Doubek, and Joost de Winter. 2017. Driver response times to auditory, visual, and tactile take-over requests: A simulator study with 101 participants. In *Systems, Man, and Cybernetics (SMC), 2017 IEEE International Conference on*. IEEE, 1505–1510.
35. Nadine B Sarter, David D Woods, and Charles E Billings. 1997. Automation surprises. *Handbook of human factors and ergonomics* 2 (1997), 1926–1943.
36. Kristin E Schaefer, Jessie YC Chen, James L Szalma, and Peter A Hancock. 2016. A meta-analysis of factors influencing the development of trust in automation: Implications for understanding autonomy in future systems. *Human factors* 58, 3 (2016), 377–400.
37. Albrecht Schmidt and Thomas Herrmann. 2017. Intervention User Interfaces: A New Interaction Paradigm for Automated Systems. *interactions* 24, 5 (Aug. 2017), 40–45. DOI: http://dx.doi.org/10.1145/3121357
38. Society of Automotive Engineers. 2014. Automated driving levels of driving automation are defined in new SAE international standard J3016. (January 2014). [Online].
39. Robert D Sorkin, Barry H Kantowitz, and Susan C Kantowitz. 1988. Likelihood alarm displays. *Human Factors* 30, 4 (1988), 445–459.
40. Richard I Thackray and R MARK TOUCHSTONE. 1989. Effects of high visual taskload on the behaviours involved in complex monitoring. *Ergonomics* 32, 1 (1989), 27–38.
41. Guy H Walker, Neville A Stanton, and Paul Salmon. 2016. Trust in vehicle technology. *International journal of vehicle design* 70, 2 (2016), 157–182.
42. Christopher D Wickens, Benjamin A Clegg, Alex Z Vieane, and Angelia L Sebok. 2015. Complacency and automation bias in the use of imperfect automation. *Human Factors: The Journal of the Human Factors and Ergonomics Society* (2015), 0018720815581940.