Proceedings of the Human Factors and Ergonomics Society Europe Chapter 2022 Annual Conference

de Waard, Dick ; Fairclough, Stephen; Brookhuis, Karel; Manzey, Dietrich; Onnasch, Linda; Naumann, Anna; Wiczorek, Rebecca; Di Nocera, Francesco; Röttger, Stefan; Toffetti, Antonella

Proceedings of the Human Factors and Ergonomics Society Europe Chapter 2022 Annual Conference

Enhancing Safety Critical Performance

Human Factors and Ergonomics Society
EUROPE CHAPTER

Edited by
Dick de Waard, Stephen Fairclough, Karel Brookhuis, Dietrich Manzey, Linda Onnasch, Anna Naumann, Rebecca Wiczorek, Francesco Di Nocera, Stefan Röttger, and Antonella Toffetti
ISSN 2333-4959 (online)

Please refer to contributions as follows:
[Authors] (2022), [Title]. D. de Waard, S.H. Fairclough, K.A. Brookhuis, D. Manzey, L. Onnasch, A. Naumann, R. Wiczorek, F. Di Nocera, S. Röttger, and A. Toffetti (Eds.) (2022). Proceedings of the Human Factors and Ergonomics Society Europe Chapter 2022 Annual Conference (pp. pagenumbers). Downloaded from http://hfes-europe.org (ISSN 2333-4959)

Available as open access

Published by HFES

Contents

Pilot's representation of dynamic situation in aviation, toward a visuospatial anticipation span

Marianne Jarry, Colin Blättler, & Vincent Ferrari
Centre de Recherche de l'Ecole de l'Air, Salon-de-Provence, France

Abstract

In aeronautical context, operators are confronted with complex dynamic situations in which they must represent the spatial state of several events and their possible evolutions. For example, a pilot in dense air traffic must be able to anticipate multiple trajectories to avoid collision. However, there is no description in the literature of a visuospatial anticipation span, i.e., how many dynamic events is an individual able to anticipate simultaneously? The objective of this study is twofold: (1) to set up a research protocol that objectively measures a visuospatial anticipation span and (2) to determine its limit. This study is based on the *representational momentum* paradigm (Freyd & Finke, 1984), which classically measures the ability to anticipate the movement of a single target (Hubbard & Bharucha, 1988) or a scene (Blättler, Ferrari, Didierjean, & Marmèche, 2011). The originality of this study is that 21 participants had to recall the position of five targets moving simultaneously in different directions. The results show for the first time that the individual is able to anticipate the trajectory of five events. This study may guide future studies aiming at further understanding the limits of human anticipation and thus improving human-vehicle collaboration through the development of adaptive autonomous systems.

Introduction

In the aeronautical context, operators (air traffic controllers, pilots) are confronted with complex dynamic situations in which they must represent the spatiotemporal state of several events and anticipate their evolution. For example, a pilot can visualize nearby traffic using his Traffic Collision Avoidance System (TCAS; Figure 1) and a controller can access traffic on his Air Traffic Control (ATC) screen. These display systems have many dynamic elements representing an air traffic situation. Since these operators perform a multi-tasking activity, they do not focus their attention on all these dynamic elements at all times. However, it seems that the operators manage to maintain a global coherence of the situation in what is commonly called Situation Awareness (SA) (Endsley, 1995). This model is described with three levels, from perception to projection (i.e., anticipation). It is the elements perceived at level 1 that

In D. de Waard, S.H. Fairclough, K.A. Brookhuis, D. Manzey, L. Onnasch, A. Naumann, R. Wiczorek, F. Di Nocera, S. Röttger, and A. Toffetti (Eds.) (2022). Proceedings of the Human Factors and Ergonomics Society Europe Chapter 2022 Annual Conference. ISSN 2333-4959 (online). Available from http://hfes-europe.org

can be projected into level 3. However, the studies on visuospatial span highlight the limited ability to integrate information, and the number of elements that operators can project in their SA can therefore be investigated. It is widely agreed that the visuospatial span of short-term memory is about 6 (Kessels, van Zandvoort, Postma, Kappelle & de Haan, 2000), but this type of measurement does not incorporate the dynamic dimension that is essential in aeronautical situations. The aim of this study is to measure the limit of the number of dynamic elements that the individual can represent to him/herself: i.e., a visuospatial anticipation span.

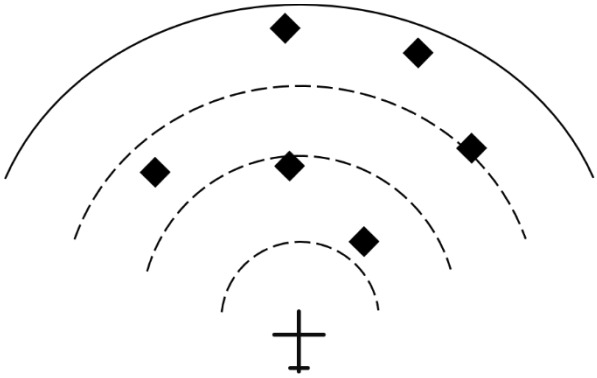


Figure 1. Simplified representation of the content of a Traffic Collision Avoidance System (TCAS) screen. The operator reading the screen is spatially symbolized by the aircraft at the bottom of the image. Each element (diamond) represents a moving object in the airspace near the aircraft. Each arc represents a distance in nautical miles from the aircraft (5, 10, 15 to 20 nautical miles).

Visuospatial anticipation has been studied for about 40 years in the Representational Momentum (RM) paradigm. RM is defined as the tendency of an individual to memorize the spatial position of a moving target or scene further away than it actually is (Freyd & Finke, 1984). Thus, when the individual observes a moving object, he or she represents its spatial position as shifted in time, i.e., ahead of time. When participants are asked to indicate the last spatial position of a moving target that disappears unexpectedly, they respond further than the actual position of disappearance (see Figure 2; Hubbard & Bharucha, 1988).

The authors explain these results by the momentum metaphor theory which suggests that the principle of momentum is incorporated in mental representations. Thus, mental representations would incorporate a component of inertia like that observed for moving objects. Indeed, a physical object in motion cannot be immediately stopped because of its momentum, in the same way a mental representation of this motion cannot be immediately stopped because of a similar momentum within the representation system. Thus, the processes underlying RM allow for the production of visuospatial anticipation. Integrating the measure of visuospatial anticipation into the model of classical SA provides an objective measure of the “projection”

component of this model. Although this study only addresses the question of how many dynamic elements an individual can anticipate, another advantage of choosing visuospatial anticipation as measured by RM study methodology is that these processes are involved in perception, action, and cognition at the same time. Approaching SA through the lens of visuospatial anticipation would allow for a very broad spectrum of SA projections to be addressed.

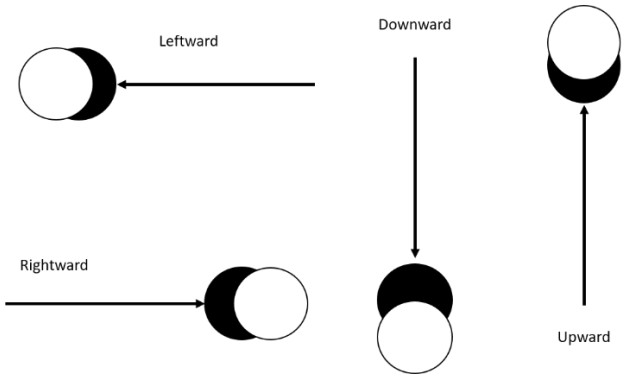


Figure 2. Materials and results adapted from Hubbard and Bharucha (1988). Target movements were horizontal leftward or rightward or vertical upward or downward. The black circles correspond to the disappearance points of the stimuli. The white circles correspond to the disappearance positions recalled by the participants on average.

Visuospatial anticipation has most often been studied through perceptual processes and attributed to a function of compensating for neural delays between retinal cell activation and the corresponding cortical activation (Nijhawan, 2002; Devalois & Devalois, 1991). Indeed, even if this delay may seem short (of the order of a hundred ms), the position of an object may have moved. Thus, an action towards this object may be delayed if no compensation is exercised. In this context, the speed of movement of an object has a direct impact on the extent of the visuospatial anticipation that is elaborated. The greater the speed of movement on the retina, the greater the extent of anticipation (Berry, Brivanlou, Jordan, & Meister, 1999). While this compensation is present at early stages of visual processing (see Hubbard 2005 for a review) other research shows that visuospatial anticipation appears at higher stages. Indeed, Hubbard and Bharucha (1988) showed that implicit knowledge of gravity force is integrated into visuospatial anticipation: a target moving downward (as if the target were falling) elicits a greater magnitude of visuospatial anticipation than if the movement is upward (see also Hubbard 2020 for a comprehensive review). Again, the implication of implicit knowledge of gravity is relevant to the interaction an individual has in, for example, grasping an object that is thrown at them (see also Hubbard, 2005 and 2006 for the influence of other types of implicit knowledge of physics such as centripetal and frictional force). Processes concerning action plans, involved during direct interaction with a moving object, have been shown to be

integral to the development of visuospatial anticipation. If an observer actively controls the target movement, then the magnitude of visuospatial anticipation is greater than if the observer passively observes the same target movement (Wexler & Klam, 2001; Blättler, Ferrari, Didierjean, & Marmèche, 2012). While the examples provided above can be considered to involve low-level processing, other research shows the involvement of higher-level processes such as conceptual knowledge (a drawing labeled "rocket" elicits greater visuospatial anticipation than if the same drawing is labelled "building", Vinson & Reed, 2002), allocation of attentional resources (the more resources are focused on the target, the less visuospatial anticipation there is, Hubbard, Kumar, & Carp, 2009) or expert knowledge (a plane landing scene seen from the pilot's point of view is anticipated more if the observer is an expert pilot than a novice, Blättler, Ferrari, Didierjean, & Marmèche, 2011).

In the aforementioned experiments and the rest of the literature regarding visuospatial anticipation, no research asks the question of how many elements the cognitive system can anticipate. Studies by Blättler et al. (2010, 2011, 2012) and Khoury, Blättler, and Fabre (2020) show that participants are able to produce visuospatial anticipation when viewing natural scenes (driving or flying from the driver/pilot's perspective). Although the scenes used have multiple elements, they all move in the same direction (approach). Finke, Freyd, and Shyi (1986) measured visuospatial anticipation of a dynamic pattern of three black dots with different directions. The induction was done by the successive presentation of 3 images. Afterwards, a fourth image was presented. The participants were then asked to answer whether "yes" or "no" the fourth image was similar to the third (Figure 3). The number of "yes" answers was higher when image 4 corresponded to the continuation of the induced movement. The authors thus observed visuospatial anticipation of the dynamics of the dot pattern. However, the authors did not assess the visuospatial representation of each dot. Thus, a strategy implemented by the participants could be to look at only one of them.

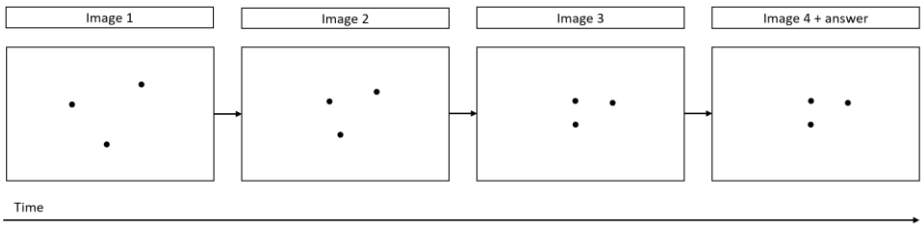


Figure 3. Material adapted from Finke, Freyd and Shyi (1986). Each image was presented successively to induce movement in the targets (image 1 to 3). Image 4 appeared and participants were asked to compare it to the previous image. Participants were then asked to answer whether "yes" or "no" image 4 and image 3 were the same. In this example, images 3 and 4 are the same.

The question remains, what happens when the individual observes elements moving in different directions? To address this shortcoming, in order to measure a visuospatial anticipation span, the present study is an adaptation of the Finke et al. (1986) protocol. Participants' response intake is modified to obtain a measure of anticipation for each element. The scene represents an aeronautical situation in the appearance of a TCAS. This display features a central triangle symbolizing the aircraft being flown and five dots representing various aircraft moving around the aircraft (no altitude information is presented as this is not a collision avoidance task). After a motion induction of the five aircraft, participants respond by pointing to the last seen position of the 5 targets on a touch screen in a specified order for each trial. Participants are expected to recall the last position of the aircrafts further along their own movement direction.

Method

Participants

A total of 21 students (21 male) in engineering school participated in the experiment (ages 21 to 23 years). Participants were recruited on a voluntary basis. They all gave their consent to participate and were told that they were free to stop the experiment at any time. All participants had normal or corrected vision.

Sample size was defined using a power test with RStudio power.t.test (power of test = 0.9, true difference mean = 0.45, standard deviation = 0.45 and significance level = 0.05).

Stimuli and apparatus

10 scenes were created on an image processing software. Each scene was composed of 5 targets (black dots of 24 pixels in diameter) in which a number is written (from 1 to 5; see Figure 4. The spatial distribution of the targets was different for each scene. The induction of movements was done by the sequential presentation of 3 images. From one image to the next each target moved 40 pixels in a straight line either on the vertical axis (up/down, down/up) or on the horizontal axis (left/right, right/left). A triangle of 62 pixels in height with a base of 58 pixels symbolized the aircraft being flown and remained stationary. The experiment was presented on a 19-inch square touch screen with a definition of 1080x720 pixels.

Procedure

Before the beginning of the experiment participant gave their consent and were told that they were free to stop the experiment at any time. Age and gender were collected for this experiment in an Excel spreadsheet along with the participant's number. They were also aware that their data were anonymised and will be kept confidential.

Participants were seated 60 cm away from the touchscreen. They were instructed to recall the last perceived position of the 5 targets using the touchscreen in a clockwise direction following the ascending order of the numbers written on each target (from 1 to 5; see Figure 4).

The experiment began with a familiarization phase (2 trials). One trial consisted of a motion induction phase of the 5 targets. The presentation time of the images was 250 ms. The interstimulus interval (ISI) was 250 ms. Each scene was presented 10 times to participants (10 times * 10 scenes = 100 trials). Once 5 responses (position of the target on the screen, x', y') were recorded, another trial was presented (Figure 5).

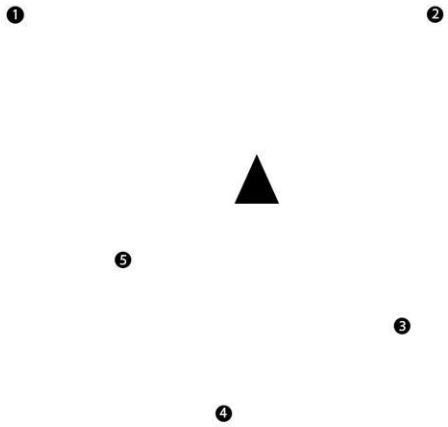


Figure 4. Example of an image presented during the experiment. The black triangle in the centre of the image symbolized the plane being flown. The black dots were the targets. A number from 1 to 5 was written in the centre of the dot.

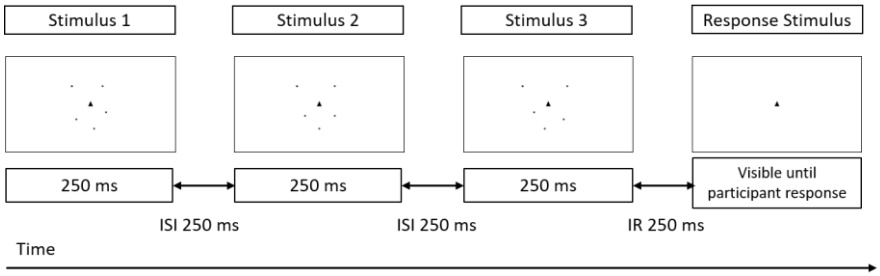


Figure 5. Diagram of a trial. Stimuli 1, 2 and 3 were used to induce a movement to each target (1 to 5). The images appeared sequentially for 250 ms. The interstimulus interval (ISI) was 250 ms. After a retention interval set at 250ms, the response image without the 5 targets appeared. Participants were then asked to point to the last perceived position of the targets (corresponding to stimulus 3) on the touch screen. Once 5 responses were recorded, another trial was presented.

Results

The difference between the vanishing point (x,y) and the recalled point (x',y') is called *displacement*. For each direction of motion, the *displacement* of the recalled point along the axis of motion is called VS-*displacement* (for Visuospatial *displacement*), and the *displacement* along the orthogonal axis, O-d*isplacement* (for Orthogonal-*displacement*) (see Figure 6).

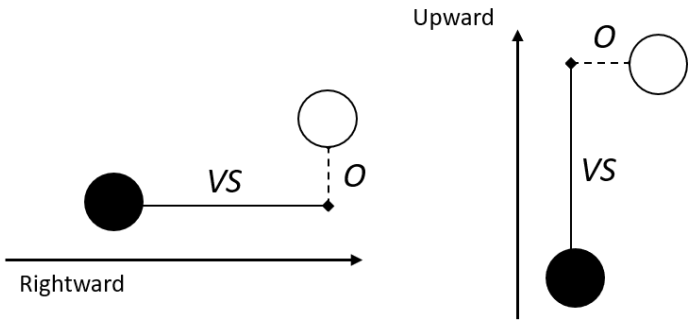


Figure 6. Example of a participant's response and the calculation of the measurements. The black circles correspond to the disappearance points of the targets. The white circles correspond to the points recalled by the participants. The displacement of the white circle along the continuous motion axis is VS-displacement and the displacement along the dashed orthogonal axis is O-displacement.

VS-displacement analysis

Thus, for each target and for each trial, a pixel value was obtained which could be positive or negative signifying the direction of the shift of the participants' visuospatial representation. Five means were calculated according to the prescribed recall number, from $M_{target1}$ to $M_{target5}$ (see representation of results Figure 7). These differences provide the magnitude of anticipation of the final position of each target. For each mean, if the value is significantly positive then a visuospatial anticipation is obtained.

Student's t tests were performed to assess whether the means were significantly different from 0. VS = + 33 pixels (*SD* = 7), $t(20) = 21.1$, $p < .001$. Tests were significant for targets 1, 2, 3, 4, and 5, $t(20) = 30.9$, $p < .001$ ($M_{target1}$ = + 42 px, *SD* = 6); $t(20) = 18.965$, $p < .001$ ($M_{target\ 2}$ = + 41 px, *SD* = 9); $t(20) = 17.437$, $p < .001$ ($M_{target3}$ = + 50 px, *SD* = 13); $t(20) = 7.45$, $p < .001$ ($M_{target4}$ = + 23 px, *SD* = 14); $t(20) = 4.122$, $p < .001$ ($M_{target5}$ = + 10 px, *SD* = 11), respectively.

A repeated measures One-Way ANOVA was conducted with target number (1-5) as a within-group factor to observe if there was an effect of response order on anticipation magnitude. The analysis revealed an effect of response order, $F(4,80)$ =

57, $p < .001$. Bonferroni post hoc tests revealed that the magnitude of anticipation for Target 1 is not different from Target 2 or Target 3; Target 1 vs. Target 2, $p = 1$; Target 1 vs. Target 3, $p = .072$. All other tests were significant; $p < .05$.

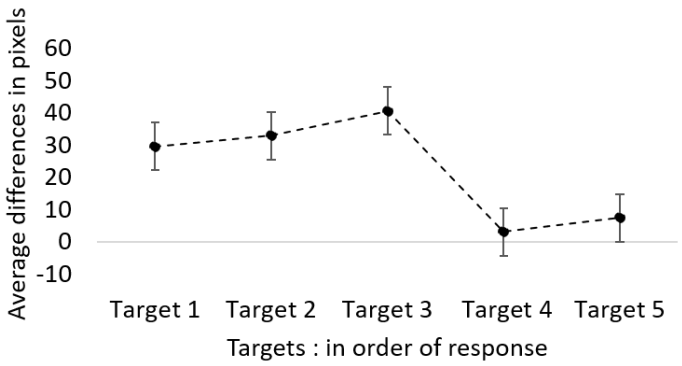


Figure 7. Graphical representation of the means of the differences between the vanishing position (at 0) and the position recalled by the participants as a function of the response order (target 1 to target 5). Variances are expressed as standard errors.

O-displacement analysis

A Student's t test was performed to assess whether the mean O-*displacement* (+78 px, $SD = 5$) was significantly different from zero. O-*displacement* is indeed different from zero, $t(20) = 63.147$, $p < .001$.

Discussion

The aim of this study was to measure the limit of the number of dynamic elements that the individual can anticipate simultaneously: i.e., a visuospatial anticipation span. For this purpose, participants were asked to recall the last position of 5 targets, with a proper motion, presented in the same scene. Participants were expected to recall the last position of the targets further in the direction of their own movements.

The results show that participants are able to anticipate a situation with five moving targets. Indeed, the overall average visuospatial anticipation of the five targets was significantly positive. This experiment allows for the first time to observe visuospatial anticipation for several elements, each of which has its own motion. Indeed, unlike Finke and Freyd (1985) and Finke et al. (1986) who evaluated a global pattern of three targets, here the visuospatial representations of each target were evaluated. Thus, the visuospatial anticipation span can be considered to be at least up to 5 items. However, a finer level of analysis puts this result into perspective.

The target-by-target analysis reveals that the magnitude of visuospatial anticipation changes with the recall order. Indeed, the magnitude of visuospatial anticipation increases until the third recalled target and then decreases rapidly. This evolution of the magnitude of visuospatial anticipation could be explained by the time that passes between the moment of target disappearance and recall. Indeed, there is a confounding variable between response time and recall order. The higher the recall number, the more the time to give the answer is important. This effect is classic in the RM literature (Freyd & Johnson, 1987; De Sá Teixeira, Kerzel & Lacquaniti, 2019). Freyd and Johnson (1987) show that visuospatial anticipation varies according to the retention interval (i.e., time between the last position seen and recall). These authors show that the magnitude of visuospatial anticipation increases up to 300 ms of IR and then decreases up to 900 ms. Thus, for these authors, this dynamic evolution of the visuospatial representation refers to the momentum metaphor. This metaphor implies that the inertia component, like the one observed for moving objects, is integrated in the visuospatial representation that participants make of the objects. This dynamic property of the visuospatial representation is therefore a *sine qua non* for showing that what is being measured is visuospatial anticipation. The pattern of results obtained in the study presented here seems to correspond to this dynamic property of the visuospatial representation. Thus, what is measured can be considered as visuospatial anticipation. Nevertheless, in this study the response times of the participants were not recorded, thus not allowing for comparative temporal analyses with previous studies (e.g., Freyd and Johnson, 1987).

Several limitations must be overcome in order to measure an anticipatory span (1) the consideration of *displacement* along the orthogonal axis, (2) the disappearance of anticipatory traces over time, and (3) the maximum number of elements that an individual is able to anticipate simultaneously. Regarding (1) it is observed here that the amount of O-*displacement* is larger than those observed in previous studies (e.g., Hubbard & Bharucha, 1988). Thus, the relationship between O-*displacement* and the number of moving elements presented in a scene should be checked in future studies. Regarding point (2), Jarry, Blättler and Ferrari (2022) have indeed shown that after a delay of 1125 ms visuospatial anticipation is no longer observed. These authors also show that after a longer delay of 2250 ms participants are behind "reality". As the method presented here asks participants to recall the position of elements in a precise order, the greater the number of elements to be recalled, the greater the recall time of the last elements. Thus, if recall occurs after 2250ms it may be that the anticipatory traces have faded because of time and not because of an exceeded span. Future studies should therefore disambiguate this point. Regarding (3), these initial results on visuospatial anticipation span are novel, but the numerical limit of visuospatial anticipation remains to be explored. Future studies will need to increase the numbers of dynamic features presented while carefully controlling the recall time in order to clearly establish a visuospatial anticipation span.

Once the visuospatial anticipation span is established, it will be possible to study its variations by the factors already identified in the visuospatial anticipation literature. The operationalization of visuospatial anticipation span assessment method will be an asset in the HMI optimization. For example, in the aviation environment, it has been observed that an experienced pilot anticipates more than a novice (Blättler et al., 2011). The same may be true regarding the visuospatial anticipation span. Indeed, studies on cognitive expertise show that the development of chunks or templates (e.g., Gobet and Simon, 1996) through the strategic links maintained by the elements present in a familiar scene considerably increases the number of elements that the individual can picture. Moreover, Ferrari, Didierjean and Marmèche (2006) have shown an anticipation component in the representations of chess experts. If chess experts do not perceive a physical movement of the pieces on the chessboard, they infer a strategic dynamism (i.e., the possible future moves from a perceived organization of pieces on the chessboard). It could be that pilots and air traffic controllers integrate visuospatial and strategic anticipation components to their representations. This research direction could reinforce the classical SA model of a generic anticipation structure based on integrative processes of perception, action and cognition.

The completion of the macroscopic model of SA by a finer understanding of the projection processes, in particular its spatiotemporal and numerical limits, will allow to improve initially the HMIs (e.g., TCAS and ATC) and the future interactions between human agents and artificial agents (e.g., autonomous vehicles).

References

Berry, M., Brivanlou, I., Jordan, T., & Meister, M. (1999). Anticipation of moving stimuli by the retina. *Nature* 398, 334–338. https://doi.org/10.1038/18678

Blättler, C., Ferrari, V., Didierjean, A., & Marmèche, E. (2011). Representational Momentum in Aviation. *Journal of Experimental Psychology: Human Perception and Performance*, 37, 1569–1577. https://doi.org/10.1037/a0023512

Blättler, C., Ferrari, V., Didierjean, A., van Elslande, P., & Marmèche, E. (2010). Can expertise modulate representational momentum? *Visual Cognition*, 18, 1253–1273. https://doi.org/10.1080/13506281003737119

Blättler, C., Ferrari, V., Didierjean, A., & Marmèche, E. (2012). Role of expertise and action in motion extrapolation from real road scenes. *Visual Cognition*, 20, 988–1001. https://doi.org/10.1080/13506285.2012.716799

De Sá Teixeira, N.A., Kerzel, D., Hecht, H., & Lacquaniti, F. (2019). A novel dissociation between representational momentum and representational gravity through response modality. *Psychological Research*, 83, 1223–1236. https://doi.org/10.1007/s00426-017-0949-4

De Valois, R.L., & De Valois, K.K. (1991). Vernier acuity with stationary moving Gabors. *Vision Research*, 31, 1619–1626. https://doi.org/10.1016/0042-6989(91)90138-U

Endsley, M.R. (1995). Toward a theory of situation awareness in dynamic systems. *Human Factors*, 37, 32–64. https://doi.org/10.1518/001872095779049543

Ferrari, V., Didierjean, A., & Marmeche, E. (2006). Dynamic perception in chess. *Quarterly Journal of Experimental Psychology*, 59, 397-410.

Finke, R.A., Freyd, J.J., & Shyi, G.C. (1986). Implied velocity and acceleration induce transformations of visual memory. *Journal of Experimental Psychology: General*, 115, 175.

Freyd, J.J., & Finke, R.A. (1984). Representational momentum. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 10, 126–132. https://doi.org/10.1037/0278-7393.10.1.126

Freyd, J.J., & Johnson, J.Q. (1987). Probing the Time Course of Representational Momentum. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 13, 259–268. https://doi.org/10.1037/0278-7393.13.2.259

Hubbard, T.L. (2005). Representational momentum and related displacements in spatial memory: A review of the findings. *Psychonomic Bulletin and Review*, 12, 822–851. https://doi.org/10.3758/BF03196775

Hubbard, T.L. (2006). Bridging the gap: Possible roles and contributions of representational momentum. *Psicologica*, 27, 1–34.

Hubbard, T.L., & Bharucha, J. J. (1988). Judged displacement in apparent vertical and horizontal motion. *Perception & Psychophysics*, 44, 211–221. https://doi.org/10.3758/BF03206290

Hubbard, T.L., Kumar, A. M., & Carp, C. L. (2009). Effects of Spatial Cueing on Representational Momentum. *Journal of Experimental Psychology: Learning Memory and Cognition*, 35, 666–677. https://doi.org/10.1037/a0014870

Jarry, M., Blättler, C., & Ferrari, V. (2022). Visuospatial Anticipation in Aeronautic: Human and Autonomous Vehicle System Interaction Optimisation. In *IHM Interaction Humain Machine 2022 proceedings.* ACM Association for Computing Machinery.

Kessels, R.P.C., van Zandvoort, M.J.E., Postma, A., Kappelle, L.J., & de Haan, E. H.F. (2000). The Corsi Block-Tapping Task: Standardization and normative data. *Applied Neuropsychology*, 7, 252–258. https://doi.org/10.1207/S15324826AN0704_8

Khoury, J., Blättler, C., & Fabre, L. (2020). Divided attention and visual anticipation in natural aviation scenes: The evaluation of pilot's experience. In D. de Waard, A. Toffetti, L. Pietrantoni, T. Franke, J-F. Petiot, C. Dumas, A. Botzer, L. Onnasch, I. Milleville, and F. Mars (2020). *Proceedings of the Human Factors and Ergonomics Society Europe Chapter 2019 Annual Conference* (pp. 15-28). http://hfes-europe.org.

Nijhawan, R. (2002). Review. In TRENDS in Cognitive Sciences (Vol. 6, Issue 9). http://tics.trends.com

Vinson, N.G., & Reed, C.L. (2002). Sources of object-specific effects in representational momentum. *Visual Cognition*, 9, 41–65. https://doi.org/10.1080/13506280143000313

How am I supposed to know? Conceptualization and first evaluation of a driver tutoring system for automated driving

Norah Neuhuber[1], Nikolai Ebinger[1], Paolo Pretto[1], & Bettina Kubicek[2]
[1]Virtual Vehicle Research GmbH, [2]University of Graz
Austria

Abstract

Drivers experience difficulties when interacting with automated driving systems. The need for driver training is generally acknowledged, but training is limited so far – leaving the driver to learn by "trial and error". We tested an Adaptive Tutoring System (ATS) in a driving simulator study. The ATS concept is based on prior research that includes a task analysis investigating the interaction with SAE level 2 systems, a re-analysis of thinking aloud data from a large-scale field study, and research on categorizing drivers. These in-depth analyses allowed us to define the tutoring content and construct a truly adaptive tutoring system. The ATS was designed to support drivers in learning how to calibrate their level of trust and reliance strategy to different driving contexts and system reliability levels. Two groups of participants drove in low- and high-risk scenarios, where one group received the tutoring (*tutoring group*), and the other group only written information (*baseline group*). Calibration of trust and reliance strategy were assessed by changes in subjective trust ratings, monitoring behaviour and system usage from low- to high-risk scenario. Results indicate that the ATS does support drivers to calibrate their interaction strategy to a changed driving context and system reliability.

Introduction and Previous Work

Vehicle automation is developing fast – in fact so fast that it is difficult for drivers to keep up with the pace of development. Current systems on the market are equipped with SAE (Society of Automotive Engineers) level 2 systems which support the driver in lateral and longitudinal task of driving (SAE, 2018). These new, often complex systems require the driver to learn a whole new set of skills and gain knowledge about a number of new topics (Heikoop, 2019). For years, human factors research has already pointed to the arising problems and risks which are associated with the introduction of vehicle automation (Endsley, 2017; Victor et al., 2018).

It is generally acknowledged that additional tutoring for advanced driver assistance systems is needed as most of the time drivers only receive a short introduction from the car dealer and are left to learn how to interact with these systems by "trial and

In D. de Waard, S.H. Fairclough, K.A. Brookhuis, D. Manzey, L. Onnasch, A. Naumann, R. Wiczorek, F. Di Nocera, S. Röttger, and A. Toffetti (Eds.) (2022). Proceedings of the Human Factors and Ergonomics Society Europe Chapter 2022 Annual Conference. ISSN 2333-4959 (online). Available from http://hfes-europe.org

error" (Boelhouwer, 2020b; Endsley, 2017). This is highly unacceptable given the potential risks resulting from this lack of training.

In response to this, researchers begin to investigate different driver training approaches which generally show positive effects on interaction performance (Boelhouwer et al., 2020a; Forster et al., 2019a; Payre et al., 2016). However, one critical skill will be to recognize when the system reaches its operational boundaries and to calibrate the level of trust and reliance strategy according to changes in system reliability (Lee, 2020). Research shows that drivers tend to do that in general (Kraus et al., 2019) but also, that this is not necessarily true for all drivers. Recent research suggests that distinctive categories of drivers are observable when it comes to this skill (Neuhuber et al., in press). Since this is such a critical skill and drivers seem to differ, it is important to adapt tutoring approaches to the respective category a driver can be assigned to. The tutoring system proposed in this study aims to specifically address this point.

Adaptive Tutoring System (ATS)

The ATS content was defined based on a task analysis describing the required knowledge and skills when interacting with SAE level 2 systems, a re-analysis of thinking aloud data from a large-scale field study (Neuhuber et al., 2020), and research on categorizing drivers based on their interaction strategy with advanced driver assistance systems (ADAS; Neuhuber et al., in press).

For the task analysis every stage of the interaction with a level 2 system was analysed. Results were the cumulative set of knowledge and skill requirements for drivers. These requirements were structured based on the theory of knowledge spaces (Heller, Steiner, Hockemeyer & Albert, 2006) to ensure a logical order of the tutoring content (figure 1).

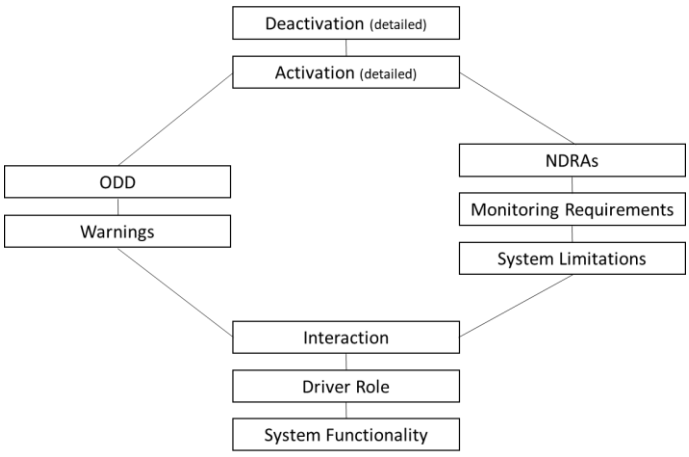


Figure 1. Knowledge-tree for the interaction with level 2 systems. NDRA = Non-Driving Related Activity, ODD = Operational Design Domain.

The scope of the study was limited so the final setup focused on a reduced number of knowledge and skill aspects. The tutoring omitted the topics of Operational Design Domain (ODD) and detailed content about warnings and system (de-)activation.

The analysis of the thinking aloud data highlighted which areas the tutoring system would need to focus on. The data is based on a field-study with 100 participants interacting with a commercially available level 2 system on a highway section. The overall results show that 28% of the participants experience some form of mode confusion. Almost half of the participants (45%) report to have difficulties to interact with the system. Most prominently, 63% of the participants made comments which suggested that they had not yet developed a correct mental model about the system functionalities and limitations. Therefore, the results highlight the need to focus on system functionality and limitations in the tutoring.

An additional aspect of the ATS is to provide adaptive tutoring lessons based on the driver's interaction strategy. This is based on previous research (Neuhuber et al., in press) which indicates that not all drivers calibrate the level of trust and reliance strategy to changes in situational risk and system reliability. Distinctive categories of drivers are observable in this regard (*under-trusting*, *over-trusting*, and *calibrated*). Under-trusting drivers tend to report low levels of trust and intensely monitor the system in low-risk situations. In contrast, over-trusting drivers show high levels of trust, tend to monitor the system less intense and are also hesitant to take-over manual control in high-risk driving situations. Calibrated drivers adapt their interaction strategy to the level of situational risk.

The initial prototype of the ATS consists of mainly three parts: i) a short video which is shown to the driver before the first interaction with the system; ii) short "reminders" to consolidate the learnt material; and iii) adaptive tutoring content which is based on a general categorization of drivers.

The ATS video combined verbal explanations to the defined topics with either schematic visualizations or short video clips (figure 2).

Figure 2. Exemplary screenshots of the ATS video. Schematic visualization of Lane Keeping Assist functionality (left) and video clip of ADAS activation process during driving (right).

The short reminders consisted of a short audio file triggered by the experimenter at the beginning of each drive. For the purpose of this study, the reminders focused on

the topic of system limitations, specifically focusing on the influence of harsh weather conditions on system reliability. The goal was to trigger a critical assessment of the situation by the participants.

Adaptive tutoring was triggered by the experimenter and was based on an assessment of the current level of trust, monitoring behaviour and whether or not participants took over manual control in relation to the current driving context (*low*- vs. *high-risk*). Participants were categorized into three main categories, "under-trusting", "over-trusting" or "calibrated/neutral" (see table 1). The specific content of the adaptive tutoring for each category is explained in table 2.

Table 1. Logic of driver categorization for low- and high-risk driving condition

	Low-risk			*High-risk*		
	Trust	*Monitoring*	*Take-over*	*Trust*	*Monitoring*	*Take-over*
Under-trusting	low	high	yes	-	-	-
Over-trusting	-	-	-	high	low	no
Calibrated / Neutral	high	low	no	low	high	yes

Table 2. Content of the adaptive tutoring for each driver category

Driver Category	*Content*
Under-trusting	Re-assurance that the system functions reliably under good weather conditions; need to re-assess the situation and decide whether intense monitoring and/or take-over of manual control is warranted.
Over-trusting	Same logic as for "under-trusting" category. Reminder that the system tends to display system limitations under harsh weather conditions; to re-assess the situation and decide whether reliance behaviour needs to be adapted.
Calibrated / Neutral	Re-assurance that the driver assesses the situation correctly.

Present study

In the present study it is investigated whether the ATS supports drivers in learning how to calibrate their level of trust and reliance strategy to different driving contexts and system reliability levels. Two groups are being compared - one group receives the ATS (*tutoring group*), the other group receives only written information (*baseline group*). It is hypothesized that the tutoring group is better able to perform the trust and reliance calibration process. It is particularly hypothesized that, from low- to high-risk driving scenario, the tutoring group shows a reduced level of trust (H1), increased monitoring (H2) and increased number of manual take-overs (H3) compared to the baseline group.

Method

Participants

A total of 20 participants (9 female), 10 in the baseline and 10 in the tutoring group, took part in the study. Participants were aged between 21 and 43 years (M = 27.75, SD = 6.21). On average, participants held their driving license for 8.3 years (SD = 6.17). All participants received a compensation of ten euros. Participants were pre-selected according to the following criteria: possession of a driving license for more than three years and no or very limited previous experience with advanced driving assistance systems. Table 3 provides mean values regarding age, driving experience and propensity to trust (assessed with the *propensity to trust* subscale, Körber, 2018) regarding the two experimental groups (*baseline*, *tutoring*). Potential group differences were checked with two-sample t-tests. No significant differences between the two groups were observable.

Table 3. Mean (M) and standard deviation (SD) for age, driving experience and propensity to trust for baseline and tutoring group

	N	*Age*		*Driving Experience*		*Propensity to Trust*	
		M	*SD*	*M*	*SD*	*M*	*SD*
Baseline	10	29.00	7.41	10.00	7.77	3.00	0.63
Tutoring	10	26.55	4.81	6.66	3.69	2.94	0.66

Experimental Design and Procedure

Participants received written instructions about the study and its goals and signed an informed consent at the beginning of the study. A short drive was undertaken to give participants the chance to familiarize with the driving simulator. Participants were randomly assigned to either the baseline or the tutoring group. Each participant drove in two experimental conditions that differed in the risk level that was present during the drive (*low-risk*, *high-risk*). Participants were asked to take a seat in the simulator were they also received either the written information (*baseline group*) or watched the tutoring video on a tablet mounted in the middle console (*tutoring group*).

A secondary task was introduced to divert drivers' attention between two tasks. Participants could watch YouTube videos on a tablet whenever they felt it was not necessary to direct attention to the automated system or the road. This task was chosen as i) it is a realistic task for drivers to engage in while using ADAS (Dunn et al., 2019), ii) it can be interrupted quickly, and iii) participants do not feel the urge to respond (like, e.g., when engaged in a SMS conversation). Subjects were instructed to drive on the left lane of a two-lane highway, speed up to 130km/h and then to activate the automated driving systems, i.e. Lane Keeping Assist (LKA) and Adaptive Cruise Control (ACC). Participants could take-over manual control whenever they deemed necessary. Each experimental drive lasted approximately 15 minutes followed by a short break and administration of a short set of questions.

Simulator Scenario

The same highway-section was used for both drives. Traffic was implemented on the right lane driving slower than 130km/h to motivate participants to drive on the left lane. The low-risk drive had good weather conditions with normal visibility and no system unreliabilities. In the high-risk drive, risk was operationalized by combining environmental factors (heavy rain and low visibility), system unreliability and a monetary incentive to perform well (in terms of driving safely) in this condition (instruction that half of the compensation depends on the performance within this experimental drive). The system unreliability was implemented as a failure to detect the lane markings correctly and therefore the vehicle swerved to the centre of the two-lane highway. The duration of the system unreliability was approximately 20 seconds, starting at 2:20 min, 4:40 min, 9:00 min and 12:50 min within the 15-minute drive (figure 3).

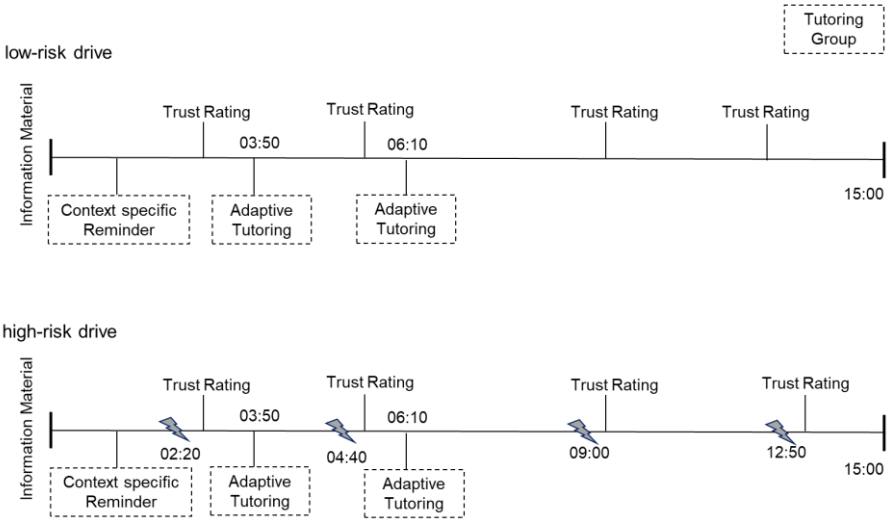


Figure 3. Experimental procedure for low- and high-risk drive. System failures are indicated by lightning bolts.

Information Material

The baseline group received written information which was similar to a driver manual describing the available systems in the vehicle. Participants in the tutoring received the ATS as described above. The content was the same for both groups.

Equipment and Data Processing

The study was conducted using a semi-static driving simulator from Vi-Grade operating with the VI-DriveSim software package. The Simulator is equipped with an automated driving system which keeps the lane and the set speed. Eye-Tracking data was collected using the DIKABLISGLASSES 3 and processed using the software D-

Lab and the statistic software R (R Core Team, 2013). As a secondary task the online video platform YouTube was used (videos were self-chosen by the participants). The application was presented on a 10.1-inch Android Tablet. The set-up is depicted in Figure 4.

Figure 4. Experimental set-up with exemplary participant wearing eye-tracking glasses and using the tablet mounted on the middle console.

Analysed Variables

Trust in automation

Trust was measured using a 3-item trust scale (adapted from Mayer et al., 1995). Answers were given on a scale ranging from 1 (*not at all*) to 7 (*completely*). The internal consistency measured with Cronbach's Alpha was high for both scenarios (α = .74 after *low-risk* scenario, and α = .80 after *high-risk* scenario). During the drive, single trust items (*How much do you trust the system?*) with a scale from 1 (*not at all*) to 7 (*completely*) were issued to assess the current level of trust. This assessment was only used for the categorization of drivers to trigger the adaptive feedback in the tutoring group.

Monitoring Behaviour

Results are reported regarding the percentage of time a participant spent monitoring the system behaviour (monitoring ratio). Fixations to the areas *street* and *dashboard* where combined to calculate an overall parameter.

Manual Take-Over

Results are reported regarding the total number of manual take-overs participants performed throughout the driving scenario.

Data Analysis

Data was analyzed using two-way ANOVAs with the factors condition (*low-risk, high-risk*) and group (*baseline, tutoring*). A significance level of .05 was used for all statistical tests.

Results

Trust

The analysis of the self-reported trust data revealed non-significant results, therefore hypothesis 1 was not supported. This was the case for the main effect between the experimental conditions (*low-risk* vs. *high risk*; $F(1,36) = 2.4$, $p = .129$, $\eta^2 = .06$) and for the main effect between the intervention groups (*baseline* vs. *tutoring*, $F(1,36) = 2.6$, $p = .113$, $\eta^2 = .06$). The interaction between the two factors indicates also a non-significant result ($F(1,36) = 0.4$, $p = .365$, $\eta^2 = .02$). A post-hoc power analysis revealed extremely low statistical power of .08 for detecting the small observed effect sizes. Mean values are shown in figure 5.

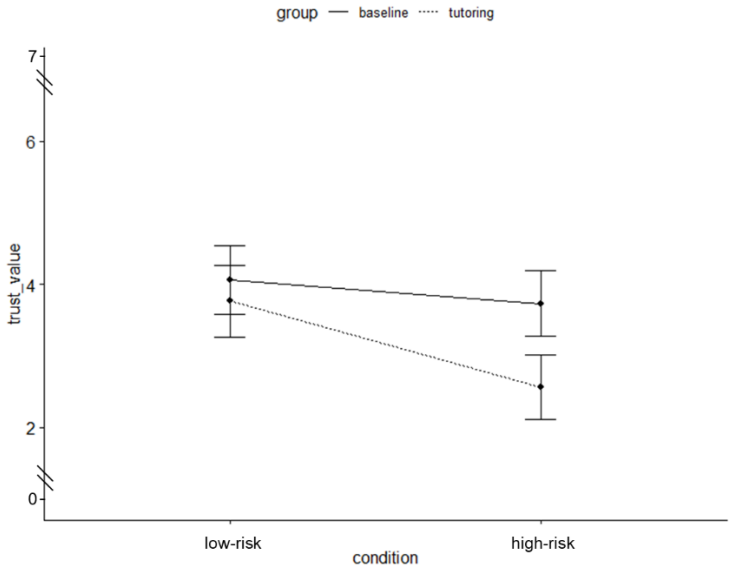


Figure 5. Mean trust values for baseline and tutoring group in the low- and high-risk driving condition. Error bars show Standard Error.

Monitoring Behaviour

The analysis regarding to monitoring behaviour partially support hypothesis 2. Results indicate a significant difference between the two driving scenarios regarding the amount of attention participants directed towards the system ($F(1,36) = 9.7$, $p = .004$, $\eta^2 = .17$). Participants generally increased the amount of monitoring from the low- to the high-risk driving scenario (figure 6). The analysis also shows a significant difference between the baseline and tutoring group ($F(1,36) = 10.1$, $p = .003$, $\eta^2 = .18$). Participants in the tutoring group generally showed an increased monitoring of the system compared to participants in the baseline group (figure 6). An interaction between the two factors is not observable ($F(1,36) = 1.1$, $p = .296$, $\eta^2 = .02$).

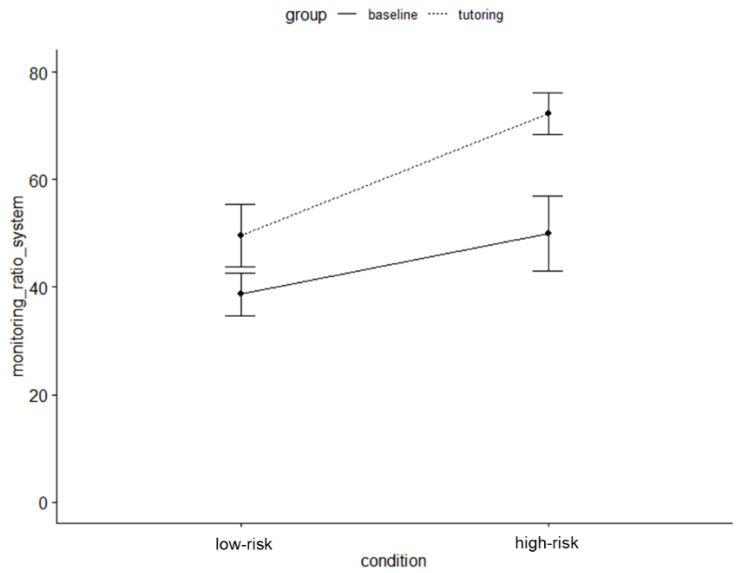


Figure 6. Mean monitoring ratio towards the system for baseline and tutoring group in the low- and high-risk driving condition. Error bars show Standard Error.

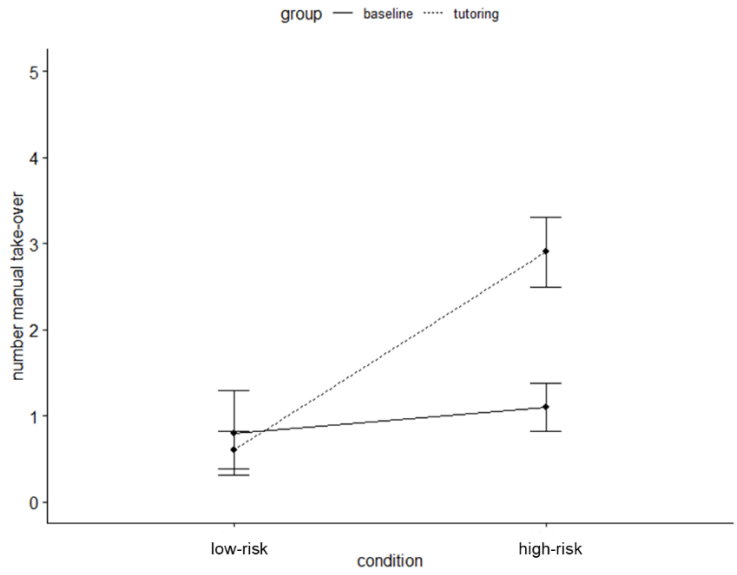


Figure 7. Mean number of manual take-overs for baseline and tutoring group in the low- and high-risk driving condition. Error bars show Standard Error.

Manual Take-Over

Results indicate a significant difference between the two driving scenarios regarding the number of manual take-overs ($F(1,36) = 4.8$, $p = .035$, $\eta^2 = .08$). Participants generally took over manual control more often in the *high-* compared to the *low-risk* driving scenario (figure 7). Furthermore, the tutoring group took over manual control more often compared to the baseline group. ($F(1,36) = 12.7$, $p = .001$, $\eta^2 = .21$). However, the results indicate a significant interaction between the factors condition and group ($F(1,36) = 7.5$, $p = .009$, $\eta^2 = .12$). The difference between the tutoring and the baseline groups becomes apparent within the high-risk drive as participants in the tutoring group tended to take-over manual control more often than participants in the baseline group, supporting hypothesis 3 (figure 7).

Discussion

The presented study investigates a first prototypical adaptive tutoring system for drivers interacting with semi-automated driving functions. Results indicate that the tutoring approach does help drivers to adapt their interaction strategy to changing driving context and changed system reliability – a skill which is crucial for a safe interaction (Lee, 2020). The tutoring group particularly showed increased monitoring towards the system and a higher tendency to take-over manual control within the high-risk driving scenario. However, this calibration process is not necessarily reflected in the self-reported levels of trust as no difference between the baseline and tutoring group and between the two experimental conditions (low- and high-risk) was observable.

The results of this study add to the previous scientific literature on the effect of driver tutoring by focusing on a truly adaptive approach which supports drivers in the calibration process (Boelhouwer et al., 2020a; Forster et al., 2019a; Payre, 2016). Similar to the study of Payre et al. (2016), it was also observable in this study that merely the amount of time interacting with an automated driving system already benefits drivers. Forster et al. (2019b) also report an increase of performance up until the fifth interaction with a system. The results of this study suggest that the ATS could speed up this process significantly. In contrast to the study by Forster et al. (2019a), the presented results support the assumption that a tutoring system is more effective than written information alone.

The tutoring seems to address particularly one problem which has been discussed previously: some drivers seem to be hesitant to take-over manual control in uncertain situations (Victor et al., 2018). Participants in the baseline group did increase the intensity with which they monitored the system in the high-risk driving scenario. This confirms previous studies reporting that drivers generally calibrate their interaction strategy when encountering system failures (Krause et al., 2019). However, at the same time participants in the baseline group were hesitant to take over manual control of the vehicle in situations where the system was clearly not functioning as intended. Compared to this, participants in the tutoring group took over manual control on average three times, almost matching the four occurrences of system failures in the scenario. These results suggest that the tutoring system successfully supported drivers in assessing changes in system reliability and consequently, to act accordingly.

Due to limited time in the study, the tutoring omitted other important topics, as for example regarding the Operational Design Domain (ODD) or a broader scope of potential system limitations (as for example when encountering construction sites or roundabouts). These topics would of course be necessary to allow the formation of a complete mental model and should be included in future studies. The limited sample size of the study introduces further limitations: First, in combination with the observed small effect size for the results on self-reported trust, the small sample size contributed to an extremely low statistical power of the analysis. The non-significant results could have stemmed from this limitation. Replications of this study are thus needed to determine the reliability of the non-significant results. Second, the small sample does not allow to generalize for a general population, but it can provide a basis to formulate new hypotheses and to inform future research. In conclusion, the results indicate that the ATS supports drivers in the calibration process and reduces uncertainty for drivers in how to act when system limitations occur. This effectively leads to a higher take-over readiness and over-all safer interactions.

Acknowledgements

This research has been funded by the Austrian Research Promotion Agency (FFG) (Grant No 27397222). The publication was written at the Virtual Vehicle Research GmbH in Graz and partially funded by the COMET K2 – Competence Centers for Excellent Technologies Program of the Federal Ministry for Transport, Innovation and Technology (bmvit), the Federal Ministry for Digital, Business and Enterprise (bmdw), the Austrian Research Promotion Agency (FFG), the Province of Styria and the Styrian Business Promotion Agency (SFG).

References

Boelhouwer, A., van den Beukel, A.P., van der Voort, M.C., Verwey, W.B., & Martens, M.H. (2020a). Supporting drivers of partially automated cars through an adaptive digital in-car tutor. *Information*, *11*(4), 185-207.

Boelhouwer, A., Van den Beukel, A.P., Van der Voort, M.C., Hottentot, C., De Wit, R.Q., & Martens, M.H. (2020b). How are car buyers and car sellers currently informed about ADAS? An investigation among drivers and car sellers in the Netherlands. *Transportation Research Interdisciplinary Perspectives*, *4*, 100103.

Dunn, N., Dingus, T., & Soccolich, S. (2019). *Understanding the impact of technology: Do advanced driver assistance and semi-automated vehicle systems lead to improper driving behavior* (Report 202-638-5944). Washington, DC, USA: AAA Foundation for Traffic Safety.

Endsley, M.R. (2017). From Here to Autonomy: Lessons Learned From Human–Automation Research. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *59*, 5–27.

Forster, Y., Hergeth, S., Naujoks, F., Krems, J., & Keinath, A. (2019a). User education in automated driving: Owner's manual and interactive tutorial support mental model formation and human-automation interaction. *Information*, *10*(4), 143-165.

Forster, Y., Hergeth, S., Naujoks, F., Beggiato, M., Krems, J.F., & Keinath, A. (2019b, June). Learning and Development of mental models during interactions with

driving automation: A simulator study. In *Proceedings of the Tenth International Driving Symposium on Human Factors in Driver Assessment, Training and Vehicle Design*, 24-27.

Heikoop, D.D., Hagenzieker, M., Mecacci, G., Calvert, S., Santoni De Sio, F., & van Arem, B. (2019). Human behaviour with automated driving systems: A quantitative framework for meaningful human control. *Theoretical Issues in Ergonomics Science*, *20*, 711–730.

Heller, J., Steiner, C., Hockemeyer, C., & Albert, D. (2006). Competence-based knowledge structures for personalised learning. *International Journal on E-learning*, *5*(1), 75-88.

Kraus, J., Scholz, D., Stiegemeier, D., & Baumann, M. (2019). The More You Know: Trust Dynamics and Calibration in Highly Automated Driving and the Effects of Take-Overs, System Malfunction, and System Transparency. *Human Factors*, *62*, 718–736.

Körber, M. (2018, August). Theoretical considerations and development of a questionnaire to measure trust in automation. In *Congress of the International Ergonomics Association* , 13-30.

Lee, J. D. (2020). Driver Trust in Automated, Connected, and Intelligent Vehicles. In *Handbook of Human Factors for Automated, Connected, and Intelligent Vehicles*. CRC Press.

Mayer, R.C., Davis, J.H., & Schoorman, F.D. (1995). An Integrative Model of Organizational Trust. *The Academy of Management Review*, *20*(3), 709-7034.

Neuhuber, N., Lechner, G., Kalayci, T.E., Stocker, A., & Kubicek, B. (2020, July). Age-related differences in the interaction with advanced driver assistance systems-a field study. In *International Conference on Human-Computer Interaction*, 363-378.

Neuhuber, N., Pretto, P., & Kubicek, B. (in press). Interaction Strategies with Advanced Driver Assistance Systems. *Transportation Research Part F: Traffic Psychology and Behaviour*.

R Core Team. (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. http://www.R-project.org/

SAE International (2018). *Taxonomy and definitions for terms related to driving automation systems for on-road motor vehicle* (Report J3016). Warrendale, PA, USA: SAE International.

Payre, W., Cestac, J., & Delhomme, P. (2016). Fully automated driving: Impact of trust and practice on manual control recovery. *Human Factors*, *58*, 229-241.

Victor, T.W., Tivesten, E., Gustavsson, P., Johansson, J., Sangberg, F., & Ljung Aust, M. (2018). Automation Expectation Mismatch: Incorrect Prediction Despite Eyes on Threat and Hands on Wheel. *Human Factors*, *60*, 1095–1116.

The influence of vestibular feedback on transitions between different levels of automation

Pia Wald[1], Laura Hiendl[2], Martin Albert[3], & Klaus Bengler[1]
[1]Technical University of Munich, [2]University of Regensburg, [3]AUDI AG, Germany

Abstract

The driver's tasks and responsibilities vary in a multi-level automated driving car. While drivers have to monitor the system and the environment in assisted and partially automated driving, they can engage in non-driving related tasks during higher levels of automation. To support drivers in their tasks and increase their mode awareness, the system should provide comprehensible feedback about it's state and intentions. Two different feedback concepts were implemented for this purpose, comparing a visual-auditory with a visual-auditory-vestibular feedback. A driving study (N=47) was conducted with a test vehicle simulating partially and highly automated motorway driving. Depending on their experience with adaptive cruise control (ACC), participants were split into three groups and experienced manual, partially and highly automated driving as well as transitions between these levels. The results revealed that both concepts generated high levels of trust and acceptance. Experience with ACC showed no significant effect. However, visual-auditory feedback with additional vehicle motions could significantly increase the predictability of the automated vehicle's behaviour. Moreover, in partially automated driving visual-auditory-vestibular feedback was perceived as more relieving than without vehicle motions.

Introduction

The driver's role is changing as automated driving functions become increasingly widespread. According to the taxonomy of the Society of Automotive Engineers (SAE, 2016), automated driving vehicles (SAE L2-L5) can perform both lateral and longitudinal vehicle guidance, with only the driver's responsibilities changing. During partially automated driving (SAE L2), the driver has to monitor the automated system and the environment. In higher levels of automation (LoA), the driver is allowed to withdraw from supervising and accomplish a non-driving related task (NDRT). Future vehicles may combine several LoA. The greatest challenges facing these multi-level systems are not only the variation in responsibility for the driving task, but also the transitions between different LoA. Literature to date has mainly considered questions regarding the time until manual control is regained and the respective influencing factors (Zhang et al., 2019). The available time budget and driver's reaction to a take-over request (TOR) mostly range from five to ten seconds (e.g., Gold et al., 2013), whereas studies regarding the mental stabilization time after a transition have shown

In D. de Waard, S.H. Fairclough, K.A. Brookhuis, D. Manzey, L. Onnasch, A. Naumann, R. Wiczorek, F. Di Nocera, S. Röttger, and A. Toffetti (Eds.) (2022). Proceedings of the Human Factors and Ergonomics Society Europe Chapter 2022 Annual Conference. ISSN 2333-4959 (online). Available from http://hfes-europe.org

that it takes the driver up to 40 seconds to regain full attention (Merat et al., 2014). This means that the cognitive processing of a take-over situation takes more time than the (reflexive) motoric response to a TOR (Zeeb, 2016) and should be further addressed (Merat et al., 2014).

The varying responsibilities (SAE, 2016) in a multi-level system and transitions leads to new challenges in the HMI design (Othersen, 2016). Regular feedback from the vehicle is important to keep the driver informed, despite the more passive role. Feedback can be provided in a variety of ways and via all modalities (Bengler et al., 2020). As state of the art, feedback is usually presented visually (Albert et al., 2015), auditorily (Forster et al., 2017) or tactilely (Petermeijer et al., 2017). The design and information content of feedback depends on the respective LoA and thus on the driver's task (Beggiato et al., 2015; Bengler et al., 2020). In this context, multimodal feedback and interfaces are advantageous, resulting in a better system awareness (Bengler et al., 2020; Wickens, 2002). Moreover, multimodal feedback can be complemented by active vehicle motions covering the vestibular modality. These motions provide a new possibility to communicate intentions of the automation before initiating a manoeuvre. Previous studies showed for partially automated driving that detecting a preceding vehicle should be announced via pitch motions (Cramer et al., 2018). Additionally, roll motions should announce lane changes (Cramer, 2019). These pitch and roll motions have been considered useful in assisting drivers regarding their mode and system awareness (Cramer et al., 2018; Cramer, 2019).

There is still some uncertainty about vestibular feedback in a multi-level automated vehicle. The aim of this study is to examine whether additional vestibular feedback can improve the driver's mode awareness. Furthermore, based on previous results showing a correlation between experience with adaptive cruise control (ACC) in years and evaluation of feedback (Wald et al., 2021), the influence of experience with ACC on the assessment of the feedback concepts will be investigated. Additionally, this study provides an overview of activation times for different transitions depending on feedback. As most of the transition studies so far took place in driving simulators, those findings have to be confirmed in a real road environment (Zhang et al., 2019) and need to include more realistic scenarios (Eriksson & Stanton, 2017). Thus, two different feedback concepts, one with active vehicle motions and one without motions, were investigated in a real-world driving study with uncritical transitions between different LoA by three groups depending on the ACC experience.

Method

Sample

Forty-seven drivers with a mean age of 32.91 (*SD* = 9.93) years, ranging from 22 to 59, participated in the experiment. The sample represented a variation of gender and technical background (23.4% technical female, 25.6% non-technical female, 31.9% technical male and 19.1% non-technical male). Participants drove an average of 14,468 km a year (*SD* = 8,888 km) before and 9,000 km a year (*SD* = 5,782 km) during the period influence by the COVID-19 pandemic situation. 72% of the sample had used lane keeping assistance and 36% partially automated driving systems before. Moreover, 66% had previous experience with ACC, 16 participants with little

experience (M = 1.29, SD = 0.77, min = 0.1, max = 2) and 15 with high experience (M = 8.73, SD = 4.53, min = 3, max = 18) with ACC.

Test setup and equipment

The driving study was conducted on the three-lane German A9 motorway between the Manching and Denkendorf exits. However, only the right and middle lane of the motorway was used for safety reasons. The test vehicle, an Audi A5 (year of construction 2012), drove at a maximum of 120 km/h. A prototypical automation system was implemented that was able to simulate partially (SAE L2) and highly (SAE L4) automated driving. The test vehicle performed lateral and longitudinal vehicle guidance.

The participant sat in the driver's seat, and there were two further experimenters in the test vehicle. The experimenter in the passenger seat acted as a safety driver. Additional equipment such as a second interior mirror, additional exterior mirrors, driving school pedals, and a monitor to display essential information about the system, assisted the safety driver. Besides triggering lane changes, this experimenter could adapt to speed limits and provoke HMI elements, pitch and roll motions (referring to Cramer et al., 2018; Wald et al., 2021). The second experimenter sat in the back seat, coordinated the questionnaires and gave the participants instructions.

Study procedure

The driving study took place during the COVID-19 pandemic situation, so a hygiene concept was developed with experts beforehand which is similar to Wald et al. (2021). Fig. 1 presents the sequence of the driving study. The experiment was conducted in German. Participants initially received verbal instruction on the procedure, the test vehicle operation, and the various transitions (Figure 1). They then practised activating the different LoA in the stationary test vehicle, which was followed by the test drives. During all driving sessions, participants drove manually on the motorway and activated the automation system in the right lane. During the first three minutes of the settling-in drive, the test vehicle performed no lane changes since the participants got familiar with the system. Drivers received neither visual nor vestibular feedback during the settling-in phase, but only basic information such as current speed and position in the instrument cluster. Subsequently, they experienced two feedback concepts consisting of four transitions in a randomized order. During L4, participants had to play a game on a tablet mounted in the centre console.

Human-Machine Interface

The human-machine interface consisted of visual elements in the instrument cluster, auditory signals, and active vehicle motions. According to literature recommendations (Beggiato et al., 2015), system's status, future and current manoeuvres, current velocity and a preceding vehicle were presented in the cluster (Wald et al., 2021). The LoA were displayed in different colours (L2 in blue, L4 in green) for this driving study. Lane changes were announced with an arrow in the cluster, the direction indicator and additional active roll motions in the vestibular concept. A degressive roll profile with an angle of 3.0° and an acceleration of -4.5°/s² was used to announce

lane changes (Cramer, 2019; Wald et al., 2021). Moreover, pitch motions announced a slower detected preceding vehicle with an angle of 1° and an acceleration of -5°/s² (Cramer, 2019).

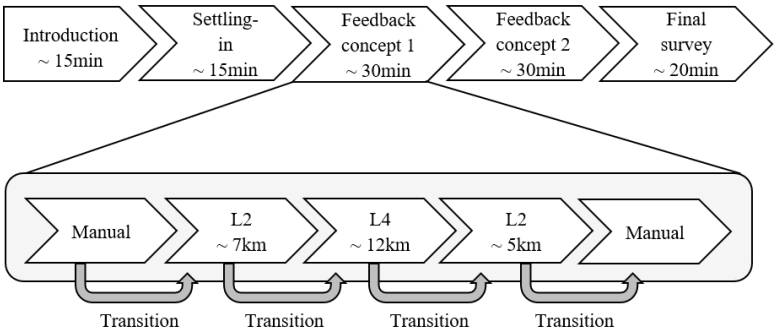


Figure 1. Sequence of the driving study and transitions.

Immediately after entering the motorway, a system suggestion to activate L2 was displayed. Based on experts' advice for an unobtrusive suggestion, no additional auditory hint was given. However, a sound announced the other transitions for the following LoA with an additional pop-up. The transition pop-up in L2 indicated that L4 was available, whereas a transition pop-up in L4 requested the driver to activate L2. A task description was shown after the transition had been accepted. Depending on the following LoA, the description indicated that performing a NDRT (in L4) was allowed or prompted the driver to fully monitor the system (in L2). A transition to manual driving (SAE L0) was announced with red symbols and an intrusive sound.

Processing and evaluation of the data

Objective data included vehicle data and internal data from the automation system, from which *activation times* were calculated. After each transition, questions were asked about mode awareness. At the end of both concept driving parts, the participants answered questionnaires regarding their subjective perception of the feedback concepts. *Trust* in automation was assessed by the questionnaire from Körber (2019) which is divided into six subscales on a five-point Likert scale ranging from 1 ("strongly disagree") to 5 ("strongly agree"). Participants rated the three subscales Reliability/Competence, Understanding/Predictability, and Trust in Automation to obtain respective trust of each feedback concept. The German version of the questionnaire by van der Laan et al. (1997) was used to evaluate *acceptance* of the feedback concepts. This survey is divided into the subscales usefulness and satisfying based on nine items on a five-point scale from -2 to 2. *Mode awareness* was measured with two questions (Othersen, 2016) after each transition for the previous mode on a 15-point scale consisting of five categories from "very little" to "very strong" with the additional opportunity "no answer". Participants were asked to orally validate their *task awareness* ("I was always aware which tasks I had and which ones the system had.") and their *monitoring behaviour* ("I have permanently monitored the system."). After each concept drive, participants were asked to rate specific statements for the *feedback characteristics* on a 7-point rating scale from 1 ("does absolutely not apply")

to 7 (“does absolutely apply”). Three statements, each for L2 and L4, stated whether the feedback was perceived as annoying, distracting and relieving. Moreover, predictability of the automated vehicle was validated after each transition with the statement “How predictable was the system behaviour in the previous mode?” (Petermann-Stock, 2015) on a 15-point scale.

The data were analysed using MATLAB, the statistics with R. For this study, a repeated measure mixed design was used combining the between-subject factor experience with ACC and depending on the dependent variable, the within-subject factors feedback concept, transition and LoA. The participants received both visual-auditory (VA) and visual-auditory-vestibular (VAV) feedback. The sample was divided by their experience with ACC (zero, little and high). An analysis was performed and interpreted, even if the Shapiro-Wilk test showed significance, as the ANOVA is considered robust against a violation of the normal distribution (Blanca et al., 2017). A significance level of $\alpha = 0.05$ was initially applied and partial eta-squared was computed as effect size statistics. Degrees of freedom were corrected when Mauchly’s test for sphericity showed significance (Greenhouse-Geisser). Homogeneity of variance was assessed by Levene’s test for equality of error variances and homogeneity of covariances was calculated by Box’s test for equality of covariance matrices. Unless otherwise stated, data was homogenous in variance and covariance. Post hoc comparisons were controlled with Benjamini-Hochberg corrected p-values (Benjamini & Hochberg, 1995).

Results

Activation times

The values for the mean (*M*), standard deviation (*SD*) as well as minimum and maximum for each transition depending on the feedback concept are presented in Table 1. Analysis of variance for activation time finds neither a significant effect of experience with ACC or feedback, nor any significant interactions ($p > .05$). However, ANOVA yielded a significant main effect for transition ($F(1.62,63.18) = 18.48$, $p < .001$, $\eta_p^2 = 0.321$). Following post-hoc analysis revealed that the activation time for transition from L4 to L2 ($M = 8.77$, $SD = 4.83$) is significantly higher compared to transition from L2 to L4 ($M = 5.33$, $SD = 2.26$, $p < .001$) and to transition from L2 to L0 ($M = 5.64$, $SD = 3.3$, $p < .001$).

Table 1. Descriptives of participants' activation times for different transitions depending on feedback concept.

	Visual-auditory			Visual-auditory-vestibular		
	M (SD)	*Min*	*Max*	*M (SD)*	*Min*	*Max*
L2 to L4	5.41 (2.66)	2.36	15.77	5.24 (1.81)	2.82	12.68
L4 to L2	8.48 (3.7)	3.54	24.99	9.07 (5.77)	2.86	31.48
L2 to L0	5.34 (3.47)	1.71	23.78	5.95 (3.13)	2.26	15.96

Trust and acceptance

Trust is presented in Figure 2. Overall, both concepts were evaluated as reliable, predictable, and generated high trust in automation. The applied ANOVA indicated no significant differences between the feedback concepts for Reliability ($F(1,44) = 1.79$, $p = .188$, $\eta_p^2 = 0.039$), Predictability ($F(1,44) < 1$, $p > .05$) and Trust in Automation ($F(1,44) = 1.27$, $p = .267$, $\eta_p^2 = 0.028$). Moreover, there was neither an effect of experience with ACC nor an interaction between experience and feedback for all three subscales ($p > .05$).

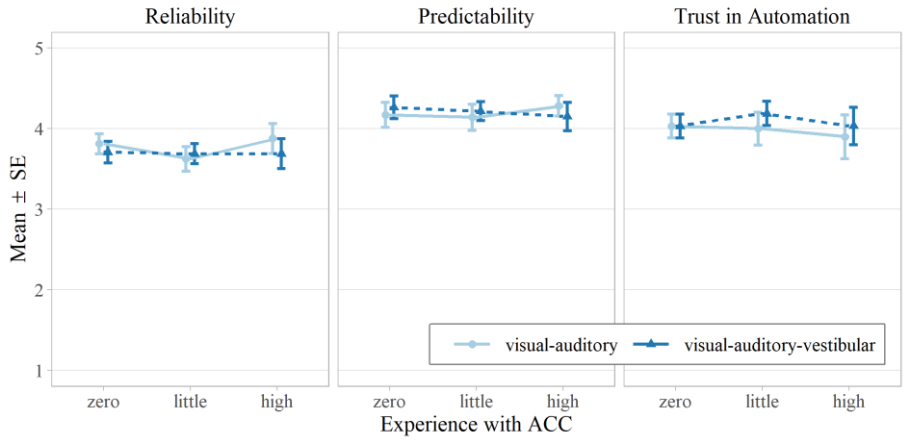


Figure 2. Participants' mean ratings of the feedback concepts for the three dimensions of the questionnaire from Körber (2018) depending on the experience with ACC.

Concerning acceptance, both concepts were rated as useful (VA: $M = 0.74$, $SD = 0.32$, VAV: $M = 0.74$, $SD = 0.37$) and satisfying (VA: $M = 1.38$, $SD = 0.48$, VAV: $M = 1.34$, $SD = 0.64$). However, analyses of variance showed neither significant differences between the feedback concepts ($F(1,44) < 1$, $p > .05$) nor between the experience with ACC ($F(2,44) < 1$, $p > .05$) for both scales. Moreover, there were no interaction effects between feedback and experience for either scale.

Mode Awareness

Results for task awareness showed that neither experience with ACC ($F(2,44) = 1.89$, $p = .163$, $\eta_p^2 = 0.079$) nor the feedback concept ($F(1,44) < 1$, $p > .05$) yielded a significant effect. However, there was a significant effect of LoA ($F(1.77,77.68) = 17.34$, $p < .001$, $\eta_p^2 = 0.283$). Post hoc tests revealed a significant higher task awareness for L0 ($M = 14.37$, $SD = 1.28$) compared to the first ($M = 12.67$, $SD = 2.67$, $p < .001$) and the second ($M = 12.9$, $SD = 2.09$, $p < .001$) L2 section. Moreover, L4 ($M = 13.85$, $SD = 1.85$) generated a higher task awareness than the two L2 sections ($p < .001$). The applied ANOVA indicated no significant interaction effects. Figure 3 presents the result.

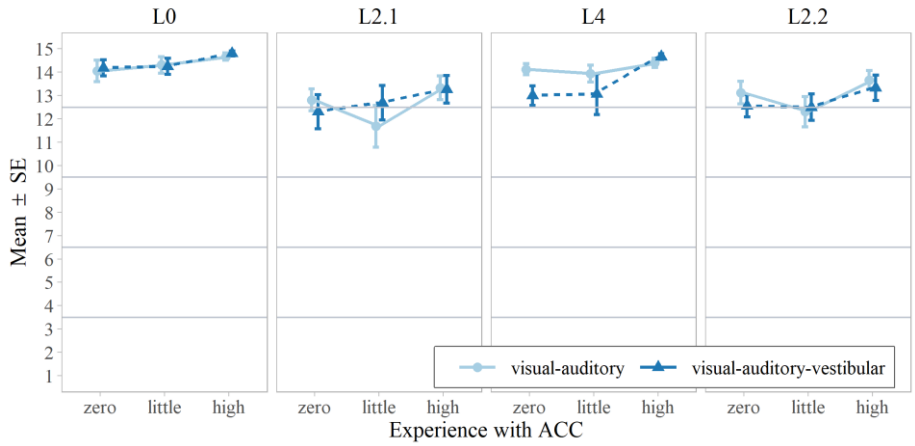


Figure 3. Participants' mean ratings for the feedback concepts of their task awareness depending on the experience with ACC.

Additionally, participants should monitor the system constantly during L2 and not at all during L4. The ANOVA for self-rated monitoring behaviour for the two L2 sections revealed neither a main effect for experience with ACC ($F(2,44) < 1$, $p > .05$) nor for feedback ($F(1,44) < 1$, $p > .05$). However, a significant effect of LoA was found ($F(1,44) = 6.73$, $p = .013$, $\eta_p^2 = 0.133$), indicating a decreasing monitoring behaviour from the first ($M = 11.4$, $SD = 2.57$) to the second ($M = 10.69$, $SD = 2.87$) L2 section. Furthermore, a significant interaction between LoA and feedback ($F(1,44) = 4.54$, $p = .039$, $\eta_p^2 = 0.093$) was noted. Subsequent post hoc analysis showed no significant differences ($p > .05$). Moreover, there were no further significant interaction effects. The applied ANOVA for L4 revealed no significant effects for experience with ACC ($F(2,44) = 1.27$, $p = .292$, $\eta_p^2 = 0.054$) and feedback ($F(1,44) < 1$, $p > .05$). The interaction between the two factors achieved statistical significance ($F(2,44) = 3.28$, $p = .047$, $\eta_p^2 = 0.013$), but post hoc tests showed no significant results.

Feedback characteristics

The mean values for distracting, annoying and relieving can be found in Table 2. Analysis of variance for annoying and distracting found neither a significant main effect nor any significant interactions ($p > .05$). On a descriptive level, VA seems to be less annoying and less distracting in L4 (cf. Table 2). The ANOVA for relieving yielded no significant differences in experience, LoA or feedback. However, the interaction between feedback and LoA reached statistical significance ($F(1,44) = 5.37$, $p = .025$, $\eta_p^2 = 0.109$). Post-hoc comparisons indicated that VAV is more relieving than VA ($p = .006$) in L2.

Table 2. Assessment of feedback characteristics for L2 and L4 depending on the feedback concept.

		Distracting		*Annoying*		*Relieving*	
		M	*SD*	*M*	*SD*	*M*	*SD*
L2							
	Visual-auditory	1.70	0.81	1.40	0.68	4.43	1.64
	Visual-auditory-vestibular	1.70	0.91	1.40	0.68	5.04	1.35
L4							
	Visual-auditory	1.81	1.48	1.49	1.04	4.53	2.14
	Visual-auditory-vestibular	2.26	1.67	1.81	1.42	4.40	2.13

Concerning predictability, analysis of variance found neither a significant effect for experience with ACC nor for LoA or any significant interaction. Feedback, however, had a significant effect on the predictability ($F(1,41) = 5.77$, $p = .021$, $\eta_p^2 = 0.123$). Participants rated VAV ($M = 11.57$, $SD = 2.49$) as more predictable than VA ($M = 11.21$, $SD = 2.73$).

Conclusion and Discussion

The aim of the current study was to examine whether additional vestibular feedback can improve driver's mode awareness. Therefore, two different feedback concepts for a multi-level system with partially and highly automated motorway driving were evaluated. In general, both concepts generated high trust and acceptance scores. These results are consistent with previous research (Cramer, 2019; Wald et al., 2021). Moreover, results for the single item predictability revealed that VAV was more predictable compared to VA, although the subscale Understanding/Predictability of the trust questionnaire showed no differences between the feedback concepts. This inconsistency may be due to the fact that the subscale considered understanding in addition to predictability, thus allowing a more precise measurement. Furthermore, VAV was more relieving in L2. These results support the findings of Cramer (2019). However, additional active vehicle motions in L4 appear to be distracting and annoying on a descriptive level. Contrary to expectations, this study did not find a significant difference between the concepts for task awareness and monitoring behaviour. Results revealed that L2 generated a lower task awareness than L0 and L4. These findings further support the idea of recent studies indicating that the driving task should either be fully undertaken by the driver or completely surrendered to the automated driving system (Petermann-Stock, 2015). Additionally, this study found that the monitoring behaviour decreased after L4. Activation times for the transition from L4 to L2 were higher compared to other transitions (L0 to L2 and L2 to L0) in uncritical situations. This result may be explained by the fact that drivers had to deflect from the NDRT and orientate themselves in the environment. Surprisingly, no differences were found between the experience in ACC which is contrary to a previous study (Wald et al., 2021).

The generalisability of these results is subject to certain limitations. Due to the real-world scenario, standardisation of the requirements is difficult since the surrounding traffic and the weather are not controllable. To ensure similar conditions, the study

proceeded on same times during the day. Moreover, participation in the study was voluntary, what might have positively influenced the results because participants were interested in automated driving. Overall, this study strengthens the idea that additional vestibular feedback can improve the human driver interaction in partially automated driving. Based on these and previous results, further research should combine different feedback strategies in a multi-level system (e.g., using vestibular feedback only in partially automated driving) to support drivers in their tasks.

Acknowledgments

We would like to thank our colleagues, in particular Stephan Bültjes and Stephanie Cramer for their assistance with the test vehicle hardware as well as the software. The Ethics Board of the Technical University Munich provided ethical approval for the hygiene concept and this study, the corresponding ethical approval code is 295/21 S.

References

Albert, M., Lange, A., Schmidt, A., Wimmer, M., & Bengler, K. (2015). Automated Driving - Assessment of Interaction Concepts Under Real Driving Conditions. In *6th International Conference on Applied Human Factors and Ergonomics and the Affiliated Conferences, AHFE 2015*.

Beggiato, M., Hartwich, F., Schleinitz, K., Krems, J., Othersen, I., & Petermann-Stock, I. (2015). What would drivers like to know during automated driving? Information needs at different levels of automation. In *7. Tagung Fahrerassistenz.*

Bengler, K., Rettenmaier, M., Fritz, N., & Feierle, A. (2020). From HMI to HMIs: Towards an HMI Framework for Automated Driving. *Information*, *11*.

Benjamini, Y., & Hochberg. Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society*, *57*, 289–300.

Blanca, M.J., Alarcón, R., Arnau, J., Bono, R., & Bendayan, R. (2017). Non-normal data: Is ANOVA still a valid option? *Psicothema*, *29*, 552-557.

Cramer, S. (2019). *Design of Active Vehicle Pitch and Roll Motions as Feedback for the Driver During Automated Driving*. PhD thesis. Technische Universität München.

Cramer, S., Kaup, I., & Siedersberger, K.-H. (2018). Comprehensibility and Perceptibility of Vehicle Pitch Motions as Feedback for the Driver During Partially Automated Driving. *IEEE Transactions on Intelligent Vehicles*, *4*, 3-13.

Eriksson, A., & Stanton, N.A. (2017). Takeover Time in Highly Automated Vehicles: Noncritical Transitions to and From Manual Control. *Human Factors*, *59*, 689-705.

Feldhütter, A., Segler, C., & Bengler, K. (2018). Does Shifting Between Conditionally and Partially Automated Driving Lead to a Loss of Mode Awareness? In Stanton N. (Eds), *Advances in Human Aspects of Transportation. AHFE 2017*. Springer

Forster, Y., Naujoks, F., & Neukum, A. (2017). Increasing anthropomorphism and trust in automated driving functions by adding speech output. In *IEEE Intelligent Vehicles Symposium 2017* (pp. 365-372).

Gold, C., Damböck, D., Lorenz, L.M., & Bengler, K. (2013). "Take over!" How long does it take to get the driver back into the loop? *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 57,* 1938-1942.

Körber, M. (2019). Theoretical Considerations and Development of a Questionnaire to Measure Trust in Automation. In Bagnara S., Tartaglia R., Albolino S., Alexander T., Fujita Y. (Eds*) Proceedings of the 20th Congress of the International Ergonomics Association (IEA 2018).*

Merat, M., Jamson, A.H., Lai, F.C.H., Daly, M., & Carsten, O.M.J. (2014). Transition to manual: Driver behaviour when resuming control from a highly automated vehicle. *Transportation Research Part F: Traffic Psychology and Behaviour, 27,* 274-282.

Othersen, I. (2016). *Vom Fahrer zum Denker und Teilzeitlenker*. PhD thesis. Technische Universität Braunschweig.

Petermann-Stock, I. (2015). *Automation und Transition im Kraftfahrzeug*. PhD thesis. Technische Universität Braunschweig.

Petermeijer, S.M., Cieler, S., & de Winter, J.C.F. (2017). Comparing spatially static and dynamic vibrotactile take-over requests in the driver seat. *Accident Analysis & Prevention, 99,* 218-227.

SAE. (2016). *Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles* (2016-09 ed.) (No. J3016).

Van der Laan, J.D., Heino, A., & De Waard, D. (1997). A Simple Procedure for the Assessment of Acceptance of Advanced Transport Telematics. *Transportation Research Part C: Emerging Technologies, 5,* 1–10.

Wald, P., Haentjes, J., Albert, M., Cramer, S., & Bengler, K. (2021). Active Vehicle Motion as Feedback during Different Levels of Automation. In *IEEE International Intelligent Transportation 2021* (pp. 1713–1720).

Wickens, C.D. (2002). Multiple resources and performance prediction. *Theoretical Issues in Ergonomics Science, 3,* 159-177.

Zeeb, K., Buchner, A., & Schrauf, M. (2016). Is take-over time all that matters? The impact of visual-cognitive load on driver take-over quality after conditionally automated driving. *Accident Analysis and Prevention, 92,* 230-239.

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

Manoeuvre design in automated driving: investigation of on-ramp situations under the variation of safety distances and traffic flow

Konstantin Felbel, Andre Dettmann, Adelina Heinz, & Angelika C. Bullinger
Chemnitz University of Technology
Germany

Abstract

When designing automated driving, defensive manoeuvre design scores higher in user ratings than dynamic manoeuvre design. Seemingly contradictory, dynamic driving manoeuvres have been shown to render the perception of the drive as more natural and understandable – in specific situations. To examine manoeuvre design in such a specific scenario, a driving simulator study on a highway was conducted with 36 participants. The participants experienced twelve on-ramp situations in which the automated vehicle reacted to a merging vehicle by either changing lanes, braking or continued driving (no reaction). Also, the distance to the merging vehicle and traffic flow were varied. In each situation, participants were asked to assess the experience using a handset control (indicating their desire to react to the situation). After each situation, participants rated their experienced trust and acceptance in the manoeuvre design. Results show that lane change was the preferred decision, resulting in higher trust, comfort and acceptance ratings. Data from speech protocols and handset control indicate that automated cars should react as early as they recognize a merging vehicle on the on-ramp. Interestingly, when traffic density was high, braking was rated comparable to lane change.

Introduction

In the near future, a mixed traffic scenario of manually driven as well as highly automated vehicles (HAVs) is expected (Ghiasi et al., 2017; Patel et al., 2017). This situation will be present until the full transition where only a small quantity of manual driven vehicles are in use, which could take decades (Altenburg et al., 2018). Till then, HAVs must be able to handle situations requiring interactions with other road users (Rasouli et al., 2017; Schwarting et al., 2019). As road traffic is a social system (Müller et al.; Rasouli et al., 2017) these interactions should be efficient, smooth, safe and predictable (ERTRAC, 2019; Felbel et al., 2021). To meet these requirements, HAVs have difficult prerequisites from both a technical as well as a human factors viewpoint. First, sensor data from a demanding environment must be captured and fused (Liu et al., 2017). Second, based on the sensory data real time motion planning is calculated and performed. This motion planning is not only dependent on safety (Artunedo et al., 2019) but also on users' trust (Dettmann et al., 2021; Kraus, 2020; Lee et al., 2004), acceptance (Detjen et al., 2021; Jian et al., 2000) and discomfort

In D. de Waard, S.H. Fairclough, K.A. Brookhuis, D. Manzey, L. Onnasch, A. Naumann, R. Wiczorek, F. Di Nocera, S. Röttger, and A. Toffetti (Eds.) (2022). Proceedings of the Human Factors and Ergonomics Society Europe Chapter 2022 Annual Conference. ISSN 2333-4959 (online). Available from http://hfes-europe.org

(Hartwich et al., 2018). Studies on automated driving manoeuvre design have shown a preference for driving trajectories, which are experienced as natural. For example, Rossner and Bullinger (2019) have shown that participants prefer a lateral shift to the right (i.e. away from the oncoming traffic) to increase the lateral distance in situation with oncoming traffic. Contrary, in situations without oncoming traffic a positioning in the middle of the lane was preferred. In a driving simulator study based on real road environments Peng et al. (2021) have shown that drivers are able to distinguish between the natural driving manoeuvres of humans and the more machine-like negotiations of an artificial controller. Natural driving manoeuvres are described as more proactive where they are not only reacting to given situations but predict future driving scenarios. Mullakkal-Babu et al. (2022) compared in a simulation-based approach a cut-in scenario a predictive and a reactive automated driving system. The predictive system resulted in a significant better performance on aspects such as temporal proximity to crash, expected crash severity and the number of aborted lane changes by human-driven vehicles. This machine prediction could be enhanced by incorporating the anticipation capabilities of human drivers and therefore, must be considered while developing automated manoeuvre designs (Dettmann et al., 2021). Drivers' anticipation for upcoming traffic situations seems to be influenced by situational characteristics. According to Muehl et al. (2020), a perceivable reasons (e.g. causal cues or target cues) support the anticipation of other driving behaviour. This indicates that not only the motion of other vehicles needs to be considered to understand the anticipation process but also the context in which the vehicles operate. In addition, context might also influence the evaluation of a performed automated manoeuvre of the own vehicle, especially if it is not in line with the user's expected manoeuvre. Therefore, context needs to be considered when designing and evaluating automated driving functions. There is yet not sufficient knowledge which manoeuvre an automated vehicle should perform. In the present study, we try to fill the identified research gap by investigating three different automated manoeuvre designs in an on-ramp-situation (i.e., highway context).

Method

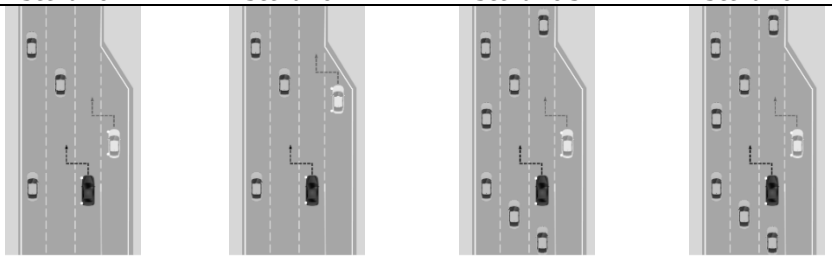
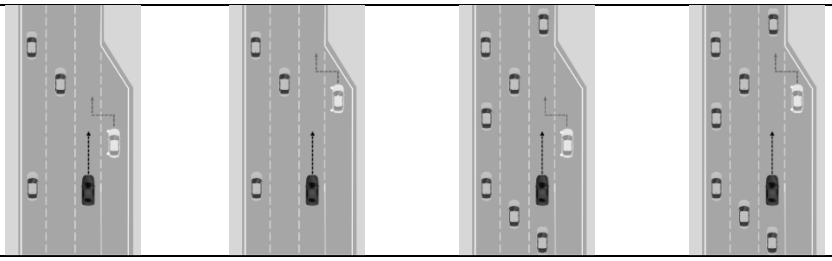
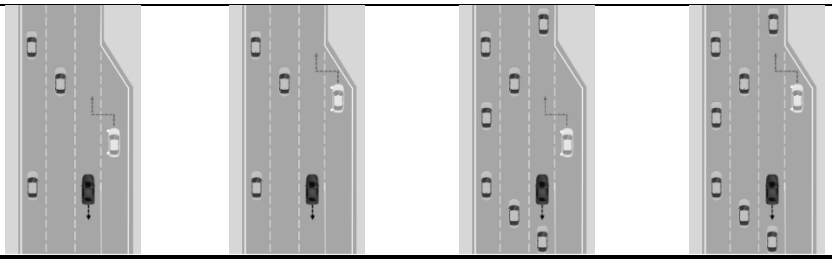
The study sample is based on 36 participants (10 female, 26 male). Their average age was 35.1 years (*SD* = 12.1 years). On average, the subjects drove approximately 59.000 km (*SD* = 42.400 km) in the last five years and had had a driver's license for 16.8 years (*SD* = 11.6). The distribution of travel time among road types was predominantly among urban traffic with 43%, followed by highway (35%) and rural roads (22%). More than half of the participants had no experience with either driving simulator studies (78%) or autonomous driving studies (94%). However, the majority considered themselves moderately to very well informed about the topic on automated driving. Participants assumed that automated vehicles would consistently obey applicable speed limits (86%), drive at a greater distance from other road users than human drivers (69%), and always act cooperatively (81%).

Design and apparatus

To investigate the manoeuvre design of a HAV under variation of distances and traffic flow in a highway on-ramp scenario, a driving simulator study based on a mixed

design was applied. Each participant experienced all driving scenarios (within-subject design with repeated measurement).

Table 1. Combinations of the investigated driving scenarios

Manoeuvre design: lane change			
Traffic density: low		Traffic density: high	
Relative position: close	Relative position: far	Relative position: close	Relative position: far
Scenario 1	**Scenario 2**	**Scenario 3**	**Scenario 4**
			
Manoeuvre design: continuous driving			
Scenario 5	**Scenario 6**	**Scenario 7**	**Scenario 8**
			
Manoeuvre design: braking			
Scenario 9	**Scenario 10**	**Scenario 11**	**Scenario 12**
			

Black vehicle: automated EGO-vehicle. White vehicle: merging vehicle. The relative position indicates the position of the merging vehicle to the EGO-vehicle.

Three implicit factors were considered in the highway scenarios: i) the manoeuvre design of the automated EGO-vehicle, ii) traffic density (low with 3 vs. high with 7 cars) and iii) the relative position of the merging vehicle (570 m vs. 660 m) to the EGO-vehicle. The manoeuvre design of the EGO-vehicle was perceived and assessed from the participants' first-person perspective. The variations of manoeuvre design involved lane changes (yes/no), continuous driving (yes/no), and braking (yes/no). This resulted in a total of twelve driving scenarios (3 x 2 x 2). The twelve experimental conditions were randomized for each participant to achieve

comparability with respect to all conceivable confounding variables as well as to exclude possible sequence effects. Table 1 shows the twelve combinations of the investigated scenarios.

Each scenario started in a highway parking lot. The automated EGO-vehicle then drove at a constant speed of 100 km/h in the right-hand highway lane. This was followed by an approximately 850 m highway on-ramp situation where the surrounding traffic and the relative speed of the merging vehicle to the EGO-vehicle were varied according to the scenarios in table 1. The merging vehicle had a speed of 80 km/h on the acceleration lane (250 m) and then increased its speed to 100 km/h as soon as it changed the lane onto the highway. The EGO-vehicle pulled into a parking lot again at the next exit. This marked the end of one driving scenario and the next situation started via a trigger by the participants. Each scenario lasted about 90 seconds. The simulator used to recreate the scenarios can be classified as a type C simulator (Rimini-Döring et al., 2004) using a projector-based vision system with a field of view of 180 degrees (Figure 1). The steering wheel and both pedals have force feedback actuators implemented and provide a realistic input to control the simulated vehicle. SILAB 7.0 was used to simulate the situations.

Figure 1. Driving simulator (left), handset controller with "desire for reaction" scale (right)

Procedure and materials

The participants were asked to complete a demographic questionnaire (sex, age, annual mileage). Furthermore, technology affinity was assessed using the ATI Scale (Franke et al., 2019) as well as sensation seeking (Hoyle et al., 2002). Additionally, momentary fatigue was queried. Subsequently, the participants were able to familiarize themselves with the driving simulator in similar situations as described above. After starting the experiment, the participants experienced 12 experimental drives in randomized order. After each drive, they filled out a questionnaire that included the evaluation of the manoeuvre design of the automated EGO-vehicle and the assessment of trust, comfort and acceptance through single items (from 1 to 10; higher is better). To gather an online assessment of subjective data for the evaluation of the manoeuvre design a handset controller was used. Participants had the opportunity to use the controller by pressing the lever to report back their desire for a reaction from the automated EGO-vehicle using a scale from 0% desire for reaction to 100% desire for reaction. More actuation of the hand controller indicated an increased desire for a reaction. This also made it possible to determine the exact location on the highway where a different or earlier vehicle reaction (i.e. manoeuvre

design) was desired. After the final experimental drive, a final questionnaire on perceived fatigue and a semi standardized interview was presented. The driving simulator study took about 75 minutes.

Results

Descriptive Analysis

Regarding the assassament of the overall attitude towards automated driving, participants' answer was predominantly positive (70 %). However, only 14% of the participants dealt with the topic of "automated driving" professionally.

Figure 2 illustrates the general subjective ratings of the three experienced manoeuvres. Table 2 gives the description and the numerical value for the ratings. The mean values across all scenarios were summarised to show the tendency which manoeuvre HAVs should perform in an on-ramp situation in relation to a merging vehicle. Lane change and braking were perceived as defensive driving, with lane changes being judged more defensively. Continuous driving was assessed as offensive driving and was perceived as more reckless and riskier compared to the lane change. Braking as well as the continuous driving irritated the participants. The profile diagram shows that lane change manoeuvres tended to be preferred in response to an oncoming vehicle, as they were rated as more predictable, comfortable and cooperative.

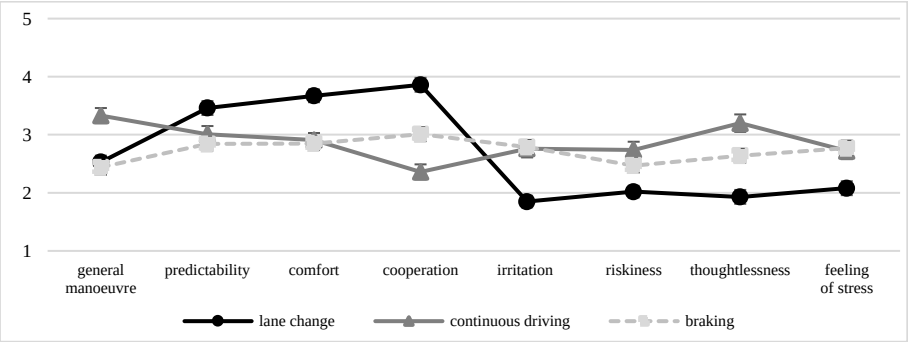


Figure 2. Subjective ratings for the lane change, continuous driving and braking manoeuvre. Error bars reflect Standard Error.

Ratings of trust, comfort and acceptance were compared performing Three-way ANOVAs with repeated measurements including "manoeuvre design" (lane change vs. continuous driving vs. braking), "traffic density" (low vs. high) and "relative position of the merging vehicle" (close vs. far). Figure 3 shows the mean values of the dependent variables for all scenarios.

Trust

A within-subject tests show significantly higher trust ratings for the lane change manoeuvre ($F(2,70) = 13.195$, $p < .001$, $\eta_p^2 = .274$) with a large effect. Furthermore, interaction effects could be identified between first, the three manoeuvre designs and relative position with a large effect ($F(2,70) = 7.687$, $p < .001$, $\eta_p^2 = .180$), and

second, between all independent variables: manoeuvre design, traffic density and relative position of the merging vehicle with a medium effect ($F(2,70) = 3.144$, $p = .049$, $\eta_p^2 = .082$).

Table 2. Description and numerical value of the eight subjective single items (see figure 2)

Item	**= 1**		**= 5**
general manoeuvre	very defensive	o o o o o	very offensive
predictability	very unpredictable	o o o o o	very predictable
comfort	very uncomfortable	o o o o o	very comfortable
cooperation	very uncooperative	o o o o o	very cooperative
irritation	very low	o o o o o	very high
riskiness	very low	o o o o o	very high
thoughtlessness	very low	o o o o o	very high
feeling of stress	very low	o o o o o	very high

Inference analysis

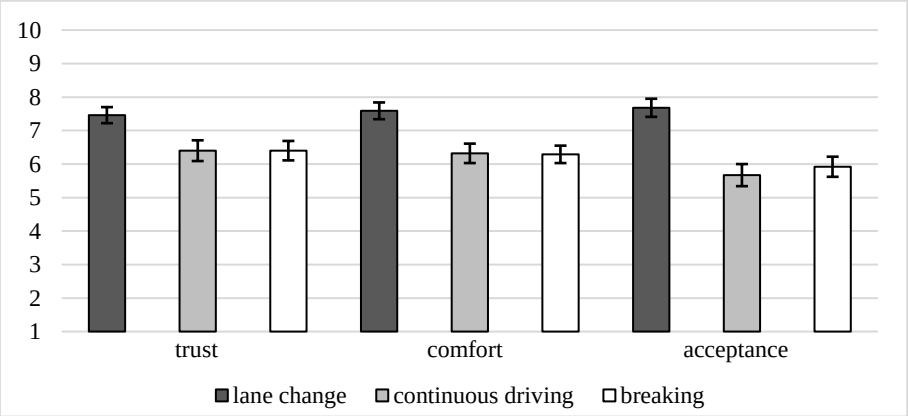


Figure 3. Mean values for trust, comfort and acceptance (single item questions from 1 – 10 after experiencing a driving scenario e.g. lane change + high traffic density + close relative position to merging vehicle). Error bars reflect Standard Error.

Comfort

Regarding comfort, a within-subject test shows a significant difference for the manoeuvre design ($F(2,70) = 23.148$, $p < .001$, $\eta_p^2 = .398$), in which the lane change was rated as most comfortable with a large effect. In addition, traffic density ($F(1,35) = 8.589$, $p = .006$, $\eta_p^2 = .197$) and relative position to the merging vehicle ($F(1,35) = 9.889$, $p = .003$, $\eta_p^2 = .220$) had each a significant impact on the comfort rating with a large effect. Furthermore, a significant interaction exists between the

manoeuvre design and the relative position ($F(2,70) = 13.811$, $p < .001$, $\eta_p^2 = .283$) with a large effect size.

Acceptance

A within-subject tests show significantly higher acceptance ratings for the lane change manoeuvre ($F(2,70) = 26.915$, $p < .001$, $\eta_p^2 = .435$) with a large effect. Traffic density ($F(1,35) = 10.045$, $p = .003$, $\eta_p^2 = .223$) and relative position to the merging vehicle ($F(1,35) = 25.213$, $p < .001$, $\eta_p^2 = .419$) significantly influence the rating with a large effect. In addition, a significant interaction effect is imminent between the manoeuvre design and the relative position to the merging vehicle ($F(2,70) = 19.914$, $p < .001$, $\eta_p^2 = .363$) with a large effect.

Handset control results

All data was cumulated over the driven distance of 1100 m of the four different scenario combinations. Figure 4 shows the cumulated values for the scenario with **low** traffic density and a **close** relative position to the merging vehicle. The desire for reaction rises approximately at the same time for all three manoeuvre designs. In the lane change manoeuvre scenario, the participants' desire to react is at a maximum about 50 m before the EGO-vehicle changes its lane and falls rapidly after the lane change is completed. In the braking and continuous driving scenario, the strongest desire for reaction is reached at about the same distance. In the lane change scenario, the desire for reaction reaches its maximum and minimum faster compared to the other manoeuvre designs.

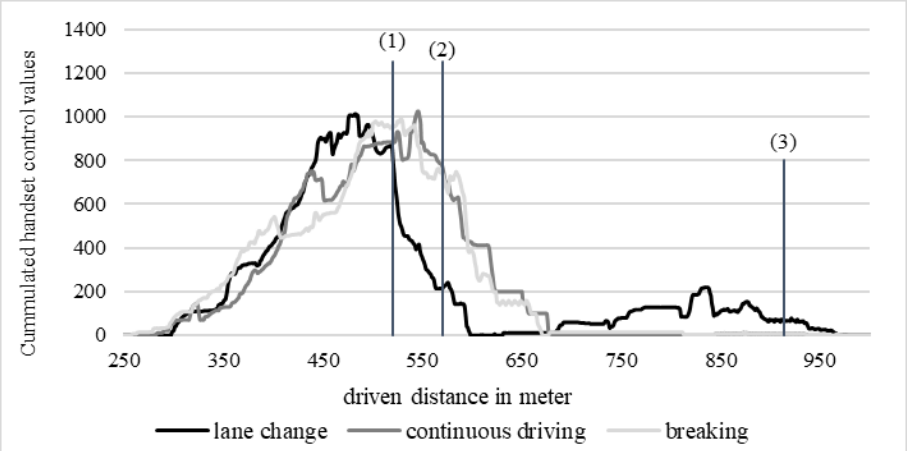


Figure 4. Cumulated handset control values over all participants: low traffic density and close relative position to the merging vehicle. (1) EGO-vehicle reacts to the merging vehicle by changing lanes or braking. (2) Merging vehicle enters the highway. (3) EGO-vehicle changes lane to the right

Figure 5 shows the cumulated values for the scenario **low** traffic density and a **close** relative position to the merging vehicle. The desire for reaction rises approximately at the same time for all three manoeuvre designs. One exception is a small peak in the

lane change scenario at about 200 m before the merging vehicle initiates its lane change. Participants "desire to react" falls as soon as a reaction of the EGO-vehicle is noticeable. Only in the continuous driving scenario it stays high till 175 m after the merging vehicle changed its lane onto the highway and falls quickly afterwards. In the lane change manoeuvre design scenario, the desire for reaction reaches its maximum and minimum faster compared to the other manoeuvre designs.

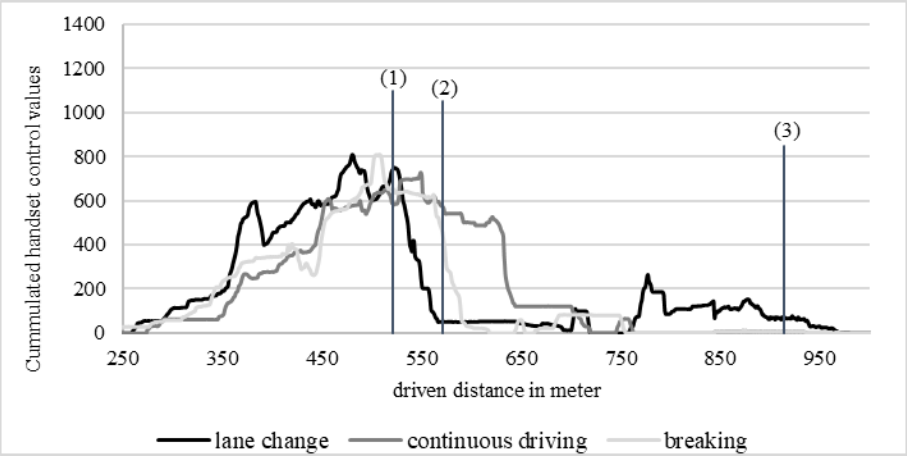


Figure 5. Cumulated handset control values over all participants: high traffic density and close relative position to the merging vehicle. (1) EGO-vehicle reacts to the merging vehicle by changing lanes or braking. (2) Merging vehicle enters the highway. (3) EGO-vehicle changes lane to the right

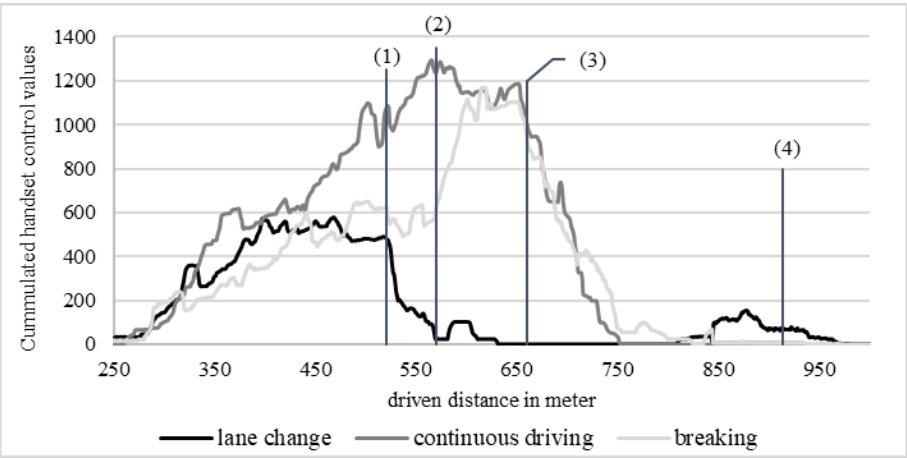


Figure 6. Cumulated handset control values over all participants: high traffic density and far relative position to the merging vehicle. (1) EGO-vehicle reacts to the merging vehicle by changing lanes. (2) EGO-vehicle reacts to merging vehicle by braking. (3) Merging vehicle enters the highway. (4) EGO-vehicle changes lane to the right

Figure 6 shows the cumulated values for the scenario **low** traffic density and a **far** relative position to the merging vehicle. The desire for reaction rises approximately at the same time for all three manoeuvre designs, but stays at a low level in the lane

change manoeuvre scenario. Interestingly, the handset control value in the braking scenario has two peaks. An initial peak, when the merging vehicle is visible and another when the EGO-vehicle breaks. In this combination, the reaction values of the manoeuvre breaking and continuous driving have the same peak hight. The lane change manoeuvre has the lowest handset control value.

Figure 7 shows the cumulated values for the scenario **high** traffic density and a close relative position to the merging vehicle. Overall, the same reaction profile as in Figure 6 can be seen. One difference is the higher handset control value for the braking manoeuvre. It reaches is highest peak slightly before the merging vehicle changes its lane from the on-ramp.

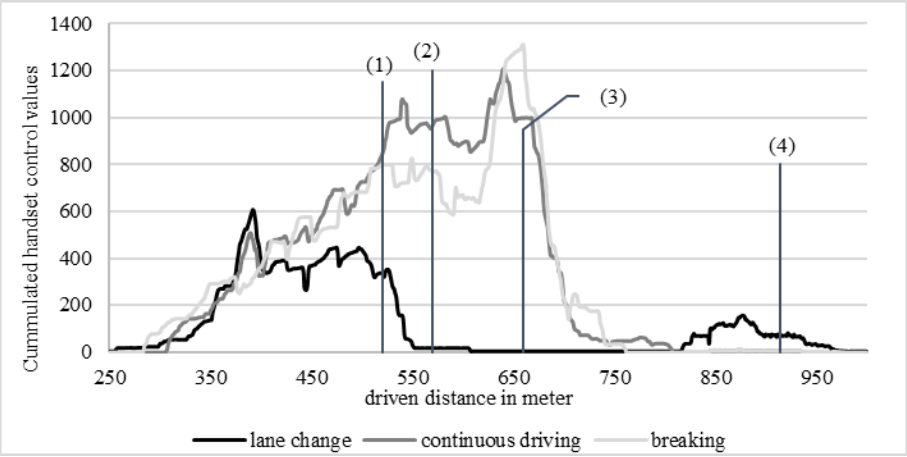


Figure 7. Cumulated handset control values over all participants: high traffic density and far relative position to the merging vehicle. (1) EGO-vehicle reacts to the merging vehicle by changing lanes. (2) EGO-vehicle reacts to merging vehicle by braking. (3) Merging vehicle enters the highway. (4) EGO-vehicle changes lane to the right

Discussion and Conclusion

The aim of the study was to investigate three different automated manoeuvre designs in an on-ramp situation on a highway. The results show that the manoeuvre design, traffic density as well as the relative position of the merging vehicle have a significant influence on the evaluation of trust, comfort, and acceptance. Under the variation of traffic density and the relative position to the merging vehicle, participants' preference for the lane change manoeuvre was identified. Trust, comfort and acceptance ratings were significantly higher. In addition, participants considered the lane change manoeuvre as more cooperative. According to Mullakkal (2022) a proactive automated driving function should be implemented over an reactive. In this case, a more cooperative manoeuvre design (i.e. lane change) can be classified as a proactive driving function. Furthermore, the handset control values for the lane change were the lowest, indicating less demand for another reaction. Over all three manoeuvre designs, participants wanted an earlier reaction, shown by the early increase in the handset control values. This is in line with Roßner and Bullinger (2019), where an early

reaction to an imminent obstacle is desired. Although, braking was mentioned positive in the subsequent interview, it was not rated significantly different than the continuous driving manoeuvre in the subjective questionnaires or the handset control data.

It could be explained by insufficient braking reaction of the EGO-vehicle, indicated by the second peak in the handset control values, as the values rises even after the breaking was initiated. While causal cues were kept the same in all scenarios, the participants described the lane change as more anticipatory. It can be explained by the inherent expectation of the participants. The majority assessed the lane change as the correct manoeuvre and may therefore be subject to hindsight bias. Hence, in subsequent studies, the evaluation of anticipation should be examined before the manoeuvre is carried out.

Based on the results, the following recommendations on manoeuvre design can be given:

- early initialisation of automated reaction to merging vehicles through e.g. early vehicle movement or HMIs
- in light traffic, a lane change is preferred. When lane changing, an early use of the indicators is advised
- in heavy traffic, braking is preferred when the adjacent lane is occupied
- continuous driving should be performed only in combination with an HMI. Otherwise there is no feedback from the EGO-vehicle to the passenger as to whether it has recognised the forthcoming situation

Acknowledgements.

The research was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation – [Project-ID 416228727 – SFB 1410]). The sponsor had no role in the study design, the collection, analysis and interpretation of data, the writing of the report, or the submission of the paper for publication.

References.

Altenburg, S., Kienzler, H.-P., & Auf der Maur, A.. (2018*). Einführung von Automatisierungsfunktionen in der Pkw-Flotte: Auswirkungen auf Bestand und Sicherheit*. ADAC. https://www.prognos.com/uploads/tx_atwpubdb/ADAC_Automatisiertes_Fahren_Endbericht_final_01.pdf

Artunedo, A., Villagra, J., & Godoy, J. (2019). Real-Time Motion Planning Approach for Automated Driving in Urban Environments. *IEEE Access*, *7*, 180039–180053. https://doi.org/10.1109/ACCESS.2019.2959432

Dettmann, A., Hartwich, F., Roßner, P., Beggiato, M., Felbel, K., Krems, J., & Bullinger, A. C. (2021). Comfort or Not? Automated Driving Style and User Characteristics Causing Human Discomfort in Automated Driving. *International Journal of Human–Computer Interaction*, *331–339.* https://doi.org/10.1080/10447318.2020.1860518

ERTRAC. (2019). *Connected Automated Ariving Roadmap.*

https://www.ertrac.org/uploads/documentsearch/id57/ERTRAC-CAD-Roadmap-2019.pdf

Felbel, K., Dettmann, A., Lindner, M., & Bullinger, A.C. (2021). Communication of Intentions in Automated Driving – the Importance of Implicit Cues and Contextual Information on Freeway Situations. In H. Krömker (Ed.), *Lecture Notes in Computer Science. HCI in Mobility, Transport, and Automotive Systems* (Vol. 12791, pp. 252–261). Springer International Publishing. https://doi.org/10.1007/978-3-030-78358-7_17

Franke, T., Attig, C., & Wessel, D. (2019). A Personal Resource for Technology Interaction: Development and Validation of the Affinity for Technology Interaction (ATI) Scale. *International Journal of Human–Computer Interaction, 35*, 456–467. https://doi.org/10.1080/10447318.2018.1456150

Ghiasi, A., Hussain, O., Qian, Z., & Li, X. (2017). A mixed traffic capacity analysis and lane management model for connected automated vehicles: A Markov chain method. *Transportation Research Part B: Methodological, 106*, 266–292. https://doi.org/10.1016/j.trb.2017.09.022

Hoyle, R.H., Stephenson, M.T., Palmgreen, P., Lorch, E.P., & Donohew, R. (2002). Reliability and validity of a brief measure of sensation seeking. *Personality and Individual Differences, 32*, 401–414. https://doi.org/10.1016/S0191-8869(01)00032-0

Kraus, J. (2020). *Psychological Processes in the Formation and Calibration of Trust in Automation* [Dissertation]. Universität Ulm, Ulm.

Lee, S.E., Olsen, E.C., & Wierwille, W.W. (2004). *A Comprehensive Examination of Naturalistic Lane-Changes. National Highway Traffic Safety Administration*. https://doi.org/10.1037/e733232011-001

Liu, S., Tang, J., Zhang, Z., & Gaudiot, J.-L. (2017). *Computer Architectures for Autonomous Driving. Computer, 50(8)*, 18–25. https://doi.org/10.1109/MC.2017.3001256

Mühl, K., Stoll, T., & Baumann, M. (2020). Look ahead: understanding cognitive anticipatory processes based on situational characteristics in dynamic traffic situations. *IET Intelligent Transport Systems, 14*, 233–240. https://doi.org/10.1049/iet-its.2018.5557

Mullakkal-Babu, F.A., Wang, M., Van Arem, B., & Happee, R. (2022). Comparative Safety Assessment of Automated Driving Strategies at Highway Merges in Mixed Traffic. *IEEE Transactions on Intelligent Transportation Systems, 23*, 3626–3639. https://doi.org/10.1109/TITS.2020.3038866

Müller, L., Risto, M., & Emmenegger, C. The social behavior of autonomous vehicles. In Lukowicz, Krüger et al. (Hg.) 2016 – *Proceedings of the 2016 ACM* (pp. 686–689). https://doi.org/10.1145/2968219.2968561

Patel, R.H., Härri, J., & Bonnet, C. (2017). Braking Strategy for an Autonomous Vehicle in a Mixed Traffic Scenario. In Proceedings of the 3rd International Conference on Vehicle Technology and Intelligent Transport Systems (pp. 268–275). *SCITEPRESS - Science and Technology Publications*. https://doi.org/10.5220/0006307702680275

Peng, C., Merat, N., Romano, R., Hajiseyedjavadi, F., Paschalidis, E., Wei, C., Radhakrishnan, V., Solernou, A., Forster, D., & Boer, E. (2021). Drivers' Evaluation of Different Automated Driving Styles: Is It both Comfortable and Natural? https://doi.org/10.31234/osf.io/26bsy

Rasouli, I. Kotseruba and J. K. Tsotsos, Agreeing to cross: How drivers and pedestrians communicate, *2017 IEEE Intelligent Vehicles Symposium (IV)*, 2017, pp. 264-269, doi: 10.1109/IVS.2017.7995730.

Rimini-Döring, M., Keinath, A., Nodari, E., Palma, F., Toffetti, A., Floudas, N., Bekiaris, E., Portouli, V., & Panou, M. (2004). *Considerations on Test Scenarios. Evaluation and Assessment Methodology,* Deliverable 2.1.3 (Project deliverables): aide – adaptive integrated driver-vehicle interface. http://www.aide-eu.org/pdf/sp2_deliv_new/aide_d2_1_3.pdf

Rossner, P., & Bullinger, A.C. (2019). Do You Shift or Not? Influence of Trajectory Behaviour on Perceived Safety During Automated Driving on Rural Roads. In *HCI in Mobility, Transport, and Automotive Systems* (Vol. 11596, pp. 245–254). Springer. https://doi.org/10.1007/978-3-030-22666-4_18 (Original work published 2019)

Schwarting, W., Pierson, A., Alonso-Mora, J., Karaman, S., & Rus, D. (2019). Social behavior for autonomous vehicles. *Proceedings of the National Academy of Sciences of the United States of America, 116(50)*, 24972–24978. https://doi.org/10.1073/pnas.1820676116

Stange, V. (2021). *Human driver and passenger reactions to highly automated vehicles in mixed traffic on highways and in urban areas* [Dissertation]. Technische Universität Carolo-Wilhelmina zu Braunschweig, Braunschweig, Germany.

User experience of a self-driving minibus - reflecting vision, state and development needs of automated driving in public transport

Annika Dreßler & Emma Höfer
German Aerospace Center (DLR)
Germany

Abstract

Beside sharing, electrification of drives, and on-demand operations, the idea of using automated vehicles (AV) in public transport is one building block that is often included in conceptions of a sustainable and efficient future mobility system. If successfully implemented, it could allow new public transportation services where this is not economically feasible at present. The success of such services will crucially depend on their use by the population, which is in turn determined by perceptions of their usefulness, ease of use, safety, and attractiveness. We provide insights on user perceptions of an urban self-driving minibus service in the project HEAT (Hamburg Electric Autonomous Transportation) from the second phase of pilot operation in 2021. Based on data from passenger surveys (n = 446) that were conducted directly after the ride, we analyse the status of progress and identify further development needs from a user perspective. Results show positive attitudes towards using driverless vehicles in public transport, but also a need to further improve system performance in order to create a viable mobility alternative. We point out and discuss measures how performance could be increased.

Introduction

Vision of automated vehicles in public transport

The development of self-driving vehicles, in combination with electrification and shared mobility, is thought of as a potential way to make public transport more efficient, flexible and needs-oriented and to enhance the environmental compatibility of mobility overall (Fulton et al., 2017). Especially, the reduction of personnel costs is hoped to allow to offer public transportation where it is not economically feasible at present (Bösch et al., 2018). Municipalities and public transport operators worldwide are interested in exploring how these technologies can be used and in understanding their opportunities and constraints (UITP, 2016).

In D. de Waard, S.H. Fairclough, K.A. Brookhuis, D. Manzey, L. Onnasch, A. Naumann, R. Wiczorek, F. Di Nocera, S. Röttger, and A. Toffetti (Eds.) (2022). Proceedings of the Human Factors and Ergonomics Society Europe Chapter 2022 Annual Conference. ISSN 2333-4959 (online). Available from http://hfes-europe.org

Current Status

While autonomous driving has to some degree been successfully established in environments where separated lanes can be implemented (e.g., metro, light rail and certain shuttle bus applications; cf. Wang, 2016), it seems to remain a challenge in urban mixed traffic environments. In a review of European pilot projects with automated bus systems, Hagenzieker and colleagues (2021) found that the buses typically operate at low speeds, with 78% of pilots below 21 km/h and the most frequent category being 12-16 km/h. They are often slower than cyclists and other surrounding traffic and tend to stop, often suddenly, when any object comes within a certain distance, regardless of the relative trajectories (e.g., also when cyclists or cars are overtaking). These characteristics lead to many overtakings and other exceptional manoeuvres by other road users. So far, stewards on board have generally been indispensable, due to constraints in legislation, on the one hand, but also for solving situations that the automation cannot handle on its own.

Research goal

Introducing autonomous driving brings along a number of changes for transportation users as well as other road users interacting with the self-driving vehicles (Dreßler et al., 2019; Heikoop et al., 2020). The successful implementation of such systems will also depend on how well they fit human requirements, including how much they are perceived as useful, easy to use, and attractive by their (potential) users. System design should therefore be user-centred from the beginning of development (Nielsen, 2009; Wickens et al., 2004). The work presented here was part of the iterative user research in the project HEAT (Hamburg Electric Autonomous Transportation). It aimed to yield a comprehensive picture of how users experienced a self-driving shuttle piloted in a public road environment and what they conceived for the future use of this technology.

Theoretical and implementation background

Facets of user experience

Theories of user acceptance identify factors that predict the use of or the intention to use a product or service based on new technology and specify how these factors shape this intention (Madigan et al., 2016, 2017; Venkatesh et al., 2012). The factors represent dimensions of user experience, i.e., “a person's perceptions and responses that result from the use or anticipated use of a product, system or service” (ISO 9241-210). They can further be specified by distinguishing perceptions of instrumental (pragmatic) qualities, perceptions of non-instrumental (hedonic) qualities, and emotional reactions (Thüring & Mahlke, 2007).

The user surveys applied aimed to assess the most important facets of user experience, based on existing evidence in the context of self-driving vehicles and public transport (cf. Madigan et al., 2016, 2017). These included pragmatic qualities (perceived usefulness, safety, reliability, ease of use), hedonic qualities (perceived comfort and

diversion•), and emotional reactions (self-assessments of valence and arousal; Russell, 1980). To qualify these assessments, the surveys contained further items to describe detailed aspects of user experience, such as perceptions of the driving style created by the autonomous driving functions and their interplay with potential actions on the part of the vehicle attendants.

Pilot operations

The project HEAT, funded by the German Federal Ministry of the Environment, Nature Conservation and Nuclear Safety, ran from 2018 to 2021. Its aim was to explore the application of electric, self-driving shuttles in urban public transport. The project included two phases of test operations with passengers in the Hamburg district of HafenCity: the first one from October to November, 2020, serving a fixed route of 800 m length with two stops; the second one from August to October, 2021, on a fixed route of 1.8 km length with five stops (Figure 1). In both operation phases, there were vehicle attendants on board who supervised the autonomous shuttle's driving with an allowed maximum speed of 25 km/h on the public roads with speed limits of 30 km/h, and 50 km/h, respectively.

Figure 1. Test operations with passengers 2020 and 2021: routing and stops.

The vehicle (2.95 t) had room for one wheelchair and was technically permitted to transport up to seven passengers (sitting and standing). Due to COVID-19, only three passengers were allowed to ride simultaneously. The shuttle was developed to drive the test route, including the crossing of traffic lights, completely automated. However, in case other vehicles parked on the lane had to be passed, the shuttle attendant had to approve this manoeuvre manually before the shuttle carried it out automatically as it involved a deviation from the defined driving lane. Before riding, passengers were required to register (including acceptance of carriage conditions and privacy policy) using the HEAT app or by filling out a paper form.

• meaning a sense of fun or entertainment, e.g., due to the novelty of the experience (cf. Madigan et al., 2017)

Figure 2. The autonomous vehicle applied in the test operations.

Methods

Structure of the user survey

The survey consisted of 30 items, covering two pages in its paper-pencil-version, with the following sections: informed consent and introduction referring to the most recent ride on the shuttle, use context (purpose, date and time, position taken in the shuttle, prior experience), physical user experience (cabin temperature, air quality), experience of the shuttle's way of driving (frequency, kind and valence of unexpected experiences) in four situation categories, overall user experience (emotional valence and arousal; perceived safety, usefulness, reliability, ease of use, comfort and diversion), qualitative feedback (aspects liked and disliked about the design; wishes for improvement), kind and assessment of information sources used, introduction of potential role of driverless shuttles in the future, respondent's intention to use such shuttles and applications deemed useful, individual characteristics (e.g. gender, prior experience with other driverless vehicles), personal technological innovativeness (based on Goldsmith & Hofacker, 1991), and a final, free-text item that asked if there was anything else the respondent would like to communicate regarding the shuttle.

Data collection

The survey existed in a paper-pencil and an online version (SoSciSurvey). One or two pollsters were present at the main shuttle stop and approached passengers with the survey after their ride. Respondents participated on a voluntary basis without compensation. On-site data collection was carried out in accordance with safety rules due to the COVID-19 situation. As an alternative to the paper-pencil version, the link to the online questionnaire was distributed via the HEAT app and postcards with a QR code available in the vehicle.

Results

For brevity, the presentation of results focuses on the second phase as the patterns of results were mostly similar in both phases while the number of participants was higher in the second phase, and the route had reached its final expansion then.

Sample characteristics

The survey was completed by 446 passengers, aged 8 to 82 years (M = 39.7, SD = 17.2). There were more male (54.7%) than female (32.5%) respondents (other gender: 0.9%, no response: 11.9%). Despite the extension of the route in comparison to the first trial phase, only 4 respondents (0.9%) reported having used the shuttle for transportation purposes ("to get from A to B"). Most (89.2%) still took the ride for curiosity, in order to try out the new technology (no response: 7.4%; "other" purpose: 2.5%).

Ride experience

Perceptions of driving style

Figure 3 shows the reported frequencies of unexpected experiences in four situation categories. Most unexpected experiences were associated with braking behaviour: Altogether, 78.4% of respondents reported at least one unexpected experience regarding braking. In accelerating and driving around bends, the shuttle's driving appeared more consistent with expectations, as only 6.0% and 8.9% of respondents indicated unexpected experiences, while 70.6 and 74.2% did not notice anything unusual, and 17.9 and 18.4% did not respond to the item. Finally, regarding any other driving situation, around 12.8% of respondents reported one or more unexpected experiences.

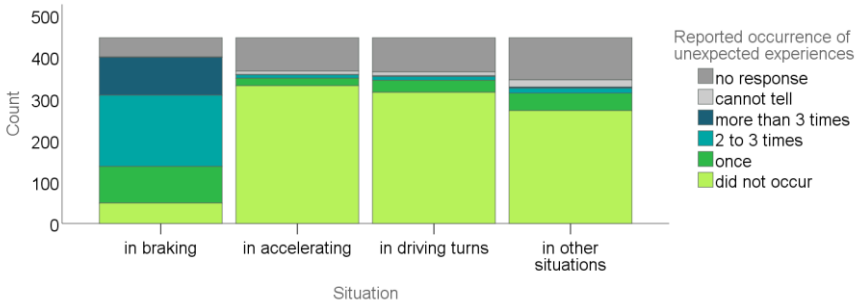


Figure 3. Reported frequency of unexpected experiences regarding the shuttle's way of driving in four manoeuvre / situation categories.

Of the respondents who reported at least one unexpected experience in braking, 62.9% also gave some qualification of what the experience was about. Many responses referred to the onset of braking which was qualified as abrupt or unexpectedly sharp in certain cases. The causes of braking were mostly recognizable to passengers and mostly involved other motorized vehicles or bicyclists coming near, e.g., in standing very near the lane and/or partly protruding into the lane (e.g., side mirror), parking out, or overtaking. In a number of cases, the cause was not obvious to the respondent.

Of all passengers who felt surprised by the shuttle's braking at least once, the majority did not classify this experience as unpleasant. However, 22.6% stated that the braking felt unpleasant to them, which corresponds to around every sixth of all passengers who took the survey. Few events were marked as unpleasant in the other three driving situation categories. The five instances reported in the acceleration category were all

associated with sudden braking, which also occurred in some of the eight instances in the turning category, while the rest involved slow cautious advancement around the bend. Of the users who reported unexpected experiences in other driving situations, twelve indicated that these also felt inconvenient to them. The experiences were about waiting due to obstacles in the lane (incorrectly parked vehicles), the behaviour of other road users (e.g., car coming too near in overtaking), and, in three cases, unexpected positioning behaviour of the shuttle (e.g., late lane change for turning, with car passing on the right).

Use experience

The distributions of user assessments concerning use experience (boxplots) are shown in Figure 4. For analysis, the coding of mirrored scales was reversed, for all scales to point in the same direction.

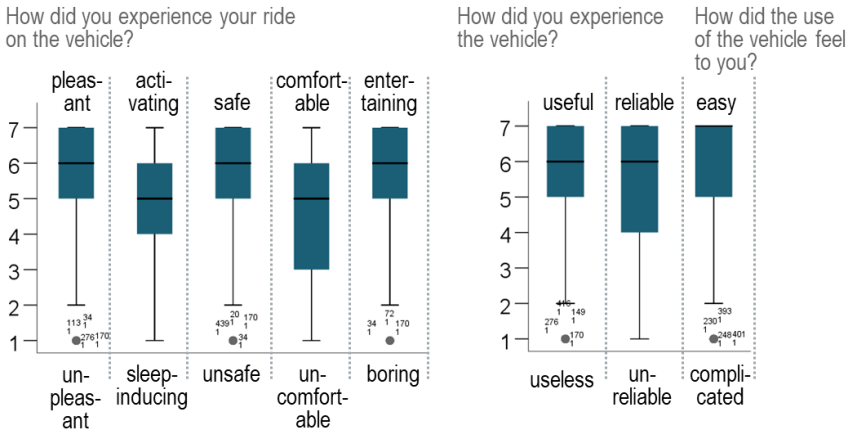


Figure 4. User assessments concerning emotional valence and arousal, perceived safety, comfort and hedonic quality of the ride (left) and perceived usefulness, reliability and ease of use of the shuttle (right).

Overall, passengers reported positive emotional experience of the ride, with a mean assessment of M = 5.9 (SD = 1.3) on the scale from 1 – unpleasant to 7 – pleasant. The arousal associated with this valence was experienced as normal to slightly activated (M = 4.8, SD = 1.5). Regarding the dimensions of pragmatic quality, passengers felt safe on the ride (M = 5.6, SD = 1.5) and perceived the shuttle as useful (M = 5.6, SD = 1.6). Perceived ease of use was high (M = 5.8, SD = 1.6). Perceived reliability (M = 5.2, SD = 1.7) was slightly lower, with more variation in the individual assessments. Concerning hedonic quality, passengers expressed high levels of fun associated with the ride (M = 5.9, SD = 1.4, from 1 – boring to 7 – entertaining). Perceived comfort was slightly above the middle of the scale on average and showed more variation in individual scores (M = 4.7, SD = 1.7).

Qualitative feedback

59.0% of respondents used the free text field to give an indication of what they liked about the design of the shuttle. The most frequent category of mentions (n = 141)

referred to aesthetic aspects of design, including topics such as clear, simple or functional design; modern, futuristic or distinctive design, attractive design, and light colours. The next most frequent category (n = 85) referred to spatial design aspects, including large windows on all sides ("without advertisement"), good panoramic view and lightness in the cabin as well as a spacious interior•. Some respondents named the compact size of the vehicle. A number of statements (n = 23) dealt with technological aspects, most frequently the electric drive, which was liked mostly for its silence and partly for the aspect of environmental protection. A smaller number of statements mentioned the aspect of autonomous driving or the monitors inside the shuttle where route information was displayed. Two further considerable categories concerned seat design (n =11; aspects: comfortable, material wood, and belts) and accessibility (n = 17), including suitability for wheelchairs and low access height.

Looking at aspects not liked about the design of the shuttle, 42.2% of respondents gave some qualitative information. The most frequent category of mentions (n = 66) addressed the design of the seats, which were characterized as hard or uncomfortable by around half of these remarks. A smaller number of mentions revolved around the safety belts that some seats were equipped with, with different foci (not available on all seats, unclear where using belt is required, unnecessary or provisional, not wide enough for users with large body girth). Individual mentions referred to a low height of the seats or the orientation. The next most frequent category involved spatial aspects of design (n = 63). Most of these referred to aspects of capacity, with topics like (too) small size, few seats or little space. A smaller number of remarks dealt with the availability of handles to hold on to (too few, or unfavourable arrangement). A number of statements (n = 33) referred to characteristics of driving (mostly braking). Individual statements involved aspects of accessibility, namely a rather narrow space for turning a wheelchair inside the shuttle, the usability of the ramp (probably referring to the steepness of the angle) and the low auditive perceptibility of the shuttle for road users.

Of the respondents 41.9% provided a statement on what they would consider desirable to improve the design of the shuttle. The most frequent category of mentions (n = 67) involved technological aspects. Around half of these concerned the further development of the driving functions, in order to, for example, reach a higher velocity, smoothen the driving by avoiding sudden breaking, or enable fully automated operation. Ideas not mentioned before include using the shuttle's connectivity to enable phased green traffic lights for the shuttle. The suggestions concerning seat design (n = 40) and spatial design (n = 47) mostly take up the criticism presented above, by proposing softer or more comfortable seats, vessels with higher capacity, more seats overall, more seats in direction of driving, and enhanced possibilities to hold on to a handle, e.g., in standing, sitting down/getting up or getting on/off the vehicle, and a bit more space for turning wheelchairs or prams inside.

• Mind the maximum number of passengers of three (plus two vehicle attendants).

Intention to use self-driving shuttles in public transport

The item to assess passengers' intention to use self-driving public shuttles in the future had three response options: besides *yes*, *definitely* and *no way*, passengers had the possibility to choose *yes if* ... and then qualify the conditions in a free text field. 68.4% of respondents indicated they would definitely use driverless shuttles. 18.8% expressed a use intention given certain conditions. Within these, the most frequent category of mentions (n = 34) involved that the technology be fully tested, developed and safe. A related topic (n = 14) concerned the further development of performance, often mentioning higher velocity, but occasionally also aspects such as a bigger fleet or network, smooth driving or higher transparency of the technology. Ten respondents (2.2%) stated they would not use driverless shuttles, and 10.5% did not respond to the item.

Discussion

Limitations

Our goal was to capture a comprehensive picture of user experience in passengers who had experienced a self-driving shuttle and could base their opinions on this. As we can suppose that most of our respondents tried the HEAT shuttle based on their own interest and motivation, the results apply to persons who are generally open to using this technology and may differ for persons who are not.

Importantly, the results must be considered in the light of the presence of vehicle attendants on board. This means that the user assessments and requirements that were collected can be applied, but certain additional requirements concerning an autonomous operation without an attendant on board did most probably not become obvious. Prior research in the HEAT context has shown that drivers of public transport vehicles fulfil additional functions from a user perspective, including that of a system expert providing helpful information, an instance of supervisory control and a contact person in case of exceptional situations (Dreßler et al., 2019). Design of autonomous shuttles must propose alternative solutions to enable the same, e.g., through proactive passenger information, safety and security measures, or the possibility to get in contact with a service or control centre (cf. Grippenkoven et al., 2019).

Moreover, the focus of the project was on piloting the technology and giving the public an opportunity to try it. Thus, the shuttle operations were not integrated in the regular public transport offers. This was reflected, e.g., in that the shuttle could be used for free and did not appear as part of travel chains proposed in public transport information systems. This trial character needs to be considered, too, when interpreting the user experience results: As most of the users tried the shuttle for curiosity, the demands and expectations were probably lower than they would be in using the shuttle as part of a regular travel chain. Finally, the trial had to be carried out under particular conditions due to the COVID-19 situation (e.g., only three passengers could use the shuttle at a time, nose-mouth covers were worn), which may have changed user experience in certain respects compared to the conceived normal operations.

Conclusions

Passengers experienced the vehicle and the ride on it in a positive way overall. The vast majority expressed their willingness to use self-driving shuttles if these were a readily available transport option. Notably, this was the case even though passengers experienced imperfections in the vehicle's way of driving that mostly concerned occasional "jerky" driving due to braking. The overall pattern of user assessments including the observation that technological aspects were rarely mentioned under dislikes, but more often under ideas for improvement shows that passengers obviously took account of a to-be-expected development status in their evaluation. They were positive about the technology overall and understanding about some current constraints, but they also expect that these be resolved in future applications to make self-driving shuttles a competitive transport option.

The most important optimization potential in the current system, both from a user perspective and with regard to the interactions of the vehicle with surrounding traffic, is the further advancement of the anticipation capability and performance of the autonomous driving functions. It is necessary in order to enable a fluent driving style and avoid abrupt braking reactions as well as waiting times of both the automated vehicle and the surrounding traffic due to mutual obstruction of the way. In addition to high-definition maps and the recognition of environment features for positioning, the current AV shuttles mostly exploit trajectory information of surrounding objects for manoeuvre planning. Some are also connected with elements of the road infrastructure to get more information, e.g., on the status of traffic lights, or additional trajectory information of surrounding objects from road infrastructure sensors, as in the case of the HEAT shuttle. However, while the systems need to interact with human-operated vehicles and vulnerable road users, the exploitation of trajectory information as it is currently done does not seem to lead to satisfactory performance. The vehicles behave rather reactive and lack the anticipatory skills that characterize expert human drivers who exploit predictions about the further course of events based on knowledge of situation categories and including additional cues, as for example the reversing lights of other vehicles. With regard to the achieved speed and autonomy, the HEAT system was already rather advanced within current trials of AV in public transport. However, user feedback concerning use intention and research findings on perceived usefulness (Madigan et al., 2016) clearly shows that efficiency needs to be further advanced if AV in public transport is to become a viable transport option for a large number of users. Beside the enhancement of autonomous driving functions, additional measures may help to improve overall system performance. One of them is the thoughtful identification and/or design of a suitable operation environment and the stopping points with regard to infrastructural conditions. In a video analysis that was also conducted in the HEAT project, much less conflicts in association with waiting (due to other vehicles or the HEAT vehicle itself representing obstacles), passing and overtaking were observed on roads with at least two lanes per direction (Wissen Bach, 2021). In addition, the quality of interactions between AV and surrounding traffic can probably be enhanced by additional means of information and communication, for example a more prominent marking of the vehicle as being autonomously driving as well as being a vehicle of public transport (associated with entry and exit of passengers, necessity of taking care in passing). Dynamic display of

the current status and the pending next manoeuvre (e.g., obstacle ahead, planning to drive around it automatically/manually; duration) and maybe also recommendations may further help other road users to understand and anticipate the vehicle's actions and interact in a safe way.

The results give hints to what users require in public transport in general and underscore the insight that perceived usefulness is the most important criterion in choosing a means of transport. While part of the user feedback referred to autonomous driving, a major portion dealt with general aspects of service quality in public transport that are independent of automation level. These include basic aspects such as availability and reliability, but also respect for user needs and human factors such as accessibility, practicability (e.g., in the transport of luggage and other items) and the need for comfort and aesthetics (see suggestions for improvement in results part).

In the first visions of using autonomous vehicles in shared transport with flexible routes, the systems were conceived to be readily available by now. The results of current pilots highlight that autonomous driving functions the way they are configured currently are not mature to blend in smoothly into mixed traffic environments. Research and development strive to conquer new ground concerning the operational design domains and find solutions for challenging environments. Given the observation that the development is progressing slower than originally thought and the urgency of transforming transportation for sustainability, municipalities, transport operators and policy makers should not wait for automated driving functions to be fully mature to implement innovative transport services in mixed-traffic environments, but in parallel develop concepts how digitization and user-centred design can be used to enhance the availability of public transport and stimulate sharing. This includes the use of automated driving in more constrained environments and the exploration of how flexible on-demand transport with human operators can be made possible today already.

References

Bösch, P.M., Becker, F., Becker, H., & Axhausen, K. W. (2018): Cost-based analysis of autonomous mobility services. *Transport Policy, 64*, 76–91. DOI: 10.1016/j.tranpol.2017.09.005.

Dreßler, A., Grippenkoven, J., Jipp, M., Ihme, K., & Drewitz, U. (2019). Secure, helpful, lovable: Incorporating user needs in the design of autonomous vehicles systems for public transport. *International Transportation, 71*(1), 22-26. ISSN 0020-9511

Fulton, L., Mason, J., & Meroux, D. (2017). *Three revolutions in urban transportation. How to achieve the full potential of vehicle electrification, automation and shared mobility in urban transportation systems around the world by 2050*. UC Davis and ITDP.

Goldsmith, R.E. & Hofacker, C.F. (1991). Measuring Consumer Innovativeness. *Journal of the Academy of Marketing Science 19*(3), 209-221.

Grippenkoven, J., Fassina, Z., König, A., & Dreßler, A. (2019). Perceived Safety: a necessary precondition for successful autonomous mobility services. In D. de Waard et al. (Eds.), d, K. Brookhuis, D. Coelho, S. Fairclough, D. Manzey, A. Naumann, L. Onnasch, S. Röttger, A. Toffetti, and R. Wiczorek (Eds.)

(2019). *Proceedings of the Human Factors and Ergonomics Society Europe Chapter 2018 Annual Conference*. ISSN 2333-4959 (online), pp. 119-133.

Heikoop, D., Nuñez Velasco, J.P., Boersma, R., Bjørnskau, T., & Hagenzieker, M.P. (2020). Automated bus systems in Europe: A systematic review of passenger experience and road user interaction. *Advances in Transport Policy and Planning, 5*, 51–71. DOI: 10.1016/bs.atpp.2020.02.001.

Madigan, R., Louw, T., Dziennus, M., Graindorge, T., Ortega, E., Graindorge, M., & Merat, N. (2016). Acceptance of Automated Road Transport Systems (ARTS): An Adaptation of the UTAUT Model. *Transport Research Arena TRA2016, 14* (Supplement C), 2217–2226.

Madigan, R., Louw, T., Wilbrink, M., Schieben, A., & Merat, N. (2017). What influences the decision to use automated public transport? Using UTAUT to understand public acceptance of automated road transport systems. *Transportation Research Part F, 50*, 55–64.

Nielsen, J. (2009). *Usability engineering*. Amsterdam: Morgan Kaufmann

Russell, J.A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology, 39*, 1161-1178.

Wissenbach, S.J. (2021). *Analyse der Interaktion eines autonomen Shuttlebusses mit anderen Verkehrsteilnehmenden*. Masterarbeit, Technische Universität Berlin.

Thüring, M. & Mahlke, S. (2007). Usability, Aesthetics and Emotion in Human-Technology Interaction. *International Journal of Psychology, 42*, 253-264

UITP (2016). *Autonomous vehicles: A potential game changer for urban mobility. Policy brief*. Available from: https://cms.uitp.org/wp/wp-content/uploads/2020/06/Policy-Brief-Autonomous-Vehicles_2.4_LQ.pdf

Venkatesh, V., Thong, James Y. L., & Xu, X. (2012). Consumer acceptance and use of information technology: Extending the unified theory of acceptance and use of technology. *MIS Quarterly*, *36*(1), 157–178.

Wickens, C.D., Gordon, S.E. & Liu, Y. (2004). *An introduction to human factors engineering*. Upper Saddle River, N.J: Prentice Hall

How media reports influence drivers' perception of safety and trust in automated vehicles in urban traffic

Mirjam Lanzer & Martin Baumann
Ulm University
Germany

Abstract

Automated vehicles are expected to bring benefits not only for their drivers, but also for traffic safety and the environment. For this to happen, drivers must be willing to use automated vehicles, which depends on whether they perceive them as safe and trust them. One source of information that influences how automated vehicles are viewed is media coverage, such as newspaper or magazine articles. To investigate the impact of media reports on the perceived safety of and trust in automated vehicles, we conducted an online experiment. After presenting the features of an SAE Level-3 automated vehicle, participants' (N = 114) initial safety perceptions and trust were measured along with other variables. Participants were then randomly assigned to read one of three newspaper articles that portrayed automated driving in the city as either positive, negative, or neutral. Perceived safety and trust were then measured again. Finally, participants experienced an automated drive through urban traffic and the dependent variables were assessed one more time. Results indicate that the information from the media report significantly influenced trust and perceived safety, especially the negative report. However, after experiencing the automated ride, trust recovered back to the initial level and perceived safety even increased.

Introduction

In March 2018, two accidents involving automated vehicles (AVs) occurred within a few days of each other. In the first accident, an Uber test vehicle in self-driving mode killed a pedestrian who was crossing the street while pushing a bicycle (NTSB, 2019). This was the first fatal accident involving an AV in which a pedestrian was killed. In the second accident, the driver of a Tesla Model X operating in 'Autopilot' mode was killed when the vehicle drifted out of its lane and crashed into a barrier (NTSB, 2020). Both accidents led to a great deal of media coverage with headlines like "Self-driving uber car kills pedestrian in Arizona, where robots roam" (Wakabayashi, 2018; New York Times). An analysis of over 1.7 million tweets before and after the aforementioned fatal accidents revealed that not only did the number of tweets about AVs increase in the days following the crashes, but so did the proportion of negative tweets and the negativity in the tweet texts themselves (Penmetsa et al., 2021). A content analysis of online and print articles on automated driving in German newspapers revealed a similar pattern. Following accidents involving AVs, which

In D. de Waard, S.H. Fairclough, K.A. Brookhuis, D. Manzey, L. Onnasch, A. Naumann, R. Wiczorek, F. Di Nocera, S. Röttger, and A. Toffetti (Eds.) (2022). Proceedings of the Human Factors and Ergonomics Society Europe Chapter 2022 Annual Conference. ISSN 2333-4959 (online). Available from http://hfes-europe.org

included the two fatal accidents described here, an increase in the frequency of reporting and a more negative tone of articles was observed (Jelinski et al., 2021). In general, i.e. not immediately after an accident, the tone of articles in German newspapers reporting on AVs as well as their headlines were mostly neutral with a slight tendency towards a positive tone for the article and a slight tendency towards a negative tone for the headline (Jelinski et al, 2021). A larger study from 2019 examined around 20,000 electronic news articles about AVs from European countries with a sentiment analysis (dos Santos et al., 2022). Overall, the predominant sentiment was negative (e.g. AVs are unreliable), followed by neutral (e.g. AVs are vehicles that drive themselves) and then positive (e.g. AVs are safe). Articles were further divided into categories, for instance market (e.g. stock markets) or test and safety (e.g. accident reports). In the test and safety category, that the authors describe as very relevant to the acceptance of AVs, the sentiment was slightly more neutral than negative, and considerably more negative than positive (dos Santos et al., 2022).

Media coverage, like newspaper or magazine articles, is an important source of information that is widely available to the lay audience and that influences the perception of AVs (Sharma & Mishra, 2022). In two large European surveys in September 2019 with over 27,000 citizens, six out of ten respondents said that they had read, seen or heard something about AVs in the previous twelve months (dos Santos et al., 2022). Consumers' expectations about AVs and expected benefits are high, e.g. increased safety (Ro & Ha, 2017; Utriainen, 2021), increased driver comfort (Hartwich et al., 2018) or reduced vehicle emissions (Stogios et al., 2019). However, the majority of respondents in a European survey said that they would not feel comfortable with AVs, both as a passenger of an AV and as another road user (e.g. a pedestrian being in the presence of an AV) (dos Santos et al., 2022). This is a crucial issue that should be addressed because as long as the majority is not comfortable or willing to use AVs, they cannot unfold their benefits. Two factors that play a decisive role in predicting whether people intend to use AVs are trust and perceived safety. When people perceive AVs as safe and have trust in them, they are more likely to use them (Zhang et al., 2021; Zoellick et al, 2019). Trust and perceived safety are in turn influenced by the information that is available to drivers (Khastgir et al., 2018; Kraus et al., 2019; Ward et al., 2017). For example, higher levels of trust were observed when drivers had a detailed explanation about AV system functions compared to none (Khastgir et al., 2018). When people have not yet had the opportunity to interact with a new technological system, they rely on second-hand information, e.g. from media reports (Miller et al., 2021). Previous experimental studies have shown that information from media reports have direct and indirect influence on people's willingness to ride an AV (Anania et al, 2018; Zhu et al., 2020). When presented with negative (positive) information, even as short as a single headline, people are less (more) willing to ride an AV (Anania et al., 2018). Media reports displaying the opportunities and risks of AVs in a neutral manner were shown to influence attributes like comfort and usability but not perceived safety or trust (Feldhütter et al., 2016).

Present study

Up to the authors knowledge, no studies have been conducted that compare the impact of positive, negative and neutral media reports on drivers' perceived safety of and

trust in AVs. Following the protocol of Feldhütter et al. (2016), perceived safety and trust were measured at the beginning of the study as a baseline, directly after reading a media report and after experiencing an automated drive. An urban traffic setting was chosen both for the media report as well as the automated drive. Previous studies have mainly focused on SAE Level-4 (Feldhütter et al., 2016) or Level-5 (Anania et al., 2018; Zhu et al., 2020) AVs. While these AVs are not commercially available yet, Mercedes-Benz now offers an approved SAE Level-3 system for series production vehicles (Mercedes-Benz Group, 2021). Now that these vehicles are on the road, they may also move more into the focus of reporting. Thus, a Level-3 AV was examined in the present study.

Method

Sample

Overall, 114 people participated in the study. Fourteen participants were excluded from data analysis because they failed to correctly answer a manipulation check question (n = 9), had technical issues (n = 3) or stated to not have answered the questionnaire truthfully (n = 2). The remaining sample (N = 100) consisted of 37 men, 62 women and one non-binary person. Participants' age ranged from 18 to 62 years (M = 28.8 years, SD = 12.4 years). About half of the participants (47%) had their driving licence for 3 to 10 years, around one third (30%) for more than 10 years and around one fourth (23%) for less than 3 years. The majority (60%) indicated that they drive either daily or at least four times a week. Around one fifth of the sample (18%) stated that they had experience with automated driving (SAE Level-2 system).

Participants were recruited via social media and university mailing lists. In order to be able to participate, people were required to be German native speakers and to hold a valid driving licence. Participants were compensated with course credits or had the chance to win one of three €22 Amazon vouchers.

Materials and experimental design

A 3x3 mixed design was applied in this study, with the between factor media report (positive vs. neutral vs. negative) and the within factor measuring time (baseline vs. after media report vs. after automated drive). Three media reports (see Figure 1 for an example) in the style of short newspaper articles were created that were all similar in length (90-94 words). The positive article described how an AV prevented a fatal accident with pedestrians, the negative article described how a pedestrian died in a crash with an AV and the neutral article described how AVs could change urban traffic in the future.

Pedestrian dies in crash with automated vehicle

Tempe/Phoenix - A pedestrian was fatally injured in a crash last night. The pedestrian was attempting to cross the street at a location without a crosswalk and was struck by an automated vehicle. Investigations to date show that the cause of the accident was probably a software error. The vehicle did not recognize the pedestrian due to poor lighting conditions, which is why it drove into the pedestrian without braking. The pedestrian succumbed to his serious injuries at the scene of the accident. It is still unclear whether the driver of the automated vehicle or the manufacturer will be held liable.

Figure 1. Example of one media report (negative) in the style of a short newspaper article created for this study (translated from German by the authors).

Trust in the AV was measured with the German version (Kraus et al., 2020) of the Trust in Automation Scale (Jian et al., 2000). The scale consists of seven items that are rated on a 7-point Likert scale with two poles (1 = "do not agree at all" and 7 = "completely agree"). Perceived safety was measured with one item ("How safe do you feel in an automated vehicle?") that was rated on a 10-point Likert scale with two poles (1 = "not safe at all" and 10 = "very safe"). Affinity for technology was assessed with the German version of the Affinity for Technology Interaction (ATI) Scale (Franke et al., 2019). To experience a drive with an AV, a 3-minute long video was created using Unity and the asset Windridge city. From the driver's perspective, participants saw a drive through a simulated city with pedestrians crossing in front of them (see Figure 2). The automated drive was similar to the baseline drive by Colley et al. (2020).

Figure 2. Automated drive through the city from the driver's perspective with a pedestrian crossing in front of the AV.

In all crossing cases, the AV decelerated in a timely fashion and came to a complete stop. No critical situations occurred during the journey. As the AV was a Level-3 system, it could reach system limits and then issue a take-over request (TOR). The participants would then have to take over control. However, this did not happen during the automated drive.

Procedure

The study was conducted online using the survey platform Unipark. After providing informed consent, participants first answered questions about demographic data, their typical driving behaviour, their experience with automated driving and their affinity for technology. Then the functionalities of a Level-3 AV were explained to them. This included that the AV takes over the complete longitudinal and lateral control, is equipped with an emergency brake assistant and issues a TOR when system limits are reached. Participants were further informed that in case of a TOR, the AV will prompt the participants to intervene (e.g. by selecting possible courses of action via a button press). Afterwards, trust and perceived safety were measured for the first time. Participants were then randomly assigned to one of the three media reports (positive, neutral, or negative). After reading the report, a comprehension check followed. They were asked whether they understood the text and had the possibility to specify any unclarities. As a manipulation check, one question on the content of the media report was included. Trust and perceived safety were then measured for the second time. Next, participants experienced the automated ride via video from the driver's perspective. After that, perceived safety and trust were measured for the third time, as well as takeover willingness. Participants then had the possibility to specify unclarities, provide additional comments regarding the study, and state whether they answered the questions honestly. Lastly, participants were debriefed and informed that the newspaper articles were created for research purposes and that the contents did not depict true events.

Data preparation and analysis procedure

For each dependent variable, a 3x3 mixed ANOVA with the between factor media report (positive vs. neutral vs. negative) and the within factor measuring time (baseline = T1, after media report = T2, after automated drive = T3) was calculated using *SPSS 27*. Since the ANOVA is considered robust to a violation of the normal distribution (Glass et al., 1972), it was also performed and interpreted when a normal distribution according to the Shapiro-Wilk test was not given. The Greenhouse-Geisser adjustment is reported in cases where the assumption of sphericity assessed by Mauchly's test was violated. The homogeneity of the error variances was violated once (trust at T2), so a Box-Cox transformation was performed and the analyses continued with the transformed variable. For simple main effects of the between factor (media report), a one-way ANOVA was calculated. For simple main effects of the within factor (measuring time), a one-way repeated measures ANOVA was calculated. Bonferroni-adjusted t-tests were used as post-hoc comparisons between two specific groups.

Results

A total of 100 participants were included in the analysis. Of those, 34 read the positive media report, 28 the neutral media report and 38 the negative media report.

Perceived safety

Ratings of perceived safety ranged from 1 to 10, higher values representing higher levels of perceived safety. Across all media report groups, perceived safety was highest after experiencing the automated drive (T3; $M = 6.70$, $SD = 1.95$) compared to after reading the media report (T2; $M = 6.13$, $SD = 1.95$) and the baseline (T1; $M = 6.12$, $SD = 1.88$). Across all measuring times, perceived safety was highest in the positive media report group ($M = 6.61$, $SD = 1.77$), followed by the neutral media report group ($M = 6.36$, $SD = 2.00$) and the negative media report group ($M = 6.02$, $SD = 1.98$).

Figure 3 shows perceived safety by media report group and measuring time. At the baseline (T1), the mean values for perceived safety were similar for all three groups. After the media report (T2), mean perceived safety values diverged. While perceived safety increased for the positive media report group, it decreased for the negative media report group and remained at the same level for the neutral media report group. After the automated drive (T3), the values of all three groups converged again and all increased, with the strongest increase for the negative media report group.

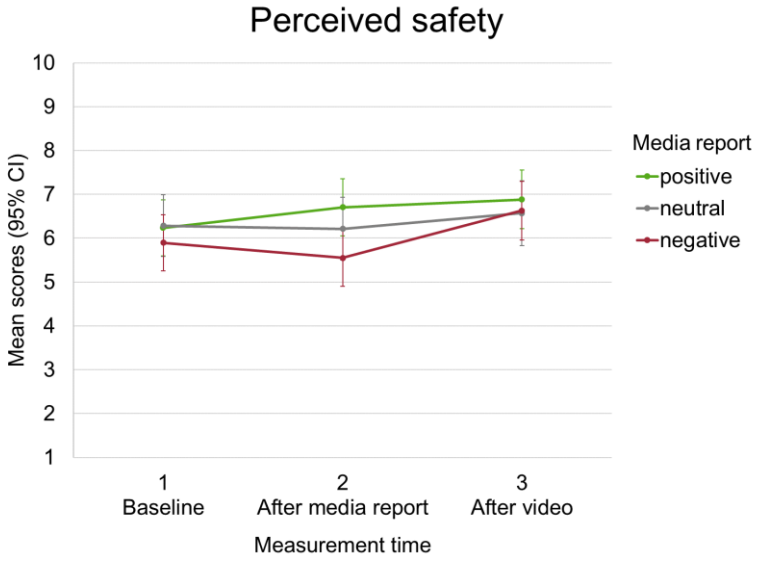


Figure 3. Mean scores of perceived safety per media report group and measuring time. Error bars reflect 95% confidence intervals (CI). Higher ratings represent higher perceived safety.

The mixed ANOVA revealed a significant main effect for measuring time, $F(1.36, 132.23) = 8.56$, $p = .002$, partial $\eta^2 = .08$. The interaction effect, $F(2.73, 132.23) = 2.32$, $p = .084$, partial $\eta^2 = .05$, and the main effect for media report, $F(2, 97) = 1.05$, $p = .355$, partial $\eta^2 = .02$, were not significant. Post-hoc tests confirmed that there were no significant differences for T1 and T3 between all three media report groups. However, after reading the media report (T2), perceived safety was significantly higher for the positive group compared to the negative group, $M_{diff} = 1.15$, $p = .036$. As for the development over time, perceived safety significantly

increased for the positive group from T1 to T2, M_{diff} = 0.47, p = .003, and for the negative group from T2 to T3, M_{diff} = 1.08, p = .002.

Trust

Trust ratings ranged from 1 to 7, higher values representing higher levels of trust. Across all media report groups, trust was highest after the automated drive (T3; M = 4.73, SD = 1.13) followed by the baseline (T1; M = 4.54, SD = 1.21) and after reading the media report (T2; M = 4.41, SD = 1.28). Across all measuring times, trust was similarly high for the neutral (M = 4.77, SD = 1.16) and the positive (M = 4.68, SD = 1.09) group and lower for the negative group (M = 4.29, SD = 1.25).

Figure 4 shows trust by media report group and measuring time. At the baseline (T1), mean trust ratings were similar for all three groups. After reading the media report (T2), trust ratings in the negative media report group decreased while they remained at the same level for the positive and the neutral group. After experiencing the automated drive (T3), trust in the negative group increased back to the baseline and was similar to that of the positive and the neutral group.

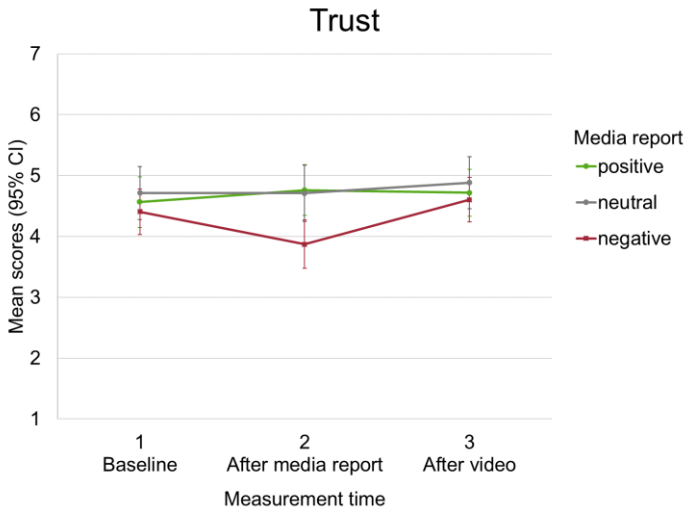


Figure 4. Mean scores of trust per media report group and measuring time. Error bars reflect 95% confidence intervals (CI). Higher ratings represent higher trust.

The mixed ANOVA resulted in a significant interaction effect, F(3.77, 182.78) = 6.99, p < .001, partial η^2 = .13. Regarding the simple main effects for measuring time, a significant effect was found between the groups at T2, F(2, 97) = 5.83, p = .004. After reading the media report (T2), trust was significantly lower for the negative group compared to the positive, M_{diff} = - 1.00, p = .009, and the neutral group, M_{diff} = -0.94, p = .022. Post-hoc tests confirmed that there were no significant differences between the three groups at T1 or T3. Regarding the simple main effects for media report group, a significant effect of measuring time was found for the negative group, F(2, 74) = 16.85, p < .001. For the negative group, trust was

significantly lower at T2 compared to T1, M_{diff} = -0.53, $p < .001$, and T3, M_{diff} = -0.73, $p < .001$. Post-hoc tests confirmed that there were no significant differences between measuring times for the positive and the neutral group.

Discussion

Information from the media report significantly influenced perceived safety and trust. In all three media report groups, participants did not differ significantly in their initial scores for perceived safety and trust. Thus, the groups had no baseline differences that needed to be considered. After reading the media report, the groups diverged. The group with the positive media report had significantly higher trust and perceived safety values than the group with the negative report. For perceived safety this is due to a significant increase in the positive group, while for trust this is due to a significant decrease in the negative group. After experiencing an automated drive through the city, the ratings for all three groups converged again. The convergence is likely due to a significant increase in perceived safety and trust in the negative group. In line with previous research, the neutral media report had no impact on perceived safety and trust (Feldhütter et al., 2016).

Similar to Anania et al. (2018) who used only headlines, the short newspaper articles used in this study led to significant differences between groups. Even relatively short media reports have the potential to influence people's perceptions of AVs. Positive media coverage of AVs can lead to AVs being seen as safe, which increases people's willingness to use them (Anania et al., 2018; Zoellick et al., 2019). Negative reporting on AVs on the other hand reduces trust in them and therefore also the willingness to use them (Anania et al., 2018). This is particularly important for two reasons. First, trust is the most critical predictor for intention to use AVs (Zhang et al., 2021). Second, negative media coverage of AVs is more prevalent (dos Santos et al., 2022) and negative messages are spread more rapidly and widely on social media compared to neutral or positive ones (Tsugawa & Ohsaki, 2015). Even ambiguous story events that could be interpreted as positive or negative, are increasingly negative connotated over multiple transmission episodes (Bebbington et al., 2017). A negative news story, e.g. about an accident involving an AV, could reach more people and have a stronger impact than a positive report. This is especially important as long as people do not yet have their own experience with AVs which would offset the negative reports. Even though vehicles with SAE Level-3 systems can be purchased now, they are very expensive and it will take some time before they are available to a larger number of drivers.

Limitations, strengths and future research

In this study, participants read only one short newspaper article. However, people consume lots of different media reports each day (Zenith, 2019). Further research is needed to examine whether prolonged media consumption exacerbates or attenuates the results shown here. In addition, it would be interesting to study how different combinations of positive, negative and neutral media reports influence people's perception of AVs. In this study, the effect of the media report was captured directly after reading it. A longer delay between reading the media report and experiencing the AV could be of interest. However, the scenario used here might be comparable to

consuming media, e.g. a short newspaper article, just before getting into the vehicle and starting a journey. Furthermore, mass media and social media seem to influence people's perceptions of AVs differently (Zhu et al., 2020). Thus, a comparison of mass media reports, as used in this study, and social media reports could yield new insights. Moreover, the automated drive was only experienced online. The findings should be verified under more realistic or immersive conditions such as a driving simulator or in a field study under real driving conditions.

Conclusion

Media reports about automated driving influence people's perception of AVs. While neutral reports have no influence on perceived safety and trust, positive and especially negative reports have. After experiencing automated driving, the decreased trust and perceived safety evoked by the negative media report recovered. Even though Level-3 AVs are available, it will take some time before a large number of people will experience these systems first hand. However, with more and more Level-3 AVs on the road, media coverage of AVs could be more about real-life events such as accidents than about other topics such as technical developments. If media coverage tends to be unbalanced and too negative, the public might distrust AVs and be unwilling to use them, negating all the benefits AVs may have not only for their drivers but also for the traffic safety of all road users as well as for the environment.

Acknowledgement

This research was funded by the Ministry of Science, Research, and Art Baden-Württemberg and the Ministry of Transport Baden-Württemberg within the Funding Program "Smart Mobility" (Project "INTUITIVER"). We would like to thank Nicole Damm, Miriam Kuhn, Kim Stucke, and Laura Waldmann for their supporting work.

References

Anania, E.C., Rice, S., Walters, N.W., Pierce, M., Winter, S.R., & Milner, M.N. (2018). The effects of positive and negative information on consumers' willingness to ride in a driverless vehicle. *Transport Policy, 72*, 218-224. https://doi.org/10.1016/j.tranpol.2018.04.002

Bebbington, K., MacLeod, C., Ellison, T.M., & Fay, N. (2017). The sky is falling: Evidence of a negativity bias in the social transmission of information. *Evolution and Human Behavior, 38*(1), 92-101. https://doi.org/10.1016/j.evolhumbehav.2016.07.004

Colley, M., Bräuner, C., Lanzer, M., Walch, M., Baumann, M., & Rukzio, E. (2020). Effect of visualization of pedestrian intention recognition on trust and cognitive load. In *12th International Conference on Automotive User Interfaces and Interactive Vehicular Applications (AutomotiveUI '20)* (181–191). New York, NY: Association for Computing Machinery. https://doi.org/10.1145/3409120.3410648

dos Santos, Fabio Luis Marques, Duboz, A., Grosso, M., Raposo, M.A., Krause, J., Mourtzouchou, A., Balahur, A., Ciuffo, B. (2022). An acceptance divergence? Media, citizens and policy perspectives on autonomous cars in the European

Union. *Transportation Research Part A: Policy and Practice, 158*, 224-238. https://doi.org/10.1016/j.tra.2022.02.013

Feldhütter, A., Gold, C., Hüger, A., & Bengler, K. (2016). Trust in automation as a matter of media influence and experience of automated vehicles. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 60* 2024-2028. https://doi.org/10.1177/1541931213601460

Franke, T., Attig, C., & Wessel, D. (2019). A personal resource for technology interaction: Development and validation of the affinity for technology interaction (ATI) scale. *International Journal of Human–Computer Interaction, 35*, 456-467. https://doi.org/10.1080/10447318.2018.1456150

Glass, G.V., Peckham, P.D., & Sanders, J.R. (1972). Consequences of failure to meet assumptions underlying the fixed effects analyses of variance and covariance. *Review of Educational Research, 42*, 237-288. https://doi.org/10.3102/00346543042003237

Hartwich, F., Beggiato, M., & Krems, J.F. (2018). Driving comfort, enjoyment and acceptance of automated driving–effects of drivers' age and driving style familiarity. *Ergonomics, 61*, 1017-1032. https://doi.org/10.1080/00140139.2018.1441448

Jelinski, L., Etzrodt, K., & Engesser, S. (2021). Undifferentiated optimism and scandalized accidents: The media coverage of autonomous driving in germany. *Journal of Science Communication, 20*, 1-25. https://doi.org/10.22323/2.20040202

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

Khastgir, S., Birrell, S., Dhadyalla, G., & Jennings, P. (2018). Calibrating trust through knowledge: Introducing the concept of informed safety for automation in vehicles. *Transportation Research Part C: Emerging Technologies, 96*, 290-303. https://doi.org/10.1016/j.trc.2018.07.001

Kraus, J.M. (2020). *Psychological processes in the formation and calibration of trust in automation* (Doctoral dissertation). Retrieved from Open Access Repositorium der Universität Ulm und Technischen Hochschule Ulm. http://dx.doi.org/10.18725/OPARU-32583

Kraus, J.M., Forster, Y., Hergeth, S., & Baumann, M. (2019). Two routes to trust calibration: Effects of reliability and brand information on trust in automation. *International Journal of Mobile Human Computer Interaction, 11*, 1-17. https://doi.org/10.4018/IJMHCI.2019070101

Mercedes-Benz Group (2021, Decembre 9). First internationally valid system approval for conditionally automated driving. Retrieved from https://group.mercedes-benz.com/innovation/product-innovation/autonomous-driving/system-approval-for-conditionally-automated-driving.html

Miller, L., Kraus, J., Babel, F., & Baumann, M. (2021). More than a feeling – Interrelation of trust layers in human-robot interaction and the role of user dispositions and state anxiety. *Frontiers in Psychology, 12*, 378. https://doi.org/10.3389/fpsyg.2021.592711

NTSB. (2019). *Collision between vehicle controlled by developmental automated driving system and pedestrian, Tempe, Arizona, March 18, 2018* (Accident

report NTSB/HAR-19/03). Washington, DC: National Transportation Safety Board.

NTSB. (2020). *Collision between a sport utility vehicle operating with partial driving automation and a crash attenuator, Mountain View, California, March 23, 2018* (Accident report NTSB/HAR-20/01). Washington, DC: National Transportation Safety Board.

Penmetsa, P., Sheinidashtegol, P., Musaev, A., Adanu, E.K., & Hudnall, M. (2021). Effects of the autonomous vehicle crashes on public perception of the technology. *IATSS Research, 45*, 485-492. https://doi.org/10.1016/j.iatssr.2021.04.003

Ro, Y., & Ha, Y. (2019). A factor analysis of consumer expectations for autonomous cars. *Journal of Computer Information Systems, 59*, 52-60. https://doi.org/10.1080/08874417.2017.1295791

Sharma, I., & Mishra, S. (2022). Quantifying the consumer's dependence on different information sources on acceptance of autonomous vehicles. *Transportation Research Part A: Policy and Practice, 160*, 179-203. https://doi.org/10.1016/j.tra.2022.04.009

Stogios, C., Kasraian, D., Roorda, M.J., & Hatzopoulou, M. (2019). Simulating impacts of automated driving behavior and traffic conditions on vehicle emissions. *Transportation Research Part D: Transport and Environment, 76*, 176-192. https://doi.org/10.1016/j.trd.2019.09.020

Tsugawa, S., & Ohsaki, H. (2015). Negative messages spread rapidly and widely on social media. In Proceedings of the 2015 ACM on Conference on Online Social Networks (COSN '15) (151-160). New York, NY: Association for Computing Machinery. https://doi.org/10.1145/2817946.2817962

Wakabayashi, D. (2018, March 19). Self-driving uber car kills pedestrian in Arizona, where robots roam. The New York Times. Retrieved from https://www.nytimes.com/2018/03/19/technology/uber-driverless-fatality.html

Ward, C., Raue, M., Lee, C., D'Ambrosio, L., & Coughlin, J.F. (2017). Acceptance of automated driving across generations: The role of risk and benefit perception, knowledge, and trust. In: Kurosu, M. (Ed.), *Human-Computer Interaction. User Interface Design, Development and Multimodality. HCI 2017. Lecture Notes in Computer Science, vol 10271* (254-266). Springer, Cham. https://doi.org/10.1007/978-3-319-58071-5_20

Zenith (2019). *Daily time spent with media per capita worldwide from 2011 to 2021, by medium.* Retrieved from https://www.statista.com/statistics/256296/distribution-of-media-time-worldwide/

Zhang, T., Zeng, W., Zhang, Y., Tao, D., Li, G., & Qu, X. (2021). What drives people to use automated vehicles? A meta-analytic review. *Accident Analysis & Prevention, 159*, 106270. https://doi.org/10.1016/j.aap.2021.106270

Zhu, G., Chen, Y., & Zheng, J. (2020). Modelling the acceptance of fully autonomous vehicles: A media-based perception and adoption model. *Transportation Research Part F: Traffic Psychology and Behaviour, 73*, 80-91. https://doi.org/10.1016/j.trf.2020.06.004

Zoellick, J.C., Kuhlmey, A., Schenk, L., Schindel, D., & Blüher, S. (2019). Amused, accepted, and used? Attitudes and emotions towards automated vehicles, their

relationships, and predictive value for usage intention. *Transportation Research Part F: Traffic Psychology and Behaviour*, *65*, 68-78. https://doi.org/10.1016/j.trf.2019.07.009

I also care in manual driving - Influence of type, position and quantity of oncoming vehicles on manual driving behaviour in curves on rural roads

Patrick Roßner, Marty Friedrich, & Angelika C. Bullinger
Chemnitz University of Technology
Germany

Abstract

There is not yet sufficient knowledge on how people want to be driven in a highly automated vehicle. Many studies suggest that automated vehicles should drive like a human driver, e.g. moving to the right edge of the lane when meeting oncoming traffic. To generate naturally looking trajectory behaviour, more detailed studies on manual driving are necessary. This is a driving simulator study investigating different oncoming traffic scenarios in curves. Fourty-six participants experienced three different oncoming traffic scenarios either on a 3.00 m or on a 3.50 m lane width in manual driving. Results show that participants react to oncoming traffic by veering to the right edge of the lane. We also found that the type of oncoming vehicles influences manual driving behaviour. Trucks lead to significantly greater reactions and hence to more lateral distance between the ego and the oncoming vehicle. From this study on manual driving, we recommend an adaptive autonomous driving style which adjusts its trajectory behaviour on type and position of oncoming vehicles. Thus, our results help to design an accepted and trusted trajectory behaviour for highly automated vehicles.

State of knowledge

Sensory and algorithmic developments enable an increasing implementation of automation in the automotive sector. Ergonomic studies on highly automated driving are essential aspects for later acceptance and use of highly automated vehicles (Banks, 2015; Elbanhawi et al., 2015). In addition to studies on driving task transfer or out-of-the-loop issues, there is not yet sufficient knowledge on how people want to be driven in a highly automated vehicle (Gasser, 2013; Radlmayr & Bengler, 2015). First insights show that preferences regarding the perception and rating of driving styles are widely spread. Many prefer their own or a very similar driving style and reject other driving styles that include e.g. very high acceleration and deceleration rates or small longitudinal and lateral distances to other road users (Festner et al., 2016; Griesche et al., 2016; Dettmann et al., 2021). Studies show that swift, anticipatory, safe and seemingly natural driving styles are prioritized (Bellem et al., 2016; Hartwich et al.; 2015; Dettmann et al., 2021). In the literature, trajectory behaviour as one part of the driving style is mostly implemented as a lane-centric

In D. de Waard, S.H. Fairclough, K.A. Brookhuis, D. Manzey, L. Onnasch, A. Naumann, R. Wiczorek, F. Di Nocera, S. Röttger, and A. Toffetti (Eds.) (2022). Proceedings of the Human Factors and Ergonomics Society Europe Chapter 2022 Annual Conference. ISSN 2333-4959 (online). Available from http://hfes-europe.org

position of the vehicle in the lane. From a technical point of view this is a justifiable and logical conclusion, but drivers show quite different preferences, especially in curves and in case of oncoming traffic (Bellem et al., 2017; Lex et al., 2017). In manual driving situations without oncoming traffic, participants drive close to the centre of the lane on straights (Schlag & Voigt, 2015; Rosey et al., 2009). In curves, participants show a different driving behaviour and move closer to the road centre in left turns and closer to the roadside in right turns (Rossner & Bullinger, 2018). Several studies report a tendency to cut the curve by hitting the apex, especially for left turns (Bella, 2005; Bella, 2013; Spacek, 2005). When meeting oncoming traffic in manual driving, participants increase their lateral safety distance by moving to the right edge of the lane, both on straights (Schlag & Voigt, 2015; Rossey et al., 2009; Triggs, 1997) as well as in left and right curves (Lex et al., 2017; Schlag & Voigt, 2015). When meeting heavy traffic, participants' reactions are even greater (Spacek, 2005; Dijksterhuis, 2012; Mecheri et al., 2017; Schlag und Voigt, 2015; Rosey et al., 2009; Räsänen, 2005). With the appearance of oncoming traffic in left curves, two manual driving strategies overlay: to hit the apex and to avoid short lateral distances to the oncoming traffic. Therefore, left curves are going to be the main focus of this study. In summary, the implementation of this natural driving behaviour into an automated driving style includes high potential to improve the driving experience in an automated car. Previous studies (Rossner & Bullinger 2018, Rossner & Bullinger 2019, Rossner & Bullinger 2020a, Rossner & Bullinger 2020b; Rossner et al. 2021) show that reactive trajectory behaviour in highly automated driving leads to significantly higher acceptance, trust and subjectively experienced driving performance on straights and in curves. In order to implement adaptive trajectories that modify trajectory behaviour on different lane widths and adjust their behaviour on type and position of oncoming vehicles, it seems most relevant to investigate manual trajectory behaviour in more detail. The aim of this study is to gain more knowledge on manual driving to implement better reactive trajectories that include less negative side effects and lead to a better driving experience. The results of the study will help to design an accepted and trustfully trajectory behaviour for highly automated vehicles.

Method and variables

A fixed-based driving simulator (Fig. 1) was used to conduct a mixed-design experiment. Forty-six participants experienced three different oncoming traffic scenarios either on a 3.00 m or on a 3.50 m lane width in manual driving.

Table 1. Participant characteristics

	Number	*Age*		*Driving licence holding (years)*		*Mileage last five years (km)*	
	N	*M*	*SD*	*M*	*SD*	*M*	*SD*
Female	13	35.2	6.9	17.3	6.9	60,300	57,000
Male	33	33.9	7.7	15.7	7.8	98,300	69,900
Total	46	34.3	7.4	16.2	7.5	87,500	68,100

Figure 1. Driving simulator with instructor centre (left) and a participant (right)

All participants were at least 25 years old and had a minimum driving experience of 2.000 km last year and 10.000 km over the last five years (see Table 1 for details). On the simulated rural road straight and curve sections were showed in alternation, so after each curve a straight section followed. The curves had a radius of 450 m and a length of 250 m. The test track consisted of 7 right and 19 left curves with none, one or two oncoming vehicles. Left curves with oncoming traffic lead to more safety critical situations and were therefore implemented in a higher number. For the same reason, two oncoming vehicles only occurred in left curves. Oncoming traffic was balanced to minimize sequence and habituation effects. The speed of the oncoming traffic was set at 80 km/h and represented either by a car or a truck. Participants were instructed to drive 100 km/h, but should feel free to reduce speed. However, all curves could be safely passed at 100 km/h (Vetters, 2012). Higher speeds of the ego vehicle were excluded by an activated limiter function at 100 km/h within the driving simulation. Consequently, the ego vehicle encountered the oncoming traffic at the apex of the curve with a very high probability (Fig. 2). Driving data, e.g. velocity or lateral position, was recorded throughout the whole experiment.

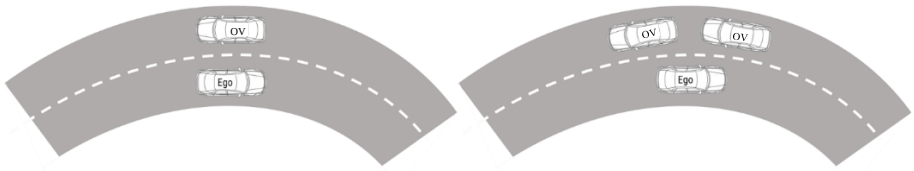


Figure 2. Ego vehicle ecountering oncoming traffic (OV) at the apex of the curve.

Results

Script-based data monitoring discovered zero invalid data recording cases, which needed to be excluded for further analysis. Each curve was splitted into 10 equal parts of 25 m. Driving data were averaged for each section (S). The analysis focused on the lateral behaviour of the ego vehicle in each sector in dependence of oncoming traffic, curve type and lane width. Lateral distance as main dependent variable was measured form the centre of the ego vehicle to the road side (Fig. 3). Left and right curves are reported separately in the following sections.

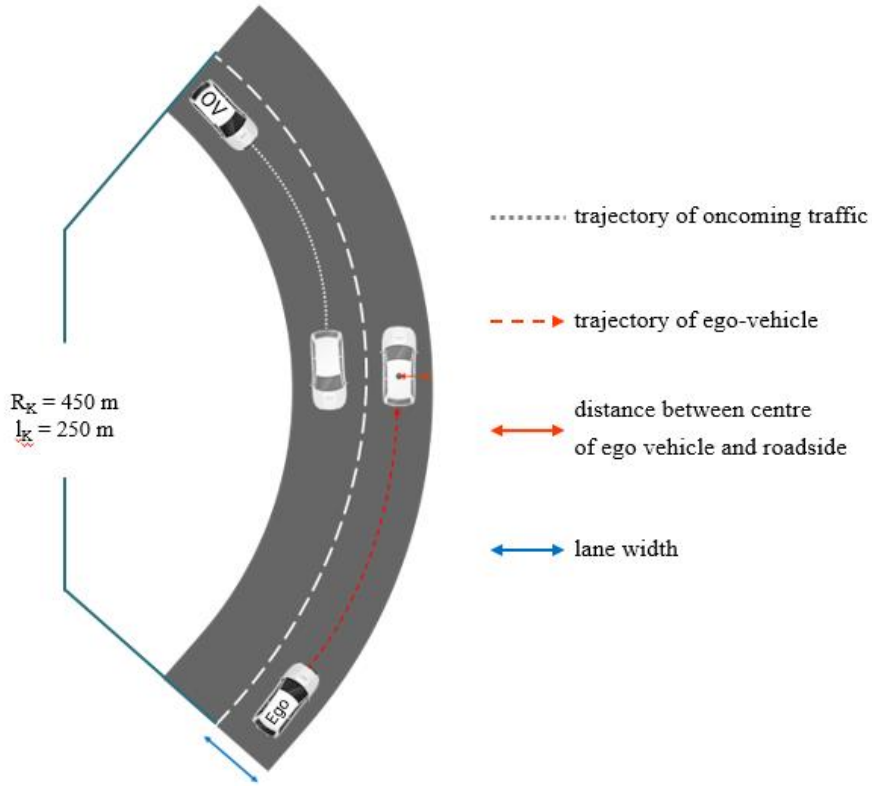


Figure 3. Measures in oncoming traffic scenario. OV = Oncoming vehicle, Ego = Ego vehicle.

Left curves

Without oncoming traffic, a similar behaviour for both lane widths can be observed. Participants enter the curve with a lateral shift to the road side. This represents a typical curve cutting manoeuvre. When passing the curve, participants increase the lateral distance to the roadside and reduce the distance to the road centre until section 9. In section 10, the opposite behaviour is shown, because participants prepared to drive out of the curve into the straight section. With oncoming traffic, a different behaviour can be determined. Between section 2 and 8 on about 150 m driven, a relocation of the trajectory in reaction to the oncoming traffic is performed. Lateral position differs most in section 5, which is the planned meeting point with the oncoming vehicles. On the lane width condition 3.00 m, lateral position without oncoming traffic differs 0.45 m from the truck and 0.27 m from the car scenario. On the lane width condition 3.50 m, participants show smaller reactions. The differences in lateral position without oncoming traffic is 0.37 m to trucks and 0.21 m to cars.

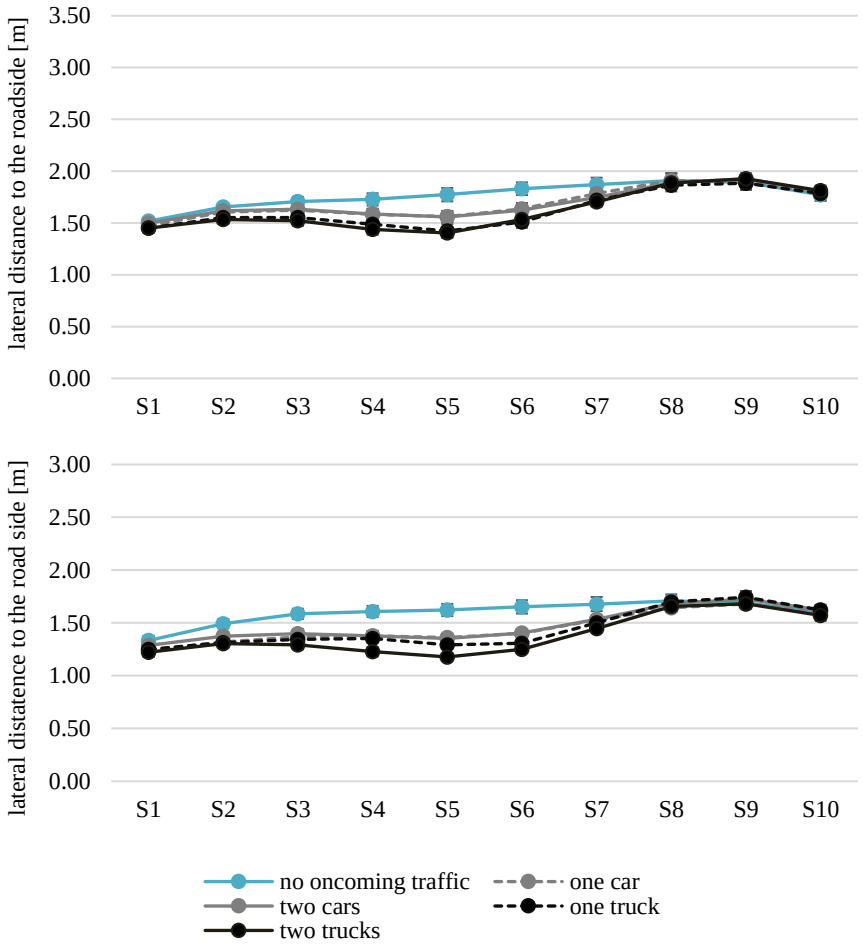


Figure 4. Mean values of lateral distance to the road side for each section in left curves

In addition, it was examined whether the between-subject factor lane width has an influence on the participant's driving behaviour. Using the rmANOVA with between-subject factor, the the lane width can be verified as an influencing factor ($F (1, 44) = 17.17$, $p < .001$, $\eta_p^2 = .28$). A post-hoc analysis according to Bonferroni also showed that the lane behaviour in the curve without oncoming traffic differed significantly from all other traffic situations ($p < .001$). The situation of oncoming traffic with one car dif-fered significantly from the scenario with two oncoming trucks ($p < .001$). The situa-tion with two oncoming cars in comparison to one oncoming truck ($p = .01$) and two oncoming trucks ($p < .001$) have significantly larger distances to the roadside. All oncoming traffic situations compared with one oncoming truck showed no significant difference ($p = .29$).

Right curves

Without oncoming traffic, a similar behaviour for both lane widths can be observed. Participants entre the curve with a lateral shift to the road centre. Again, this represents a typical curve cutting manoeuvre When passing the curve, test participants decrease the lateral distance to the roadside und increase the distance to the road centre until Section 8. In section 9 and 10, the opposite behaviour is conducted, because participants prepared to drive out of the curve into the straight section.

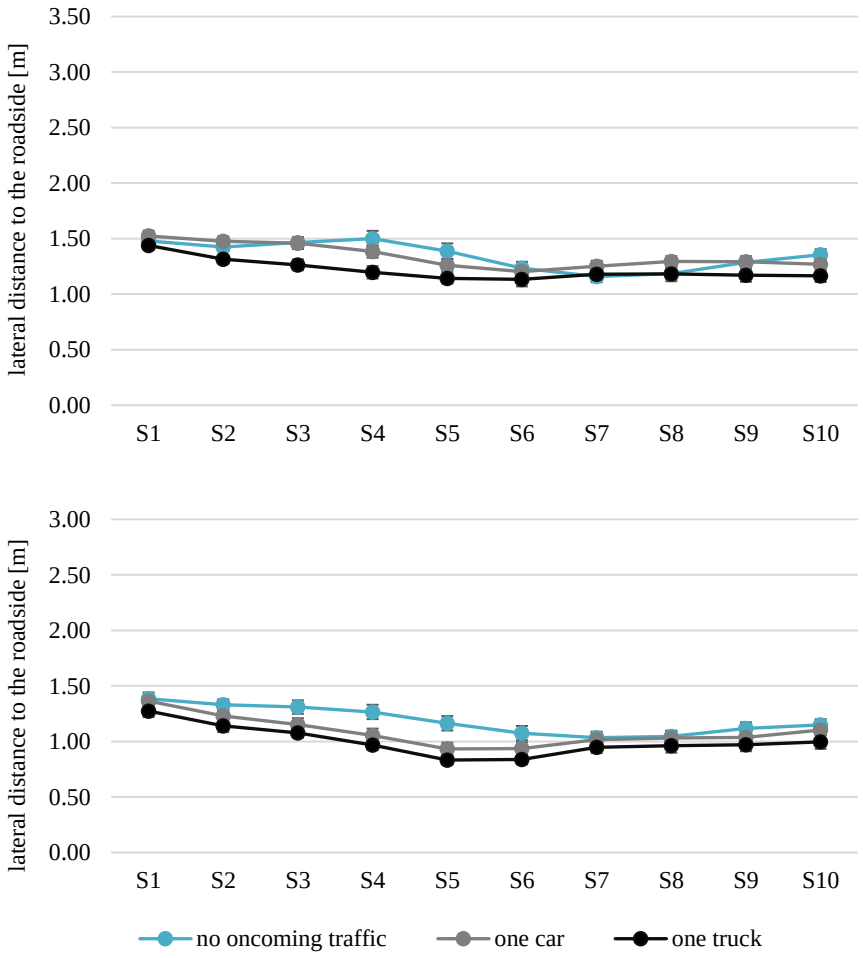


Figure 5. Mean values of lateral distance to the road side for each section in right curves

With oncoming traffic, a different behaviour can be determined. Between section 2 and 7 on about 125 m driven, a relocation of the trajectory because of the oncoming traffic is performed. As seen in left curves, lateral position differs most in section 4

respectively 5, which is the planned meeting point with the oncoming vehicles. On the lane width condition 3.00 m, lateral position without oncoming traffic differs 0.33 m from the truck and 0.23 m from the car scenario. On the lane width condition 3.50 m, participants show smaller reactions. The differences in lateral position without oncoming traffic is 0.25 m to trucks and 0.13 m to cars.

As for the left curves, the factors oncoming traffic ($F(2, 88) = 29.92$, $p < .001$, $\eta_p^2 = .41$) and the position in the curve ($F(3.76, 166.11) = 25.18$, $p < .001$, $\eta_p^2 = .36$) were identified as significant main effects. The lane width also led to significantly different lateral distances ($F(1, 44) = 25.17$, $p < .001$, $\eta_p^2 = .36$). The post-hoc analysis according to Bonferroni showed that the type of oncoming traffic differs for each one (no oncoming traffic vs. car $p < .05$, no oncoming traffic vs. truck $p < .001$, car vs. truck $p < .001$).

Conclusion and outlook

The aim of this study was to gain more knowledge on manual driving to implement better reactive trajectories that include less negative side effects, e.g. passing oncoming traffic with too small distances to the OV or to the road side, and that lead to a better driving experience. The use of manual drivers' trajectories as basis for implementing highly automated driving trajectories shows high potential to increase perceived safety on straights and curves (Rossner & Bullinger 2019; Rossner & Bullinger 2020a; Rossner & Bullinger 2020b; Rossner et al. 2021). Results of the study show in left curves without oncoming traffic, that participants gradually increase the lateral distance to the roadside. This indicates that the participants try to reduce the curve radius and minimize centrifugal forces that would potentially occur in a real world driving environment (Spacek 2005; Schlag & Voigt, 2015). Participants then show the opposite lateral behavior in right curves, but the strategy of passing the curve follows the same scheme. When considering the oncoming traffic situations, a distinction can be made with regard to the type of oncoming traffic. In both curve types and on both lane widths, significant differences in lateral position are found comparing none oncoming traffic, oncoming cars and oncoming trucks. When meeting a car, the lateral safety distance should be increased by moving about 0.20 m to the roadside based on the trajectory without oncoming traffic. If the oncoming vehicle is a truck, the safety distance should be increased by about 0.35 m to the roadside based on the trajectory without oncoming traffic. In contrast to this, the number of oncoming vehicles has no significant effect on the reaction of the participants. It is found that the car situations (one car vs. two cars) and truck situations (one truck vs. two trucks) do not differ significantly from one another. These results amplify the need of adaptive trajectories for highly automated vehicles to generate a positive driving experience and, therefore, higher acceptance rates of highly automated vehicles (Siebert, 2013; Hartwich et al., 2015). In all use cases, a safe driving performace has to be guaranteed during the whole drive. The overall safety Finally, the limitations of studies in fixed-based driving simulators depict the transfer of the results to real world driving situations. Bella (2009) arguments that specific use cases can be researched and various parameters (e.g. speed, lateral distance, angle of the brake pedal) can be recorded in driving simulator studies. Of course, no movement forces are perceptible, but the visual impression has a great

influence on the perception of the oncoming traffic situations and the perceived lateral distances. Curves with a radius of 450 m and a length of 250 m can be passed safely with 100 km/h, so that the absence of movement forces is not that important. Nevertheless, it is very recommeded to conduct a similar study in a real world environment. The results also cover only a small part of the existing use cases. Other factors, such as meeting oncoming traffic at the beginning respectively the end of the curve, curves with additional horizontal course or the influence of additional traffic on the ego vehicle's lane are further topics to be investigated.

Acknowledgements

This research was partially supported by the German Federal Ministry of Education and Research (research project: KomfoPilot, funding code: 16SV7690K) as well as by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project-ID 416228727 – SFB 1410. The sponsor had no role in the study design, the collection, analysis and interpretation of data, the writing of the report, or the submission of the paper for publication. We are very grateful to Maximilian Hentschel for his assistance in driving simulation programming.

Reference

Banks, V.A., & Stanton, N.A. (2015). Keep the driver in control: Automating automobiles of the future. *Applied Ergonomics, 53B,* 389-395. https://doi.org/10.1016/j.apergo.2015.06.020

Bella, F. (2005). Speeds and Lateral Placements on Two-Lane Rural Roads: Analysis at the Driving Simulator. In 13th International Conference "Road Safety on Four Continents".

Bella, F. (2009). Can Driving Simulators Contribute to Solving Critical Issues in Geometric Design? *Transportation Research Record*: Journal of the Transportation Research Board, 2138, 120–126. https://doi.org/10.3141/2138-16.

Bella, F. (2013). Driver perception of roadside configurations on two-lane rural roads: Ef-fects on speed and lateral placement. *Accident Analysis and Prevention, 50*, 251–262. https://doi.org/10.1016/j.aap.2012.04.015.

Bellem, H., Schönenberg, T., Krems, J.F., & Schrauf, M. (2016). Objective metrics of comfort: Developing a driving style for highly automated vehicles. *Transportation Research Part F: Traffic Psychology and Behaviour, 41*, 45-54 .

Bellem, H., Klüver, M., Schrauf, M., Schöner, H.-P., Hecht, H. & Krems, J.F. (2017). Can We Study Autonomous Driving Comfort in Moving-Base Driving Simulators? A Validation Study. *Human Factors, 59*, 442–456. doi:10.1177/0018720816682647.

Dettmann, A., Hartwich, F., Roßner, P., Beggiato, M., Felbel, K., Krems, J., & Bullinger, A.C. (2021). Comfort or not? Automated Driving Style and User Characteristics Causing Human Discomfort in Automated Driving. *International Journal of Human–Computer Interaction, 4*, 331–339. https://doi.org/10.1080/10447318.2020.1860518.

Dijksterhuis, C., Stuiver, A., Mulder, B., Brookhuis, K.A. & de Waard, D (2012).: An adaptive driver support system: user experiences and driving performance in a simulator. *Human Factors, 54*, 772–785. https://doi.org/10.1177/0018720811430502.

Elbanhawi, M., Simic, M., & Jazar, R. (2015). In the Passenger Seat: Investigating Ride Comfort Measures in Autonomous Cars. *IEEE Intelligent Transportation Systems Magazine, 7(3)*, 4–17. https://doi.org/ 10.1109/MITS.2015. 2405571.

Festner, M., Baumann, H., & Schramm, D. (2016). Der Einfluss fahrfremder Tätigkeiten und Manöverlängsdynamik auf die Komfort- und Sicherheitswahrnehmung beim hochautomatisier-ten Fahren. 32nd VDI/VW- Gemeinschaftstagung Fahrerassistenz und automatisiertes Fahren, Wolfsburg.

Gasser, T.M. (2013). Herausforderung automatischen Fahrens und Forschungs-schwerpunkte. 6. Tagung Fahrerassistenz, München.

Griesche, S., Nicolay, E., Assmann, D., Dotzauer, M., & Käthner, D.:(2016). "Should my car drive as I do? What kind of driving style do drivers prefer for the design of automated driving func-tions?" Contribution to 17th Braunschweiger Symposium Automatisierungssysteme, Assis-tenzsysteme und eingebettete Systeme für Transportmittel (AAET), ITS automotive nord e.V., pp. 185-204, ISBN 978-3-937655-37-6.

Hartwich, F., Beggiato, M., Dettmann, A., & Krems, J.F.(2015). Drive me comfortable: Custom-ized automated driving styles for younger and older drivers. 8. VDI-Tagung „Der Fahrer im 21. Jahrhundert".

Köhler, B. (2017). *Auswirkungen der Wahrnehmung von Markierungskonstellationen auf das Fahrverhalten in Arbeitsstellen auf Bundesautobahnen*. Dissertation, Karlsruher Instituts für Technologie. Karlsruhe, Germany.

Leutzbach, W., Maier, W. & Döhler, M. (1981): Untersuchung des Spurverhaltens von Kraftfahrzeugen auf Landstraßen durch Verfolgungsfahrten. Forschungsgesellschaft für Straßen und Verkehrswesen. Straße und Autobahn, Heft 8.

Lex, C., Schabauer, M., Semmer, M., Magosi, Z., Eichberger, A., Koglbauer I., Holzinger, J. & Schlömicher, T. (2017). Objektive Erfassung und subjektive Bewertung menschlicher Trajek-toriewahl in einer Naturalistic Driving Study. VDI-Berichte Nr. 2311, pp. 177-192.

Mecheri, S., Rosey, F., & Lobjois, R. (2017). The effects of lane width, shoulder width, and road cross-sectional reallocation on drivers' behavioral adaptations. *Accident Analysis and Prevention, 104*, 65–73. https://doi.org/10.1016/j.aap.2017.04.019.

Radlmayr, J., & Bengler, K. (2015) Literaturanalyse und Methodenauswahl zur Gestaltung von Systemen zum hochautomatisierten Fahren. FAT-Schriftenreihe, vol. 276. VDA, Berlin.

Räsänen, M. (2005). Effects of a rumble strip barrier line on lane keeping in a curve. *Accident Analysis and Prevention, 37*, 575–581. https://doi.org/10.1016/j.aap.2005.02.001.

Rosey, F., Auberlet, J.-M., Moisan, O. & Dupré, G. (2009). Impact of Narrower Lane Width: Comparison Between Fixed-Base Simulator and Real Data.

Transportation Research Record: Journal of the Transportation. Research Board, 2138(1), 112–119. https://doi.org/10.3141/2138-15.

Rossner, P. & Bullinger, A.C.: Drive me naturally: Design and evaluation of trajectories for highly automated driving manoeuvres on rural roads. Technology for an Ageing Society, Postersession Human Factors and Ergonomics Society Europe Chapter 2018 Annual Conference, Berlin (2018). https://www.hfes-europe.org/posters-2018/

Rossner P. & Bullinger A.C. (2019) Do You Shift or Not? Influence of Trajectory Behaviour on Perceived Safety During Automated Driving on Rural Roads. In: Krömker H. (eds) HCI in Mobility, Transport, and Automotive Systems. HCII 2019. Lecture Notes in Computer Science, vol 11596. Springer, Cham.

Rossner P. & Bullinger A.C. (2020a). Does driving experience matter? Influence of trajectory behaviour on drivers' trust, acceptance and perceived safety in automated driving. Understanding Human Behaviour in Complex Systems, In D. de Waard, A. Toffetti, L. Pietrantoni, T. Franke, J-F. Petiot, C. Dumas, A. Botzer, L. Onnasch, I. Milleville, and F. Mars (2020). Proceedings of the Human Factors and Ergonomics Society Europe Chapter 2019 Annual Conference, pp 73-84. https://www.hfes-europe.org/understanding-human-behaviour-complex-systems/ ISSN 2333-4959.

Rossner P. & Bullinger A.C. (2020b) I Care Who and Where You Are – Influence of Type, Position and Quantity of Oncoming Vehicles on Perceived Safety During Automated Driving on Rural Roads. In Krömker H. (Ed.) HCI in Mobility, Transport, and Automotive Sys-tems. Driving Behavior, Urban and Smart Mobility. HCII 2020. Lecture Notes in Computer Science, vol 12213. Springer, Cham. https://doi.org/10.1007/978-3-030-50537-0_6.

Rossner, Friedrich, & Bullinger (2021). Hitting the Apex Highly Automated? – Influence of Trajectory Behaviour on Perceived Safety in Curves. In: HCI International 2021 - Late Breaking Papers: HCI Applications in Health, Transport, and Industry: 23rd HCI International Conference, HCII 2021, Virtual Event, July 24–29 2021, Proceedings, pp. 322–331. https://doi.org/10.1007/978-3-030-90966-6_23.

Schlag, B., & Voigt, J. (2015) Auswirkungen von Querschnittsgestaltung und längsgerichteten Markierungen auf das Fahrverhalten auf Landstrassen. Berichte der Bundesanstalt für Straßenwesen. Unterreihe Verkehrstechnik, (249).

Siebert, F., Oehl, M., Höger, R., & Pfister, H.R. (2013). Discomfort in Automated Driving – The Disco-Scale. In: Proceedings of HCI International 2013, Communications in Computer and Information Science, vol. 374, pp. 337-341, Las Vegas, USA.

Spacek, P. (1999). Fahrverhalten und Unfälle in Kurven - Fahrverhalten in Kurvenbereichen (Straßenverkehrstechnik, Bd. 2). Zürich: VSS. Available at https://books.google.de/books?id=rCD1HgAACAAJ.

Spacek, P. (2005). Track Behavior in Curve Areas: Attempt at Typology. *Journal of Transportation Engineering, 131*, 669–676. https://doi.org/10.1061/(ASCE)0733-947X(2005)131:9(669).

Triggs, T.J. (1997). The Effect of Approaching Vehicles on the Lateral Position of Cars Travelling on a Twolane Rural Road. *Australian Psychologist, 32(3)*, 159–163. https://doi.org/10.1080/00050069708257375.

Vetters, A. (2012). Die neuen "Richtlinien für die Anlage von Landstraßen" RAL - Stand 2012. Zugriff am 28.10.2019. Verfügbar unter http://www.vsvi-mv.de/fileadmin/Medienpool/Seminarunterlagen/Seminare_2012/Vortrag_1_-_neue_RAL_Frau_Vetters.pdf.

Voß, G. & Schwalm, M. (2017). Bedeutung kompensativer Fahrerstrategien im Kontext automatisierter Fahrfunktionen. Berichte der Bundesanstalt für Straßenwesen, *Fahrzeugtechnik Heft F 118*, ISBN 978-3-95606-327-5.

Sleeping during highly automated driving – target groups and relevant use cases of an in-car sleeping function

Markus Tomzig & Christina Kaß
Würzburg Institute for Traffic Sciences (WIVW)
Germany

Abstract

In highly automated driving (SAE Level 4), the driver will be no longer responsible for driving and can sleep during the ride. This opportunity is likely to change user needs. Our work focuses on the first phase of the user-centred design approach and aims to identify the target groups who are willing to use the sleep function and the relevant use cases. First, we conducted an online survey with $N = 264$ participants to investigate the characteristics that describe the future users of the sleep function. To derive relevant use cases, $N = 7$ participants of the online survey with a high intention to sleep during automated driving were invited to a subsequent interview study. The online survey identified predictors for a high intention to use a sleep function, such as young age as well as a high frequency and duration of sleeping as a passenger in public transport or cars. However, the results showed that there is no distinct target group. The interviews revealed that the wish to sleep during automated driving is related to the individual's current mobility behaviour and the personal desire to enhance comfort during inconvenient trips. We derived exemplary use cases. Future research should identify requirements for comfortable sleep during highly automated driving.

Introduction

Nowadays, sleeping as a car driver is still a vision of future mobility. Previous studies found relaxing, napping and sleeping as very popular non-driving related task (Becker et al., 2018; Kyriakidis, Happee & de Winter, 2015). Whereas in today's cars it is neither possible nor legal to sleep as a car driver, this scenario could become real with the advancing automation of vehicles. The Society of Automotive Engineers (SAE) differs between six levels of automation (SAE, 2021). While in Levels 0 to 3, the driver has to be at least ready to take over the driving task, in Levels 4 and 5 users do not need to supervise the driving task. Sleeping as a driver would, thus, be legitimate initially in Level 4, called "high driving automation" (SAE, 2021).

Up to now, it is unclear how such an in-car sleeping function must be designed to fit user requirements. Before specifying user's requirements and designing and realising solutions, the user-centred design approach demands to specify the context of use (DIN, 2010). Therefore, it is important to specify the target group and potential use

In D. de Waard, S.H. Fairclough, K.A. Brookhuis, D. Manzey, L. Onnasch, A. Naumann, R. Wiczorek, F. Di Nocera, S. Röttger, and A. Toffetti (Eds.) (2022). Proceedings of the Human Factors and Ergonomics Society Europe Chapter 2022 Annual Conference. ISSN 2333-4959 (online). Available from http://hfes-europe.org

cases of the function to be developed. The use cases do not only depend on the technical feasibility but also on the requirements of the target group.

To identify the target group of a Level 4 sleeping function, we conducted an online survey. The goal of this survey was to exploratively reveal characteristics which are associated with the intention to sleep during highly automated driving and to outline potential use cases. As a survey cannot explain the personal motives contributing to the wish to sleep during automated driving, a subsequent interview study aimed to gain a deeper understanding about the target group's motives and needs that lead to their wish to sleep in the car. The interviews ought to reveal distinct and attractive usage scenarios that meet the needs of the target group.

Methods

Online survey

Sample and procedure

The participants were recruited from the test driver panel of the Wuerzburg Institute for Traffic Sciences (WIVW), via project partners and social media. In total, $N = 264$ ($n = 79$ female, $n = 185$ male, $n = 0$ diverse) aged between 18 and 83 years ($M = 47.1$, $SD = 15.7$) took part in the survey.

The survey was conducted in German language with the software LimeSurvey (LimeSurvey GmbH). Completing the survey took about ten minutes. At the beginning, the participants were informed that their participation was voluntary, not financially compensated and that personal data would only be collected if the participant was interested in taking part in the subsequent interview study. At the end, participants could optionally provide their name and email to be invited in the interview study. All data was collected and processed in accordance to the European General Data Protection Regulation.

The survey was conducted between December 2020 and January 2021. At this time, it had to be assumed that the Covid-19 pandemic significantly affected the mobility of large parts of the society. For this reason, participants were asked to answer all questions about their mobility as they had experienced it prior to the onset of the pandemic, i.e., during a period with fewer or no home office and travel restrictions.

Measures

The main goal of the survey was the identification of characteristics describing a target group for an in-car sleeping function. The criterion determining the target group was defined by the intention to use such a future sleeping function. The intention to use was measured by the level of agreement with the statement "In the future, I would sleep during the automated drive if I had the opportunity to do so". The approval was indicated on a seven-point Likert-scale from "fully disagree" to "fully agree".

To statistically describe the target group, several potential predictors have been included in the survey. The investigated predictors addressed the following aspects: socio-demographic characteristics (age, gender, educational level, amount of weekly working hours, urbanisation of the residential environment), mobility (frequency of

drives in the city, on rural roads, on motorways, at night, and in traffic jams as well as the annual mileage, commuting time, the frequency of private and business trips), sleeping behaviour (frequency and duration of naps as passenger in cars, in means of public transport, and at home, as well as the feeling to get sufficient sleep in everyday life), attitude towards driving (association of driving a car with fun, strain, stress, and discomfort as well as the prior knowledge about automated driving).

The second goal of the survey was to explore the context, in which a sleeping function is likely to be used by members of the target group. Therefore, all subjects who had indicated at least slight agreement to the criterion were conceived as members of the target group and were asked additional questions about their desired use cases of an in-car sleeping function. The participants were asked about their preferred time of day to use a sleeping function. Furthermore, they were asked about the estimated frequency and duration of use. Additional items explored the desired occasion of the drive and the types of road. A detailed list of all items and the used scales can be found in table 5 in the appendix.

Data analysis
As there were no assumptions about a higher level model as well as a hierarchical order of the predictors, all predictors were analysed by simple linear regressions.

Due to the large number of statistical tests, the *p*-values were adjusted according to Bonferroni-Holm (Holm, 1979) within each category of predictor (socio-demographic characteristics, mobility, sleeping behaviour, attitude towards driving). The level of significance is $\alpha = .05$ for all adjusted *p*-values.

Interview study

Sample and procedure
$N = 7$ participants ($n = 3$ female, $n = 4$ male, $n = 0$ diverse) aged between 32 and 63 years ($M = 40.4$, $SD = 10.9$) took part in the interview study. All participants had previously participated in the online survey and indicated that they would use a future in-car sleeping function. Therefore, they were conceived as members of the target group (answer to the item “In the future, I would sleep during the automated drive if I had the opportunity to do so” with at least “rather agree”). On average, the sample had a high intention to use ($M = 6.4$, $SD = 0.8$, $Min = 5$ “rather agree”, $Max = 7$ “fully agree”).

After giving informed consent to the study procedure and data collection in accordance to the European General Data Protection Regulation, the interview was conducted as individual online video conference. The interview was conceptualised as semi-structured, qualitative survey with an open answering form. Conducting the interview took about 45 minutes.

Measures
The interview was structured in four thematic parts. The first part was dedicated to better understand members of the target group. Participants were asked about the reasons why they wished to sleep during highly automated driving, and which benefits they saw in an in-car sleeping function. In the second block of questions, the

participants were asked to describe in detail a desired use case of the in-car sleeping function. The interviewer asked further questions, e.g., about the occasion, destination, and duration of the drive, to what time of day it may take place and how frequently these drives would occur. The third part was about the actual mobility of the participants. The participants were asked how regularly and how frequently they used a car and other means of transport (e.g., trains and busses) and also asked about the trips' occasions, durations and times of day. At the end of the interview, the interviewer asked whether the participants' mobility would change if there was the possibility to sleep during the trip.

Data analysis
The participants' qualitative verbal responses were tagged and clustered by content, meaning that responses were grouped across different questions if the participant provided related information herein. Because of the small sample size, inference statistical tests are not indicated, and all analyses were carried out qualitatively and descriptively.

Results

Online survey

Relationship between the investigated predictors and the intention to use an in-car sleeping function
Of the $N = 264$ surveyed participants, $n = 106$ agreed at least slightly to the criterion ($n = 47$ "rather agree", $n = 42$ "agree", $n = 17$ "fully agree"). The proportion of these as target group conceived persons in the entire sample was thus 40.2%. Of the remaining participants, $n = 147$ rejected the idea of sleeping during the drive ($n = 44$ "rather disagree", $n = 41$ "disagree", $n = 57$ "fully disagree"). The remaining $n = 16$ participants indicated "neither nor".

Among the socio-demographic characteristics, age significantly predicted the intention to use the sleeping function, indicating that younger participants had a higher intention to use than older ones. Further, the educational level significantly predicted the intention to use: A higher educational degree was associated with a higher intention to use. The gender, amount of weekly working hours and urbanisation of the residential environment could not predict the intention to use. The statistical results of the socio-demographic variables can be found in table 1.

Table 1. Results of linear regressions for socio-demographic variables

Variable	*β*	*Adj.R²*	*F*	*df*	*p*	*Adj. p*
Age	-.035	.071	21.130	1, 261	< .001	< .001
Gender	.224	< .001	0.703	1, 262	.402	.402
Highest educational level	.295	.085	25.450	1, 262	< .001	< .001
Amount of weekly working hours	.008	< .001	1.022	1, 262	0.313	.626
Urbanisation of the residential environment	.189	.011	3.934	1, 262	0.048	.144

After Bonferroni-Holm adjustment, the participants' mobility did not significantly predict the intention to use. There were tendencies that the intention to use was predicted by the frequency of night drives and the frequency of traffic jams, indicating that people may have a lower intention to use if they experience frequent night drives or traffic jams. However, after adjusting the *p*-values, these models are conceived as not significant. The statistical results of these and the other mobility variables are listed in table 2.

Table 2. Results of linear regressions for mobility variables

Variable	*β*	*Adj.R²*	*F*	*df*	*p*	*Adj. p*
Frequency of drives in the city	-.162	.005	2.191	1, 262	.140	.840
Frequency of drives on rural roads	-.171	.005	2.282	1, 262	.132	.924
Frequency of drives on motorways	-.075	< .001	0.496	1, 262	.482	> .999
Frequency of drives at night	-.247	.011	3.881	1, 262	.050	.398
Frequency of drives in traffic jams	-.410	.016	5.220	1, 262	.023	.207
Annual mileage	-.038	< .001	0.123	1, 262	.726	> .999
Commuting time (single way)	.089	< .001	0.550	1, 187	.459	> .999
Frequency of business trips longer than 2 hours	.119	< .001	0.721	1, 262	.397	> .999
Frequency of private trips longer than 2 hours	.057	< .001	0.115	1, 262	.734	.734

In the domain of sleeping behaviour, the frequency and duration of taking a nap as passenger in cars in the present significantly predicted the intention to use. Passengers who usually sleep longer and more often had a higher intention to use an in-car sleeping function. The same applies for persons who sleep regularly and for long periods in public means of transport. These variables also significantly predicted the intention to use. In contrast, the frequency of taking a nap at home did not significantly predict the wish to use a sleeping function. However, participants who perceived their regular nocturnal sleep as insufficient had a higher intention to use than people with sufficient sleep. The statistical results of all variables concerning sleeping behaviour are listed in table 3.

Table 3. Results of linear regressions for variables concerning sleeping behaviour

Variable	*β*	*Adj. R²*	*F*	*df*	*p*	*Adj. p*
Frequency of naps as passenger in cars on trips longer than 40 minutes	.663	.102	27.020	1 ,228	< .001	< .001
Duration of naps as passenger in cars on trips longer than 40 minutes	.630	.096	25.850	1, 233	< .001	< .001
Frequency of naps as passenger in public means of transport on trips longer than 40 minutes	.802	.175	44.760	1, 205	< .001	< .001
Duration of naps as passenger in public means of transport on trips longer than 40 minutes	.486	.121	29.270	1, 205	< .001	< .001
Frequency of naps at home up to 40 minutes	.057	< .001	0.144	1, 259	.704	> .999
Feeling of getting sufficient sleep	-.202	.021	6.517	1, 259	.011	.022

All measured variables about the attitude towards driving significantly predicted the intention to use an in-car sleeping function. Participants who associated driving with strain, stress, or discomfort and did not associate driving with fun had a higher intention to use. Prior knowledge about automated driving correlated positively with the appreciation of an in-car sleeping function. The statistical results about the attitude towards driving are listed in table 4.

Table 4. Results of linear regressions for variables concerning attitudes towards driving

Variable	*β*	*Adj. R²*	*F*	*df*	*p*	*Adj. p*
Association of driving a car with fun	-.241	.026	7.932	1, 262	.005	.021
Association of driving a car with strain	.212	.029	8.868	1, 262	.003	.016
Association of driving a car with stress	.216	.025	7.619	1, 262	.006	.012
Association of driving a car with discomfort	.203	.015	5.130	1, 262	.024	.024
Prior knowledge about automated driving	.166	.025	7.753	1, 262	.006	.017

Characteristics of use cases for an in-car sleeping function
The *n* = 106 participants with an intention to use an in-car sleeping function were asked the further questions about the circumstances in which they could imagine sleeping during the drive. The majority of the subsample wanted to use the function occasionally (*n* = 44) or rarely (*n* = 35). *N* = 14 imagined using the sleeping function frequently, *n* = 4 very frequently. *N* = 9 would (almost) never sleep during automated driving. The minimum imagined travel time to make use of the in-car sleeping function was on average *M* = 77.5 minutes (*SD* = 76.3, *min* = 10, *max* = 480).

The majority of the target group could imagine sleeping on drives to their holiday destination (*n* = 93) or on recreational trips (*n* = 63). For daily routine trips (e.g., shopping), a sleeping function does not seem to be a promising feature (*n* = 5). In the domain of job-related occasions, *n* = 61 wanted to sleep on their way from work, followed by *n* = 52 who wished to sleep on vocational drives, and *n* = 38 on their way to work.

Sleeping during automated driving was conceivable at all times of day with a preference of the early morning and the night (cf. figure 1). A sleeping function would be used mainly on motorways (*n* = 95) and during traffic jams (*n* = 90). The vision of sleeping on federal roads (*n* = 57) or straight rural roads (*n* = 55) found partial appeal. Sleeping on urban roads (*n* = 17) or on bendy rural roads (*n* = 12) does not seem to be promising.

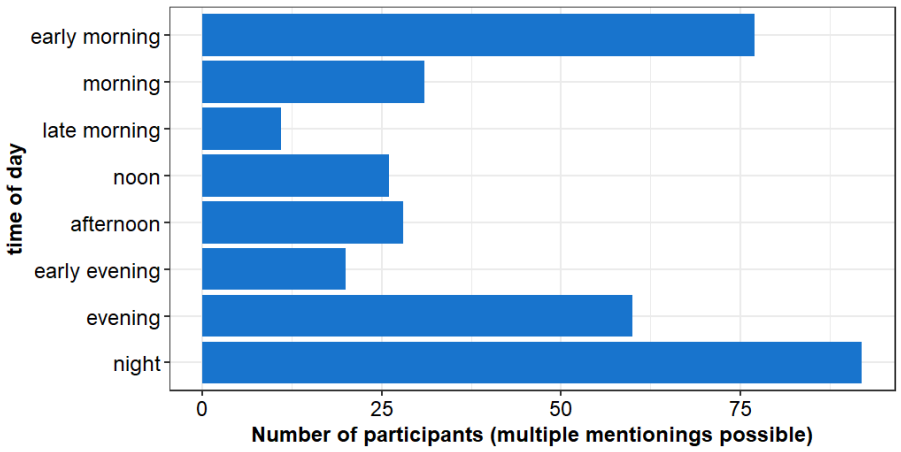


Figure 1. Intended times of day to use an in-car sleeping function.

Interview study

Reasons for the wish to sleep during automated driving
When asked about the reasons for their intention to use an in-car sleeping function, participants mentioned both, problems they associate with manual driving and solutions they expect from an in-car sleeping function. With regard to today's

problems, participants indicated that longer trips often evoke strain, stress, and persistent concentration. As a result, participants reported that they quickly become tired during long or overnight car trips and need to plan their driving times according to their fitness level. One participant added feeling uncomfortable before long trips.

Consistent with the aforementioned problems, on the benefits side, participants reported that sleeping during the trip offered the opportunity to relax and recover. The vision of an in-car sleeping function was further associated with higher driving comfort. It was expected that this would make one arrive at the destination recovered instead of fatigued. In addition to the aspect of comfort, the participants also mentioned a gain in time due to the in-car sleeping function. The time of day could be utilized better and a lack of sleep e.g., from the previous night, could be caught up during the trip. Furthermore, it would be more attractive to travel at off-peak times than it is nowadays.

Derived use cases of an in-car sleeping function

The scenarios described by the participants were clustered into three use cases. Most participants named more than one imaginable use case. The majority (n = 6) wished to sleep on drives to visit their families and/or friends who do not live nearby. The described trips took between one and five hours and were typically made on weekends (e.g., as weekend commuters). The mentioned drives began in the late afternoon and ended in the early or late evening, depending on the driven distance. With regard to the taken sleep, parts of the nocturnal sleep would be translocated into the vehicle.

N = 5 participants wished to sleep (also) on shorter trips that take between 30 and 45 minutes and start either in the early morning (outbound) or in the late afternoon (return). As occasions, participants named destinations for one day, mostly to get to and from work (n = 4) but also recreational trips (n = 1). In this use case, the participants would rather nap instead of sleeping deeply.

Third, n = 3 participants described the wish to use an in-car sleeping function during long drives to their holiday destination. The participants indicated that they already partially do these trips at night today. Traveling time is supposed to be between 8 and 10 hours. With an in-car sleeping function, the participants anticipated to sleep through the entire night and to arrive recovered.

Anticipated changes in mobility due to an in-car sleeping function

N = 3 participants indicated that their mobility behaviour would change due to the possibility to sleep during the trip. In terms of changes, the participants named to make the most of holiday trips, to travel long distances more often, and to generally travel more often and at other times. N = 2 participants felt that their mobility would change only slightly. They expected to have to plan less before trips to be fit to drive. Furthermore, they wanted to reschedule their holiday journeys into the night. The remaining n = 2 participants anticipated no changes in their mobility behaviour due to an in-car sleeping function. Their driving routes and driving times were already fixed and could not be changed easily (e.g., way to work).

Discussion

The development of high driving automation will enable the driver to completely refrain from the driving task and to sleep during the trip. The presented studies focused on the first phase of the user-centred design approach and aimed at identifying the target groups who are willing to use an in-car sleeping function and to reveal the relevant use cases with an exploratory approach. An online survey with N = 264 participants and a subsequent interview study with N = 7 participants have been conducted.

Implication of results

The results of the online survey demonstrate that an in-car sleeping function is perceived as an attractive feature. The intention to use cannot be ascribed to a distinct target group but is associated with several predictors such as young age, high education, regular naps in means of transport, and aversion to manual driving.

The online survey examined the attractiveness of different scenarios of use (duration and occasion of the journey, road conditions, and the time of day). Especially long trips and trips to recreational and holiday destinations found agreement. Furthermore, sleeping during automated driving is imaginable at any time of day, with focus on the night and early morning. However, the online survey cannot explain the personal motives contributing to the wish to sleep during automated driving. The survey's results reveal characteristics for future use cases, but they cannot explain how these characteristics are related to one another.

Targeting on these issues, the results of the subsequent interview study showed that members of the target group expect an eased and restorative instead of a stressful and straining drive. Members of the target group are unified by the common wish to improve recurring straining drives they are already experiencing today. The interviews showed how the characteristics of the expected use cases matched to one another: The participants named day trips taking between 30 and 45 minutes starting in the early morning and/or late afternoon, weekend trips taking between one and five hours starting in the late afternoon, and long holiday trips taking between eight and ten hours. These results clarify that an in-car sleeping function must address a wide application area. Members of the target group expect short naps as well as long, deep sleep. This should be considered in the further ergonomic development of the sleeping function. A sleeping function should be able to be used at all times of day. In contrast, the interviews provided evidence that long sleeping periods during the day are as unlikely as short powernaps during night trips.

A considerable part of the interview sample expected changes in their mobility due to the availability of an in-car sleeping function. However, we consider the indicated changes as rather small. Essentially, the participants expected to travel longer distances and more often. Major changes in mobility, e.g., changes in their place of residence or job have not been mentioned.

Methodological limitations

The results are supposed to provide a forecast about future target groups and use cases of an in-car sleeping function. The accuracy of this forecast may be limited by the fact that the surveyed sample is not representative to the general population and may have been subject to a self-selection. It is, thus, possible that the participants had a higher interest to the survey's topic than the general population which may lead to an overestimated interest in the sleeping function. The actual proportion of the target group by the entire population may, thus, be overestimated. Further, the interviews examined only a very small subsample and gathered qualitative results and perspectives of few persons. Conclusions to the general population must therefore be drawn very carefully. However, we consider that the subsample fairly represents the target group as the subsample's results in the interviews fitted the results of the entire sample in the online survey. Moreover, the interviews provided complementary impressions about needs and motives of the target group.

The quality of the forecast may further be reduced due to the fact that to this point of time, high driving automation is not generally available and sleeping as driver is not possible yet. Therefore, the results base on today's needs which can be solved by a future in-car sleeping function. This may explain that the use cases reflect mobility behaviour that is rather typical for today's time. It must be considered that needs and mobility may change with emerging new technologies. It cannot be excluded that high driving automation and the possibility to sleep during the trip encourage new patterns of mobility which can hardly be discovered by an early user-centred approach. Likewise, up to this point, little is known about the technical conditions, possibilities, and limitations of a sleeping function. It is therefore unclear whether the user requirements are technically feasible (e.g., highly automated driving for eight hours to holiday destination).

Conclusion

The online survey showed that there is not one distinct target group that would like to use an in-car sleeping function, but that there are several indicators that predict the intention to use. The interviews explored conceivable use cases. A sleeping function is appreciated if recurring drives are perceived as inconvenient (e.g., associated with fatigue). Further research is now required to examine the user requirements towards the design of an in-car sleeping function. To present design solutions, technical development must also be taken into account.

References

Becker, T., Herrmann, F., Duwe, D., Stegmüller, S., Röckle, F., & Unger, N. (2018). *Enabling the value of time. Implications for the interior design of autonomous vehicles.* Stuttgart, Germany: Fraunhofer IAO.

DIN. (2010). *Ergonomics of human-system interaction - Part 210: Human-centred design for interactive systems* (ISO 9241-210:2019); German version. Berlin, Germany: Deutsches Institut für Normung.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics, 6*, 65-70.

Kyriakidis, M., Happee, R., & de Winter, J.C. (2015). Public opinion on automated driving: Results of an international questionnaire among 5000 respondents. *Transportation Research Part F: Traffic Psychology and Behaviour, 32*, 127-140.

SAE. (2021). *Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles* (Vol. J3016). SAE International.

Appendix

Table 5. Items used in the online survey

Category	*Item*		*Answer*
Socio-demographics	1.	Age	Numerical input [years]
	2.	Gender	Single choice • female • male • diverse
	3.	Highest educational level	Single choice • no graduation • lower secondary graduation • secondary graduation • advanced technical certificate • higher education entrance qualification • bachelor's degree • master's degree/diploma • doctoral degree
	4.	Amount of weekly working hours	Numerical input [hours per week]
	5.	Urbanisation of the residential environment	Single choice • rural • rather rural • rather urban • urban • metropolitan
Mobility	6.	Frequency of drives in the city	Single choice • fewer than 1x per week • 1-2x per week • 3-5 times per week • (almost) daily
	7.	Frequency of drives on rural roads	cf. item 7
	8.	Frequency of drives on motorways	cf. item 7
	9.	Frequency of drives at night	cf. item 7
	10.	Frequency of drives in traffic jams	cf. item 7
	11.	Annual mileage	Single choice • up to 5000 km • 5000 – 15000 km • 15000 – 25000 km • 25000 – 35000 km • more than 35000 km

Category	*Item*		*Answer*
	12.	Commuting time (single way)	Single choice • up to 15 minutes • 15 – 30 minutes • 30 – 45 minutes • 45 – 60 minutes • more than 60 minutes
	13.	Frequency of business trips longer than 2 hours	Single choice • never or rarely • multiple times per year • multiple times per month • weekly or more frequently
	14.	Frequency of private trips longer than 2 hours	cf. item 13
Sleeping behaviour	15.	Frequency of naps as passenger in cars on trips longer than 40 minutes	Single choice • I am never a passenger • (almost) never • rarely (3-4x of 10 trips) • occasionally (5-6x of 10 trips) • often (7-8x of 10 trips) • very often (9-10x of 10 trips)
	16.	Duration of naps as passenger in cars on trips longer than 40 minutes	Single choice • I am never a passenger • not at all • 0-10 minutes • 10-20 minutes • 20-30 minutes • longer than 30 minutes
	17.	Frequency of naps as passenger in public means of transport on trips longer than 40 minutes	cf. item 15
	18.	Duration of naps as passenger in public means of transport on trips longer than 40 minutes	cf. item 16
	19.	Frequency of naps at home up to 40 minutes	Single choice • (almost) never • 1-2x / week • 3-5x / week • (almost) every day
	20.	Feeling of getting sufficient sleep	Single choice • 1 - fully disagree • 2 - disagree • 3 - rather disagree • 4 - neither nor • 5 - rather agree • 6 - agree • 7 - fully agree

Category	*Item*		*Answer*
Attitude towards driving	21.	Association of driving a car with fun	cf. item 20
	22.	Association of driving a car with strain	cf. item 20
	23.	Association of driving a car with stress	cf. item 20
	24.	Association of driving a car with discomfort	cf. item 20
	25.	Prior knowledge about automated driving (self-assessment)	cf. item 20
Use case	26.	Time of day to use an in-car sleeping function	Multiple choice • early morning (5-7 a.m.) • morning (7-10 a.m.) • late morning (10-12 a.m.) • noon (12 a.m. – 2 p.m.) • afternoon (2-6 p.m.) • early evening (6-8 p.m.) • evening (8-11 p.m.) • night (11 p.m. – 5 a.m.)
	27.	Minimum duration of trip for using sleeping function	Numerical input [minutes]
	28.	Anticipated frequency of use	Single choice • (almost) never • rarely (3-4x of 10 trips) • occasionally (5-6x of 10 trips) • often (7-8x of 10 trips) • very often (9-10x of 10 trips)
	29.	Acceptable road conditions to use a sleeping function	Multiple choice • motorways • federal roads • straight rural roads • bendy roads • cities • traffic jams
	30.	Anticipated driving occasions for using sleeping function	Multiple choice • vocational trips • way to work • way home from work • daily routines (e.g. shopping) • recreational trips • holidays

Do you bike virtually safe? An explorative VR study assessing the safety of bicycle infrastructure

Marc Schwarzkopf, André Dettmann, Jonas Trezl, & Angelika C. Bullinger
Chemnitz University of Technology,

Abstract

Driven by the mobility transition towards a more ecological modal split, bicycles are becoming more popular as a means of transportation in cities. Therefore, bicycle infrastructure should become an increasing focus of urban planners. When designing infrastructure measures for cyclists, user acceptance, especially subjective safety, and comfort experience, are important reasons for the usage. To evaluate such factors in advance and derive the corresponding design requirements for urban planners at an early stage, the use of virtual reality (VR) can help to evaluate planned infrastructure measures. This paper presents an experimental design to evaluate infrastructure measures for cyclists in a VR study with 20 participants in an urban context. Subjects were presented 19 infrastructure measures in VR which were previously evaluated by an expert focus group for objective safety and divided into three safety categories. The images were randomized, and subjects were asked in a structured think-aloud procedure to provide statements about the subjective assessment, as well as reasons for their decision. In this paper, we present the study design and the results regarding reasons for or against specific infrastructure measures and will conclude with a methodological discussion regarding infrastructure assessment via virtual reality to aid urban planners and authorities.

Background

In the wake of societal and technological trends such as demographic change (Buffel & Phillipson, 2012) and advancing digitization (Kramers et al., 2014), almost all areas of daily life are changing. For the future design of cities, this means adjustments in urban development as well as the (re)design of inner-city mobility (Loorbach & Shiroyama 2006; Burns, 2013). In this context, influencing the mobility behaviour of citizens has been the focus of many research projects for some time. This was amongst others done through gamification (Torres-Toukoumidis et al., 2022), financial (cash credits; Thøgersen, 2009; Bamberg & Schmidt, 2001) or political marketing measures to increase the frequency of use of alternative modes of transport (walking, cycling, public transport).

All measures used to increase the attractiveness of alternative means of transportation, can generally be divided into hard and soft policies (Gärling et. al, 2009). Soft policies

In D. de Waard, S.H. Fairclough, K.A. Brookhuis, D. Manzey, L. Onnasch, A. Naumann, R. Wiczorek, F. Di Nocera, S. Röttger, and A. Toffetti (Eds.) (2022). Proceedings of the Human Factors and Ergonomics Society Europe Chapter 2022 Annual Conference. ISSN 2333-4959 (online). Available from http://hfes-europe.org

include strategies that focus on increasing information access, increasing motivation and (digital) communication aspects. Hard measures attempt to influence the use of transport modes through physical changes to infrastructure, increased costs for car use, or control of road space. Regarding soft policies, research indicates a mostly short-term change in behaviour (Bamberg & Schmidt, 2001) while hard policies suggest that urban design measures have a higher potential to stimulate a long-term behavioural change. Wardman, Tight and Page (2007) studied factors influencing cycling behaviour and found that cycle lanes separated from the road can increase the frequency of cycling trips by 55% and lead to a slight decrease in car commuting. Here, when designing infrastructure measures for cyclists, user acceptance, especially subjective safety, and comfort experience, are an important reason for the usage (Springer et al., 2021). Besides safety and comfort, a high satisfaction with the implemented infrastructure is another important indicator for the frequency of use. For example, a correlation could be found between satisfaction with the availability of parking spaces for bicycles and actual bicycle use (Martens, 2007).

Such evaluations of bicycle infrastructure and their effectiveness on safety, comfort and overall satisfaction are mostly carried out via onsite field observations of road, traffic junctions and the surrounding areas. Problematic in this regard is the often resource-intensive effort to assess the infrastructure as well as accompanying circumstances such as daytime, weather, season, traffic situation and actual access to participants for interviewing purposes. Moreover, only a momentary observation of the initial structural condition and situation is possible. While this assessment of infrastructure measures in their current state is very important for urban planners, future improvements also need to be assessed with regard to safety and comfort. Here, the difficulties of a beforehand assessment or a longitudinal study design is difficult due to the complex, resource-intensive and lengthy implementation of new measures.

To overcome both difficulties in assessing bicycle infrastructure we propose the usage of virtual reality studies. Based on the findings of and Higuera-Trujillo et al. (2017), who compared different display formats such as photographs, 360° panoramas, and virtual reality to a physical environment, the usage of 3D technologies are the most promising for usage in controlled conditions. The visual evaluation of urban spaces is easily achievable in terms of accessible technology for capturing or creation the necessary 360° images and presenting them to participants under laboratory conditions. Following the studies of Mouratidis & Hassan (2020), who examinded architecture and urban design in virtual environments, this approach seems useful as VR proved to be useful for assessment of public spaces. In addition, cyber sickness appeared to be unproblematic, making this method applicable to a wide range of people. In the following section, we present an explorative study to examine infrastructure measures for cyclists in terms of safety and comfort in a VR environment. We try to answer the question if the study design is suitable for this kind of evaluation and if the subject's responds regarding reasons for or against specific infrastructure measures are valid. We will conclude with a methodological discussion regarding infrastructure assessment via virtual reality to aid urban planners and authorities.

Method

Participants

N = 20 subjects participated in the study. The subjects were randomly recruited via a mailing list and various student groups in messenger services. The only restriction was that subjects had to own a bicycle. Participation was remunerated with 15 €. The mean age was 29.7 years (*SD* = 7.4). 14 subjects reported using a bicycle at least once a week, no subject reported never cycling. Bicycle use was mainly for leisure activities (*n* = 17) and commuting to work (*n* = 14). 11 respondents describe themselves as experienced cyclists, 9 respondents would rather describe themselves as little or not at all experienced cyclists. According to the ATI technology affinity scale (Franke et al., 2019), five respondents are not technology affine, 13 respondents have a low technology affinity and one respondent has a high technology affinity. 14 subjects had already participated in another VR study.

Material

The 360° recordings of the infrastructure measures were made with a GoPro Max in the cities of Chemnitz and Dresden in Germany. The images were then evaluated in a focus group of four experts from the fields of infrastructure planning and bicycle safety and divided into three groups (low, medium and high subjective safety) using 2D normalised pictures. For the evaluation of the 360° images, 19 images were presented (2D normalised examples can be seen in Figure 1). To present the images in a virtual environment a Unity application was programmed. Unity v2019.4.12f1 was used in combination with the SteamVR asset to implement the app on the HTC Vive Pro and the required head tracking. The used 360° photos were declared as skyboxes in Unity, as these are rendered around the entire scene, giving the impression of a complex landscape on the horizon. To control the experimental sequence within the application, a script was written in C#. In addition, a questionnaire was created that included questions about demographics, cycling experience and the ATI (affinity for technology interaction) scale (Franke et al., 2019).

Figure 1. 2D normalised sample images for the three safety categories

Procedure

The method was based on the findings of Mouratidis & Hassan (2020) and Higuera-Trujillo et al. (2017). Each study run had a time window of only 60 minutes for health protection reasons. At the beginning of the study, an initial questionnaire was completed. Afterwards, the VR headset was put on the subject by the experimenter. In a demo image, the subject was able to adjust the headset and familiarise him- or herself with the environment and the controls. Afterwards, a total of 19 further 360° photos were presented in a randomized order. For each photo, subjects were asked by the experimenter to assess whether s/he would consider the presented infrastructure to be subjectively safe for cyclists. In addition, s/he was asked to give reasons for her/his assessment. A prepared think-aloud protocol was used (Smelser & Baltes, 2001). Furthermore, the test person was asked whether s/he knew the infrastructure. At the end of each image, the test person was asked to rate the subjective security on a Likert scale of 1 (low subjective security) and 5 (high subjective security). During the study, subjects were asked about their well-being. All statements were recorded and transcribed. The experimenter was able to follow the subject's actions and movements on his computer (Figure 2). The evaluation of an image was not subject to a time limit. The experimenter was responsible for presenting the next picture.

Figure 2. Setup during the experiment

Results

Since, to the authors' knowledge, there is no comparable study, an essential issue is the validity of the study design. This chapter is divided into two parts. In the first part, the statements and data of the test persons are considered under the aspects of validity and reliability. In the second part, the evaluation of the infrastructure measures by the test persons is analysed.

Validity and reliability

The experimental design presented is a mixed-method approach. However, the qualitative contents predominate. For this reason, no quantitative criteria can be used to assess validity. However, indications for the validity of the design can be derived from the statements of the test persons. Four hypotheses were formed for this purpose:

H1: The interindividual difference in the evaluation of subjectively perceived safety between the subjects is small ($SD \leq 1$).

H2: Subjects use the same criteria to evaluate subjective safety.

H3: The affinity for technology has no significant influence on the evaluation of subjective safety.

H4: Subjective self-assessment of cycling experience has no influence on the evaluation of subjective safety.

The subjects' evaluation of subjective safety has a low standard deviation (Fig. 3). The maximum value of the standard deviation is 1.04 and the minimum value is 0.54.

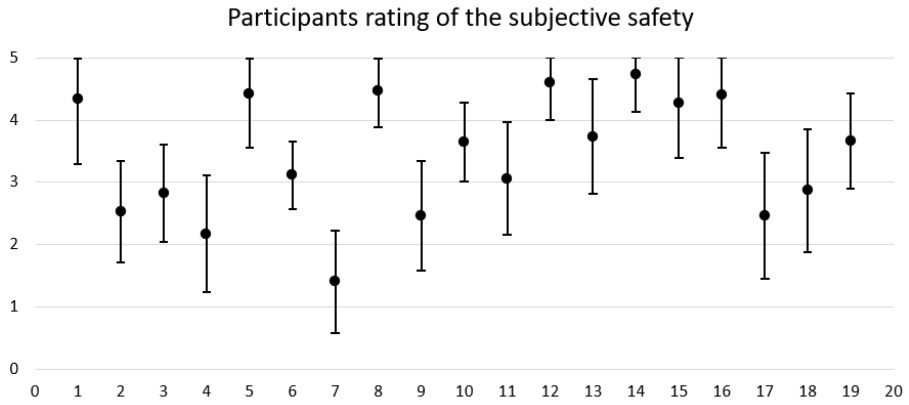


Figure. 3. Evaluation of the subjective safety of infrastructure measures for cyclists; error bars reflect Standard Error (Rating; Scale: 1 – not safe at all; 2 – not safe; 3 – neutral; 4 – safe; 5 – very safe, numbers on the x-axis represent the images shown to the particpants)

The think-aloud protocol is used to identify the strategies or heuristics used by the subjects. (Smelser & Baltes, 2001). Consistent statements by the test persons indicate comparable strategies or approaches to solving a problem (Smelser & Baltes, 2001). Applied to the VR study, this means that the use of the same or comparable criteria for assessing subjective safety across all subjects is an indication of high validity of the test procedure. The assessment in VR would therefore be carried out across all subjects using a comparable set of criteria. To verify this, the think-aloud protocols were evaluated using qualitative content analysis according to Mayring (Mayring & Fenzl, 2019) and the statements made were clustered according to the underlying criteria. In addition, the images were categorised into three categories based on the

subjects' subjective safety ratings: high subjective safety (subjective safety rating ≥ 3.8, n_{images} = 6), medium subjective safety (subjective safety rating 2.26 to 3.7, n_{images} = 8) and low subjective safety (subjective safety rating ≤ 2.25, n_{images} = 5). The details are presented in Table 1.

Table 1. Number of frequencies of mentioning the evaluation criteria according to the evaluation of subjective safety. Numbers in brackets indicate the mean value per criteria per user for each image (but just for images where the criteria is visible, e. g. the mean value for parking cars was only calculated for pictures with parking cars).

	High	*Medium*	*Low*
Separation	78 (13)	79 (11,3)	73 (14,6)
Width	41 (6,83)	49 (7)	25 (5)
Signage/marking	51 (8,5)	44 (6.3)	52 (10,4)
Condition of the surface	19 (3,17)	32 (4,6)	34 (6,8)
Clarity	15 (2.5)	27 (4,5)	30 (6)
Parking cars	3 (3)	25 (12,5)	36 (12)
Other	6 (1)	4 (2)	7 (1)

Another potential factor influencing the validity of the experimental design is the affinity for technology. This was determined using ATI (Franke et al., 2019). The sample was divided into two groups (low vs. medium affinity for technology). Four test persons had a low affinity for technology. No difference could be determined for the factor technology affinity on the evaluation of subjective safety (Mann-Whitney U-test, two tailed, $p < 0.05$). The details are presented in figure 4.

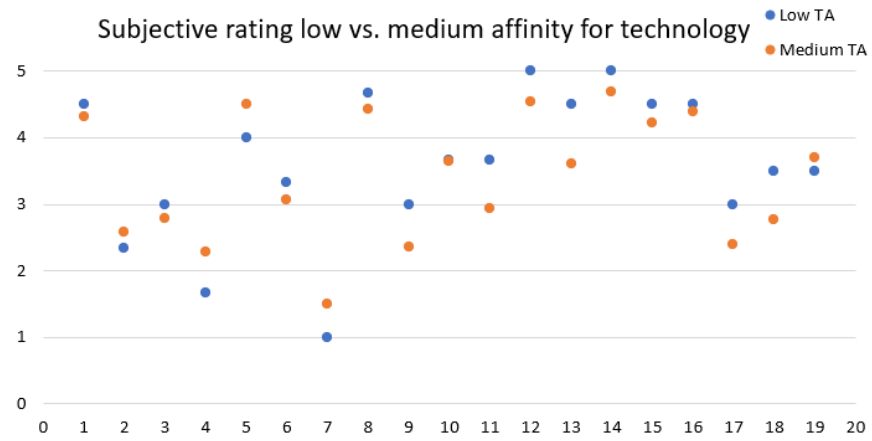


Figure. 4. Evaluation of subjective safety grouped by technology affinity (TA) (Rating; Scale: 1 – not safe at all; 2 – not safe; 3 – neutral; 4 – safe; 5 – very safe)

Another potential factor influencing the validity of the experimental design is the cycling experience of the subjects. The sample was divided into two groups (low vs. high experience). Seven test persons had low experience. No difference could be

determined for the influence factor experience on the evaluation of subjective safety (Mann-Whitney U-test, two tailed, $p < 0.05$). The details are presented in figure 5.

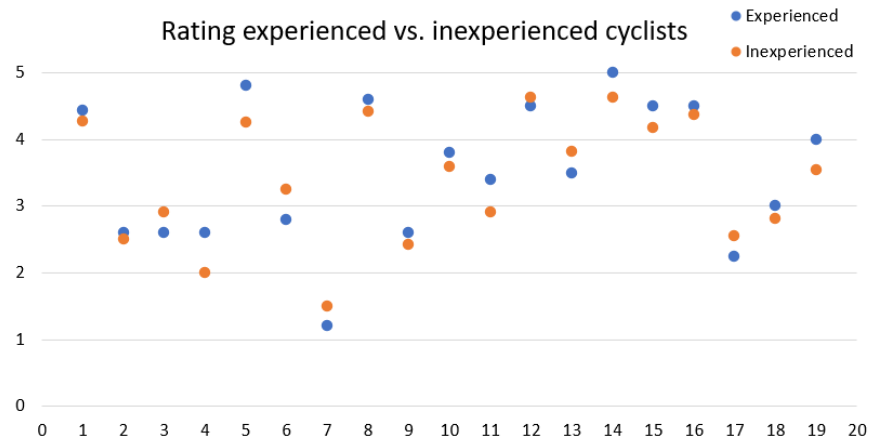


Figure. 5. Evaluation of subjective safety grouped by level of cycling experience (Rating; Scale: 1 – not safe at all; 2 – not safe; 3 – neutral; 4 – safe; 5 – very safe)

Furthermore, the subjects were asked about the ease of use of the VR headset and whether there were factors that had a (perceptible) technical influence on the evaluation of the infrastructure measures. All subjects found the implementation intuitive and stated that there were no perceptible technical factors that interfered with the evaluation. The investigators were also unable to observe any operating errors in the subjects' use of the VR peripherals. In addition, the subjects were also regularly asked about their well-being; no subject complained of nausea or dizziness.

The greatest influence on the evaluation of the subjective safety of the infrastructure measures appears to be the factor of familiarity. Subjects who already knew the infrastructure shown tended to include contextual influencing factors in their evaluation (e. g. density of road traffic, accessibility of the infrastructure measure, disruptive factors on the way to the infrastructure, etc.). The subjects expressed this accordingly in the think-aloud procedure and were asked by the test leaders to omit this information from their evaluation of subjective safety to make sure, that the participants only evaluate what they see on the images. We address this issue in the next iteration of the method.

Evaluation of the infrastructure

The experts' assessment of the subjective safety for cyclists of the infrastructure measures was carried out in a focus group of four people. The experts were from the field of infrastructure planning and road safety. The experts' evaluations for the sample are comparable to the evaluation by the test persons (Tab. 2).

Table 2. Comparison of the results of the expert evaluation and the test persons' evaluation of subjectively perceived safety (ER = Experts Rating; Scale: 1 – not safe at all; 2 – not safe; 3 – neutral; 4 – safe; 5 – very safe)

#	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
ER	H	L	M	L	H	M	L	H	M	M	M	H	M	H	H	H	L	M	M
M_P	4.33	2.42	2.8	2.15	4.47	3.17	1.4	4.52	2.45	3.72	3.05	4.61	3.77	4.66	4.17	4.47	2.27	2.83	3.66

Discussion

The data collected in the study served exploratory purposes only. The number of test persons as well as the division of the groups (experienced vs. inexperienced cyclists; technology-savvy vs. non-technology-savvy) do not create a basis for consideration from the point of view of inferential statistics. The presented study only served as a proof of concept for the study design and will be expanded in further research projects. Furthermore, due to corona restrictions, the implementation time per respondent was limited to 45 minutes. For this reason, questionnaires on presence in VR or in-depth evaluation of subjective safety and perceived comfort experience could not be used.

Although the presentation of the infrastructure measures was randomised, sequence effects could not be eliminated. The subjects tended to compare infrastructure measures with measures that had already been evaluated.

Nevertheless, the findings presented suggest that the methodology presented is suitable for evaluating different infrastructure measures for cyclists in an experiment. It was shown that the intersubjective evaluation in the presented sample has a low standard deviation. Furthermore, subjects use comparable criteria for the evaluation, regardless of individual cycling expertise, which is in line with previous findings (BMVI, 2017; Götschi et al., 2018; FixMyCity, 2021). This suggests that the use of VR peripherals does not significantly influence the underlying mental models used to evaluate subjective safety. Furthermore, the results have shown that the subjects arrive at comparable assessments of subjective safety as a focus group consisting of experts from the fields of infrastructure planning and cycling safety. The use of an additional group of experts from these fields is therefore not considered necessary for future studies. However, none of the experts were from public planning authorities, so a comparison of the subjective safety of cycling infrastructure with the objective safety defined by national regulations could not be made.

Considerations for future work

The presented methodology for evaluating the subjective safety of infrastructure for cyclists will be further developed in future projects. Besides a larger sample and a group balance (experienced vs. inexperienced, young vs. old), the technique will be further adapted. For example, the method will be expanded with images from a 3D stereoscopic camera. Furthermore, the use of 360° videos as well as auditory stimuli are planned. The methodology could also to be used for the evaluation of other infrastructure areas (e. g. footpaths, bus stops, crossings, etc.). The aim is to provide urban authorities with a methodology for the cost-effective and simple evaluation of

implemented infrastructure measures, which can also be used in direct citizen participation.

References

Bamberg, S. & Schmidt, P. (2001). Theory-driven subgroup-specific evaluation of an intervention to reduce private car use. *Journal of Applied Social Psychology, 31*, 1300-1329

Buffel, T, Phillipson, C, & Scharf, T. (2012) Ageing in urban environments: Developing 'age-friendly' cities. *Critical Social Policy, 32*, 597-617. doi:10.1177/0261018311430457

Burns, L. (2013) A vision of our transport future. *Nature 497*, 181-182. https://doi.org/10.1038/497181a

BMVI (2017). Fahrrad-Monitor Deutschland 2017. *Ergebnisse einer repräsentativen Online-Befragung*.https://kreisverbaende.adfc-nrw.de/uploads/media/fahrradmonitor -2017-ergebnisse.pdf

FixMyCity (2020). *Studie zur subjektiven Sicherheit im Radverkehr - Ergebnisse und Datensatz einer Umfrage mit über 21.000 Teilnehmenden*. https://fixmyberlin.de/research/subjektive-sicherheit

Franke, T., Attig, C., & Wessel, D. (2019). A personal resource for technology interaction: development and validation of the affinity for technology interaction (ATI) scale. *International Journal of Human–Computer Interaction, 35*, 456-467.

Gärling, T., Bamberg, S., Friman, M., Fujii, S., & Richter, J. (2009). Implementation of Soft Transport Policy Measures to Reduce Private Car Use in Urban Areas.

Higuera-Trujillo, J. L., Maldonado, J.L.T., & Millán, C.L. (2017). Psychological and physiological human responses to simulated and real environments: A comparison between Photographs, 360 Panoramas, and Virtual Reality. *Applied Ergonomics, 65*, 398-409.

Kramers, A., Höjer, M., Lövehagen, N., & Wangel, J. (2014). Smart Sustainable Cities – Exploring ICT Solutions for Reduced Energy Use in Cities, *Environmental Modelling & Software*, 52-62.

Krukowicz, T., Firląg, K., Sobota, A., Kołodziej, T., & Novacko, L. (2021). The relationship between bicycle traffic and the development of bicycle infrastructure on the example of Warsaw. *Archives of Transport, 60*. 187-203. 10.5604/01.3001.0015.6930.

Loorbach, D. & Shiroyama, H. (2016): The Challenge of Sustainable Urban Development and Transforming Cities. In: *Governance of Urban Sustainability Transitions*, (pp. 3-12).

Martens, K. (2007) Promoting bike-and-ride: The Dutch experience, *Transportation Research Part A: Policy and Practice, 41* , 326-338, https://doi.org/10.1016/j.tra.2006.09.010.

Mouratidis, K. & Hassan, R. (2020). Contemporary versus traditional styles in architecture and public space: A virtual reality study with 360-degree videos. *Cities, 97*, 102499.

Mayring, P. & Fenzl, T. (2019). Qualitative inhaltsanalyse. *In Handbuch Methoden der empirischen Sozialforschung* (pp. 633-648). Springer VS, Wiesbaden.

Smelser, N. J., & Baltes, P.B. (Eds.). (2001). *International encyclopedia of the social & behavioral sciences* (Vol. 11). Amsterdam: Elsevier.

Springer, S., Kreußlein, M., Krems, J. F. (2021). Shedding Light on the Dark-Field of Cyclists' Safety Critical Events: A Feasibility Study in Germany. *Proceedings of 9th International Cycling Safety Conference ICSC 2021*. Lund, Sweden.

Thøgersen, J. (2009). Promoting public transport as a subscription service: Effects of a free month travel card. *Transport Policy, 16*, 335-343.

Torres-Toukoumidis, A., Vintimilla-Leon, D., De-Santis, A. & López-López, & P.C. (2022). Gamification in Ecology-Oriented Mobile Applications—Typologies and Purposes. *Societies, 12*. 1-12. 10.3390/soc12020042.

Wardman, M., Tight, M., & Page, M. (2007). Factors influencing the propensity to cycle to work. *Transportation Research Part A: Policy and Practice, 41*, 339-350.

Towards fast human-centred contouring workflows for adaptive external beam radiotherapy

Nicolas F. Chaves-de-Plaza[1,4], Prerak Mody[2,4], Klaus Hildebrandt[1], Marius Staring[2], Eleftheria Astreinidou[2], Mischa de Ridder[3], Huib de Ridder[1], & René van Egmond[1]
[1]Delft University of Technology, [2]Leiden University Medical Center, [3]University Medical Center Utrecht, [4]HollandPTC Delft, The Netherlands

Abstract

Delineation of tumours and organs-at-risk permits detecting and correcting changes in the patients' anatomy throughout the treatment, making it a core step of adaptive external beam radiotherapy. Although auto-contouring technologies have sped up this process, the time needed to perform the quality assessment of the generated contours remains a bottleneck, taking clinicians between several minutes and an hour to complete. The authors of this article conducted several interviews and an observational study at two treatment centres in the Netherlands to identify challenges and opportunities for speeding up the delineation process in adaptive therapies. The study revealed three contextual variables that influence contouring performance: usable additional information, applicable domain-specific knowledge, and available editing capabilities in contouring software. In practice, clinicians leverage these variables to accelerate contouring in two ways. First, they use domain-specific knowledge and relevant clinical features such as the proximity of the organs-at-risk to the tumour to enable targeted inspection of the delineation. Second, clinicians modulate editing precision depending on the effect they anticipate the edit will have on the patient outcome. By implementing these acceleration strategies in guidelines and contouring tools, developers and workflow builders could increase contouring efficiency and consistency without affecting the patient outcome.

Introduction

External Beam Radiotherapy (EBRT) is the most common form of RT and has become one of humanity's main tools against cancer, together with surgery and systemic treatment. In EBRT, ionizing radiation is directed at the patient's tumour to destroy the malignant cells. Over the last decades, significant technological improvements have been made in treatment planning and delivery, which increased the precision of EBRT. For instance, proton beam therapy (PT) can harness the ability of protons to deposit all their energy at a specific spot (Newhauser & Zhang, 2015; Wilson, 1946). This capability permits PT more precisely shape the radiation dose to the tumor, minimizing the dose to the surrounding healthy tissue and reducing side

In D. de Waard, S.H. Fairclough, K.A. Brookhuis, D. Manzey, L. Onnasch, A. Naumann, R. Wiczorek, F. Di Nocera, S. Röttger, and A. Toffetti (Eds.) (2022). Proceedings of the Human Factors and Ergonomics Society Europe Chapter 2022 Annual Conference. ISSN 2333-4959 (online). Available from http://hfes-europe.org

effects (Langendijk et al., 2013; Lundkvist et al., 2005; Simone et al., 2011; Thomas & Timmermann, 2020).

Harnessing the precision increase of dose delivery technology requires adapting the patient's treatment plan to the anatomy of the day. Figure 1 presents the general workflow of this treatment paradigm known as adaptive EBRT. Adaptive EBRT imposes severe time constraints on online treatment planning processes (orange boxes in Figure 1) because longer within fraction times can lead to new anatomical changes, offsetting the value of the adaptation. Also, an increase in the footprint of treatment planning processes would reduce patient throughput, compromising the viability of adaptive EBRT.

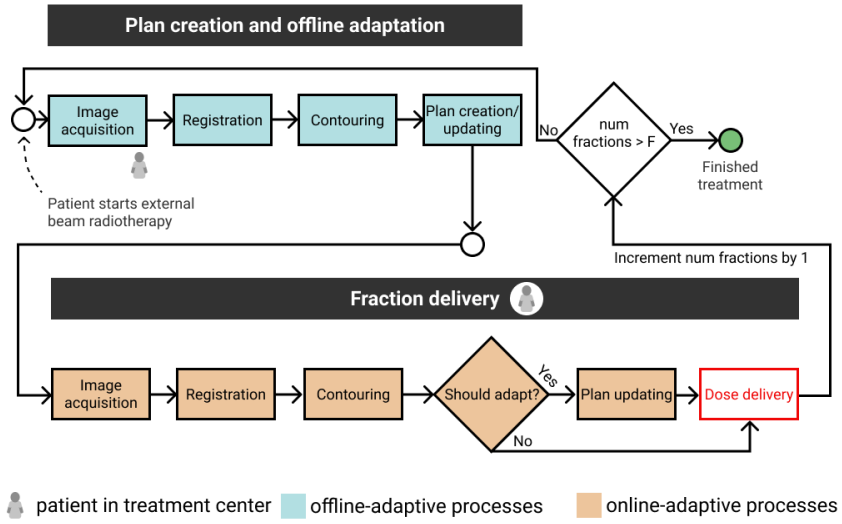


Figure 1. Schematic of external beam radiotherapy (EBRT) dose delivery pipeline. Each box corresponds to one process, and the diamonds to decisions in the workflow. The goal is to deliver the prescribed dose to the patient (red box) in F fractions spread over several days. Adaptive strategies help mitigate dose deviations due to changes in the patient's anatomy during the treatment. Adaptation can be online within a fraction (orange boxes) or offline between fractions (blue boxes).

The present study investigates the challenges that the contouring process poses to the implementation of adaptive EBRT. Despite the availability of auto-contouring technologies, contouring remains human-centred because clinicians need to perform an extensive quality assessment of the generated delineations to ensure that they do not contain inaccuracies (Cardenas et al., 2019; Nikolov et al., 2020; van Dijk et al., 2020; Vandewinckele et al., 2020). Therefore, to reduce the footprint of the contouring process, it is necessary to understand human factors that impact its duration.

This study extends prior works in two ways. First, it focuses on the time dimension of contouring performance, uncovering factors that influence it. Traditionally, researchers have directed their attention to analysing the effect of different image modalities, guidelines, contouring software, and experience on output-based performance metrics like accuracy and inter-observer contouring variability

(Bekelman et al., 2009; Brouwer et al., 2014; Steenbakkers et al., 2005, 2006; Vinod et al., 2016). This focus makes sense considering the influence that these metrics have on patient safety (Karsh et al., 2006; Njeh, 2008). Nevertheless, factors that affect time can also impact accuracy, motivating the need to study them. On the one hand, other things equal, accuracy degrades in time-constrained scenarios (Chignell et al., 2014; Pew, 1969). On the other, if clinicians perform demanding tasks for extended periods, they can become fatigued and lose situation awareness, which will also impact accuracy (Endsley, 2021; Evans et al., 2019).

Second, this work studies the contouring process in its clinical context. Prior works have investigated the effect of input devices and user interfaces on contouring time using experiments in highly controlled environments (Multi-Institutional Target Delineation in Oncology Group, 2011; Ramkumar, 2017; Steenbakkers et al., 2005). These studies' findings hold for the general contouring case. Nevertheless, this needs not to be the case in the time-constrained phase of adaptive EBRT (orange boxes in Figure 1). This study follows a qualitative context-driven approach to uncover factors that affect contouring performance in adaptive EBRT and discusses potential context-aware strategies to mitigate them. Adopting an ecological approach to researching human factors that affect contouring performance can help designing representative experiments and evaluations for contouring in time-critical scenarios (Flach et al., 2018). Furthermore, the findings from this study represent the initial step of methodologies like Ecological Interface Design, which aims to develop systems that promote adaptive performance (Vicente, 2002).

To summarize, the present study investigates factors that affect the duration of the contouring process and discusses potential mitigation strategies. It complements and extends prior studies that analysed human factors of contouring performance (Aselmaa et al., 2014; Ramkumar et al., 2017), providing an updated account of the process workflow in the time-critical context of adaptive EBRT. Finally, the present study contributes to the state-of-the-art of clinical contouring workflows in adaptive EBRT in two ways:

1. It reports the results of an observational study in two cancer treatment centres in the Netherlands. The study of the Contouring Workflow provided a situated account of the current contouring workflows in the context of adaptive EBRT, together with factors that can affect its performance.
2. It discusses acceleration strategies based on the context of adaptive EBRT that tool developers and clinicians can leverage to adapt the contouring workflow to time-constrained scenarios.

The Contouring Activity

An exploratory literature review was performed to establish baseline knowledge about the contouring activity and its role in adaptive therapies. The query used for the search (Scopus, PubMed, and Google Scholar) included the keywords: adaptive, adaptation, proton therapy, radiotherapy, contouring, automatic, semi-automatic, workflow, and head-and-neck. The latter term was relevant since the study's participants (next section) were specialists in this region. The search yielded around 50 articles with publishing years ranging between 2008 and 2021.

As Figure 2 depicts, the main inputs of the contouring activity are 3D images (stacks of hundreds of 2D images) that describe the patient anatomy. Among these, there is an image to contour, usually a Computerized Tomography (CT), and supporting information such as previous contours of the patient and other image modalities such as Magnetic Resonance Imaging (MRI) and Positron Emission Technology CT (PET-CT). Using available information, contouring consists of drawing the boundaries of anatomical structures relevant to the patient's cancer in the image to contour. The two main anatomical groups are the target volumes (TVs), which correspond to areas affected by tumoral cells, and the organs at risk (OARs), which correspond to healthy tissue.

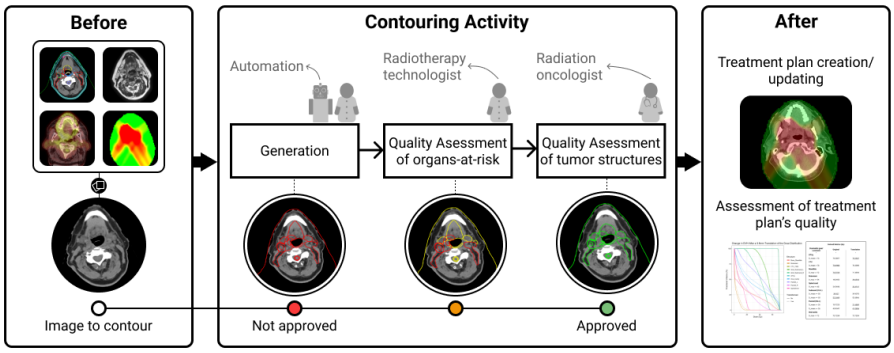


Figure 2. Components of the contouring activity. The inputs (left) are the image to contour and, optionally, other three-dimensional datasets like MRI and PET-CT scans and dose distribution volumes. The contouring activity has two main processes that several actors perform: generation of contours and its quality assessment. After approving the contours, clinicians can use them to create/update the patient's treatment plan and assess its quality.

As the right panel of Figure 2 indicates, the goal of the contouring activity is to produce contours suitable for creating or updating the patient treatment plan and assessing its quality. Several actors participate in this workflow in the clinic, distributing contouring tasks based on the anatomical structures' groups. In general, radiotherapy technologists (RTTs) start by delineating the OARs. After this, the radiation oncologists (ROs), who are directly responsible for the patient's outcome, assess the quality of the OARs contours and draw the boundaries of the TVs, the structures with the highest priority. The study described in the next section was designed based on this understanding of the contouring activity.

Study of the Contouring Workflow

A study of the contouring workflow was conducted to identify characteristics of adaptive EBRT affecting contouring performance and to identify context-dependent strategies that tool developers can leverage to improve it. The following subsections detail the study's design and describe the methodology used for analysing the resulting data.

Study design

Participants

Two radiation oncologists (RO) and two radiotherapy technologists (RTT) from two cancer treatment centres in the Netherlands specializing in the head-and-neck area joined the study. Table 1 summarizes the participants' information. One of the institutes, the Leiden University Medical Center (LUMC), offers photon-based volumetric modulated arc therapy (VMAT) treatments. The second, the Holland Proton Therapy Centre (HollandPTC), offers proton therapy (PT). Despite the differences in dose delivery technology, both institutions have a similar workflow, performing offline adaptations. The latter means that the patient's treatment plan is updated sparsely during treatment (entails re-executing blue boxes in Figure 1). The Institutional Review Board at the Delft University of Technology approved this research. Each participant provided informed consent to be part of the study.

Procedure

The study had three sessions. The first one, a one-hour-long semi structured interview, permitted establishing rapport with the participants and validated the initial understanding of the EBRT workflow. In the second and third sessions, the participants performed their contouring duties while being recorded. As Table 1 shows, these meetings lasted between one and two hours, depending on the participants' time. In the second session, clinicians performed initial contouring. The third focused on adaptive contouring, where clinicians perform a quality assessment of automatically generated contours. Given the limited clinicians' time to participate, they contoured a subset of anatomical including the tumours and organs close to them that could affect the patient outcome.

Table 1. Participants of the qualitative sessions. Two radiation oncologists (RO) and two radiotherapy technologists (RTT) from two institutions in the Netherlands participated. In some cases, due to their tight schedules, they could not attend all the sessions.

ID	*Institution*	*Role*	*Session*	*Time (hours)*
P1	LUMC	RO	1, 2, 3	5
P2	LUMC	RTT	2, 3	2
P3	HollandPTC	RO	1, 2	3
P4	HollandPTC	RTT	1, 2, 3	5

Materials

For the observational sessions, clinicians at each centre had access to the data of two previously treated head and neck patients. Each patient file included initial treatment planning data such as CT, PET-CT, and MRI scans and daily images such as CBCT and CT, relevant for sessions 2 and 3, respectively. For session 3, starting delineations could have been generated by another clinician or automated methods like deformable or rigid registration and deep learning-based contouring. For inspecting and editing the contours, clinicians used their routine software.

Data Analysis

Table 2. The first column presents the themes that emerged during the Thematic Analysis of the transcripts of the semi-structured interviews and observational sessions of the Study of the Contouring Workflow. The second column presents the coarser codes obtained after several grouping iterations finer ones. Lastly, the third column displays, for each theme, a representative example from the transcribed data.

Theme	*Codes*	*Example*
Adaptive contouring context	Clinical workflow, standardization, physical and clinical artifacts, training, institution specific considerations, EBRT technology	"Now it takes one day to do the whole plan. So, we have to make a new calculation and it has to go into the the LINAC so it has to get another check." [P2]
Structure priority and effect of inaccuracies on patient's treatment	Anatomical knowledge, downstream effects, characteristics of different anatomical structures, clinical priorities, tumour-related considerations	"I guess if it's an inner region where for instance the cheek region here. Those are minor [edits], but if we see this region where you have the parotid gland. There it could influence dose to the OARs quite significantly. So there. Then I would say it's a major [edit]." [P1]
Dealing with uncertain regions in the image-to-contour	Anatomical knowledge, image modalities, papers and guidelines, information required for certainty	"With the nasopharyngeal cancers, then I will take an MRI and then I will draw on the MRI. So, then I know exactly where the brainstem is." [P4]
Editing capabilities of contouring software	Characteristics of contouring software, experience with the tools, use of automation	"It seems to me that it's a model based one [automatically generated contour] because the model based one always has trouble here at the head of the mandible at the joint." [P3]
Distribution of labour and clinicians experience	Experience with the contouring task, collaboration, task distribution, protocols	"When an RTT does it [a contour]? Sometimes it's very nice and when a not so experienced RTT does it it's not a very good delineation and then it costs me either a lot of time to adjust every slice or I just start again and that's most of the time." [P3]

The recordings of the three sessions were transcribed and analysed using Thematic Analysis (Braun & Clarke, 2006). The coding process was bottom-up, first labelling patterns in the transcripts and then grouping the resulting fine-grained codes into coarser ones based on their similarity. Table 2 displays the underlying coarser codes, the resulting themes, and sample data excerpts. The screen recordings of sessions 2 and 3 were also relevant as they showcased the way clinicians interact with the user

interface during the contouring process. The interactions were mapped onto a timeline like the one that Figure 4 depicts. For the y-axis, the authors drew inspiration from the literature on contouring tasks (Aselmaa et al., 2017) but grouped them into four categories to simplify the coding process and the analysis. These are direct and indirect manipulation, navigation, and non-contouring interactions.

Initial Contouring

Results

Initial contouring (IC) occurs when executing the plan creation and offline adaptation process in Figure 1 for the first time. At LUMC and HollandPTC, initial contouring (IC) takes two to six hours for head-and-neck (HN) cancers, requiring delineating more than twenty structures. The following paragraphs group the observations about the IC workflow into three characteristics, finishing with a discussion on how these can affect contouring performance.

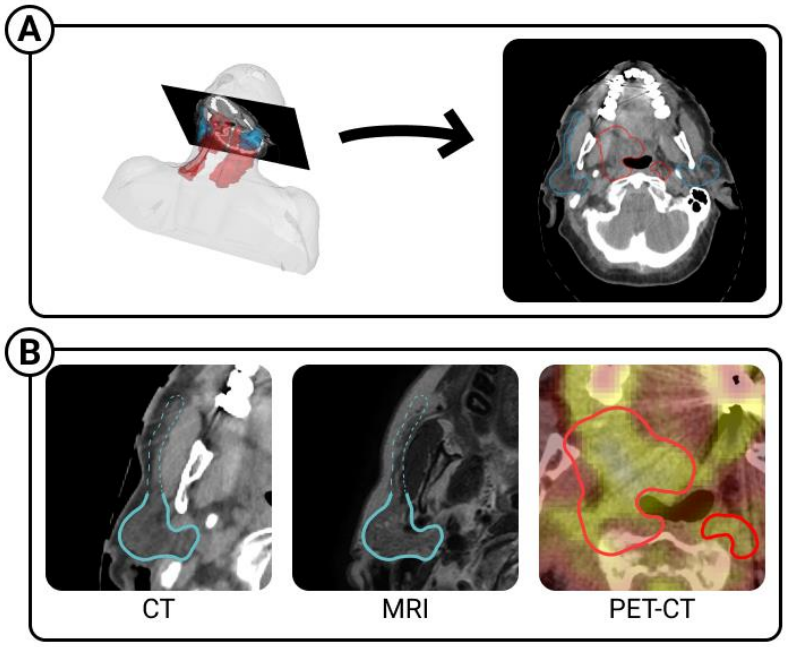


Figure 3. Available information available at contouring. The central input is the image to contour which, as panel A depicts, is a three-dimensional image made from several 2D slices. Other three-dimensional images available at the surveyed centres are magnetic resonance imaging (MRI) and positron imaging technology CT (PET-CT) scans. As panel B shows, MRI helps differentiate soft tissue, and PET-CT aids in detecting and delineating tumours.

Usable Additional Information

At IC, no pre-existing contours of the patients exist, given that this process occurs after they have started treatment. Instead, clinicians use information from multiple image modalities acquired beforehand. The main image modality in radiotherapy, CT,

usually does not provide enough boundary information when the contrast between adjacent tissues is not enough or when there is noise or artifacts in the image acquisition process. In these cases, clinicians rely on Magnetic Resonance Imaging (MRI) and Positron Emission Technology-CT (PET-CT) scans, acquired for most patients at HollandPTC and LUMC. As Figure 3 shows, MRI helps differentiate soft tissue structures: ”MRI makes it easier for us to delineate the parotid glands because you can see them very good at an MRI.”. For PET-CT, this modality permits clinicians to locate tumours and estimate their boundaries with higher precision: ”We actually scan all of our head and neck patients [with PET-CT] because it makes our delineations so much accurate, so that is now standard.” [P1].

In practice, clinicians align additional images to the CT before using them for contouring. This process, known as image registration, can take several minutes per image pair and requires the clinician’s intervention to verify the alignment’s quality. Registering the images allows clinicians to scroll through them in parallel using the contouring software, enabling direct comparison of the structures in both scans.

Applicable Domain-Specific Knowledge

In some cases, the information in the images is not enough. At IC, this happens when MRI and PET-CT scans are not available and moreover there are no pre-existing contours of the patients (they just started the treatment). In these cases, clinicians rely on domain-specific knowledge they access in two ways. First, they leverage guidelines (Brouwer et al., 2015) and atlases that describe and indicate what the contours should look like, respectively. Second, they draw on their experience. Experienced clinicians know what areas can be challenging to delineate given the available data. They use this domain-specific anatomical knowledge to direct their attention and estimate contours over unclear image boundaries. An example of this dynamic occurs when the radiation oncologists (ROs) review the delineations created by the radiotherapy technologists (RTTs): “We [ROs] think that it [delineating the swallowing muscles] is too hard for RTTs, need quite a bit of anatomical knowledge to know where they are exactly. And in this case, this patient doesn't have a very big tumour in the throat, but most of the time patients have quite a big tumour here. And you can't see the swallowing muscles that good. So, then you need to know exactly where they run from to delineate them.” [P1].

Editing Capabilities of Contouring Software

In practice, at IC, clinicians create the contours from scratch. As the timeline on the top section of Figure 4 depicts, this entails starting with an empty delineation and gradually building the contours through a series of interactions. At the surveyed institutions, clinicians favoured a semi-automatic workflow, which consisted of two phases. First, they generated initial contours using the between-slice interpolation tool. This tool requires clinicians to manually delineate a subset of the slices spanning the structure, after which the rest of the structure's contours will be interpolated (this autocompletion corresponds to the indirect editing interaction around the second eighty in Figure 4). Finally, revert to the manual brush tool to correct inaccuracies. As the timeline shows, the generation of contours takes more time than the refinement, and clinicians spend most of the time directly editing the delineations with the brush.

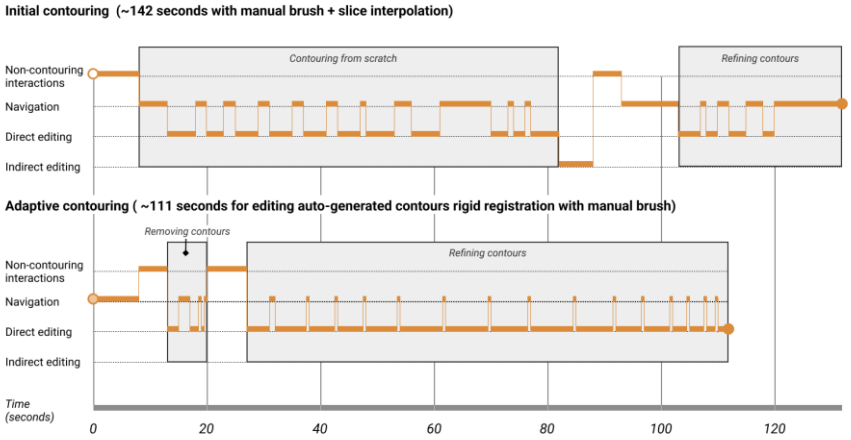


Figure 4. Interaction timelines for initial and adaptive contouring. In both cases, a radiotherapy technologist from LUMC (P2 in Table 1), delineated the right submandibular gland of a head and neck cancer patient. The x-axis encodes time, and the y-axis differentiates the principal interaction categories. Non-contouring interactions correspond to changes in the interface that do not affect the contours, like changing the layout or visualization parameters. Navigation refers to changing the current slice of the image to contour. Finally, direct and indirect manipulations entail altering the delineations in the 2D slice or through a button in the menu, respectively. Note how initial contouring starts from scratch (empty circle) while adaptive contouring starts with pre-generated delineations (partially filled circle).

Discussion

Clinicians use contours produced at IC to create the patient's treatment plan. Therefore, they seek maximal accuracy, often at the expense of longer task durations. The three characteristics of the IC context described before affect contouring time in several ways. First, extra image modalities reduce the task difficulty, which can result in reduced dwelling times to determine where the contour should go. Nevertheless, additional images need to be registered to the main one, a time-consuming process that could offset the performance benefits gains that the process offers. Second, domain-specific knowledge can reduce the extent of the contouring task by letting clinicians direct their attention to where it is needed. Yet, following the accuracy directive, they still must go through the whole volume to ensure no inaccuracy remains. Finally, the semi-automatic between slice interpolation tool spares clinicians from needing to edit several slices but still requires significant manual effort to initialize the method.

Adaptive Contouring

Results

LUMC and HollandPTC implement an offline-adaptive dose delivery pipeline, which entails updating the treatment plan several times during treatment by repeating the plan creation and offline adaptation process between fractions. Adaptive contouring (AC) occurs in this setting and differs from initial contouring (IC) in that the time is

more critical and the resources scarcer. At the surveyed institutions, AC takes one to two hours for head and neck cancer patients. Like the previous section, the following paragraphs detail the AC context and discuss how it affects the process' performance.

Usable Additional Information
In contrast with IC, at AC, no extra images of the patient are acquired. Therefore, clinicians have access to the image to contour, a CT at LUMC and HollandPTC, the images acquired for IC, and the approved IC contours. In practice, clinicians only use the latter and do so in two ways. First, because IC contours document all the clinical decisions made for the current patient, they use them as a patient-specific atlas to resolve complex contouring tasks. Regarding having an atlas for contouring, P4 mentioned that "it's always nice to have it [the atlas] like a verification. Because the brainstem isn't that difficult, but like if you have the swallowing muscles or something, that's really something. If you have the atlas side by side, it really can come in handy." [P4] Second, clinicians use approved IC contours to create an initial segmentation. For this, they align, or register, the IC and AC images and then "propagate" the contours from the former to the latter.

Applicable Domain-Specific Knowledge
In addition to general anatomical knowledge, at AC, clinicians use knowledge about dosimetry and the patient tumour to structure and guide the contouring process. On the one hand, it can help them direct their attention to critical areas. On the other, it lets them modulate the contouring based on the structure's relevance to the patient's treatment plan. For instance, P2 mentioned that while some contours require maximal attention and precision: "…with this type of organs, as with all the nervical organs, as in optical nerves and brain stem and spinal cord, when it's critical, so when the PTV is nearby, then it's very important that we draw this very precise." Others accept rougher contours as they will not significantly impact the patient's outcome: "this submandibular gland, it gets too much dose, so it won't work. After irradiation, this one is gone. So, at that point, we can decide to delineate, but it isn't, it's OK if it isn't quite perfect."

Editing Capabilities of Contouring Software
As mentioned before, clinicians do not start delineating from scratch at AC. Instead, they generate a starting point by propagating the contours from the initial scan to the current one. Therefore, the goal at AC is to perform a quality assessment (QA) of these delineations. The timeline in the bottom section of Figure 4 exemplifies the series of interactions that clinicians usually perform during the QA process. In the timeline, it is possible to see how starting from partial delineations, they reach the final ones after a series of relatively long direct editing interactions interleaved with brief navigation operation ones. Between slice interpolation, the tool clinicians use for contouring from scratch does not work for contour refinement. Therefore, for extensive errors across multiple slices like the one Figure 5 depicts, clinicians face two options. Either manually fix the contour on every slide or delete the delineation and re-do it from scratch using between-slice interpolation.

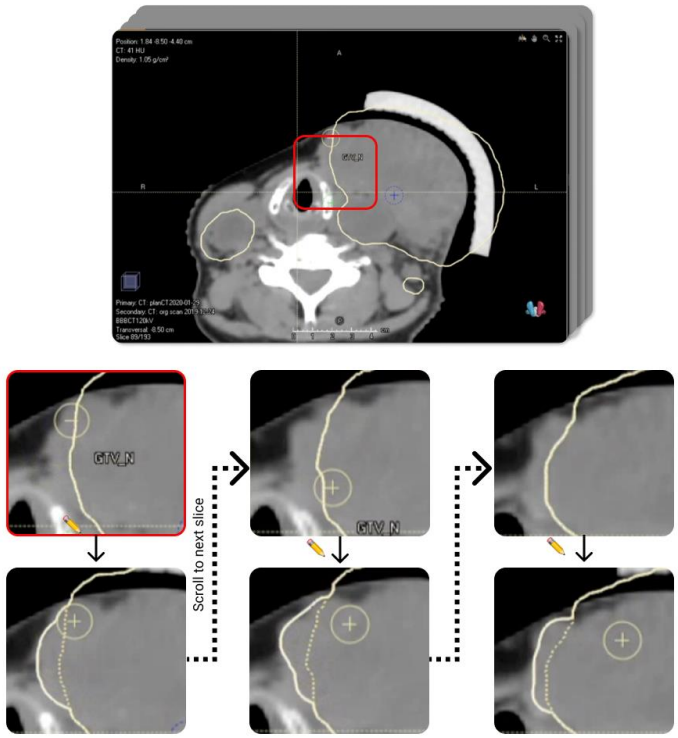


Figure 2. Editing faulty delineations often entails redundant interactions. The top image presents an inaccurate auto-generated contour of a tumoral structure. As can be observed, the internal side of the contour fails to include the whole structure, which causes an error that spans three slices. The images below present the sequence of steps that P1 followed to amend the inaccuracy.

Discussion

While clinicians use IC contours for creating the treatment plan, they use AC contours to update the plan. For this reason, at this stage, their primary concern therefore seemed to be to faithfully translate IC contours to the current patient anatomy. The identified contextual characteristics affect AC performance in several ways. First, having information about the role that each structure plays in the patient's treatment helps direct clinicians' attention to delineations that can affect the patient outcome. A potential pitfall of the current prioritization approach is that it is purely heuristic and based on clinicians' experience instead of available information such as the planned dose. Second, by using IC-approved contours, clinicians can reduce the time for analysing and editing complex or large regions by propagating them via registration. Nevertheless, same as with other image modalities at IC, the time it takes to perform the registration might offset the time gains. Finally, although contouring is overall faster at AC due to the contours being pre-generated, there is no tool to efficiently perform QA, requiring clinicians to invest significant manual effort.

Discussion

The Study of the Contouring Workflow provided an understanding of several characteristics that affect contouring duration in adaptive EBRT. This section takes these observations as input and lays down several ways of accelerating the adaptive contouring activity, which is increasingly time-pressured due to clinics implementing more responsive adaptative workflows. The discussion differentiates between the inspection, navigation, and editing tasks, which account for most of the delineation time. Figure 6 summarizes the study's findings and the resulting context-dependent acceleration strategies.

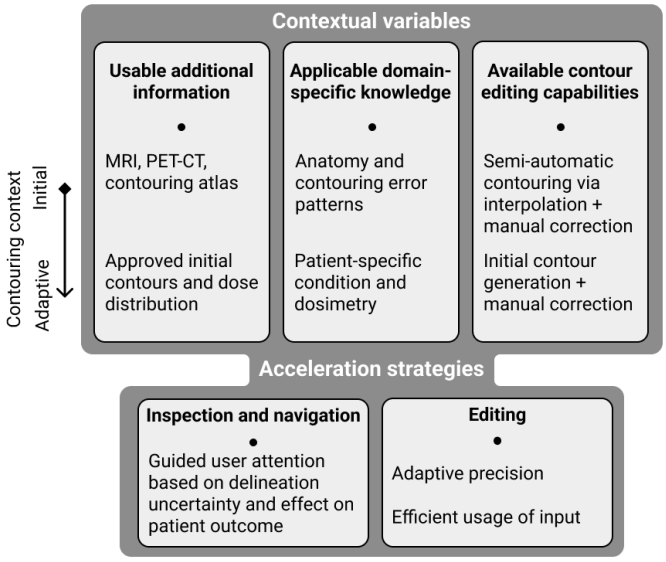


Figure 3. Schematic of the approach that the present study followed. First, it identified three variables that influence contouring performance and described their roles in the initial and adaptive contouring contexts. These variables were then mapped to strategies for accelerating the inspection, navigation, and editing tasks.

Inspection and Navigation

In adaptive contouring, clinicians prioritized inspection of tumour contours because an error could result in overexposure of surrounding organs to radiation or, worse, in underexposure of the cancerous tissue (Aliotta et al., 2019). This observation suggests that patient-specific treatment-level information provides a valuable signal to define the contouring priority of anatomical structures. Heuristics based on dose information allow clinicians to decide faster (Marewski & Gigerenzer, 2012). Nevertheless, problems like cognitive bias, loss of situation awareness, or varying levels of experience can introduce inconsistencies in a heuristic-based contouring process, which could risk patient safety (Graber et al., 2002; Tversky & Kahneman, 1974). Protocols and checklists could be implemented to enable effective heuristics usage while mitigating their pitfalls (Chan et al., 2012; Chera et al., 2012; Marks et al., 2011). These could be based on metrics like Normal Tissue Complication Probability

(NTCP) that have been shown to affect the patient outcome (Brouwer et al., 2014). Figure 7 presents an example of prioritization based on the local characteristics of the dose distribution. As can be observed, while a potential inaccuracy in the tumour delineation has a high priority, errors in the parotid glands are less urgent due to their lower impact on the patient's treatment.

Before prioritizing errors, clinicians need to detect them. Several methods have been proposed in the literature for assisting this task. They vary in the information and the mechanism used to perform the search. As for the former, it is possible to compute shape (Heimann & Meinzer, 2009; Hermann & Klein, 2015) and image or appearance-related (Gao et al., 2010) characteristics of the contours, e.g. the surface area or the intensity histogram, respectively. Another possible indicator of the contours' quality is their uncertainty or variability, which can come from historical patient data (Chu et al., 2013), the auto-contouring algorithm (LaBonte et al., 2020; Mody et al., 2021), or directly from the image-to-contour (Top et al., 2011). After gathering all these sources of information, available techniques identify potential errors in two ways. Firstly, by letting a classifier automatically find data-based rules for separating inaccurate from the accurate regions (Altman et al., 2015; Chen et al., 2015; Hui et al., 2018; Kalpathy-Cramer & Fuller, 2010; McIntosh et al., 2013; Rhee et al., 2019; Sandfort et al., 2021). Secondly, they delegate the search task to the users, presenting them with the traditional two-dimensional image and contour slices together with informative overlays such as uncertainty iso-lines (Al-Taie et al., 2014; Prassni et al., 2010) and contour box plots (Whitaker et al., 2013). These two-dimensional visualizations have been augmented by adding three-dimensional views (Lundström et al., 2007; Raidou et al., 2016) and letting the user interact with the data by filtering and sorting mechanisms (Furmanová et al., 2021; Saad et al., 2010).

Two challenges that existing error detection tools face are maintaining users' trust in the system and lowering the cognitive load they impose. As to the former, a system failing to spot inaccuracies that affect the patient's treatment (false negatives) would erode the users' trust (Asan et al., 2020; White et al., 2011). This might explain the limited adoption of automatic error detection systems in clinical practice. Regarding cognitive load, abrupt context changes when guiding clinicians' attention to different parts of the 3D image can build up fatigue, potentially leading to errors like classifying a true positive the system suggested as a false positive (Allnutt, 1987; Persson et al., 2019). Visualization methods like 3D views complementing attention guidance mechanisms could help mitigate this issue.

Editing

Currently, clinicians use mostly manual tools when fixing an inaccuracy. For errors that occupy a large portion of the volume, like the example in Figure 5, this often means that the user will perform similar edits across slices. Existing semi-automatic interactive contouring techniques mitigate this issue by extrapolating rough feedback provided by the clinician. Their general workflow consists of two steps. First, the clinician provides a rough indication of the change to be made or the area to update via coarse inputs such as scribbles, points, or a bounding box. Based on this input, the algorithm proceeds to update the segmentation. Traditionally Markov Random Field-based algorithms are being used (Kato & Zerubia, 2012; Rother et al., 2004). Recently,

deep learning-based implementations have appeared that offer more sophisticated suggestions based on the clinician's input (Dai et al., 2015; Lin et al., 2016; Maninis et al., 2018).

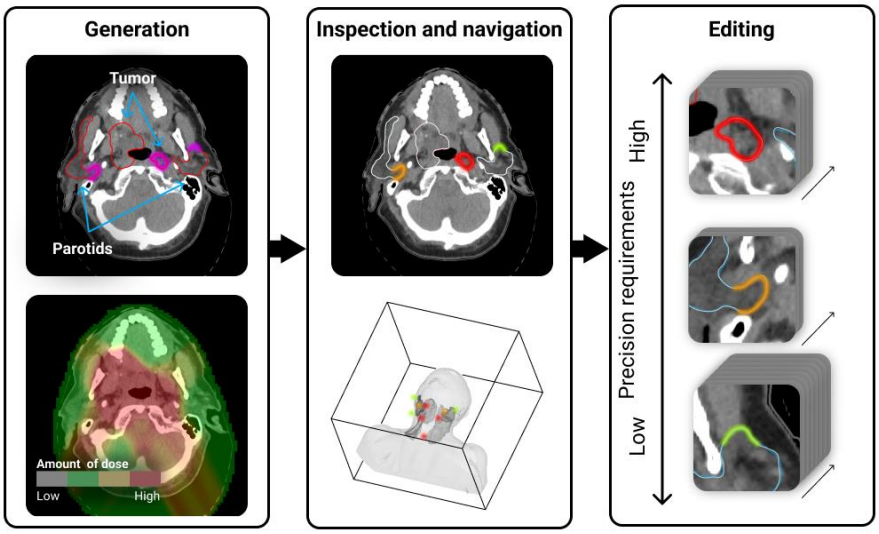


Figure 4. Components for accelerating the inspection, navigation, and editing tasks. The first step (leftmost column) is to generate the contours and gather extra information like delineation variability and the dose distribution. Based on these sources, potential errors can be flagged and categorized depending on their effect on the patient outcome. In the example, an error in the tumour's delineations was flagged as high priority (red) because it can significantly change the treatment plan. As for the parotid glands, the orange inaccuracy is in a region where the dose distribution varies more quickly than in the case of the green one. Therefore, subsequent processes (like treatment plan updating) that rely on the orange contours could be more sensitive to changes in these contours.

The adoption of these semi-automatic interactive editing tools in the clinic remains challenging. Based on discussions with clinicians, the reason for their resistance to these interactive editing tools seems to be that they perceive scribbles as a blunt tool for communicating to the algorithm what they want. Therefore, more research is needed to determine which type of input mechanism the clinicians prefer and how the algorithm should respond (Amrehn et al., 2016; Hebbalaguppe et al., 2013). For instance, do they prefer coarse inputs like scribbles? Or would they be more comfortable with high precision inputs such as selecting a contour from an ensemble of candidates (Ferstl et al., 2016)? With editing being the most time-consuming QA operation, obtaining a synergy between humans and AI is paramount.

Limitations and Future Work

A limitation of this work is the reduced number of treatment centres and clinicians surveyed in the study, which might have led to weighting heavily on custom institutional practices and personal preferences. As a promising solution, questionnaires like the one reported in (Bertholet et al., 2020) could be prepared to

validate the conclusions with a larger pool of participants. Another limitation is the qualitative nature of the timelines used to illustrate the dynamics between the clinicians and the contouring software. In further studies, we plan to use keystroke logging software to include more fine-grained actions and more accurate timings. The latter would be especially valuable for comparing different segmentation tools.

In terms of future work, we will translate the findings of this study into a practical human-centred contouring protocol that clinicians can adapt to their institution-specific adaptive EBRT capabilities and constraints. In addition to the clinician-level considerations that the present article considered, such protocol will also account for team dynamics, which also emerged as a performance factor in the surveyed institutions.

Conclusion

This study characterized the contouring workflows in adaptive EBRT. An observational study at two treatment centres in the Netherlands revealed several context-dependent characteristics that influence delineation performance. Based on these observations, strategies for accelerating inspection, navigation, and editing tasks were discussed. By applying these when developing and commissioning tools, tool builders and clinicians can decrease the delineation time and thus increase the suitability of this process for time-critical therapies like online-adaptive EBRT.

Acknowledgement

The authors of this work are grateful for the assistance and collaboration of the personnel at both Holland Proton Therapy Center and Leiden University Medical Center. The research for this work was funded by Varian, a Siemens Healthineers Company, through the HollandPTC-Varian Consortium (grant id 2019022), and partly financed by the Surcharge for Top Consortia for Knowledge and Innovation (TKIs) from the Ministry of Economic Affairs and Climate.

References

Aliotta, E., Nourzadeh, H., & Siebers, J. (2019). Quantifying the dosimetric impact of organ-at-risk delineation variability in head and neck radiation therapy in the context of patient setup uncertainty. *Physics in Medicine & Biology*, *64*(13), 135020. https://doi.org/10.1088/1361-6560/ab205c

Allnutt, M. F. (1987). Human factors in accidents. *British Journal of Anaesthesia*, *59*, 856–864. https://doi.org/10.1093/bja/59.7.856

Al-Taie, A., Hahn, H. K., & Linsen, L. (2014). Uncertainty estimation and visualization in probabilistic segmentation. *Computers & Graphics*, *39*, 48–59. https://doi.org/10.1016/j.cag.2013.10.012

Altman, M. B., Kavanaugh, J. A., Wooten, H. O., Green, O. L., DeWees, T. A., Gay, H., Thorstad, W. L., Li, H., & Mutic, S. (2015). A framework for automated contour quality assurance in radiation therapy including adaptive techniques. *Physics in Medicine and Biology*, *60*, 5199–5209. https://doi.org/10.1088/0031-9155/60/13/5199

Amrehn, M., Glasbrenner, J., Steidl, S., & Maier, A. (2016). Comparative Evaluation of Interactive Segmentation Approaches. In T. Tolxdorff, T. M. Deserno, H. Handels, & H.-P. Meinzer (Eds.), *Bildverarbeitung für die Medizin 2016* (pp. 68–73). Springer. https://doi.org/10.1007/978-3-662-49465-3_14

Asan, O., Bayrak, A. E., & Choudhury, A. (2020). Artificial Intelligence and Human Trust in Healthcare: Focus on Clinicians. *Journal of Medical Internet Research*, *22*(6), e15154. https://doi.org/10.2196/15154

Aselmaa, A., Goossens, R., Rowland, B., Laprie, A., Song, W., & Freudenthal, A. (2014, July 19). Medical Factors of Brain Tumor Delineation in Radiotherapy for Software Design. *Proceedings of the 5th International Conference on Applied Human Factors and Ergonomics*.

Aselmaa, A., van Herk, M., Laprie, A., Nestle, U., Götz, I., Wiedenmann, N., Schimek-Jasch, T., Picaud, F., Syrykh, C., Cagetti, L. V., Jolnerovski, M., Song, Y., & Goossens, R. H. M. (2017). Using a contextualized sensemaking model for interaction design: A case study of tumor contouring. *Journal of Biomedical Informatics*, *65*, 145–158. https://doi.org/10.1016/j.jbi.2016.12.001

Bekelman, J. E., Wolden, S., & Lee, N. (2009). Head-and-Neck Target Delineation Among Radiation Oncology Residents After a Teaching Intervention: A Prospective, Blinded Pilot Study. *International Journal of Radiation Oncology*Biology*Physics*, *73*, 416–423. https://doi.org/10.1016/j.ijrobp.2008.04.028

Bertholet, J., Anastasi, G., Noble, D., Bel, A., van Leeuwen, R., Roggen, T., Duchateau, M., Pilskog, S., Garibaldi, C., Tilly, N., García-Mollá, R., Bonaque, J., Oelfke, U., Aznar, M. C., & Heijmen, B. (2020). Patterns of practice for adaptive and real-time radiation therapy (POP-ART RT) part II: Offline and online plan adaption for interfractional changes. *Radiotherapy and Oncology*, *153*, 88–96. https://doi.org/10.1016/j.radonc.2020.06.017

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, *3*, 77–101. https://doi.org/10.1191/1478088706qp063oa

Brouwer, C.L., Steenbakkers, R.J.H.M., Bourhis, J., Budach, W., Grau, C., Grégoire, V., van Herk, M., Lee, A., Maingon, P., Nutting, C., O'Sullivan, B., Porceddu, S.V., Rosenthal, D.I., Sijtsema, N.M., & Langendijk, J.A. (2015). CT-based delineation of organs at risk in the head and neck region: DAHANCA, EORTC, GORTEC, HKNPCSG, NCIC CTG, NCRI, NRG Oncology and TROG consensus guidelines. *Radiotherapy and Oncology: Journal of the European Society for Therapeutic Radiology and Oncology*, *117*, 83–90. https://doi.org/10.1016/j.radonc.2015.07.041

Brouwer, C.L., Steenbakkers, R.J.H.M., Gort, E., Kamphuis, M.E., van der Laan, H.P., van't Veld, A.A., Sijtsema, N.M., & Langendijk, J.A. (2014). Differences in delineation guidelines for head and neck cancer result in inconsistent reported dose and corresponding NTCP. *Radiotherapy and Oncology*, *111*, 148–152. https://doi.org/10.1016/j.radonc.2014.01.019

Cardenas, C.E., Yang, J., Anderson, B.M., Court, L.E., & Brock, K.B. (2019). Advances in Auto-Segmentation. *Seminars in Radiation Oncology*, *29*, 185–197. https://doi.org/10.1016/j.semradonc.2019.02.001

Chan, A.J., Islam, M.K., Rosewall, T., Jaffray, D.A., Easty, A.C., & Cafazzo, J.A. (2012). Applying usability heuristics to radiotherapy systems. *Radiotherapy and Oncology*, *102*, 142–147. https://doi.org/10.1016/j.radonc.2011.05.077

Chen, H., Zhang, S., Chen, W., Mei, H., Zhang, J., Mercer, A., Liang, R., & Qu, H. (2015). Uncertainty-Aware Multidimensional Ensemble Data Visualization and Exploration. *IEEE Transactions on Visualization and Computer Graphics*, *21*, 1072–1086. https://doi.org/10.1109/TVCG.2015.2410278

Chera, B.S., Jackson, M., Mazur, L. M., Adams, R., Chang, S., Deschesne, K., Cullip, T., & Marks, L.B. (2012). Improving Quality of Patient Care by Improving Daily Practice in Radiation Oncology. *Seminars in Radiation Oncology*, *22*, 77–85. https://doi.org/10.1016/j.semradonc.2011.09.002

Chignell, M., Tong, T., Mizobuchi, S., & Walmsley, W. (2014). Combining Speed and Accuracy into a Global Measure of Performance. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *58*, 1442–1446. https://doi.org/10.1177/1541931214581301

Chu, C., Oda, M., Kitasaka, T., Misawa, K., Fujiwara, M., Hayashi, Y., Nimura, Y., Rueckert, D., & Mori, K. (2013). Multi-organ segmentation based on spatially-divided probabilistic atlas from 3D abdominal CT images. *Medical Image Computing and Computer-Assisted Intervention: MICCAI ... International Conference on Medical Image Computing and Computer-Assisted Intervention*, *16*(Pt 2), 165–172. https://doi.org/10.1007/978-3-642-40763-5_21

Dai, J., He, K., & Sun, J. (2015). *BoxSup: Exploiting Bounding Boxes to Supervise Convolutional Networks for Semantic Segmentation*. 1635–1643. https://openaccess.thecvf.com/content_iccv_2015/html/Dai_BoxSup_Exploiting_Bounding_ICCV_2015_paper.html

Endsley, M.R. (2021). Situation Awareness. In *Handbook of Human Factors and Ergonomics* (pp. 434–455). John Wiley & Sons, Ltd. https://doi.org/10.1002/9781119636113.ch17

Evans, S.B., Cain, D., Kapur, A., Brown, D., & Pawlicki, T. (2019). Why Smart Oncology Clinicians do Dumb Things: A Review of Cognitive Bias in Radiation Oncology. *Practical Radiation Oncology*, *9*, e347–e355. https://doi.org/10.1016/j.prro.2019.03.001

Ferstl, F., Kanzler, M., Rautenhaus, M., & Westermann, R. (2016). Visual Analysis of Spatial Variability and Global Correlations in Ensembles of Iso-Contours. *Computer Graphics Forum*, *35*(3), 221–230. https://doi.org/10.1111/cgf.12898

Flach, J.M., Hancock, P.A., Caird, J., & Vicente, K.J. (2018). *Global Perspectives on the Ecology of Human-Machine Systems*. CRC Press.

Furmanová, K., Muren, L. P., Casares-Magaz, O., Moiseenko, V., Einck, J. P., Pilskog, S., & Raidou, R. G. (2021). PREVIS: Predictive visual analytics of anatomical variability for radiotherapy decision support. *Computers & Graphics*, *97*, 126–138. https://doi.org/10.1016/j.cag.2021.04.010

Gao, X., Su, Y., Li, X., & Tao, D. (2010). A Review of Active Appearance Models. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, *40*(2), 145–158. https://doi.org/10.1109/TSMCC.2009.2035631

Graber, M., Gordon, R., & Franklin, N. (2002). Reducing Diagnostic Errors in Medicine: What's the Goal? *Academic Medicine*, *77*, 981–992.

Hebbalaguppe, R., McGuinness, K., Kuklyte, J., Healy, G., O'Connor, N., & Smeaton, A. (2013). How interaction methods affect image segmentation: User experience in the task. *2013 1st IEEE Workshop on User-Centered Computer Vision (UCCV)*, 19–24. https://doi.org/10.1109/UCCV.2013.6530803

Heimann, T., & Meinzer, H.-P. (2009). Statistical shape models for 3D medical image segmentation: A review. *Medical Image Analysis*, *13*(4), 543–563. https://doi.org/10.1016/j.media.2009.05.004

Hermann, M., & Klein, R. (2015). A visual analytics perspective on shape analysis: State of the art and future prospects. *Computers & Graphics*, *53*, 63–71. https://doi.org/10.1016/j.cag.2015.08.008

Hui, C.B., Nourzadeh, H., Watkins, W.T., Trifiletti, D.M., Alonso, C.E., Dutta, S.W., & Siebers, J.V. (2018). Quality assurance tool for organ at risk delineation in radiation therapy using a parametric statistical approach. *Medical Physics*, *45*, 2089–2096. https://doi.org/10.1002/mp.12835

Kalpathy-Cramer, J., & Fuller, C.D. (2010). Target Contour Testing/Instructional Computer Software (TaCTICS): A Novel Training and Evaluation Platform for Radiotherapy Target Delineation. *AMIA Annual Symposium Proceedings*, *2010*, 361–365.

Karsh, B.-T., Holden, R.J., Alper, S.J., & Or, C.K.L. (2006). A human factors engineering paradigm for patient safety: Designing to support the performance of the healthcare professional. *BMJ Quality & Safety*, *15* (suppl 1), i59–i65. https://doi.org/10.1136/qshc.2005.015974

Kato, Z., & Zerubia, J. (2012). Markov Random Fields in Image Segmentation. *Foundations and Trends® in Signal Processing*, *5*(1–2), 1–155. https://doi.org/10.1561/2000000035

LaBonte, T., Martinez, C., & Roberts, S. A. (2020). We Know Where We Don't Know: 3D Bayesian CNNs for Credible Geometric Uncertainty. *ArXiv:1910.10793 [Cs, Eess]*. http://arxiv.org/abs/1910.10793

Langendijk, J. A., Lambin, P., De Ruysscher, D., Widder, J., Bos, M., & Verheij, M. (2013). Selection of patients for radiotherapy with protons aiming at reduction of side effects: The model-based approach. *Radiotherapy and Oncology*, *107*, 267–273. https://doi.org/10.1016/j.radonc.2013.05.007

Lin, D., Dai, J., Jia, J., He, K., & Sun, J. (2016). *ScribbleSup: Scribble-Supervised Convolutional Networks for Semantic Segmentation*. 3159–3167. https://openaccess.thecvf.com/content_cvpr_2016/html/Lin_ScribbleSup_Scribble-Supervised_Convolutional_CVPR_2016_paper.html

Lundkvist, J., Ekman, M., Ericsson, S.R., Jönsson, B., & Glimelius, B. (2005). Proton therapy of cancer: Potential clinical advantages and cost-effectiveness. *Acta Oncologica*, *44*, 850–861. https://doi.org/10.1080/02841860500341157

Lundström, C., Ljung, P., Persson, A., & Ynnerman, A. (2007). Uncertainty Visualization in Medical Volume Rendering Using Probabilistic Animation. *IEEE Transactions on Visualization and Computer Graphics*, *13*, 1648–1655. https://doi.org/10.1109/TVCG.2007.70518

Maninis, K.-K., Caelles, S., Pont-Tuset, J., & Van Gool, L. (2018). *Deep Extreme Cut: From Extreme Points to Object Segmentation*. 616–625. https://openaccess.thecvf.com/content_cvpr_2018/html/Maninis_Deep_Extreme_Cut_CVPR_2018_paper.html

Marewski, J.N., & Gigerenzer, G. (2012). Heuristic decision making in medicine. *Dialogues in Clinical Neuroscience*, *14*, 77–89. https://doi.org/10.31887/DCNS.2012.14.1/jmarewski

Marks, L.B., Jackson, M., Xie, L., Chang, S.X., Burkhardt, K.D., Mazur, L., Jones, E. L., Saponaro, P., LaChapelle, D., Baynes, D.C., & Adams, R.D. (2011). The challenge of maximizing safety in radiation oncology. *Practical Radiation Oncology*, *1*, 2–14. https://doi.org/10.1016/j.prro.2010.10.001

McIntosh, C., Svistoun, I., & Purdie, T. G. (2013). Groupwise Conditional Random Forests for Automatic Shape Classification and Contour Quality Assessment in Radiotherapy Planning. *IEEE Transactions on Medical Imaging*, *32*, 1043–1057. https://doi.org/10.1109/TMI.2013.2251421

Mody, P., Chaves-de-Plaza, N., Hildebrandt, K., van Egmond, R., de Ridder, H., & Staring, M. (2021). Comparing Bayesian Models for Organ Contouring in Head and Neck Radiotherapy. *ArXiv:2111.01134 [Cs, Eess]*. http://arxiv.org/abs/2111.01134

Multi-Institutional Target Delineation in Oncology Group. (2011). Human-computer interaction in radiotherapy target volume delineation: A prospective, multi-institutional comparison of user input devices. *Journal of Digital Imaging*, *24*, 794–803. https://doi.org/10.1007/s10278-010-9341-2

Newhauser, W.D., & Zhang, R. (2015). The physics of proton therapy. *Physics in Medicine and Biology*, *60*(8), R155–R209. https://doi.org/10.1088/0031-9155/60/8/R155

Nikolov, S., Blackwell, S., Zverovitch, A., Mendes, R., Livne, M., De Fauw, J., Patel, Y., Meyer, C., Askham, H., Romera-Paredes, B., Kelly, C., Karthikesalingam, A., Chu, C., Carnell, D., Boon, C., D'Souza, D., Moinuddin, S.A., Consortium, D.R., Montgomery, H., … Ronneberger, O. (2020). Deep learning to achieve clinically applicable segmentation of head and neck anatomy for radiotherapy. *ArXiv:1809.04430 [Physics, Stat]*. http://arxiv.org/abs/1809.04430

Njeh, C.F. (2008). Tumor delineation: The weakest link in the search for accuracy in radiotherapy. *Journal of Medical Physics*, *33*(4), 136. https://doi.org/10.4103/0971-6203.44472

Persson, E., Barrafrem, K., Meunier, A., & Tinghög, G. (2019). The effect of decision fatigue on surgeons' clinical decision making. *Health Economics*, *28*, 1194–1203. https://doi.org/10.1002/hec.3933

Pew, R.W. (1969). The speed-accuracy operating characteristic. *Acta Psychologica*, *30*, 16–26. https://doi.org/10.1016/0001-6918(69)90035-3

Prassni, J., Ropinski, T., & Hinrichs, K. (2010). Uncertainty-Aware Guided Volume Segmentation. *IEEE Transactions on Visualization and Computer Graphics*, *16*, 1358–1365. https://doi.org/10.1109/TVCG.2010.208

Raidou, R.G., Marcelis, F.J.J., Breeuwer, M., Gröller, E., Vilanova, A., & van de Wetering, H.M.M. (2016). *Visual Analytics for the Exploration and Assessment of Segmentation Errors*. The Eurographics Association. https://doi.org/10.2312/vcbm.20161287

Ramkumar, A. (2017). *HCI in interactive segmentation: Human-computer interaction in interactive segmentation of CT images for radiotherapy*. https://repository.tudelft.nl/islandora/object/uuid%3A0f0259f1-0c33-442f-b851-86a846e736fc

Ramkumar, A., Stappers, P. J., Niessen, W. J., Adebahr, S., Schimek-Jasch, T., Nestle, U., & Song, Y. (2017). Using GOMS and NASA-TLX to Evaluate Human–Computer Interaction Process in Interactive Segmentation. *International Journal of Human–Computer Interaction*, *33*, 123–134. https://doi.org/10.1080/10447318.2016.1220729

Rhee, D. J., Cardenas, C.E., Elhalawani, H., McCarroll, R., Zhang, L., Yang, J., Garden, A S., Peterson, C.B., Beadle, B.M., & Court, L.E. (2019). Automatic detection of contouring errors using convolutional neural networks. *Medical Physics*, *46*, 5086–5097. https://doi.org/10.1002/mp.13814

Rother, C., Kolmogorov, V., & Blake, A. (2004). 'GrabCut': Interactive foreground extraction using iterated graph cuts. *ACM Transactions on Graphics*, *23*, 309–314. https://doi.org/10.1145/1015706.1015720

Saad, A., Hamarneh, G., & Möller, T. (2010). Exploration and Visualization of Segmentation Uncertainty using Shape and Appearance Prior Information. *IEEE Transactions on Visualization and Computer Graphics*, *16*, 1366–1375. https://doi.org/10.1109/TVCG.2010.152

Sandfort, V., Yan, K., Graffy, P.M., Pickhardt, P. J., & Summers, R.M. (2021). Use of Variational Autoencoders with Unsupervised Learning to Detect Incorrect Organ Segmentations at CT. *Radiology: Artificial Intelligence*, *3*(4), e200218. https://doi.org/10.1148/ryai.2021200218

Simone, C.B., Ly, D., Dan, T. D., Ondos, J., Ning, H., Belard, A., O'Connell, J., Miller, R. W., & Simone, N. L. (2011). Comparison of intensity-modulated radiotherapy, adaptive radiotherapy, proton radiotherapy, and adaptive proton radiotherapy for treatment of locally advanced head and neck cancer. *Radiotherapy and Oncology*, *101*, 376–382. https://doi.org/10.1016/j.radonc.2011.05.028

Steenbakkers, R.J.H.M., Duppen, J.C., Fitton, I., Deurloo, K.E.I., Zijp, L.J., Comans, E.F.I., Uitterhoeve, A.L.J., Rodrigus, P.T.R., Kramer, G.W.P., Bussink, J., De Jaeger, K., Belderbos, J.S.A., Nowak, P.J.C.M., van Herk, M., & Rasch, C.R. N. (2006). Reduction of observer variation using matched CT-PET for lung cancer delineation: A three-dimensional analysis. *International Journal of Radiation Oncology*Biology*Physics*, *64*, 435–448. https://doi.org/10.1016/j.ijrobp.2005.06.034

Steenbakkers, R.J.H.M., Duppen, J.C., Fitton, I., Deurloo, K.E.I., Zijp, L., Uitterhoeve, A.L.J., Rodrigus, P.T.R., Kramer, G.W.P., Bussink, J., Jaeger, K. D., Belderbos, J.S.A., Hart, A.A.M., Nowak, P J.C.M., van Herk, M., & Rasch, C.R.N. (2005). Observer variation in target volume delineation of lung cancer related to radiation oncologist–computer interaction: A 'Big Brother' evaluation. *Radiotherapy and Oncology*, *77*, 182–190. https://doi.org/10.1016/j.radonc.2005.09.017

Thomas, H., & Timmermann, B. (2020). Paediatric proton therapy. *The British Journal of Radiology*, *93*(1107), 20190601. https://doi.org/10.1259/bjr.20190601

Top, A., Hamarneh, G., & Abugharbieh, R. (2011). Active Learning for Interactive 3D Image Segmentation. In G. Fichtinger, A. Martel, & T. Peters (Eds.), *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2011* (pp. 603–610). Springer. https://doi.org/10.1007/978-3-642-23626-6_74

Tversky, A., & Kahneman, D. (1974). Judgment under Uncertainty: Heuristics and Biases. *Science*, *185*(4157), 1124–1131. https://doi.org/10.1126/science.185.4157.1124

van Dijk, L.V., Van den Bosch, L., Aljabar, P., Peressutti, D., Both, S., Steenbakkers, R.J.H.M., Langendijk, J.A., Gooding, M.J., & Brouwer, C.L. (2020). Improving automatic delineation for head and neck organs at risk by Deep Learning Contouring. *Radiotherapy and Oncology*, *142*, 115–123. https://doi.org/10.1016/j.radonc.2019.09.022

Vandewinckele, L., Claessens, M., Dinkla, A., Brouwer, C., Crijns, W., Verellen, D., & Elmpt, W. van. (2020). Overview of artificial intelligence-based applications in radiotherapy: Recommendations for implementation and quality assurance. *Radiotherapy and Oncology*, *153*, 55–66. https://doi.org/10.1016/j.radonc.2020.09.008

Vicente, K.J. (2002). Ecological Interface Design: Progress and Challenges. *Human Factors*, *44*, 62–78. https://doi.org/10.1518/0018720024494829

Vinod, S.K., Min, M., Jameson, M.G., & Holloway, L.C. (2016). A review of interventions to reduce inter-observer variability in volume delineation in radiation oncology. *Journal of Medical Imaging and Radiation Oncology*, *60*, 393–406. https://doi.org/10.1111/1754-9485.12462

Whitaker, R.T., Mirzargar, M., & Kirby, R.M. (2013). Contour Boxplots: A Method for Characterizing Uncertainty in Feature Sets from Simulation Ensembles. *IEEE Transactions on Visualization and Computer Graphics*, *19*, 2713–2722. https://doi.org/10.1109/TVCG.2013.143

White, M.P., Cohrs, J.C., & Göritz, A.S. (2011). Dynamics of Trust in Medical Decision Making: An Experimental Investigation into Underlying Processes. *Medical Decision Making*, *31*, 710–720. https://doi.org/10.1177/0272989X10394463

Wilson, R.R. (1946). Radiological Use of Fast Protons. *Radiology*, *47*, 487–491. https://doi.org/10.1148/47.5.487

Annoyance by Alarms in the ICU: A Cognitive Approach to the Role of Interruptions by Patient Monitoring Alarms

Idil Bostan[1,2], Elif Özcan[1,2], Diederik Gommers[2], & René van Egmond[1]
[1]Delft University of Technology,
[2]Department of Adult Intensive Care, Erasmus Medical Center
The Netherlands

Abstract

Nurses rely on patient monitoring systems for care delivery in ICUs. Monitoring systems communicate information to nurses and alert them through audiovisual alarms. However, excessive numbers of alarms often interrupt nurses in their tasks, and desensitize them to alarms. The affective consequence of this problem is that nurses are annoyed and feel frustration towards monitoring alarms. This situation leads to stress on nurses and threatens patient safety. Literature on sound annoyance distinguishes between annoyance induced by bottom-up (perceptual) and top-down (cognitive) processing. Extensive research on perceptual annoyance informs us on how to alleviate the problem by better sound design. However, addressing the cognitive aspect requires a broader understanding of annoyance as a construct. To this end, in this paper we distinguish between the annoyance induced by sensory unpleasantness of alarm sounds, and annoyance induced by frequent task interruptions. We present a conceptual framework in which we can interpret nurses' annoyance by monitoring alarms. We further present descriptive analysis of the occurrence frequency of patient monitoring alarms in a neonatal ICU to illustrate the current state with regards to alarms. We aim to support nurses' organizational well-being by providing an alternative hypothesis to explaining nurses' affective states caused by auditory alarms. Future research can benefit from this paper through understanding of the context and familiarizing with the cognitive processes relevant to processing of patient monitoring alarms.

Introduction

Intensive care unit (ICU) nurses deliver care to patients by observing and evaluating patients' condition, assisting doctors in their assessments, administering treatment, and supporting all-round recovery. Through their workflow nurses rely on patient monitoring systems to observe the vital parameters and changes in patients' status. Rapidly advancing technologies have allowed us to monitor an increasing number of

In D. de Waard, S.H. Fairclough, K.A. Brookhuis, D. Manzey, L. Onnasch, A. Naumann, R. Wiczorek, F. Di Nocera, S. Röttger, and A. Toffetti (Eds.) (2022). Proceedings of the Human Factors and Ergonomics Society Europe Chapter 2022 Annual Conference. ISSN 2333-4959 (online). Available from http://hfes-europe.org

parameters. Patient monitoring systems display vital parameters visually. Information about emerging medical and technical conditions, such as vital parameters exceeding thresholds or sensors getting detached, are communicated to nurses through audio-visual alarms. Consequently, with the increase of the number of measured parameters, the number of alarms in the ICU has also increased (e.g., monitoring blood oxygenation rate, ventilating patients, connecting patients to dialysis machines). Alarms are designed to attract attention and prompt action. However, up to 90% of alarms have been identified as false or non-actionable (Cvach, 2012; Deb & Claudio, 2015; Siebig et al., 2010). Consequently, they often interrupt the workflow without benefiting care delivery. This situation can result in desensitization; inducing stress in nurses and posing threats to patient safety (Lewandowska et al., 2020; Özcan & Gommers, 2020; Wilken et al., 2017). As a result, the affective outcomes are annoyance and frustration towards alarms (Cho et al., 2016; Sowan et al., 2015). Despite the research on solution strategies to mitigate problems related to alarms, there has not been a gratifying improvement until now (Sowan et al., 2016; Yue et al., 2017). In this paper, we present a conceptual framework of cognitive annoyance supported by a data collection from eight patient monitors in the Erasmus Medical Center ICU. With this framework, we aim to inform system design to support organizational well-being of nurses.

Research to support healthcare industry in challenges related to alarms has been ongoing for several decades. Studies mostly focus on improving (psycho)acoustic characteristics of alarms to make them less annoying (Foley et al., 2020; Sreetharan et al., 2021). Indeed, psychoacoustic characteristics such as sharpness, roughness, loudness, and tonalness have been shown to influence sensory (un)pleasantness of alarm sounds (Zwicker & Fastl, 1999). However, research indicates that acoustic characteristics only explain a small portion of variance in annoyance ratings. In fact, several psychological and contextual factors, such as noise sensitivity or time of day, have been shown to play larger roles in annoyance by sounds (Janssen et al., 2011; Paunović et al., 2009; Pierrette et al., 2012). Consequently, research indicates that there are two aspects to annoyance; perceptual and cognitive (Guski et al., 1999; Sreetharan et al., 2021). The perceptual aspect of annoyance relates to (psycho)acoustic characteristics of sounds, which induce annoyance in a bottom-up processing manner. On the other hand, influences by top-down processing are categorized as cognitive annoyance and relate to the disturbing effects, such as frequent repetitions or task interruptions (Zimmer et al., 2008). As stated, an inventory of knowledge on perceptual predictors of annoyance exists; however, mechanisms of cognitive annoyance remain unexplored. We believe the persistence of the alarm annoyance problem, despite all the efforts and extensive research, stems from the knowledge gap in understanding of nurses' cognitive needs during interaction with the monitoring system. Sounds may be well designed but poorly positioned within the workflow, therefore causing annoyance.

In this paper, we aim to identify the mechanism underlying nurses' annoyance of patient monitoring alarms. We argue that alarms are annoying to nurses on a cognitive level due to the conflict they pose in their information processing; rather than simply being unpleasant sounds. To support this hypothesis, we present data of an IC unit that captures the current situation of interruptions that nurses experience.

Cognitive Annoyance

In our approach, we frame cognitive annoyance as the negative feeling induced by a sound that is the result of the cognitive processing of the sound; rather than its perceptual qualities. In the following section, we present a series of cognitive processes that take place during nurses' interaction with patient monitoring alarms, and attempt to explain the potential reasons to nurses' annoyance of them. We consider the interruptions caused by alarms as a form of conflict in information processing, which is a well-established theory in the field of cognition (Botvinick et al., 2001).

Alarms in Human Information Processing

While tending to alarms is an essential part of nurses' workflow, the excessive number of alarms limits the time and attention for other clinical tasks. Furthermore, high rates of false alarms burden the cognitive load without requiring immediate action. In the field of cognitive science, the negative impact of task interruptions is well known. Interruptions are highly costly to performance and cognition: they increase reaction time, error rates, anxiety, annoyance, and perceived task difficulty (Bailey et al., 2000). This can be interpreted using the Human Information Processing model (HIP) (Figure 1, adapted from Wickens), which explains how the mind receives and processes physical stimuli (Wickens et al., 1992). The first stage of HIP is *perception* in which incoming physical stimuli are received by the senses, and formed into basic perceptual elements. In the case of patient monitoring alarms, this is when sound waves are received by the ears and turned into electrochemical signal for further processing. Within this stage, basic features of sounds (e.g., frequency, amplitude) are detected as perceptual elements that gives rise to psychoacoustical evaluation of alarms (e.g., sensory unpleasantness caused by loud or sharp tones). The second stage is *cognition*, in which meaning is attributed to perceptual elements. This stage involves evaluation of current information against prior knowledge, and decision making on the basis of meaning within the context. Attention is engaged to selectively direct resources to relevant stimuli and task related motor functions. For the processing of alarms, this stage involves an evaluation of the alarm to determine its source (e.g., oxygen saturation, or device such as mechanical ventilator), meaning (e.g., too much oxygenation), and actions it requires (e.g., reduce the oxygen intake by adjusting the dosage). Finally, the *response* stage is when a reaction to the physical phenomenon occurs. For alarms this can involve a physical action (such as tending to the patient or to the device for adjusting settings), or simply deciding the alarm is not relevant and therefore ignoring it.

Conflict in Human Information Processing

Attention is a limited resource, as also exemplified in the HIP model (Figure 1). When multiple stimuli are competing for the same resource, a conflict occurs. Resources must be shared between competing stimuli, limiting availability and therefore impairing performance. Different modalities engage different resources, so the degree of overlap between the competing stimuli influences the loss in performance (Wickens, 2008). Monitoring alarms initially engage visual and auditory resources for perception, then cognitive resources for processing, and finally motor resources for

response. Within the workflow, nurses are often engaged in various clinical tasks, to which alarms add competition with ongoing tasks. It might often be the case that several alarms are generated within one unit at the same time, inducing further conflict to information processing.

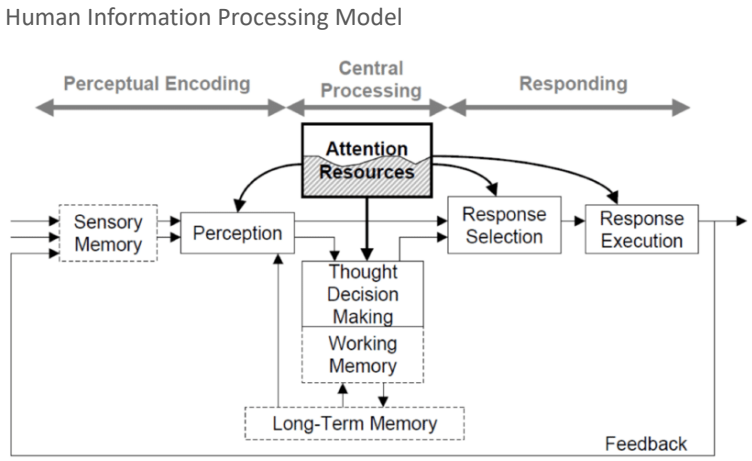


Figure 1. Human Information Processing Model (adapted from Wickens) demonstrating the processing of physical stimuli by the human mind. Consists of main stages: perception, cognition, and response. Note that attention is depicted as a limited resource.

Conflict in information processing is most commonly demonstrated by Stroop task (Stroop, 1935). In this paradigm, participants are asked to name the colour of the ink a word is written in aloud, disregarding the word itself. In congruent trials, ink colour matches the semantic meaning ("blue" written in blue); while incongruent trials demonstrate a mismatch ("blue" written in red). Incongruent trials involve higher error rates and increased reaction times. This is due to the competition between the response of reading the word and the response of verbalizing the ink colour. Both responses demand resources, resulting in a conflict.

Conflicts signals are well established to be instrumental for cognitive functioning. The mind monitors the degree of conflict in the environment, modulating level of cognitive control to match the demands (Botvinick et al., 2001). Remarkably, research indicates that conflicts are further registered as aversive signals (Dignath et al., 2020; Dreisbach & Fischer, 2012). Meaning that even in neutral and arbitrary conflicts such as the Stroop task, where the conflict holds no personal or emotional significance, people perceive it as negative affect. Therefore, the mind can be thought to keep count of conflicts in information processing and registering them as negative signals on a micro scale.

Conflict Resolution

Tasks competing for the same resources create bottlenecks in information processing (Broadbent, 1958). In order to complete both tasks, one must either multi-task or switch task. Mechanism underlying multitasking is modelled by the Threaded

Cognition Theory, which draws the analogy of a thread for each 'train of thought', or task-related processing (Salvucci & Taatgen, 2008). The theory posits that multi-tasking, even when seemingly concurrent such as talking and writing at the same time, is actually a serial process in which processing related to both tasks are sequentially alternated on a range of milliseconds. According to this view, threads are executed by favouring the least recently processed thread to balance performance outcomes. However, more recent research indicates that people have personal preferences in task prioritization (Jansen et al., 2016). When multiple arbitrary tasks are competing for resources, individual preferences influence which task is prioritised for serial processing. By rapidly alternating between multiple tasks, bottlenecks in information processing are resolved with minimal loss in performance.

Despite the efforts to attenuate the loss in performance, switching between tasks is still costly. Task switching is well known to increase error rates and reaction times (Monsell, 2003), and multi-tasking increases stress levels (Appelbaum et al., 2008). Remarkably, performance costs are less during voluntary task switching compared to involuntary task switching (Douglas et al., 2017; Vandierendonck et al., 2010). This phenomenon is thought to be due to anticipation of approaching conflict in the case of voluntary switching, in which elevated cognitive control alleviates the loss. This means frequent task-switches and periods of multi-tasking threaten the efficiency of workflow while burdening the cognition.

Annoyance by patient monitoring alarms

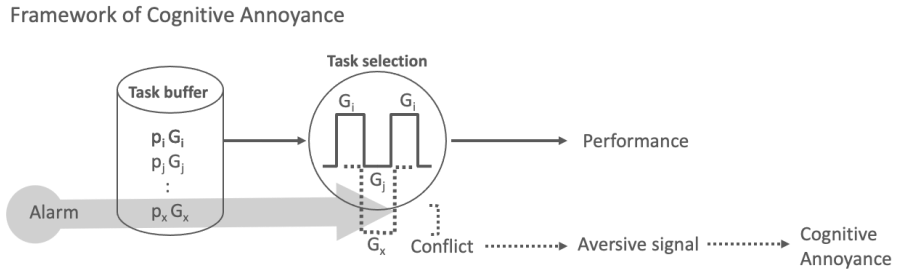


Figure 2. Schematic representing alarms inducing new tasks into the central processor. Each alarm adds a new, involuntary task (Gx), and over burdens cognitive resources. (p) is the prioritization coefficient of each task. While people have personal preferences on task prioritization, alarms, by design, override other tasks. Alarms have varying priority weights based on whether they are high, medium, or low level of priority. More task demands than available resources create conflict. Conflicts are registered as aversive signals. Accumulation of conflict signals is experienced as annoyance. Adapted from Jansen et al. (2016) with permission.

In light of the series of cognitive functions presented above, in this section we will attempt to describe how nurses might get annoyed by patient monitoring alarms. In the ICU, each new alarm imposes a new task for the nurse. Even a false alarm still requires re-allocation of perceptual and cognitive resources to identify them as false alarms, and potentially motor resources to silence the alarm. Each alarm induces a new thread to the multi-tasking processor. Therefore, alarms interfere with ongoing

tasks and require frequent task switching. Discrepancy between available resources and demands induced by multitude of tasks induces conflict in information processing, and triggers aversive signals. Since task-switches are not voluntary but imposed by alarms, they are more detrimental to cognition and performance. Consequently, we hypothesize that nurses' annoyance of monitoring alarms is an accumulation of aversive conflict signals in information processing. A schematic explanation is portrayed in Figure 2.

Data Collection

In order to quantify the frequency of alarms in the ICU and establish a description and understanding of the context, a data collection was conducted in Erasmus Medical Center, Rotterdam in the Netherlands between March and April 2022. All output from the patient monitoring system was recorded in a neonatal intensive care unit (NICU). Monitoring system automatically logs all events, so we accessed the logs to draw the data set. This study focused on alarms generated solely by the patient monitoring system. All other devices that generate audiovisual alarms, such as infusion pump or ventilation device, were not included in the analysis.

The neonatal unit contains eight patient beds in an open layout; where all beds are located close to each other and facing towards a central nurse station. This means all the alarms generated within the unit are audible to all the health care providers and patients in the unit. Nurses work in three shifts: morning, afternoon, and midnight.

Results

In a span on one month, 25 different patients were registered to the unit over different periods of time. Distribution of number of alarms per patient through the month is indicated in Figure 3. During this period, 83.023 alarms were recorded in total. Mean number of alarms per day in the unit was 2594.69, SD = 866.15. Minimum daily alarm count was 1296, and maximum was 4451. Median number of alarms generated by one patient was 1460, with a minimum of 100 and maximum of 13405.

Number of generated alarms fluctuated throughout the day. An hourly distribution of number of alarms summer over the month is presented in Figure 4a. On average, there were 111.45 alarms an hour, SD = 49.24. Minimum number of alarms per hour was 2, while maximum was 332. A frequency distribution of alarm counts per hour is presented in Figure 4b. While approximately 100 alarms per hour was the most commonly observed case, it was possible to observe over 300 alarms per hour.

Number of alarms peaked between 8:00-9:00. This period is known to be patient handover and the start of the morning routine. Patients are cleaned and daily check-ups are performed, in which sensors may get detached and trigger alarms. This is further exemplified by examining the condition that generates the alarm. Alarms were categorized into medical (those triggered by vital parameter measurements, e.g., blood oxygenation threshold exceeded, asystolie), and technical alarms (those related to the monitoring system and devices, e.g., sensor detached). Overall, 82.62% (68593) of alarms were of medical events, and 17.38% (14430) were technical events. Zooming into the time window of 8:00-9:00; 77.08% were of medical events (3377) while

22.92% (1004) were technical. This indicates more device related technical alarms were generated during the morning rounds compared to the daily averages.

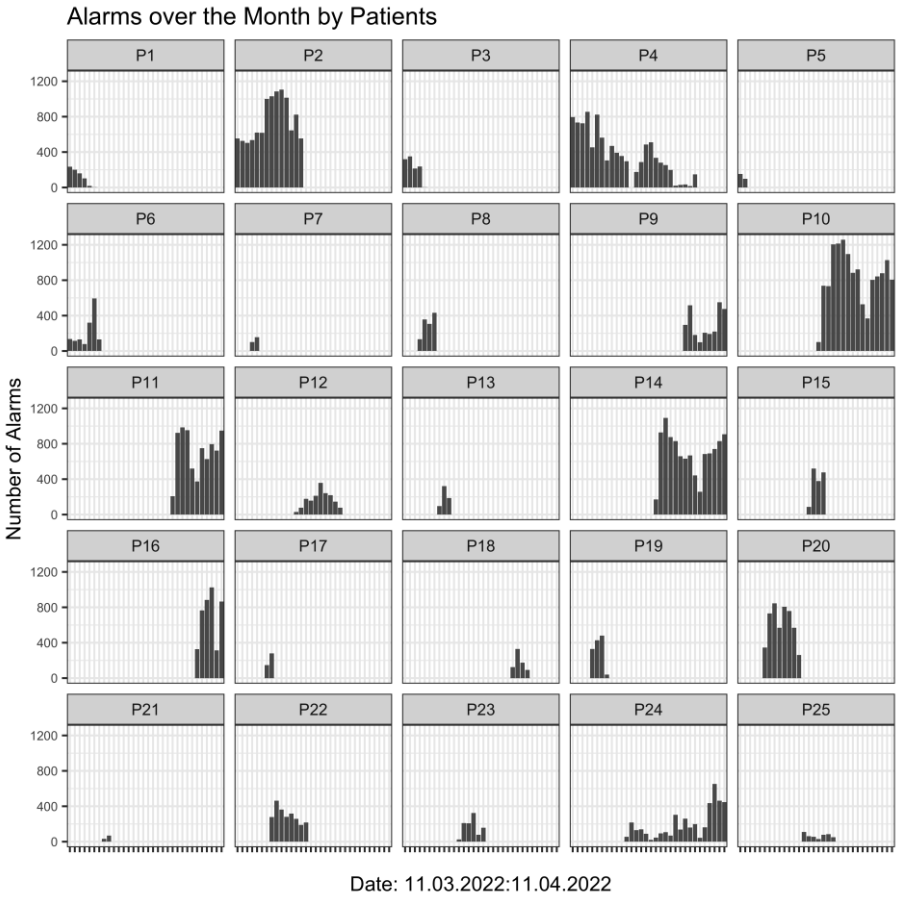


Figure 3. Number of alarms generated by each patient over the month. While some patients stay for longer periods of time; some are discharged quicker, as can be observed from the number of bars representing one day per patient.

We investigated the differences between morning, afternoon, and midnight shifts. Summarised over the patients, a daily average of 1151.84 (35.81%) alarms were generated during morning shifts, 1135.16 (32.67%) during afternoon shifts, and 1033.92 (31.52%) during midnight shifts. The number of alarms per shifts was converted into proportions for each shift and patient. These proportions were analysed by a within-subjects General Linear Model with shifts as the within-subjects factor of 3 levels. Wilk's Lambda was used a multivariate test, $F(2, 23) = 2.71$, $p = .087$. However, contrasts between levels showed a significant difference, in which there were more alarms in the morning shift (.36) compared to the midnight shift (0.32), $F(1, 24) = 5.65$, $p = .026$.

By medical standards, alarms are categorized into high, medium, and low levels of priority. Exploring the output from the patient monitor, majority of the alarms were medium priority (76.91%), while 12.97% were low priority, and only 10.05% were high priority alarms. High and medium priority alarms were often originated by medical conditions, while low priority alarms were often due to technical conditions (Table 1).

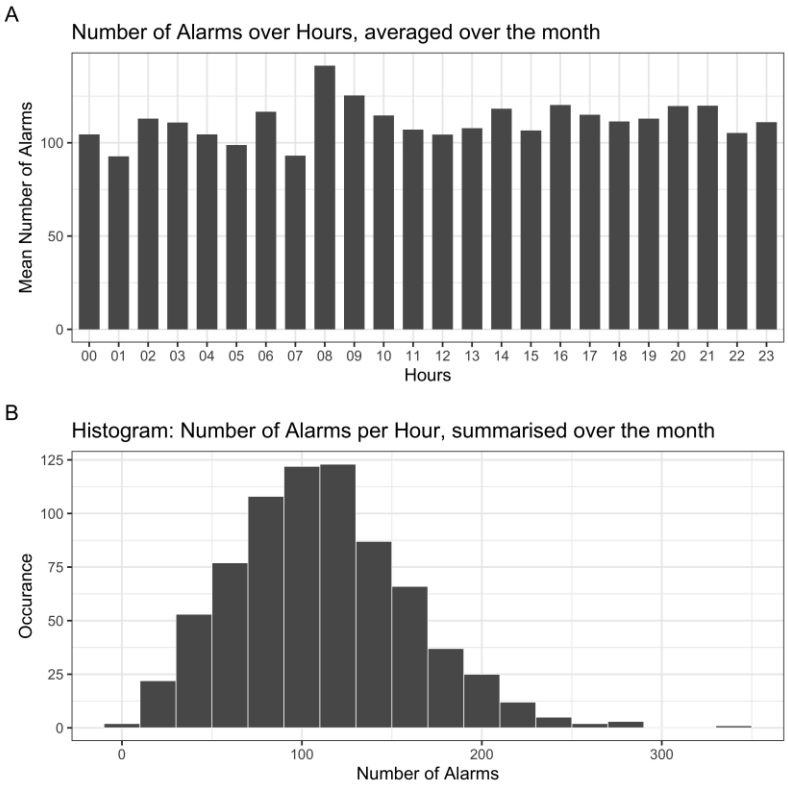


Figure 4. (A) Hourly distribution of mean alarm counts over the day. Number of alarms generated peaks around 8:00-9:00. (B) Frequency distribution of count of alarms per hour. While approximately 100 alarms per hour is commonly observed, it was possible to observe over 300 alarms per hour.

Table 1. Number of alarms per level of priority and alarming condition. Numbers are presented along with the percentage of the condition within one level of priority.

	Alarming Condition	
Level of Priority	*Medical*	*Technical*
Low	224 (2.08%)	10547 (97.92%)
Medium	60215 (94.22%)	3697 (5.78%)
High	8154 (97.77%)	186 (2.23%)

We investigated the variation among individual patients. Proportions of alarm priority levels and causes of alarms varied by patients. Distribution of priority levels for each patient is displayed in Figure 5a, and distribution of alarming condition is displayed in Figure 5b. Figure 5a demonstrates that the majority of alarms were medium priority for most of the patients. However, more low priority alarms were generated by certain patients (e.g., P6, P18, P19). Figure 5b illustrates that these patients also generate relatively high proportion of technical alarms. This indicates that these patients are relatively more mobile than others, resulting in more sensors getting detached and therefore generating more technical alarms.

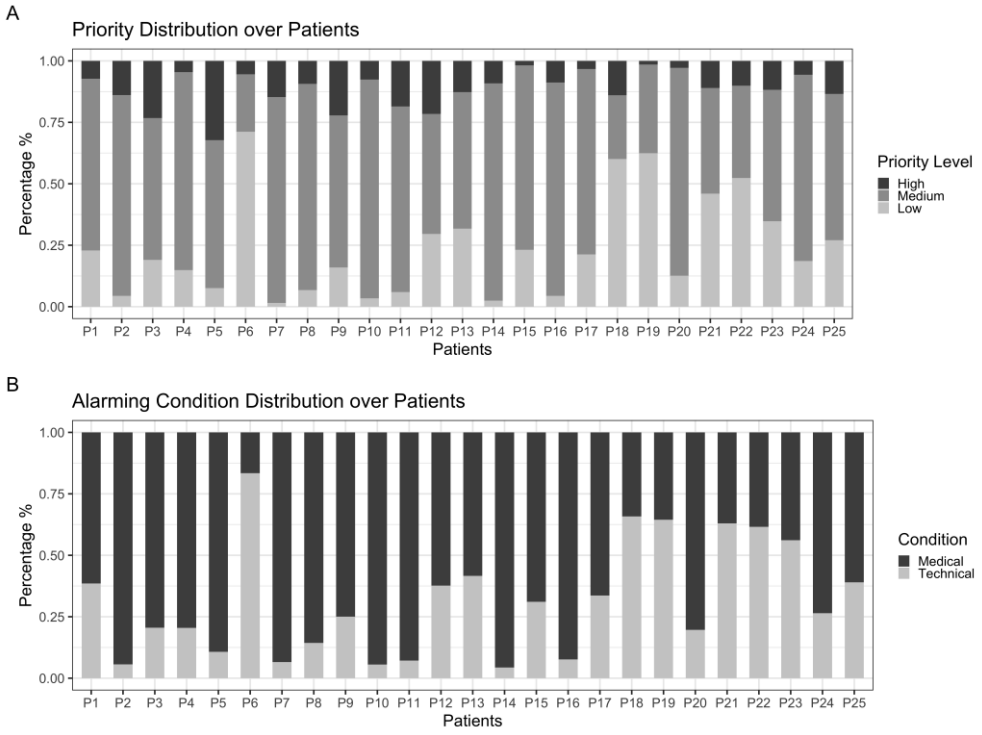


Figure 5. (A) Stacked chart of alarm priority level distribution per patient. (B) Stacked chart of alarming condition distribution per patient.

Vital parameters that generate the alarms were analysed to investigate which medical and technical conditions were most relevant for this IC unit. The patient vital parameter that generated most of the alarms was oxygen saturation level (SpO_2, 56.81%), followed by electrocardiogram (ECG, 10.43%) technical alarms. A breakdown of number of alarms by alarming vital parameter is presented in Table 2.

Table 2. Breakdown of vital parameters which trigger the alarms. Threshold refers to the alarm being triggered by vital parameter exceeding the set threshold; while technical refers to technical alerts such as artifacts, or sensor being detached. Parameters that occur less than 1% of the time are aggregated as 'other'. SpO2: oxygen saturation, ECG: electrocardiogram, HR: heart rate, RRi: impedance respiratory rate.

	Percent	*Count*
SpO_2 threshold	56.81%	47162
ECG technical	10.43%	8658
SpO_2 desaturation	9.19%	7631
HR threshold	5.41%	4491
SpO_2 technical	5.35%	4444
HF threshold	5.21%	4327
RRi threshold	3.21%	2665
Temperature threshold	2.05%	1699
RRi technical	1.43%	1186
Other	1.00%	760

While the number of alarms presented so far represent the alarming instances, alarms are often audible for longer periods of time. Therefore, the auditory stimuli present in the IC unit is in fact more prevalent than the number of alarms indicate. To capture this, we analysed the duration of alarms. Excluding the outliers where alarm duration was greater than 180 seconds, median alarm duration was 10 seconds, mean was 22.81, and SD = 30.85. A histogram of alarm durations is presented in Figure 6a. Alarm durations differed for levels of priority. High priority alarms had a mean duration of 14.15 seconds, medium alarms had mean of 25.54, and low priority alarms had a mean of 13.77 seconds (Figure 6b). Alarm durations also varied by the vital parameter that generates the alarm. Mean duration in seconds per parameter is presented in Figure 6c.

Cumulative number of alarms in the unit represent the total auditory stimuli in the environment. While the alarms are audible within the whole unit, each nurse is responsible for tending to the alarms generated by the patient assigned to them. To capture the demand of responsibility, we analysed the number of alarms generated by each patient during one shift. Averaged over the month and patients, mean number of alarms generated by one patient during one shift was 123.90, SD = 78.71.

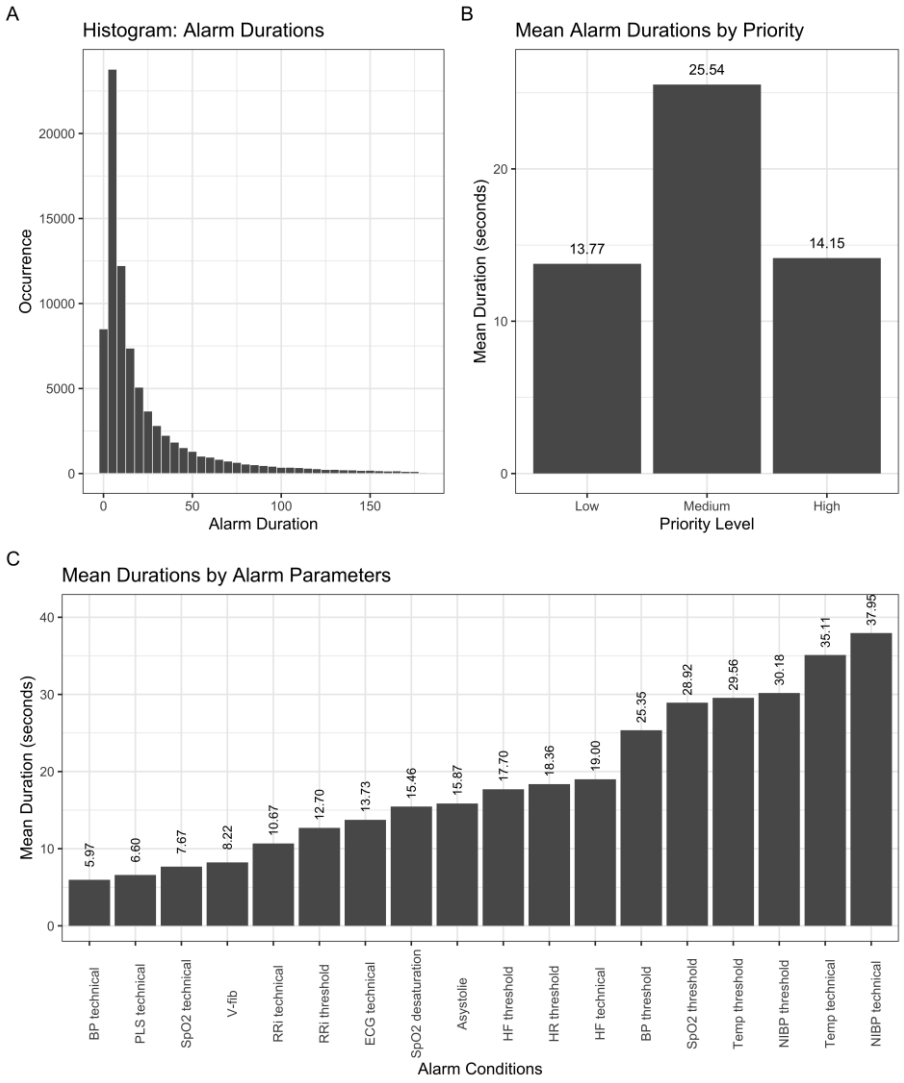


Figure 6. Alarm durations in seconds, outliers greater than 180 excluded. (A) Histogram of alarm durations. (B) Mean alarm durations by alarm priority levels. (C) Mean alarm durations in seconds by vital parameters.

Discussion

Our results of the output from patient monitors demonstrate the prevalence of alarms in the ICU. Realizing the excessive number of alarms helps us understand the experiences of ICU nurses within their workflow. Our results show that almost two alarms per minute were generated in the unit, and one patient generated an alarm every 3.22 minutes. Average duration of alarms was over 20 seconds, indicating that alarms are almost constantly audible in the IC unit. These results paint a clear picture of the auditory stimuli present in the unit as experienced by nurses and patients. The majority

of studies aiming to improve patient monitoring alarms has focused on the acoustic characteristics of alarm sounds (Edworthy et al., 2018; Foley et al., 2020; Schlesinger et al., 2018; Sreetharan et al., 2021). While efforts to improve the sound design of alarms will benefit the sensory experience, our results make it clear that the main cause of the problem is the excessive number of alarms. This number indicates the frequency by which nurses are interrupted in ongoing tasks. Consequently, we argue that the understanding of the cognitive mechanisms of the processing of alarm sounds is more important to explain the experienced annoyance. In our framework, each interruption burdens the cognitive resources by creating conflict in information processing. As conflicts are experienced as aversive signals by the mind, each interruption adds to the feeling of annoyance towards patient monitoring alarms in nurses. Therefore, we argue that efforts to improve nurses' organizational well-being requires an approach beyond enhancing the alarm sounds. Consideration of nurses' cognitive needs, capabilities, and preferences is needed to improve the information communication between patient monitoring systems and nurses.

More specifically, our analysis of the generated alarms reveals potential points to improve system design. Results demonstrate that high priority alarms are the least occurring alarms, which is the only type of alarm that requires immediate action. Low and medium priority alarms constitute the majority of alarms. These can be reduced in number by human interventions (such as customizing alarm limits), or by improvements in the system design (such as smart algorithms to prioritize and eliminate alarms). Most commonly observed cause for alarms was related to blood oxygen saturation level, which is typical for neonatal patients. Interventions that target the optimization of blood oxygen saturation monitoring can yield considerable improvement in the number of generated alarms.

We found that there is a large variation in the number and type of alarms generated by each patient. Currently, the settings of the monitoring system remain similar for each patient. However, the distribution of vital parameters that generate the alarms varies over patients. This can be explained either by the patients' medical status (relatively stable or critical), or by the frequency of movements. Patients who move around frequently cause sensors to become detached more often, leading to more technical alarms. The same effect is also visible in the reduced number of alarms during night shifts. Patients are more likely to be sleeping during the night; and there is a reduced number of lights, sounds, and general activity during night time; leading to fewer alarms generated. Such differences in patient characteristics, and conditions surrounding the patient could be an input for the monitoring system to supress non-actionable alarms based on current needs.

For essential events that do need to be notified to the nurse, literature has suggested methods to minimize the negative consequences of task interruptions. These involve methods to design smart algorithms to prioritise alarms. This can be achieved by context aware computing that supress notifications based on location signals or certain periods of time, and user aware computing that generates notifications based on attentional cues from the user (Ansari et al., 2016; Bailey & Konstan, 2006; Welch, 2011). These methods aim at notifying the user to system conditions on particular moments where the interruption is thought to yield the minimum negative effect on

performance and cognition. By understanding the cognitive mechanisms that make patient monitoring alarms annoying to nurses, we can employ design strategies in a targeted manner to minimize these effects.

In this paper, we suggested a framework in which accumulation of aversive conflict signals caused by interruptions are experienced as annoyance towards monitoring alarms (Figure 2). Our theoretical framework opens up new directions for future research. One of these is to measure annoyance when task interruption is induced by another modality, since different modalities require different resources. Another intriguing direction would be to build up on the research suggesting increased costs for involuntary task switches compared to voluntary switches. This difference is thought to be caused by anticipation of conflict (Vandierendonck et al., 2010). Endsley (1995) indicates that anticipation is an important factor in Situation Awareness. This aspect is often overlooked in the interaction between nurses and patient monitoring systems. Investigating the role of anticipation on annoyance ratings can present insights into how nurses handle (un)expected information presented through alarms. This knowledge would then inform design of the interaction between the system and the nurse as a user. These aspects will form the basis of our future research activities on cognitive annoyance in ICUs.

References

Ansari, S., Belle, A., Ghanbari, H., Salamango, M., & Najarian, K. (2016). Suppression of false arrhythmia alarms in the ICU: A machine learning approach. *Physiological Measurement, 37*(8), 1186–1203. https://doi.org/10.1088/0967-3334/37/8/1186

Appelbaum, S. H., Marchionni, A., & Fernandez, A. (2008). The multi-tasking paradox: Perceptions, problems and strategies. *Management Decision*, *46*, 1313–1325. https://doi.org/10.1108/00251740810911966

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

Bailey, B.P., Konstan, J.A., & Carlis, J.V. (2000). Measuring the effects of interruptions on task performance in the user interface. *Proceedings of the IEEE International Conference on Systems, Man and Cybernetics*, *2*, 757–762. https://doi.org/10.1109/ICSMC.2000.885940

Botvinick, M.M., Braver, T.S., Barch, D.M., Carter, C.S., & Cohen, J.D. (2001). Conflict monitoring and cognitive control. *Psychological Review*, *108*(3), 624–652. https://doi.org/10.1037//0033-295x.108.3.624

Cho, O.M., Kim, H., Lee, Y.W., & Cho, I. (2016). Clinical alarms in intensive care units: Perceived obstacles of alarm management and alarm fatigue in nurses. *Healthcare Informatics Research*, *22*(1), 46–53. https://doi.org/10.4258/hir.2016.22.1.46

Cvach, M. (2012). Monitor alarm fatigue: An integrative review. *Biomedical Instrumentation and Technology*, *46*(4), 268–277. https://doi.org/10.2345/0899-8205-46.4.268

Deb, S., & Claudio, D. (2015). Alarm fatigue and its influence on staff performance. *IIE Transactions on Healthcare Systems Engineering*, *5*(3), 183–196. https://doi.org/10.1080/19488300.2015.1062065

Dignath, D., Eder, A. B., Steinhauser, M., & Kiesel, A. (2020). Conflict monitoring and the affective-signaling hypothesis—An integrative review. *Psychonomic Bulletin and Review*, *27*(2), 193–216. https://doi.org/10.3758/s13423-019-01668-9

Douglas, H.E., Raban, M.Z., Walter, S.R., & Westbrook, J.I. (2017). Improving our understanding of multi-tasking in healthcare: Drawing together the cognitive psychology and healthcare literature. *Applied Ergonomics*, *59*, 45–55. https://doi.org/10.1016/j.apergo.2016.08.021

Dreisbach, G., & Fischer, R. (2012). Conflicts as aversive signals. *Brain and Cognition*, *78*(2), 94–98. https://doi.org/10.1016/j.bandc.2011.12.003

Edworthy, J.R., McNeer, R.R., Bennett, C.L., Dudaryk, R., McDougall, S.J.P., Schlesinger, J.J., Bolton, M.L., Edworthy, J.D.R., Özcan, E., Boyd, A D., Reid, S.K.J., Rayo, M.F., Wright, M.C., & Osborn, D. (2018). Getting Better Hospital Alarm Sounds Into a Global Standard. *Ergonomics in Design*, *26*(4), 4–13. https://doi.org/10.1177/1064804618763268

Endsley, M.R. (1995). Toward a theory of situation awareness in dynamic systems. *Human Factors*, *37*, 32–64. https://doi.org/10.1518/001872095779049543

Foley, L., Anderson, C.J., & Schutz, M. (2020). Re-Sounding Alarms: Designing Ergonomic Auditory Interfaces by Embracing Musical Insights. *Healthcare*, *8*(4), 389. https://doi.org/10.3390/healthcare8040389

Guski, R., Felscher-Suhr, U., & Schuemer, R. (1999). The concept of noise annoyance: How international experts see it. *Journal of Sound and Vibration*, *223*(4), 513–527. https://doi.org/10.1006/jsvi.1998.2173

Jansen, R.J., Van Egmond, R., & De Ridder, H. (2016). Task prioritization in dual-tasking: Instructions versus preferences. *PLoS ONE*, *11*(7). https://doi.org/10.1371/journal.pone.0158511

Janssen, S.A., Vos, H., Eisses, A.R., & Pedersen, E. (2011). A comparison between exposure-response relationships for wind turbine annoyance and annoyance due to other noise sources. *The Journal of the Acoustical Society of America*, *130*(6), 3746–3753. https://doi.org/10.1121/1.3653984

Lewandowska, K., Weisbrot, M., Cieloszyk, A., Mędrzycka-Dąbrowska, W., Krupa, S., & Ozga, D. (2020). Impact of alarm fatigue on the work of nurses in an intensive care environment—a systematic review. *International Journal of Environmental Research and Public Health*, *17*(22), 1–14. https://doi.org/10.3390/ijerph17228409

Monsell, S. (2003). Task switching. *Trends in Cognitive Sciences*, *7*(3), 134–140. https://doi.org/10.1016/S1364-6613(03)00028-7

Özcan, E., & Gommers, D. (2020). Nine Nurse-Recommended Design Strategies to Improve Alarm Management in the ICU: A Qualitative Study. *ICU Management & Practice*, *20*(2), 129–133.

Paunović, K., Jakovljević, B., & Belojević, G. (2009). Predictors of noise annoyance in noisy and quiet urban streets. *Science of the Total Environment*, *407*(12), 3707–3711. https://doi.org/10.1016/j.scitotenv.2009.02.033

Pierrette, M., Marquis-Favre, C., Morel, J., Rioux, L., Vallet, M., Viollon, S., & Moch, A. (2012). Noise annoyance from industrial and road traffic combined noises: A survey and a total annoyance model comparison. *Journal of Environmental Psychology*, *32*(2), 178–186. https://doi.org/10.1016/j.jenvp.2012.01.006

Salvucci, D.D., & Taatgen, N.A. (2008). Threaded Cognition: An Integrated Theory of Concurrent Multitasking. *Psychological Review*, *115*(1), 101–130. https://doi.org/10.1037/0033-295X.115.1.101

Schlesinger, J.J., Baum Miller, S.H., Nash, K., Bruce, M., Ashmead, D., Shotwell, M.S., Edworthy, J.R., Wallace, M.T., & Weinger, M.B. (2018). Acoustic features of auditory medical alarms—An experimental study of alarm volume. *The Journal of the Acoustical Society of America*, *143*(6), 3688–3697. https://doi.org/10.1121/1.5043396

Siebig, S., Kuhls, S., Imhoff, M., Langgartner, J., Reng, M., Schölmerich, J., Gather, U., & Wrede, C.E. (2010). Collection of annotated data in a clinical validation study for alarm algorithms in intensive care - a methodologic framework. *Journal of Critical Care*, *25*(1), 128–135. https://doi.org/10.1016/j.jcrc.2008.09.001

Sowan, A.K., Gomez, T.M., Tarriela, A.F., Reed, C.C., & Paper, B.M. (2016). Changes in Default Alarm Settings and Standard In-Service are Insufficient to Improve Alarm Fatigue in an Intensive Care Unit: A Pilot Project. *JMIR Human Factors*, *3*(1), e1. https://doi.org/10.2196/humanfactors.5098

Sowan, A.K., Tarriela, A.F., Gomez, T.M., Reed, C.C., & Rapp, K.M. (2015). Nurses' Perceptions and Practices Toward Clinical Alarms in a Transplant Cardiac Intensive Care Unit: Exploring Key Issues Leading to Alarm Fatigue. *JMIR Human Factors*, *2*(1), e3. https://doi.org/10.2196/humanfactors.4196

Sreetharan, S., Schlesinger, J.J., & Schutz, M. (2021). Decaying amplitude envelopes reduce alarm annoyance: Exploring new approaches to improving auditory interfaces. *Applied Ergonomics*, *96*, 103432. https://doi.org/10.1016/j.apergo.2021.103432

Stroop, J.R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology*, *18*(6), 643.

Vandierendonck, A., Liefooghe, B., & Verbruggen, F. (2010). Task Switching: Interplay of Reconfiguration and Interference Control. *Psychological Bulletin*, *136*(4), 601–626. https://doi.org/10.1037/a0019791

Welch, J. (2011). An evidence-based approach to reduce nuisance alarms and alarm fatigue. *Biomedical Instrumentation and Technology*, *45*(SPRING), 46–52. https://doi.org/10.2345/0899-8205-45.s1.46

Wickens, C.D., Helton, W.S., Hollands, J.G., & Banbury, S. (1992). *Engineering psychology and human performance*. London: Routledge

Wickens, C.D. (2008). Multiple resources and mental workload. *Human Factors*, *50*, 449–455. https://doi.org/10.1518/001872008X288394

Wilken, M., Hüske-Kraus, D., Klausen, A., Koch, C., Schlauch, W., & Röhrig, R. (2017). Alarm fatigue: Causes and effects. *Studies in Health Technology and Informatics*, *243*, 107–111. https://doi.org/10.3233/978-1-61499-808-2-107

Yue, L., Plummer, V., & Cross, W. (2017). The effectiveness of nurse education and training for clinical alarm response and management: a systematic review. *Journal of Clinical Nursing*, *26*(17–18), 2511–2526. https://doi.org/10.1111/jocn.13605

Zimmer, K., Ghani, J., & Ellermeier, W. (2008). The role of task interference and exposure duration in judging noise annoyance. *Journal of Sound and Vibration*, *311*(3–5), 1039–1051. https://doi.org/10.1016/j.jsv.2007.10.002

Zwicker, E., Fastl, H. (1999). Sharpness and Sensory Pleasantness. In *Psychoacoustics* (239–246). Springer Series in Information Sciences, vol 22. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-662-09562-1_9

A new approach to sound design in automated vehicles

Soyeon Kim[1], Tarek Kabbani[2], Duygu Serbes[3], Riender Happee[1], Ahu Ece Hartavi[2], & René van Egmond[1]
[1]Delft University of Technology, Delft, The Netherlands
[2]University of Surrey, Guildford, United Kingdom
[3]Ford Otosan, Istanbul, Turkey

Abstract

Human-Machine Interfaces (HMIs) aim to support the interaction between automated vehicles and drivers to improve safety and driver experience. With the development of automated vehicles, drivers interact with vehicles in new scenarios. In addition to visual modality, sound is the other modality often used in vehicles. Previously, sounds were mainly used for alarms, but they can be used in other ways in automated vehicles. Therefore, a new approach to sound design is needed. We proposed an interactive approach for sound design to improve driver safety and user experience in automated vehicles. In this study, we suggested that the driver's interaction with automated vehicles should be analyzed based on the user and contextual understanding, and the sound should be designed to consider the appropriateness of situation matching and alert levels. This study showed that the approach supports designing sounds that enhance vehicle and driver interaction.

Introduction

Although alarm sounds remain important in user interface design, sounds offer more possibilities that are currently not used. One of the advantages of sounds is capturing omnidirectional attention (Siwiak & Jame, 2009). The auditory interface can help increase visual attention for potential risk situations on the road (Beattie et al., 2014). At the same time, sound can annoy drivers, which is a significant concern in designing auditory displays (Edworthy, 1998). In addition, sounds in an auditory interface often convey a sense of urgency, which is different from the design intention (Edworthy, 1994). Despite some of these wrong implementations of sounds in user interfaces, sounds have important information capacities and advantages. In conclusion, sounds should not only be designed on their perceptual quality, but more importantly, the interaction of drivers with the vehicle and its context should be the determinant factors in the design of the sounds.

Drivers interact with the vehicle in several situations during driving. As automated driving becomes possible, new scenarios of driver-vehicle interaction have been developed such as a transition of the control. In these scenarios, sounds can be used not only to warn the user about the actual transition but to support the entire transition process.

In S. Kim, T. Kabbani, D. Serbes, R. Happee, A.E. Hartavi, and R. van Egmond (2022). Proceedings of the Human Factors and Ergonomics Society Europe Chapter 2022 Annual Conference. ISSN 2333-4959 (online). Available from http://hfes-europe.org

In several studies on automated driving, the focus was on the impact of the type of modality. Take-over requests with only sound or sounds combined with visual or tactile modality were compared in reaction time, situational awareness, and acceptance (Politis et al., 2015, Petermeijer et al., 2017, Roche et al., 2019). The function of sounds was evaluated by providing an automation information (König & Neumayr, 2017) or transition-related information (van den Beukel et al., 2016). In addition, the effect of different sounds on ADAS (Advanced driving assist system) function activation (Larsson & Västfjäll, 2013) or take-over requests (Jeon, 2019) were studied. These studies aim to identify the effect of sounds in certain situations and provide important insights into sound design for automated vehicles. Previous studies mainly provided single information in a specific scenario by sound, mainly in the form of beeping. There has been a lack of consideration of the sound design process. Designing sound should be considered how users perceive the context information by sounds or what experiences could be delivered (Özcan & van Egmond, 2008).

We suggest an approach for sound design to provide contextual information during driving. First, designing sounds is based on understanding user interaction and driving situations. Next, to design sounds to match the situation and alert level. This new approach will allow the design sounds to be better accepted and keep their functionality.

Design Methodology

The sound design process consists of two steps as follows. 1. Understanding the driver context and analyzing vehicle-driver interaction, and 2. Designing the sounds

Understanding the driver context and analyzing vehicle-driver interaction

Interaction between drivers and automated vehicles and its context is analyzed, and the situation where the sound should be provided is selected. The abstract sound is difficult to convey narrative information. Speech may be used, but this may quickly increase drivers' annoyance (Forster et al., 2017). Auditory interfaces in vehicles are commonly used to draw drivers' attention to a visual display. A driver may be aware of sounds and checks the details through the visual interface. When sounds are provided in an integrated way with visual information, sounds do not have to contain information itself. However, only function as means to attract attention. In consideration of the need for interaction, sounds need to be added in essential situations. If the same sounds are overused, the use of sounds becomes counterproductive.

Based on the interactions, the sound can be classified as an Indication or Alert. The indication provides information which is not urgent, such as informing automation state or feedback notification of drivers' input. It helps drivers recognize changes and draws attention to the visual interface. An alert warns drivers to be aware of a situation. If an indication sound is not properly designed, it may evoke a sense of urgency for an alarm sound, and vice versa.

Designing the sounds

Sounds are designed in consideration of the appropriateness of matching and alert level.

Appropriateness of matching
Sounds characteristics should be adjusted to make a sound match the situation. For example, when the pitch rises, something starts, and when the pitch decreases, something is turned off. In urgent situations, a sound is designed to discriminate from other sounds in such a way that the driver should know the situation accurately only by sound. Like in the design of warning symbols, the elements of legibility, conspicuity, discriminability, and urgency mapping are required for users to comprehend symbols (Edworthy, 1998). Analogously, these factors should be considered when designing alarm sounds.

Alert level
Sounds should be designed according to the alert level of a situation. According to ANSI (American National Standards Institute, 1991) standards, there are four stages of alert level: notice, caution, warning, and danger. An alert level is assigned in consideration of the purpose classified in the previous step. In a critical situation, the urgency level can be changed based on a reaction of a driver. For example, the urgency of a take-over request sound can be increased if a driver does not react appropriately. Changes in sound elements such as pulse interval or decibels significantly impact the perceived urgency (Hellier et al., 1993). Unpleasantness can be used to draw the user's attention (Özcan & van Egmond, 2012). Contexts of high alert levels requiring an immediate reaction use the unpleasant parameters as a partial solution. However, sounds of non-critical context consider acceptance, such as pleasantness or appropriateness of matching situations with sounds, rather than drawing immediate attention through sound.

Study 1 - Designing Sounds to Support Visual HMI for Autonomous Truck Drivers

This interactive approach to sound design was applied in a study to inform truck drivers in an autonomous driving situation. Truck drivers are one of the personas within HADRIAN a Horizon, 2020 project. The interaction scenarios developed in Hadrian capture transitions between manual and autonomous driving and driver attention, which are regarded as critical events in automated vehicles (AV).

Truck drivers are professionals and drive longer periods of time than passenger car drivers. (Belman et al., 2004) has indicated that a truck driver on average drives 8.4 hours. (Horberry et al., 2022) have recommended increasing the loudness of the auditory message in trucks in order to prevent them of being masked by background noise. Considering that truck drivers are already exposed to noise over a long period of time it was decided to keep the sound design simple in nature. Sounds normally used in this context are simple beeps. Therefore, we have made sure that the new sound design is inherited from this tradition.

A visual interface wireframe based on the HADRIAN interaction scenarios was first developed in a previous study (Kabbani et al., 2022). This wireframe guided the sound design. A list of the designed sounds is presented in Table 1 indicated by their name. The sounds have two apparent functionalities. First, the sounds indicated in Table 1 by '*Hands-on steering wheel'* and '*Ask attention'* are adaptive to the driver's state. If a driver does not respond to the warning appropriately, then the urgency level increases stepwise. Second, the sounds that support a mode change are *AV available, Driver confirmation* and *AV start* and do not change the urgency level over time.

Table 1. list of designed sounds

Situation	*Sound Names*	*Purpose*	*Expected Alert level*
AV on	AV available	Indication	Notice
	Driver confirmation	Indication	Notice
	AV start	Indication	Caution
AV off	Take-over has started	Alert	Warning
	Hands-on steering wheel	Alert	Warning – Danger
Driver distraction	Ask attention	Alert	Warning - Danger
	Parking maneuver	Indication	Caution

In SAE levels 3, 4 and 5 an automated vehicle has a role in monitoring and controlling. This allows drivers to take their hands off the steering wheel and perform non-driving related tasks (SAE International, 2018). A notification that indicates that the automation mode has changed has a positive effect on usability and safety positively (Nadri et al., 2021). For a transition to autonomous driving ('*AV on'* in Table 1), three sounds are designed to support the steps of this transition. First, a vehicle informs a driver that autonomous driving is possible (*AV available*). Second, when a driver confirms the change to the autonomous driving mode, the vehicle gives a feedback sound (*Driver confirmation*). Thirdly, a sound is provided when AV mode is started (*AV start*).

Several studies (Politis et al., 2015, Petermeijer et al., 2017) have shown that take-over requests using sounds are more advantageous than take-over requests using only a visual interface. (van der Heiden et al., 2017) found that providing a sound before a take-over request made the take-over situations safer. In our study, when a scheduled take-over request occurs from the automation to the manual mode, a driver receives a sound thirty seconds in advance (*Take-over has started*). Fifteen seconds later, another take-over request is provided (*Hands-on steering wheel*). The urgency level will then gradually increase over time if a driver does not react. There is no need for designing an AV deactivation sound, because a driver will notice that manual driving is started when the warning sound is off. In an emergency transition, a '*Take-over request*' is directly provided without the '*Take-over has started'*.

In the HADRIAN project, a detection system for monitoring a driver state was developed. This system can provide a warning when a driver is distracted. A vehicle will present the sound '*Ask attention'* when the driver is not capable to drive. It also includes situations in which a driver is not adequately responding to a transition

request. The Inter-Onset-Intervals between sounds were reduced to generate higher urgency levels. If drivers do not react to the '*Hands-on steering wheel*' and '*Ask attention*' warning, a minimum risk maneuver to protect drivers will be started indicated with the sound (Parking maneuver)

Method

The sounds in Table 1 were validated in a video simulation created by modifying the simulator truck scenario of the Hadrian project, as shown in Figure 1 to confirm whether the design intention matched users' understanding.

Figure 1. Video capture of a driver distraction situation

Participants

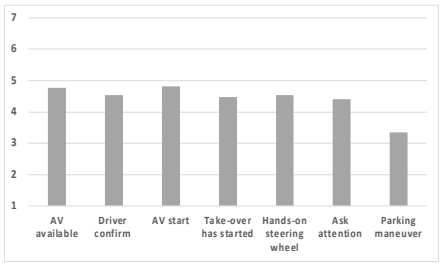
Seventeen drivers participated in the validation test. All subjects were male, and the average age was 41 years. All were professional truck drivers who had 17.1 years of experience in truck driving.

Procedure

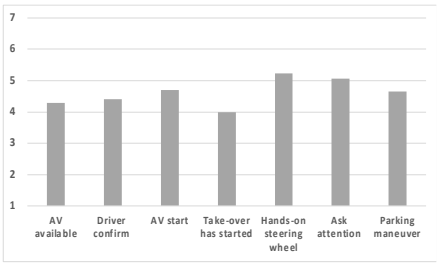
The drivers were explained the purpose of the study and their demographic data were collected. The main procedure consisted of two parts. In the first part, a *pleasantness* rating for each sound without any context was obtained using a 7-point Likert scale. We did not use a context in order to only measure *perceptual pleasantness*. In the second part, a video based on the Hadrian truck driving scenario was used to test the sounds of Table 1 in context. In this part, a participant rated the *alertness* based on the 4-level ANSI (American National Standards Institute, 1991) scale (*Notice, Caution Warning,* and *Danger*) and a 7-point Likert scale questionnaire with terms on *appropriateness, annoyance*, and *intention of use*.

Results

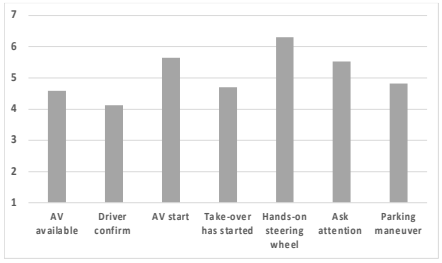
Figure 2 consists of four sub-figures, all addressing the perspective of the driver. In the top left figure, the participants' perception of pleasantness is shown. The overall perceived pleasantness scored more than 4 points. The lowest ranked pleasantness amongst all sounds were '*Parking maneuver'*. Moreover, a t-test indicated that '*AV available*', '*AV start*', and '*Take-over has started*' were rated significantly higher than the midpoint (4, *p*-value <0.05) of the scale. There was no correlation between pleasantness and the other attributes (*appropriateness, annoyance,* and *intention of use*).



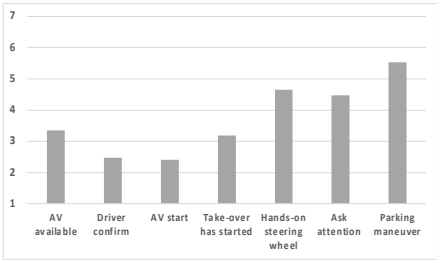
(1) Mean *pleasantness* ratings as a function of Sound Type (ratings were obtained without visual context)



(2) Mean *appropriateness* ratings as a function of Sound Type



(3) Mean *intention of use* ratings as a function of Sound Type



(4) Mean *annoyance* ratings as a function of Sound Type

Figure 2. Bar chart of the mean ratings for (1) *Pleasantness*, (2) *Appropriateness, (3) Intention of use,* and (4) *Annoyance*

The top right figure shows the appropriateness level of how well a sound matches a certain situation. Participants answered more than 4 points in all events, although only the *'Take-over has started', 'Hands-on steering wheel',* and *'Ask attention'* were rated significantly higher than 4 points from the t-test (p-value <0.05). In the bottom left figure, how necessary a sound is in a certain situation is shown. Participants also rated more than 4 points in all events, especially the average of *'AV start', 'Hands-on steering wheel',* and *'Ask attention'* were significantly higher than 4 points from the t-test (p-value <0.05). In the bottom right figures, an answer of annoyance level is shown. The '*Parking maneuver*' was rated with the highest annoyance points.

In Figure 3 the proportion of choice for the ANSI alert levels as a function of *Sound Name* is shown. More than 80% of participants responded that the sounds '*AV*

available', '*Driver confirmation'*, and '*AV start'* evoked the alertness levels *notice* or *caution*. These sounds were designed to function as an indication. Thus, this finding corresponds to the designed intention. However, around 80% of the participants indicated that the '*Take-over has started*' sound was a *notice* or a *caution* sound, although it was designed as a higher-level alert sound. No participant indicated that the alert level of the sound was dangerous. More than 60% of the participants indicated that the sounds '*Hands-on steering wheel'* and '*Ask attention'* were '*warning* or *danger'* alarms. This finding corresponds to the designed intention of the alert level. More than 60% of the participants indicated *Parking maneuver* sound was a *warning* or *danger*. However, the sound was designed for the function of indication.

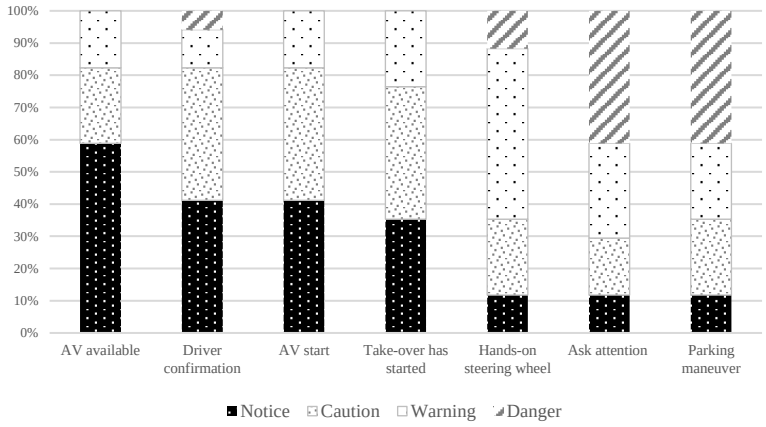


Figure 3. *Alert level responses*

Discussion

Our main finding is that it is possible to design sounds that are pleasant and of which the designed intention for a specific context matches their function and proper corresponding alert level. Seven sounds indicated in Table 1 were designed for interaction in an autonomous truck. Based on the validation results, five sounds (*'AV availble', "driving confirmation', 'AV start', 'Hands-on steering wheel',* and *'Ask attention'*) were appropriately designed for their driving context. The *'take-over has started'* and *'parking maneuver'* sounds were revised. The '*take-over has started'* sound was modified to raise the alert level, and the '*parking maneuver'* sound was modified to reduce the urgency for less annoyance and lower the alert level. The remaining sounds and the newly designed sounds will be applied in a FORD truck and evaluated in a road demonstration for the Hadrian project. This research has to be conducted.

Study 2 - Exploratory Study to Design Ambient Sounds for Highly Automated Vehicles

The above-described sound design methodology was used in a Master course, Interactive Audio Design, at the Faculty of Industrial Design Engineering. The study

purpose is to design context-relevant soundscapes including feedback sounds, with the aim of creating a better user experience. A soundscape is an acoustic environment perceived by listeners in contexts (Schafer, 1976), and it can provide context information (Aletta et al., 2016). Soundscape should orally describe the context in which a driver is and what kinds of actions are to be expected.

Students received the following scenario description on which they had to base their sound design: “The vehicle is a B-segment-sized vehicle, commonly described as a small car and the largest segment volume in Europe (i.e., Toyota Yaris, Renault Clio). The driver is around 35 years old and runs a startup company. The vehicle is highly automated, allowing the driver to work such as sending an email or writing a document during the automated driving mode". A final deliverable was a movie clip including sounds and their context. Students made a persona based on the scenario description. Next, a theme of sounds was decided based on the persona’s characteristics in a conceptualisation phase. After, they analysed driver-automated vehicle scenarios where sounds would be provided as shown in Figure 4. Next, students designed sounds using a sound design tool. Six groups of twenty-five students created movies, including at least two automation mode transition situations, and three experts evaluated the movies.

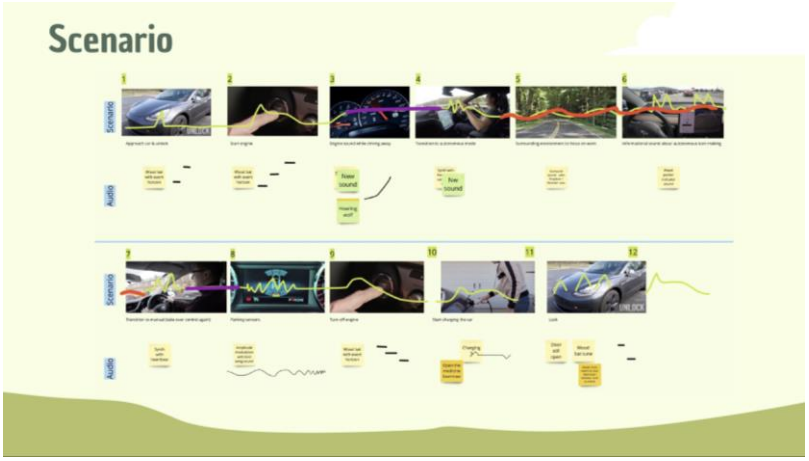


Figure 4. Analyzing interaction scenarios example of one group

The soundscape has been designed to enhance user experience and provide information to a driver with less perceived annoyance. The deliverables were evaluated by two experts in sound design and one expert in UI design. The evaluation was based on whether a sound was appropriate for a context, applicable for interaction, and highly completable. Although the soundscapes were not evaluated in a rigid experimental setting the deliverables showed that soundscape could be used in various ways in highly automated vehicles. For example, it is possible to provide automation system information using a soundscape, such as the external vehicle information (Gang, et al., 2018), lane-keeping or a round-about maneuver of an ego-vehicle (Beattie et al., 2014). In addition, soundscapes may allow drivers to be aware of scheduled take-over situations.

Discussion

Our first findings are somewhat at an informal level. However, the design brought forward the possibility of informing drivers about contextual information and the ego-vehicle actions without having to look at the road. This is one of the strengths of the use of sound in a more advanced way that enables situation awareness of the driver. This may be an essential factor in the increase of safety when a car is in SAE levels 4 and 5. Furthermore, driver-vehicle interaction through soundscapes can provide a new way of designing user experiences to drivers. Further research will be conducted on the impact of soundscape on drivers' trust, situational awareness and safety, as well as user experience.

General Discussion

This study suggested a new approach to sound design and contributed to improving drivers' experience. In Study 1, sounds were designed considering several interaction scenarios that were validated with the purpose of enhancing functionality and user experience. Study 2 showed that soundscape could be used in automated vehicles, contributing to less perceived annoyance, and therefore enhancing the user experience.

However, the limitations of the design approach need to be taken into consideration. First, it is important to note that the outcome is dependent on the capability of the interaction designer. The interaction designer needs to understand the driving context and consider the driving experience in its design. In addition, the interaction designer is required to have adequate technical sound design skills or work in close collaboration with a sound engineer to produce the sounds. Furthermore, a validation phase does require an investment in both time and cost.

In future research, there is a need to evaluate the designed sound's impact on situational awareness, trust, workload as well as user experience in the automated vehicle context.

Acknowledgements

This work has received funding from the European Union’s Horizon 2020 research and innovation program for the project HADRIAN under grant agreement no. 875597. We appreciate the help of Willem, who worked together as a sound engineer in the studies.

References

Aletta, F., Kang, J., & Axelsson, O. (2016). Soundscape descriptors and a conceptual framework for developing predictive soundscape models. *Landscape and Urban Planning*, *149*, 65-74.

American National Standards Institute. (1991). *Warning colors, signs, symbols, labels, and tag standards*. Z535.1-5. National Electrical Manufacturers Association, Washington DC.

Beattie, D., Baillie, L., Halvey, M., & McCall, R. (2014). What's around the corner? Enhancing driver awareness in autonomous vehicles via in-vehicle spatial auditory displays. *Proceedings of the 8th Nordic Conference on Human-Computer Interaction*, (pp. 189-98). New York, NY, USA, Association for Computing Machinery.

Belman D., Monaco K., & Brooks T. (2004). *Sailors of the Concrete Sea: A Portrait of Truck Drivers'*. Work and Live, Michigan State University Press.

Edworthy, J. (1994). The design and implementation of non-verbal auditory warnings. *Applied Ergonomics*, *25* (4), 202-210

Edworthy, J. (1998). Does sound help us to work better with machines?. *Interacting with Computers*, *10*, 401–409.

Forster, Y., Naujoks, F., & Neukum. A. (2017). Increasing anthropomorphism and trust in automated driving functions by adding speech output. *28th IEEE Intelligent Vehicles Symposium* (pp. 365-372). New York, USA, Institute of Electrical and Electronics Engineers,

Gang, N., Sibi, S., Michon, R., Mok, B., Chafe, C., & Ju, W. (2018). Don't Be Alarmed: Sonifying Autonomous Vehicle Perception to Increase Situation Awareness. *Proceedings of the 10th Acm International Conference on Automotive User Interfaces and Interactive Vehicular Application*s (pp.237-46). New York, USA, Institute of Electrical and Electronics Engineers.

Hellier, E.J., Edworthy J., & Dennis, I. (1993). Improving auditory warning design: Quantifying and predicting the effects of different warning parameters on perceived urgency. *Human Factors*, 35(4), 693-706.

Horberry, T., Mulvihill, C., Fitzharris, M., Lawrence, B., Lenne, M.m Kuo, J., & Wood, D. (2022). Human-Centered Design for an In-Vehicle Truck Driver Fatigue and Distraction Warning System. *IEEE Transactions on Intelligent Transportation Systems, 23* (6), 5350-5359.

Ioannis, P., Brewster, S., & Pollick, F. (2015). Language-based multimodal displays for the handover of control in autonomous cars. *In Proceedings of the 7th International Conference on Automotive User Interfaces and Interactive Vehicular Applications* (pp. 3-10). New York, USA, Institute of Electrical and Electronics Engineers.

Jeon, M. (2019). Multimodal displays for take-over in level 3 automated vehicles while playing a game. *In Conference on Human Factors in Computing Systems Proceedings* (pp. 1-6)

Kabbani, T., Kim, S., Serbes, D., Ozan, B., & van Egmond, R., and Hartavia, A. E. (2022). Improved Trucker-Vehicle Dialogue under Critical Scenarios through fluid-HMI. *Transport Research Arena (TRA) 2022*.

König, M., & Neumayr, L. (2017). Users' resistance towards radical innovations: The case of the self-driving car. *Transportation Research Part F: Traffic Psychology and Behaviour*, 44, 42–52.

Larsson, P., & Västfjäll, D. (2013). Emotional and behavioural responses to auditory interfaces in commercial vehicles. *International Journal of Vehicle Noise and Vibration*. 9, 75-95

Nadri, C., Ko, S., Colin, D., Winters, M., & Jeon, M. (2021). Novel Auditory Displays in Highly Automated Vehicles: Sonification Improves Driver Situation Awareness, Perceived Workload, and Overall Experience. *Proceedings of the*

Human Factors and Ergonomics Society Annual Meeting 65 (pp. 586-990). Washington DC, USA, Human Factors & Ergonomics Society.

Petermeijer, S., Doubek, F., & De Winter. J. (2017). Driver response times to auditory, visual, and tactile take-over requests: A simulator study with 101 participants. *IEEE International Conference on Systems, Man, and Cybernetics*, (pp. 1505-1510). New York, USA, Institute of Electrical and Electronics Engineers.

Roche, F., Somieski, A., & Brandenburg, S. (2019). Behavioral Changes to Repeated Takeovers in Highly Automated Driving: Effects of the Takeover-Request Design and the Nondriving-Related Task Modality. *Human Factors*, *61*, 839-849. Washington DC, USA, Human Factors & Ergonomics Society.

SAE International (2018). *Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles*. Warrendale, USA, SAE International.

Schafer, R.M. (1976). *Exploring New Soundscape. Unesco Courier*, 4-8, UNESCO

Siwiak, D., & Jame, F. (2009). Designing interior audio cues for hybrid and electric vehicles. *Audio Engineering Society Conference: 36th International Conference: Automotive Audio*. New York, USA, Audio Engineering Society

Özcan, E., & van Egmond, R. (2008). Product Sound Design: An Inter-Disciplinary Approach?, *Design Research Society Conference 2008*. London, UK, Design Research Society.

Özcan, E., & van Egmond, R. (2012). Basic Semantics of Product Sounds. *International Journal of Design*, 6, 41-54.

Van den Beukel, A.P., Van der Voort, M.C., & Eger, A.O. (2016). Supporting the changing driver's task: Exploration of interface designs for supervision and intervention in automated driving. *Transportation Research Part F: Traffic Psychology and Behaviour*, 43, 279-301.

Ten seconds to go! – Effects of feedback systems in human-robot collaboration

Franziska Legler[1], Dorothea Langer[1], Sebastian Glende[2] & Angelika C. Bullinger[1]
[1]Chemnitz University of Technology, Germany
[2]YOUSE GmbH, Berlin, Germany

Abstract

Human-robot collaboration (HRC) aims to increase efficiency and flexibility in production sites. The implementation in factories is, however, accompanied by risks of physical contact with robots and resulting injuries in case of system failures or workers' misconduct. One assumed reason for such safety-critical behaviour is over-trust in systems' capabilities. The question remains if feedback systems can optimize trust levels and enhance workers' safety and productivity. In the paper, we present a study in the industrial context examining the effects of a user-evaluated feedback system for fenceless HRC based on LED lighting and an information display. In the experiment, 48 participants performed a realistic collaboration task with a heavy-load robot in a pseudo real-world test environment. Dependent variables were assembling time, recognition of system failures and trust in automation. Independent variables were varied: robot feedback between groups, occurrence of system failures during collaboration and time pressure within groups in a balanced design. Results showed that the feedback system did not affect assembling time. Furthermore, system failures were more frequently detected, and (over)trust was reduced if the feedback system was applied. We discuss the potentials of feedback systems for workers' safety enhancement and the development of an appropriate trust level in HRC.

Introduction

Human robot collaboration (HRC) becomes more and more relevant in industry as it is expected to enable the flexibility of increasingly complex production sites (Oubari et al., 2018). Although the number of applications in the manufacturing sector has risen in the last years (Matheson et al., 2019), especially applications of HRC with heavy-load robots remain in niche and pilot studies (Grüling, 2014).

Efficient and safe collaborative work with robots in production sites majorly depends on trust (Freedy et al., 2007). Automation psychology studies the concept of trust in automation which in the subarea of human-robot interaction is specified as *trust in robots* or *human-robot trust* (Hancock, Billings, Schaefer, Chen et al., 2011). It is defined as "the reliance by an agent that actions prejudicial to their well-being will not be undertaken by influential others" (Hancock, Billings & Schaefer, 2011, p.24). In HRC, the robot stands for the influential other.

In D. de Waard, S.H. Fairclough, K.A. Brookhuis, D. Manzey, L. Onnasch, A. Naumann, R. Wiczorek, F. Di Nocera, S. Röttger, and A. Toffetti (Eds.) (2022). Proceedings of the Human Factors and Ergonomics Society Europe Chapter 2022 Annual Conference. ISSN 2333-4959 (online). Available from http://hfes-europe.org

Despite the principal risks involved in human-automation interaction, e.g., resulting from following a false advice of an automated system or getting injured by a robot, people often highly trust automated systems (e.g., Wickens & Xu, 2002; Dzindolet al., 2003; Legler et al., 2020; Manchon et al., 2021). During interactions with reliable systems, trust even rises in course of time (e.g., Manchon et al., 2021; Legler et al., 2020). This very high trust ('positivity bias'; Dzindolet et al., 2003) can result in over-trust in automation, associated with reduced situation awareness (Hancock, Billings, Schaefer, Chen et al., 2011) and reduced monitoring of the system (Hergeth et al., 2016). So, although trust is highly essential for a successful HRI (Aroyo et al., 2021), 'too much' trust potentially arises risks like low product quality or safety-critical behaviour of workers due to unnoticed technical malfunction. Contrary, an appropriate trust level should ensure high situation awareness and system monitoring. The trust calibration approach (e.g., Lee & See, 2004; Hancock, Billings, Schaefer, Chen et al., 2011) implies that also very high trust in automation represents the correct calibrated level of trust in case of highly reliable systems. Still, the absence of system failures is never assured. Therefore, high monitoring should always be targeted and over-trust should be prevented. Recently, over-trust in robots was described as a situation where humans misunderstand the risk of their own actions because they underestimate the probability that a robot performs its functions ineffectively or unsafely (Wagner et al., 2018). This adequately fits research on 'machine heuristics', showing that people trust machine-like, non-anthropomorphic devices and robots as a result of overreliance on technical functioning (Aroyo et al., 2021). Still, research regarding the consequences of 'too much' trust in robots is rare (Aroyo et al., 2021).

Relevant factors in industrial HRC affecting trust are system failures and time pressure. Research on automated systems found that *system failures* reduce trust and safety-critical behaviour after failure occurrence due to higher monitoring and situation awareness (e.g., Wickens et al., 2015) but also heighten the risk of workers rejecting the system. Most studies focussed on the breakdown of decision-aid systems or on providing wrong information - resulting in a reduction of trust (e.g., Dzindolet et al., 2003; Onnasch et al., 2014; Wickens et al., 2015). Trust reduction caused by system failures was also found during interactions with heavy-load robots by simulating a potentially safety-critical technical malfunction (Legler et al., 2020). But overall, research on effects of trust violations in robots' physical functioning (referred to as 'pragmatic trust'; Aroyo et al., 2021) like technical malfunctions is rare. In case of over-trust, failures of robots could even be not perceived at all (Aroyo et al., 2021) resulting from missing monitoring of robots' actions. Overall, system failures show a potential to adjust over-trust to a more appropriate level, but only if users actually perceive the failure.

Also, pacing in industrial settings can cause *time pressure* which was found to increase mental workload during human-automation interactions and this workload is therefore compensated by increased trust. While interacting with automation, time pressure increases subjective workload (Liu, Peterson et al., 2016; Wang et al., 2016). Definitions of trust, equally interpersonal and automation trust, show that trust is used to reduce complexity and to manage uncertainty (Luhmann, 1979). Also, it was argued that working under time pressure increased heuristic information processing (e.g., Rieger & Manzey, 2022). As mentioned before, the 'machine heuristics' is a specific

heuristics that was associated with over-trust. It involves attributions of machine-like objectivity, capability and infallibility to the robot and especially occurs during interactions with non-anthropomorphic robots (Aroyo et al., 2021). As industrial robots are mostly non-anthropomorphic, the heuristics is likely to be applied under time pressure, reducing workload and increasing trust. Therefore, highly reliable automated systems can enhance task performance under time pressure as users trust and rely on the automated system instead of taking incorrect actions resulting from high workload (Rieger & Manzey, 2022). On the other hand, the above-mentioned negative effects of 'too much' trust are probably increased under time pressure, especially in case of system malfunctions. In an experimental study, experts showed a tendency of over-trust (following wrong advices of a system) and automation bias while being under time pressure (van der Waa et al., 2021). Still, research regarding effects of time pressure on trust in industrial robots is missing today.

Visual feedback that gives information about the current state of an automated system, like robots, shows the potential to reduce workload, increase situation awareness and hence, affect trust and safety-critical behaviour. Visual feedback systems transfer information by using shape and colour as cognitive cues (Andersen et al., 2016). Colours carry important information, for example to signal danger and force human attention (Goldstein, 2010). If colour coding of information is corresponding to evolutionary or internalised associations, feedback systems can significantly reduce subjective workload (Blundell et al., 2020). Also, visual feedback is implemented to increase situation awareness of users (Maurtua et al., 2017; Palmarini et al., 2018; Schaefer et al., 2017). Furthermore, a meta-analysis showed that feedback affects trust in automation (Schaefer et al., 2016). In addition, trust mediates the relationship between feedback and actual reliance on the system (Dzindolet et al., 2003). Industrial robots on the shopfloor are mostly stationary within a robot cell and consist of a robot arm with several degrees of freedom and an end effector, e.g., a robotic hand or gripper (Bendel, 2020). Therefore, the robots are able to move from a fixed base or home position towards an intended goal position by manipulating the available degrees of freedom. In case of HRC, workers are situated inside the robot cell without fences and the robot cell is divided into robot zones with autonomous movements and a collaboration-zone (e.g., Bdiwi, Krusche et al., 2017). A typical industrial task with a heavy-load robot contains

- the autonomous grasping and transporting of components by the robot towards the worker,
- a collaborative handover of the component or workers' assembling at a component while the robot acts as a 'third hand' holding the component inside the collaboration-zone (e.g., Bdiwi, Pfeiffer et al., 2017),
- and robot's autonomous component storage or direct return to a home position.

Involvement of users during the design process of a visual feedback system for a fenceless HRC revealed that the following information are important for human collaborators: status information (e.g., current operation of the robot), warnings and explanations of errors, goal position of the robot to enhance the anticipation of subsequent robot actions and mode of operation (like autonomous versus collaborative phase) (Hoecherl et al., 2018). Research with industrial robots showed that workers would also prefer prospective information about future actions (Andersen et al., 2016; Liu, Kinugawa & Kosuge, 2016) and advised the usage of a

countdown for remaining collaboration time (Brending et al., 2016). All these additional information could enhance a smooth interaction during HRC and enable workers to adjust own's assembling speed, resulting in lower assembling times. Contrary, it could increase stress in workers, resulting in assembling hectically or incorrectly. In a study, it was equally found that feedback did not influence assembling time at all (Sadrfaridpour & Wang, 2018). So, effects of feedback on performance remain unclear.

To sum up current research, first, high *trust* is associated with less monitoring and situation awareness and *system failures* reduce trust. Second, *time pressure* results in high workload, reduced situation awareness and a tendency towards over-trust in HRC. Third, *visual feedback systems* can reduce workload and increase situation awareness in HRC. The question arises if feedback in HRC can support an appropriate trust level and enhance workers' safety as well as performance in situations of system failures, especially under time pressure. With industrial robots, feedback about remaining collaboration time could lower workers' assembling time (hypothesis 1). Due to increased situation awareness, feedback is expected to reduce safety-critical behaviour (hypothesis 2) and lower trust (hypothesis 3). Also, system failures should reduce trust (hypothesis 4). Time pressure is supposed to increase trust (hypothesis 5) due to increased workload.

To gain answers to the research question, first an innovative feedback system had to be designed, integrated within the control system of a heavy-load robot, and evaluated by users with regards to usability. It is described in the following chapter. Second, an experiment was conducted in a pseudo real-world test environment, varying the application of the feedback system, the occurrence of system failures and time pressure. Methodology and experimental results are described subsequently and the paper closes with conclusions for industrial applications.

Design and usability of the feedback system

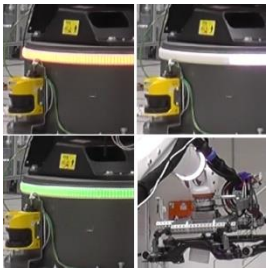
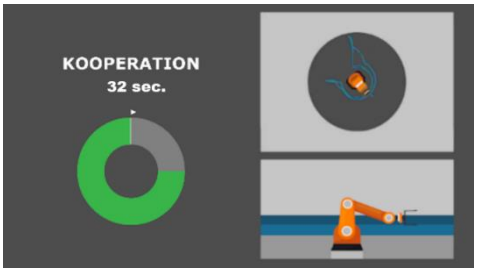
A  B 

Figure 1. A) KUKA robot with implemented LED lighting system showing differing status of base and flange lights and B) feedback display during cooperation mode.

The feedback system was designed in accordance with the human-centred design process (DIN e.V., 2020). First, demands were derived with assembly workers from the automotive industry in several workshop formats. Then visual feedback was designed by iterative prototyping using standards of ISO 9241-110:2020 (International Organization for Standardization, 2020). Finally, the visual feedback

system had an LED lighting system and an information display (see Figure 1), visualising all information highlighted by Hoecherl and colleagues (2018).

CODED INFORMATION	IMPLEMENTATION IN		VISUALISATION
	LED lighting system	Information display	
status information	**type of glow at base** · contant glow = component admission and storage, · oscillating = collaboration, · circumferential = spinning movement of robot	**dynamic topview of robot cell** · showing current position of robot and start- and endpoint of robot path	
goal position of robot	**circumferential lightings** at base showing direction of robot's movements: · approaching or · moving away relative to worker		
warnings for errors	**colour coding** · orange oscillating at base and flange *(see Figure 1A)*	–	
explanation of errors	–	**written message** · "Failure! Gesture control stopped"	
mode of operation	**colour coding** · white = automated, · green oscillating = collaboration and gesture control *(see Figure 1A)*	**colour of torus** · green = during collaboration, · grey = automated	
time information	–	**countdown** · for collaboration time (counting down seconds and fill level of torus) · for remaining time till collaboration in seconds	
safety information	–	**dynamic sideview of robot cell** · showing current height of robot flange (head level filled dark)	

Figure 2. Feedback information implemented in LED lighting system and information display.

The visual feedback system was evaluated regarding usability by 24 participants. They completed a user test with a heavy-load industrial robot with the implemented LED lighting system and information display. Participants performed an industrial assembling task with the fenceless robot, covering several recurrent assembling cycles and lasting around 25 minutes. The Post-Study System Usability Questionnaire (PSSUQ; Lewis, 2002; German translation from Schaub et al., 2012 and Kaminski,

2018; 7-point Likert-scale) was applied separately for LED lighting system and display. Figure 3 compares the scores of subscales to inversed standard values from Sauro and Lewis (2016; high values show better usability). The ratings showed strong accordance with standard values. Interface quality (general liking and pleasantness) was little below average. Overall, the LED lighting system was evaluated better than the display. Additionally, participants used a scale from '0-never' to '4-always' to rate how often they were aware of and used either part of the feedback system (both times $Mdn_{LED} = 3$, $MAD_{LED} = 1.48$; $Mdn_{display} = 1$; $MAD_{display} = 1.48$), showing greater usage of the LED lighting system. Usage of LED lighting system ($\tau = .505$, $p = .002$) and display ($\tau = .330$, $p = .041$) were positively related to usability evaluations.

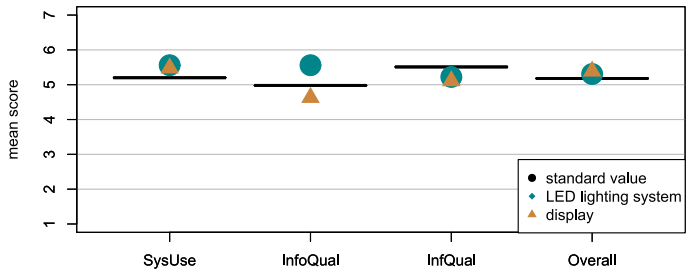


Figure 3. Subscales of PSSUQ for LED lighting system and display compared to standard values (SysUse = System Usability, InfoQual = Information Quality, IntQual = Interface Quality).

Method of experiment

Test environment

An industrial KUKA robot (Quantec prime KR 180), classified as heavy-load robot, was used as a test bed (Figure 4) with implemented feedback system (see Figure 2).

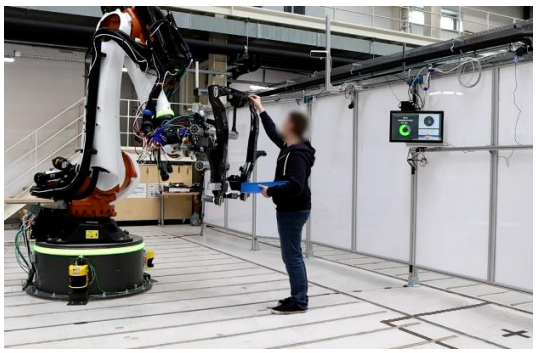


Figure 4. View of test environment with KUKA robot and participant during collaboration.

The robot cell had a collaboration-, a robot- and a safety-zone (see Figure 5A). Participants always remained inside the robot cell during the experiment, waiting inside the safety zone until the collaboration time started (otherwise activating an emergency stop). The speed of the robot was associated with particular zones, moving

at a maximum speed of 2300 mm/s outside the collaboration-zone. The robot slowed down to 500 mm/s when entering the collaboration-zone. The robot was able to support a collaborative assembling task modelled similarly to a real workplace in the automotive industry. Table 1 describes tasks of the human and the robot within a complete assembling cycle as well as provided information by the feedback system.

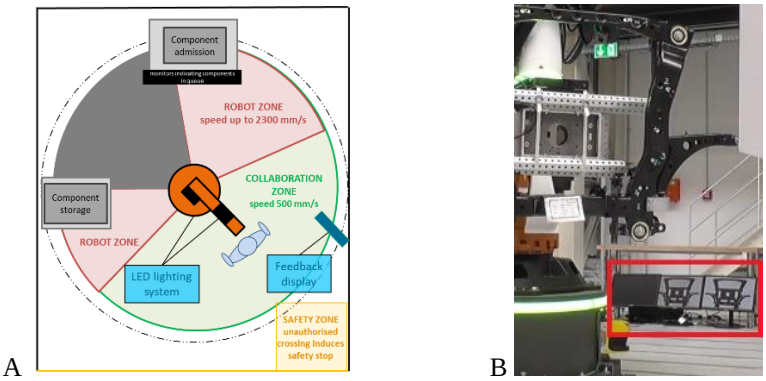


Figure 5. A) Schematic presentation of the robot cell containing different zones and associated robot speeds, B) monitors for simulating an assembling line with two components in the queue and green LED lighting system indicating collaboration phase.

Table 1. Robot and human tasks during one complete assembling cycle of the HRC task

Robot Task	*Human Task*	*Feedback system*
Phase 1: Component transportation		
admission of component (front axle carrier), transport towards and stop inside collaboration-zone	-	mode of operation (autonomous), current and future position of robot, countdown till collaboration
Phase 2: Collaboration		
support human's input by gesture control (speed 250 mm/s) holding component during assembling	(1) entering collaboration-zone (2) gesture control: setting an ergonomic height of the component located at the robot flange (3) assembling of eight hook-and- pile tapes on the component (4) entering safety-zone	mode of operation (collaborative), countdown till collaboration ends, warning and explanation in case of error
Phase 3: Component storage		
transport and storage of component, returning to component admission	-	mode of operation (autonomous), current and future position of robot

For the height setting, the robot arm reacted to up-/downward movements of the human's palm accordingly with 250 mm/s and with minimal physical distance to the

human (HRC level 3 according to Bdiwi, Pfeiffer et al., 2017). Although functional, this gesture control could be manually controlled by experimenters. Participants were not able to notice the manual control (wizard-of-oz), allowing experimenters to simulate a failure of gesture control. For the simulation of an assembling line, the test environment further had three monitors located behind the robot (see Figure 5B). If a monitor was turned on via wireless remote control, a picture of a front axle carrier (component) appeared. Experimenters were able to turn off a monitor to simulate component admission by the robot and turn on a monitor to simulate a new component in the queue. During the complete collaborative assembling task, all information described in Figure 2 were available to participants simultaneously via the LED lighting system and the feedback display.

Experimental design

A 2(feedback) x2(system failure) x2(time pressure) mixed design was conducted. ***'Feedback'*** was a between factor (control vs. feedback group), therefore, in the control group the LED lighting system and information display remained turned off. Within both groups, a balanced design of two within-factors ***'system failure'*** (none vs. occurrence of failures) and ***'time pressure'*** (none vs. time pressure) was applied. ***System failures*** were simulated in gesture control during the height setting task. For about 10 seconds the robot saliently jerked in the opposite direction of participants' palm movement. Afterward, gesture control was correctly supported. ***Time pressure*** was realised by adding a component on the assembling line monitors (see Figure 5B) after the run-off of an individual cycle time. Therefore, assembling times were measured during participants' baseline cycles and the fastest assembling time was used as individual reference for each participant. The constant robot period and individual reference were summed and used as individual cycle time. In time pressure conditions, individual cycle time only had half of the individual reference assembling time.

Measures

Demographic information like sex, age, as well as experience with industrial robots and production work, were captured in pre-survey. Additionally, Affinity for Technology Interaction (ATI; Franke et al., 2018; 6-point Likert-scale; $\alpha = .85$) was assessed for sample specification.

Further, several measures for dependent variables were applied. ***Assembling time*** was read out from log file data of the robotic system (time from finishing gesture control until robot started moving again). ***Safety-critical behaviour*** was operationalised as disregard of system failures. Without failures, times of gesture control across all scenarios showed a 5% percentile of 0.5 seconds. Hence, in failure scenarios a minimum gesture control time of 10.5 seconds should result (10 seconds failure simulation plus 0.5 seconds for actual height adjustment). Due to manual control of failures and according to expectable inaccuracies of failure time, a tolerance period of 2 seconds was set. So, in failure scenarios a time less than 8.5 seconds between the start of collaboration mode and starting assembling mode was defined as disregard of failures. De facto, participants did not finish the height setting task and started assembling during robot movement which was not allowed by participants'

instruction. ***Trust in automation*** was measured subjectively via a German translation (Pöhler et al., 2016) of Jian-Scale, after each of the five scenarios. Of its two subscales, only subscale trust (6 items; 7-point Likert scale) was used because both subscales were highly correlated and previous work suggested a two-factor structure (Pöhler et al., 2016; Legler et al., 2020). Across the scenarios, mean reliability for trust was $\alpha = .82$.

Sample

In the experiment, 48 subjects participated and were randomly assigned to two groups of equal size: control group and feedback group. Both groups were gender balanced. Participants' mean age was 26.2 years ($SD = 7.54$) in the control and 25.6 years ($SD = 7.84$) in the feedback group. Overall, affinity for technology interaction was on medium level and not differing between groups ($M_{control} = 3.79$, $M_{feedback} = 3.88$, $Z = -0.454$, $p = .650$). Slightly more participants of the feedback group (29%) had interacted with an industrial robot before (vs. 20%) and were currently or had ever worked in production sector before (29% vs. 25%). Participants received financial compensation.

Procedures

Participants got the participant information, signed a declaration of consent and filled in the pre-survey. Subsequently, participants watched two videos showing the real workplace with a handling device and an equivalent task with the robot in the test environment. Afterwards, participants were instructed about the collaborative task consisting of height setting via gesture control and assembling task (see Table 1). Participants within the feedback group additionally received a short introduction to the feedback system. All participants were instructed to

- stay within the marked safety-zone before the robot stops inside the collaboration-zone; leaving the zone would result in an emergency stop,
- use gesture control for height setting in each assembling cycle,
- stop assembling and enter the safety-zone as soon as they notice events seeming abnormal or critical
- and seek low assembling time, as compensation would depend on it.

All participants performed a baseline condition with a minimum of ten assembling cycles (see Table 1) to become familiar with the assembling task, learn gesture control and get to know the feedback system in the feedback group. After, each participant performed four interactions with the robot in randomised order. Each experimental scenario had five assembling cycles that together lasted for around four minutes. For all participants, system failures were simulated in the first and third or second and forth assembling cycle within the respective experimental scenario. Each scenario was followed by a short post-scenario survey to measure outcomes. After, participants received compensation regardless of assembling time. Overall, an experiment lasted around 60 minutes.

Data Analysis

Statistic Software R (R Core Team, 2018) was used for data analysis. Due to the nonsymmetric distribution of data, mainly nonparametric data analysis was applied. If not specified otherwise, independent or dependent Wilcoxon Signed-Rank Test was used for mean comparison and the Friedman Test for analysing variances across conditions. Nonparametric effect size *r* was calculated according to Tomczak and Tomczak (2014). All data showed successful manipulation of independent variables.

Results and discussion

This section describes and discusses the experimental results grouped by dependent variables, followed by study limitations.

Assembling time

Neither feedback nor time pressure influenced assembling time. It remained constant between different scenarios, both for the control group ($X^2(3) = 4.42$, $p = .220$) and the feedback group ($X^2(3) = 6.30$, $p = .098$) (Figure 6), showing no main effect of the independent variables *system failure* or *time pressure* for assembling time. Mean assembling time was ~20 seconds for control group and ~21 seconds for feedback group, showing no significant difference ($Z \geq -0.87$, $p \geq .386$, $r \leq .125$). The result is in line with Sadrfaridpour and Wang (2018) and opposes hypothesis 1. Still, remaining collaboration time was only shown at the information display. As the evaluation of the feedback system has shown that participants rarely used the information display compared to the LED lighting system, a direct display of remaining collaboration time at the robot flange could have enhanced performance. This should be considered in further research and industrial applications.

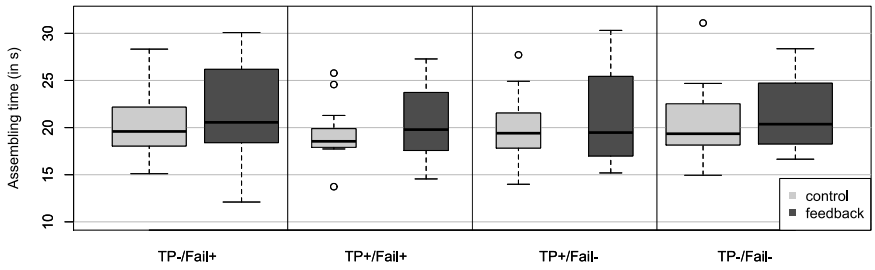


Figure 6. Assembling time dependent on scenario and group. Narrow bars indicate time pressure (TP+), wide bars no time pressure (TP-). F+ indicates scenarios with failure, F- without failure.

Additionally, the higher variance of assembling time in the feedback group cannot be explained by sample composition and should be replicated.

Safety-critical behaviour

Duration for gesture control during assembling cycles with system failures was higher for the feedback (M = 9.42, SD = 4.77) than control group (M = 12.68, SD = 3.05), resulting in a significant difference between groups (Z = -3.73, p < .001, r = .305). Additionally, Figure 7A shows that in the feedback group, durations hardly fell below the defined cut-off value for the 'disregard of system failures' (set to 8.5 seconds). Across all assembling cycles with system failures, 47% of occurring system failures were disregarded by the control group while only 5% were disregarded by the feedback group (Z = -4.28, p < .001, r = .495) (see Figure 7B).

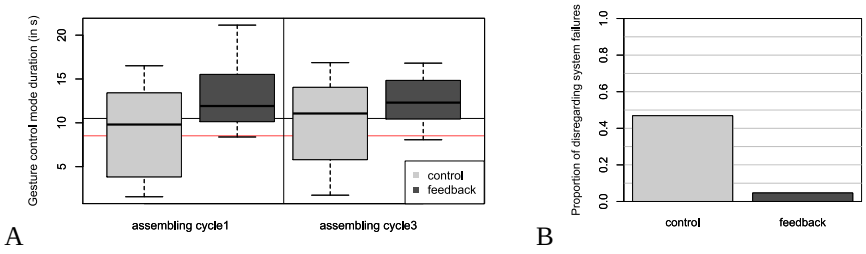


Figure 7. A) Duration of gesture control mode during assembling cycles with failures dependent on group, black line shows expected minimum duration, red line shows cut-off value for operationalisation of disregarding system failures, B) proportion of disregarding system failures across the experiment dependent on group.

In line with hypothesis 2, the visual feedback system reduced safety-critical behaviour during HRC. The evaluation of the feedback system has shown that users frequently pay attention to the LED lighting system but neglect the information display. Therefore, increased perception and correct interpretation of system failures are attributable to the lighting system. Additionally, its information quality was rated high. Hence, applying a colour coded lighting system within the attention zone of workers could enhance safety during industrial HRC.

Trust in automation

Trust was significantly decreased in the feedback compared to the control group, except for scenario TP-/Fail- (see Figure 8) which operationalizes a 'normal' assembling condition without abnormalities. This is in line with hypothesis 3 (feedback lowers trust).

After completing the baseline condition, trust was similar for both groups ($M_{control}$ = 4.92, $M_{feedback}$ = 4.74, Z = -0.95, p = .341, r = .137). Within the control group, none of the four experimental conditions significantly differed from each other ($X^2(3)$ = 5.95, p = .114). In contrast, in the feedback group trust differed across scenarios ($X^2(3)$ = 17.88, p < .001). Lowest trust values were shown during scenario TP-/Fail+ (M = 3.72, SD = 0.90) which was significantly lower than all other scenarios (Z > -2.03, p < .042, r > .415). Because of missing nonparametric test, ANOVA with repeated measures was calculated. It resulted in a significant main effect for system failure ($F(1)$ = 8.80, p < .001, η_P^2 = 0.09), a nonsignificant main effect for time

pressure ($F(1) = 3.59$, $p = .061$, $\eta_P^2 = 0.04$) and a nonsignificant interaction effect ($F(1) = 0.520$, $p = .473$, $\eta_P^2 = 0.01$). These results were in line with hypothesis 4 (system failures reduce trust) but did not support hypothesis 5 (time pressure increases trust).

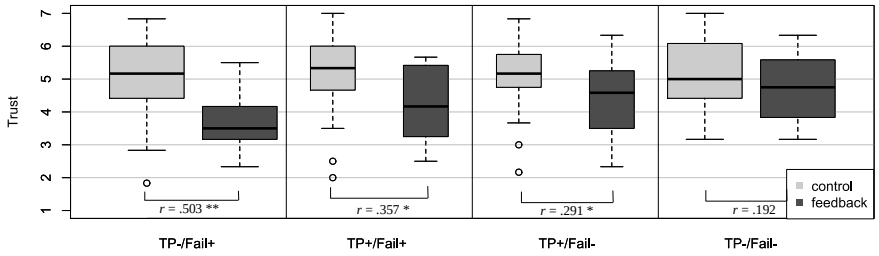


*Figure 8. Trust dependent on scenario and group. Narrow bars indicate time pressure (TP+), wide bars no time pressure (TP-). F+ indicates scenarios with failure, F- without failure. * p < .05, ** p < .01*

In the control group, mean trust was high across scenarios while trust was reduced in the feedback group after system failure occurrence. Concluding, along with the result that participants in the feedback group significantly less disregarded system failures, it can be assumed that the recognition of system failures due to feedback reduced trust and safety-critical behaviour. Trust remained at a medium level and participants did not show signs of discomfort. Thus, it can be assumed that the trust level in the control group was higher than necessary for a safe HRC, indicating a tendency towards over-trust in the robot. Also assembling time remained equally constant in the feedback group. So, negative effects due to fear or distress after failures were not seen, showing potential for intentional integration/simulation of 'safe' failures to keep trust within an appropriate level and attention focussed on the robot system without increasing assembling time. This is in line with HRI research suggesting to familiarise users with potential failures of robots (Wagner et al., 2018) and apply 'deceptive practices' of robots (Aroyo et al., 2021) to prevent over-trust. In conclusion, results point towards the occurrence of over-trust in robots. Nevertheless, today there are no specified criteria to determine how much is 'too much' trust. While in this study participants had no parallel second assembling task during robot tasks, this is not probable in production sites as it would lower production efficiency. To ensure that participants also notice system failures during parallel tasks, simple haptical devices like vibrating straps (Scheggi et al., 2014) could force attention in critical situations and visual distraction.

Study limitations

The defined cut-off value for the disregard of system failures has a huge impact on the result regarding safety-critical behaviour. Still, the difference between groups for disregarding system failures was large enough for the effect to remain when deducting the defined tolerance. Each scenario consisted of five assembling cycles that together lasted for around 4 to 5 minutes while trust develops over longer time periods. In the control group, participants could have missed failures or judged them as non-critical, and later disregarded them. Behavioural observation during the experiment and

experiencing correct gesture control during baseline supported the latter explanation. As a result of post-scenario measurement and laboratory setting, it could be assumed that participants were aware of experimental variations and expected some sort of manipulation. Flook et al. (2019) summarized that ecological validity, especially in case of error simulation, is low in laboratory settings as participants perceive the setting as artificially, controlled and therefore safe. It can still be assumed that workers also believe workplaces to be safe due to occupational safety examinations prior to workplace release. Nevertheless, long-term effects of system failures on trust and possible habituation effects to 'safe' system failures cannot be implied from this study.

Conclusion

In this scenario-based study, effects of a feedback system on operators' trust, performance operationalised by assembling time and safety-critical behaviour were examined. A heavy-load robot and an industry-oriented assembling task served as a test environment. The visual feedback system consisted of a LED lighting system on the robot and an information display. Time pressure and the occurrence of system failures were varied within groups. Assembling time was not influenced by experimental variations. Time pressure did not have a significant effect on trust or safety-critical behaviour. In contrast, the combination of system failures and feedback significantly reduced trust to a still tolerable level while not causing distress on participants but significantly increasing proper reactions to failure events. Without feedback, trust remained on a high level even after system failures, indicating a lack of awareness which at least could enhance reaction times in safety-critical situations. Especially the LED lighting system was often used by participants as a source of information, indicating the potential of a simple colour-coded feedback to calibrate trust and reduce safety-critical behaviour in industrial HRC with heavy-load robots.

Acknowledgements

This research took place within the scope of project "3DIMiR" (project number 03ZZ0459D) supported by German Federal Ministry of Education and Research. The authors acknowledge the financial support. We thank Mohamad Bdiwi, Lena Winkler and Shuxiao Hou from Fraunhofer IWU for the realisation of the test environment. Additionally, we thank design:lab weimar GmbH for the development of concept ideas for the feedback system.

References

Andersen, R.S., Madsen, O., Moeslund, T.B., & Amor, H.B. (2016). Projecting Robot Intentions into Human Environments. In IEEE (Ed.), *RO-MAN 2016 - Proceedings of the 25th IEEE International Symposium on Robot and Human Interactive Communication* (pp. 294-301). Institute of Electrical and Electronics Engineers. https://doi.org/10.1109/ROMAN.2016.7745145

Aroyo, A.M., de Bruyne, J., Dheu, O., Fosch-Villaronga, E., Gudkov, A., Hoch, H., Jones, S., Lutz, C., Sætra, H., Solberg, M., & Tamò-Larrieux, A. (2021). *Overtrusting robots: Setting a research agenda to mitigate overtrust in automation*. *Paladyn, Journal of Behavioral Robotics*, *12*(1), pp. 423-436. https://doi.org/10.1515/pjbr-2021-0029

Bdiwi, M., Krusche, S., & Putz, M. (2017). Zone-Based Robot Control for Safe and Efficient Interaction between Human and Industrial Robots. In IEEE (Ed.), *HRI '17: Proceedings of the Companion of the 2017 ACM/IEEE International Conference on Human-Robot Interaction* (pp. 83-84*)*. Association for Computing Machinery. https://doi.org/10.1145/3029798.3038413

Bdiwi, M., Pfeifer, M., & Sterzing, A. (2017). A new strategy for ensuring human safety during various levels of interaction with industrial robots. *CIRP Annals, 66*, 453-456. doi:10.1016/j.cirp.2017.04.009

Bendel, O. (2020). Die Maschine an meiner Seite. In H.J. Buxbaum (ed.), *Mensch-Roboter-Kollaboration*. Springer Gabler, Wiesbaden. https://doi.org/10.1007/978-3-658-28307-0_1

Blundell, J., Scott, S., Harris, D., Huddlestone, J., & Richards, D. (2020). Workload benefits of colour coded head-up flight symbology during high workload flight. *Displays, 65,* Article 101973. https://doi.org/10.1016/j.displa.2020.101973

Brending, S., Khan, A. M., Lawo, M., Müller, M., & Zeising, P. (2016). Reducing anxiety while interacting with industrial robots. In IEEE (Ed.), *ISWC '16: Proceedings of the 2016 ACM International Symposium on Wearable Computers* (pp. 54-55). Association for Computing Machinery. https://doi.org/10.1145/2971763. 2971780

DIN e.V. (2020). Ergonomics of human-system interaction - Part 210: Human-centred design for interactive systems, German version. EN ISO 9241-210:2019. https://dx.doi.org/10.31030/3104744

Dzindolet, M. T., Peterson, S. A., Pomranky, R. A., Pierce, L. G., & Beck, H. P. (2003). The role of trust in automation reliance. *International Journal of Human–Computer Studies, 58*, 697–718. https://doi.org/10.1016/S1071-5819(03)00038-7

Flook, R. Shrinah, A., Wijnen, L., Eder, K., Melhuish, C., & Lemaignan, S. (2019). On the impact of different types of errors on trust in human-robot interaction: Are laboratory-based HRI experiments trustworthy? *Interaction Studies, 20*, 455–486. https://doi.org/10.1075/is.18067.flo

Franke, T., Attig, C., & Wessel, D. (2019). A Personal Resource for Technology Interaction: Development and Validation of the Affinity for Technology Interaction (ATI) Scale. *International Journal of Human-Computer Interaction, 35*, 456-467.https://doi.org/10.1080/10447318.2018. 1456150

Freedy, A., DeVisser, E., Weltman G., & Coeyman, N. (2007). Measurement of trust in human-robot collaboration. In IEEE (Ed.), *2007 International Symposium on Collaborative Technologies and Systems* (pp. 106-114). Institute of Electrical and Electronics Engineers. https://doi.org/ 10.1109/CTS.2007.4621745

Grüling, B. (2014, June 15). *Neue Fertigungsstraßen im Autobau: Mein Kollege, der Roboter*. Spiegel Wissenschaft. Retrieved from https://www.spiegel.de/wissenschaft/technik/roboter-sollen-menschen-an-fertigungsstrassen-arbeit-abnehmen-a-974088.html

Goldstein, E.B. (2010). *Sensation and perception* (8th ed.). Wadsworth Cengage Learning.

Hancock, P.A., Billings, D.R., & Schaefer, K.E. (2011). Can You Trust Your Robot? *Ergonomics in Design*, 19(3), 24–29.

https://doi.org/10.1177/1064804611415045

Hancock, P.A., Billings, D.R., Schaefer, K.E., Chen, J.Y.C., de Visser, E.J., & Parasuraman, R. (2011). A meta-analysis of factors affecting trust in human-robot interaction. *Human Factors, 53*, 517–527. https://doi.org/10.1177/0018720811417254

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

Hoecherl, J., Schmargendorf, M., Wrede, B., & Schlegl, T. (2018). User-Centered Design of Multimodal Robot Feedback for Cobots of Human-RobotWorking Cells in Industrial Production Contexts. In VDMA, VDE e. V., VDE ITG and IFR (Eds.), *ISR 2018: 50th International Symposium on Robotics* (pp. 1-8). VDE Verlag. https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8470632

International Organization for Standardization (2020). *ISO 9241-110:2020 Ergonomics of human-system interaction — Part 110: Interaction principles*. ISO.

Kaminski, C.P. (2018). *Gebrauchstauglichkeitsanalyse zur Qualitätssicherung im medizinischen Kontext* [Doctoral dissertation, Eberhard-Karls-Universität Tübingen]. https://publikationen.uni-tuebingen.de/xmlui/handle/10900/80733

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

Legler, F., Langer, D., Dittrich, F., & Bullinger, A.C. (2020). I don't care what the robot does! Trust in automation when working with a heavy - load robot. In D. de Waard, A. Toffetti, L. Pietrantoni, T. Franke, J - F. Petiot, C. Dumas, A. Botzer, L. Onnasch, I. Milleville & F. Mars (Eds.), *Proceedings of the Human Factors and Ergonomics Society Europe Chapter 2019 Annual Conference: Understanding Human Behaviour in Complex Systems* (pp. 239-253). Human Factors and Ergonomics Society. https://www.hfes-europe.org/largefiles/proceedingshfeseurope2019.pdf

Lewis, J. R. (2002). Psychometric Evaluation of the PSSUQ Using Data from Five Years of Usability Studies. *International Journal of Human-Computer Interaction, 14*, 463–488. https://doi.org/10.1080/10447318.2002.96691

Liu, D., Kinugawa, J., & Kosuge, K. (2016). A projection-based making-human-feel-safe system for human-robot cooperation. In IEEE (Ed.), *2016 IEEE International Conference on Mechatronics and Automation* (pp. 1101–1106). Institute of Electrical and Electronics Engineers. https://doi.org/10.1109/ICMA.2016. 7558716

Liu, D., Peterson, T., Vincenzi, D., & Doherty, S. (2016). Effect of time pressure and target uncertainty on human operator performance and workload for autonomous unmanned aerial system. *International Journal of Industrial Ergonomics, 51*, 52–58. https://doi.org/10.1016/j.ergon.2015.01.010

Luhmann, N. (1979). *Trust and Power*. Wiley.

Manchon, J.B., Bueno, M., & Navarro, J. (2021). Calibration of Trust in Automated Driving: A Matter of Initial Level of Trust and Automation Driving Style? [Unpublished manuscript]. VEDECOM Institute. https://psyarxiv.com/bpna2/

Matheson, E., Minto, R. Zampieri E.G.G., Faccio, M., & Rosati, G. (2019). Human–Robot Collaboration in Manufacturing Applications: A Review. *Robotics, 8*(4), 100. https://doi.org/10.3390/robotics8040100

Maurtua, I., Ibarguren, A., Kildal, J., Susperregi, L., & Sierra, B. (2017). Human–robot collaboration in industrial applications: Safety, interaction and trust. *International Journal of Advanced Robotic Systems, 14*(4), 1-10. https://doi.org/10.1177/1729881417716010

Onnasch, L., Wickens, C.D., Li, H., & Manzey, D. (2014). Human Performance Consequences of Stages and Levels of Automation: An Integrated Meta-Analysis. *Human Factors, 56*, 476–488. https://doi.org/10.1177/0018720813501549

Oubari, A., Pischke, D., Jenny, M., Meißner, A., & Trübswetter, A. (2018). Mensch-Roboter-Kollaboration in der Produktion: Motivation und Einstellungen von Entscheidungsträgern in produzierenden Unternehmen. *ZWF Zeitschrift für wirtschaftlichen Fabrikbetrieb, 113*, 560–564. https://doi.org/10.3139/104.111971

Palmarini, R., Fernandez del Amo, I., Bertolino, G., Dini, G., Erkoyuncu, J. A., Roy, R., & Farnsworth, M. (2018). Designing an AR interface to improve trust in Human-Robots collaboration, *Procedia CIRP, 70*, 350-355. https://doi.org/10.1016/j.procir.2018.01.009

Pöhler, G., Heine, T., & Deml, B. (2016). Itemanalyse und Faktorstruktur eines Fragebogens zur Messung von Vertrauen im Umgang mit automatischen Systemen. *Zeitschrift Für Arbeitswissenschaft, 70*(3), 151–160. https://doi.org/10.1007/s41449-016-0024-9

R Core Team (2019). *R: A language and environment for statistical computing* (3.6.2) [Computer Software]. R Foundation for Statistical Computing https://www.R-project.org/

Rieger, T., & Manzey, D. (2022). Human Performance Consequences of Automated Decision Aids: The Impact of Time Pressure. *Human Factors, 64*, 617–634. https://doi.org/10.1177/0018720820965019

Sadrfaridpour B., & Wang Y. (2018). Collaborative Assembly in Hybrid Manufacturing Cells: An Integrated Framework for Human–Robot Interaction. *IEEE Transactions on Automation Science and Engineering, 15*, 1178-1192. https://doi.org/10.1109/TASE.2017.2748386

Sauro, J., & Lewis, J. (2016). *Quantifying the User Experience Practical Statistics for User Research*. Elsevier.

Schaefer, K. Chen, J., Szalma, J., & Hancock, P. (2016). A Meta-Analysis of Factors Influencing the Development of Trust in Automation: Implications for Understanding Autonomy in Future Systems. *Human Factors, 58*, 377-400. https://doi.org/10.1177/0018720816634228

Schaefer, K., Straub, E.R., Chen, J.Y.C., Putney, J., & Evans, E.W. (2017). Communicating intent to develop shared situation awareness and engender trust in human-agent teams. *Cognitive Systems Research, 46*, 26-39. https://doi.org/10.1016/j.cogsys.2017.02.002.

Schaub, F., Deyhle, R., & Weber, M. (2012). Password entry usability and shoulder surfing susceptibility on different smartphone platforms. In ACM (Ed.), *MUM '12: Proceedings of the 11th International Conference on Mobile and*

Ubiquitous Multimedia (pp. 1-10). Association for Computing Machinery. https://doi.org/10.1145/2406367.2406384

Scheggi, S., Morbidi, F., & Prattichizzo, D. (2014). Human-Robot Formation Control via Visual and Vibrotactile Haptic Feedback. *IEEE Transactions on Haptics*, *7*, 499-511. https://doi.org/10.1109/TOH.2014.2332173

Tomczak, M.T., & Tomczak, E. (2014). The need to report effect size estimates revisited. An overview of some recommended measures of effect size. *Trends in Sport Sciences, 21*, 19–25. http://www.tss.awf.poznan.pl/files/3_Trends_Vol21_2014__no1_20.pdf

van der Waa, J., Verdult, S., van den Bosch, K., van Diggelen, J., Haije, T., van der Stigchel, B., & Cocu, I. (2021). Moral Decision Making in Human-Agent Teams: Human Control and the Role of Explanations. *Frontiers in Robotics and AI*, *8*, 640-647. https://doi.org/10.3389/frobt.2021.640647

Wagner, A.R., Borenstein, J., & Howard, A. (2018). Overtrust in the robotic age. *Communications of the ACM*, 61 (9), 22–24. https://doi.org/10.1145/3241365

Wang, L., He, X., & Chen, Y. (2016). Quantitative relationship model between workload and time pressure under different flight operation tasks. *International Journal of Industrial Ergonomics, 54*, 93–102. http://dx.doi.org/10.1016/j.ergon.2016.05.008

Wickens, C. D., Clegg, B. A., Vieane, A. Z., & Sebok, A. L. (2015). Complacency and Automation Bias in the Use of Imperfect Automation. *Human Factors, 57*, 728–739. https://doi.org/10.1177/0018720815581940

Wickens, C.D., & Xu, X. (2002). *Automation trust, reliability and attention* (Tech. Rep. AHFD-0214/MAAD-02-2). Savoy: University of Illinois, Aviation Research Lab.

Non-technical skills in firefighting – development, implementation, and evaluation of a team training for enhancing safety critical performance

Lena Heinemann[1], Fabienne Aust[2], Maik Holtz[3], Corinna Peifer[2], & Vera Hagemann[1]
[1]University of Bremen
[2]University of Lübeck
[3]Cologne Fire Department, Institute for Security Science and Rescue Technology Germany

Abstract

During firefighting operations, critical situations and accidents caused by human errors or poor teamwork occur repeatedly. For making operations safer and less stressful for firefighters, a target group specific team training based on scientific standards was developed. The team training is specifically adapted to the target group of trainees in the fire service and is divided into five modules: communication, (shared) situational awareness & shared mental models, cooperation/support, decision-making, and leadership. Team skills are trained through practical exercises and case studies. The objective of this study was to evaluate the team training regarding its effectiveness for the improvement of non-technical skills, exemplified in this study by communication and situational awareness, as well as the participants' assessment of the training outcomes and design. A non-technical skills rating system was developed and applied in all groups to assess the behaviour, as well as self-report questionnaires. A pre-post-control group-design was used with trainees in the fire service (n = 97). Mixed ANOVAs showed no significant effects for situational awareness and communication. Descriptive results partly supported the positive development of situational awareness and communication from baseline to post in the training group compared to the control group. In addition, results indicate that the training is perceived as useful and understandable by the participants.

Introduction

In 2005, Tübingen, Germany was the scene of a devastating fire killing two members of the fire service. During this operation, the attack squad went into the attic of a burning building without informing the incident commander, breathing protection monitoring, and other team members. In the further course, a *Mayday situation* occurred because the attack squad's air supply of the breathing apparatus was exhausted. However, the support came too late, also due to the lack of information about the exact location of the attack squad (Unfallkommission "Tübingen", 2006). The attack squad in this operation did not communicate according to the standards. This meant that the rest of the team could not understand how they were proceeding and where they were in the building.

In D. de Waard, S.H. Fairclough, K.A. Brookhuis, D. Manzey, L. Onnasch, A. Naumann, R. Wiczorek, F. Di Nocera, S. Röttger, and A. Toffetti (Eds.) (2022). Proceedings of the Human Factors and Ergonomics Society Europe Chapter 2022 Annual Conference. ISSN 2333-4959 (online). Available from http://hfes-europe.org

During a firefighting operation in 2016 in the United States, the first attack squad went forward to locate, contain and extinguish a spreading fire in a residential building. A second attack squad laid a hose line to the other side of the building at the same time. While doing so, they observed a large amount of fire and smoke coming from inside the building. The second attack squad informed the incident command of the situation. It was decided that the second attack squad would direct water into the building from their position to fight the fire. Several attempts were made to contact the first attack squad, but without success. The second attack squad began adding water anyway, without making sure that the first attack squad was informed of the plan and was in a protected area. Due to the heat and water vapour generated by the second attack squad's water, the first attack squad eventually had to withdraw (Fire Near Miss, 2016). The second attack squad should have made sure that the first attack squad was out of the building or at least in a protected area before they released their water. In addition, the incident commander failed to ensure that all relevant persons were aware of the plan before implementing such a change of operational tactics.

There were no technical difficulties in either operation, but the dangerous development was due to human error. The fatal consequences in the first example and the danger to the first attack squad in the second example might have been avoided if the teamwork had worked better. Numerous studies have shown that the quality of teamwork positively influences the work performance (for an overview, see e.g. Rosen et al., 2018).

High Responsibility Teamwork

A team is defined as two or more people who, among other things, are jointly responsible for certain results, have common goals and work in mutual dependence. The members often have specialized roles and responsibilities and together they solve tasks that they could not do alone (Cannon-Bowers & Bowers, 2011, Sundstrom et al., 1990).

Baker et al. (2006) described how teamwork is critical for the delivery of health care and the same applies for the fire service. Firefighters in firefighting operations have the common goals to fight the fire, potentially rescue people and minimise the property damage. For that each team member or squad has specialized responsibilities, e.g. setting up the water supply as water squad, which are interdependently connected. The attack squad cannot fight the fire if the water squad hasn't provided for adequate water supply. The team members must coordinate their actions so that they can reach their common goal safely. The basis for being able to carry out these actions is taskwork. This means that the team members must have the position-related knowledge of specific tasks to be done and be familiar with the tools, equipment and procedures (Morgan et al., 1993). However, taskwork alone is not enough to meet the requirements of a firefighting operation; teamwork is also necessary (Cannon-Bowers & Salas, 1998). For teamwork, it is important that team members are able to anticipate each other's needs, adapt to each other's actions and have a common understanding of how a process should proceed (e.g. how to proceed with human rescue) (Baker et al., 2006).

Teams from medicine, police or, as in this study, the fire service, are named High Responsibility Teams (HRTs; Weick & Sutcilff, 2003). Their working environment is characterised by a high potential of risk and consequently the team members have to act in a very reliable manner. HRTs are exposed to physical and psychological stress at times during their work, bear responsibility for the lives of others and/or themselves, and often experience external pressure from the public or the media. In addition, activities they perform are usually not reversible and an interruption of the work situation in the form of a break or a short pause is usually not possible. HRTs also often work in changing team compositions and sometimes with people they do not know (Hagemann, 2011). Despite these dynamic conditions, in which teams have to act quickly and safely, smoothly functioning teamwork processes are particularly important for successful cooperation (Badke-Schaub, 2012), because a mistake or a misunderstood agreement can cause a lot of damage (Kluge et al., 2009). Marks et al. (2001) divide the processes of teamwork into transition and action phases. These run sequentially and can be further divided into single processes. Transition phases involve processes like the specification of goals, while processes such as monitoring progress towards the defined goals take place in action phases. Non-technical skills help perform those processes successfully. Non-technical skills are, for example, communication, coordination, decision-making, leadership, and development of a shared mental model (Branlat et al., 2009; Cannon-Bowers et al., 1995; Omodei et al., 2005; Salas et al., 2005).

In this study we focus on communication, which is defined as "the process by which information is clearly and accurately exchanged between two or more team members" (Cannon-Bowers et al., 1995, p. 345). Communication is positively associated with performance in problem-solving tasks (Nieva et al., 1985) and is essential for workplace efficiency and safety (Flin et al., 2008). In the medical context about half of communication-related errors are preventable (Leape et al., 1993). Furthermore, in the aviation sector, many accidents are partly due to communication problems (Molesworth & Estival, 2015). Second, situational awareness (SA) is studied as another non-technical skill. It consists of three components: perception of elements in current situation, comprehension of current situation and projection of future (Endsley, 1995a). Poor quality of situational awareness is the main cause of accidents when it comes to human error in aviation (Endsley, 1995b). The overview of the situation on site is particularly important for a successful operation (Wilke, 2006).

In order to lay the foundation for successful teamwork, specific team training is needed (Cannon-Bowers & Salas, 1998; Flin et al., 2002). Training of the non-technical skills described is expected to enhance teamwork processes so that the skills acquired can be applied in stressful or new situations and the team can work effectively (Cannon-Bowers et al., 1993). Crew Resource Management (CRM) is the approach used as the basis for much of this type of training today. It is a training approach that originated in aviation and represents a milestone in the development of trainings as it focuses on teamwork competencies that help teams overcome challenges that they would not be able to overcome individually (Salas et al., 1999b). Meanwhile, CRM trainings or similar approaches are especially applied in medicine. There has been considerable research on how to adapt concepts from aviation to the needs of people working in medicine (e.g. Bohmann et al., 2021; Hagemann et al.,

2015; Paige et al., 2009). The Anesthesia Crisis Resource Management (ACRM) has been the main tool in this context (e.g. Gaba et al., 2001). In the fire service, CRM approaches, referred to as Team Resource Management (TRM) trainings, are rarely used. And research on how existing concepts can be adapted to the working context of the fire service is sparse (Hagemann, 2011; Hagemann & Kluge, 2013).

Application to the fire service

For teams to be effective and successful in firefighting operations, they need training in both technical (taskwork) and non-technical (teamwork) skills (McIntyre & Salas, 1995). The current apprenticeship in the fire service concentrates mostly on technical skills, e.g. in the sections 'vehicle knowledge', 'equipment knowledge', and 'operational theory'. However, the apprenticeship does not yet include content that addresses non-technical skills. Of course, the trainees work together in teams in the operational exercises which are part of the apprenticeship. But the focus there is still on the technical skills. The non-technical skills are neither discussed in detail nor trained specifically. But team trainings which are based on CRM, have been and are successfully used in aviation (e.g. Flin et al., 2002; Salas et al., 2006) and the medical field (e.g. Bohmann et al., 2021; Gaba et al., 2001; Paige et al., 2009). For the success of those trainings it is important not to simply adopt an existing concept from aviation or medicine, but to adapt it specifically for the application context, in this case the fire service, to explicitly address the specific needs and work requirements of the firefighters (Hunt & Callaghan, 2008). This enables the application of the newly learned skills in their specific work area (Hagemann, 2011). Thus, for this study, training modules were newly developed with the specific target group of trainees in the fire service in mind and evaluated through realistic field research.

Training development for teams in the fire service

There is little empirical evidence for how to structure team training concepts for the fire service successfully. So, in order to identify the work context and target group specific non-technical skills for firefighting operations, a qualitative and quantitative requirements analysis was conducted in advance. For the qualitative part, 27 interviews were conducted with experienced firefighters from professional fire departments, volunteer fire departments, and plant fire departments. They were questioned about their firefighting operations and the positive and negative aspects of teamwork they experienced. In addition, operational reports from the web portals firefighternearmiss[1], atemschutzunfälle.eu[2] and FUK CIRS[3] were analysed regarding critical situations or accidents which were caused by failures in teamwork. From the interviews and the operational reports, a system of categories was developed into which the non-technical skills identified were classified. Based on this, an online questionnaire on positive and negative aspects of teamwork in firefighting operations was developed, which was filled in by over 700 firefighters throughout Germany between January and May 2021. The results showed which aspects of teamwork are

[1] http://firefighternearmiss.com/Reports
[2] http://atemschutzunfälle.eu/
[3] https://www.fuk-cirs.de/fallbeispiele.html

experienced frequently and/or intensively in a positive and in a negative way in firefighting operations. Based on the most frequent and most intense aspects, learning objectives for the team training were derived. Based on the learning objectives the most appropriate methods and tools were deduced. The resulting training modules communication, (shared) situational awareness & shared mental model, cooperation/support, decision making, and leadership were developed according to the learning objectives.

Hypotheses

The authors are not aware of any literature regarding the development of a team training for non-technical skills especially for trainees in the fire service. Since the newly developed team training is designed to train non-technical skills and literature confirms the positive effect of team trainings for HRTs (e.g. Hagemann et al., 2017; Salas et al., 1999a), it is expected that these non-technical skills improve in the training group after participating in the team training. Therefore, the development of these skills from before to after the training is examined in the control and training group. In Hypothesis 1, situational awareness and communication, which were assessed through self-reports immediately after an operational exercise, are used as an example of successful teamwork.

Hypothesis 1) Subjectively experienced situational awareness and communication improve in the training group compared to the control group from baseline to post.

In addition, observations were used to substantiate the self-reports with more objective data. For Research Question 1, communication and situational awareness are used as examples for successful teamwork.

Research Question 1) How do participants' communication and situational awareness, assessed by external raters, develop over the course of the study?

It is important for the success of a training that the participants evaluate it positively. Learning success is increased, for example, by the fact that participants enjoy the training and perceive it as useful. Among other things, training design also correlates with a higher subjective knowledge gain (Ritzmann et al., 2014). Research Question 2 therefore looks at how the participants themselves evaluate the training in relation to two dimensions.

Research Question 2) How did participants evaluate the team training in relation to the outcome dimensions and the training design?

Materials and Methods

Sample

A total of 97 trainees in the fire service (4 women) from professional fire departments of two German cities participated in this study. Their mean age was $M = 27.30$ ($SD = 4.79$). 67 participants (1 woman) belonged to the control group (CG), 30 participants (3 women) to the training group (TG). The mean age in the control group was $M =$

27.53 (*SD* = 5.10) and in the training group *M* = 26.81 (*SD* = 4.11). The training group is smaller than the control group, because the trainings are currently still running and not all groups could be trained and evaluated yet.

Procedure

A pre-post-control group design was used. Each of the two measurement points consisted of an operational exercise, which was evaluated by external observers with regard to non-technical skills. Directly after the operational exercise the participants filled in a questionnaire in which the general work in the team during the operational exercise was queried. The measurement points were approximately one week apart. In the meantime, the participants continued with their regular schedule, which consisted mostly of operational exercises during that phase. In the training group a two-day team training took place between the two measurement points (for an overview see Figure 1).

All participants completed the questionnaire at the first measurement point (baseline) and at the second measurement point (post, 1 week later, with (TG) or without (CG) training). In total, 97 participants, divided into 13 teams, took part in the operational exercise at both measurement times (9 teams in the control group). The teams consisted of 6 - 8 persons. There were a few changes in the team composition due to illness or absence. Otherwise, the composition of the teams including the positions in the operational exercise was kept constant. For the observation during the operational exercises, each team was divided into two sub teams. One sub team (2-3 persons) belonged to the attack squad and is referred to as “inner team” as it mainly works inside the training building. The “outer team” (3-5 persons) consisted of the water squad, machinist, and hose squad (if present). They mainly work outside the training building. 30 trainees completed the team training.

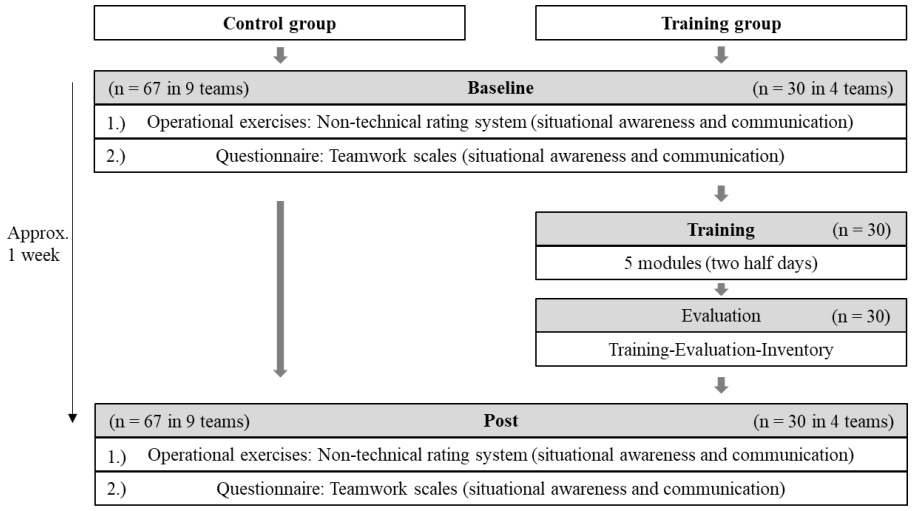


Figure 1. Experimental design.

Measures

The measurement instruments were a questionnaire answered directly by the trainees and a non-technical skills rating system used by observers to evaluate the participants' performance during an operational exercise.

Teamwork
A modified subscale of the Anti-Air Teamwork Observation Measure (Hagemann, 2011; Smith-Jentsch et al., 1998) was used to capture situational awareness. The following four items were used (1) *I used all available sources to gather information during the operational exercise.* (2) *I identified potential or anticipated problems during the operational exercise.* (3) *I identified deviations from normal conditions of a situation and informed others.* (4) *When the situation called for it, I acted without being prompted by my team members.* Participants rated these items on a 6-point-Likert scale from 1 = *never* to 6 = *always*. The Cronbach's α for the scale is .81 at baseline and .87 at post.

Modified subscales of the Anti-Air Teamwork Observation Measure (Hagemann, 2011; Smith-Jentsch et al., 1998) were also used to capture communication. The following six items were used (1) *I gave important information to the appropriate persons at the right time before being asked.* (2) *I regularly gave information of the situation to team members in order to maintain the overall picture.* (3) *I used proper phraseology*. (4) *I avoided excess chatter.* (5) *I spoke clearly.* (6) *I reported fully with all relevant information in the correct order.* Participants rated these items on a 6-point-Likert scale from 1 = *never* to 6 = *always*. The Cronbach's α for the scale is .77 at baseline and .82 at post.

Non-technical skills rating system
During the operational exercises, the firefighter trainees were evaluated by two trained observers in the five categories communication, situational awareness, decision-making, cooperation, and leadership. In addition to the direct observation on site (one observer per sub team (i.e. inner and outer)), video recordings made during the operational exercises were used to evaluate the operational exercises. This made it possible to evaluate the other sub team afterwards. For this paper, the categories communication and situational awareness are evaluated. The communication behaviour was observed on the basis of three subcategories (form of information exchange, content of information exchange, building common understanding) as well as situational awareness (gathering information, recognizing and understanding, anticipating). Spearman's ρ as an indicator of interrater reliability was .77, which can be interpreted as a strong agreement (Cohen, 1988).

Evaluation of the training
The Training-Evaluation-Inventory (TEI; Ritzmann et al., 2014) was used to evaluate the training. Based on 45 items (e.g., "Learning was fun"), the trainees rated both different outcome variables (i.e., subjective enjoyment, perceived usefulness, perceived difficulty, subjective knowledge gain, attitude towards training) and the training design (i.e., problem-based learning, activation of prior knowledge, demonstration of learning objectives and content, application of the contents in training, integration of the contents into daily work routine). The items of the training

design scale are based on Merrill (2002) who developed those as five first principles of instruction. A 5-point-Likert scale was used for this purpose (1 = *strongly disagree* to 5 = *strongly agree*). Cronbach's α is between .75 and .94 for the different subscales.

Results

Self-evaluated teamwork

Two 2x2 mixed ANOVAs with within-factor Time (baseline, post) and between-factor Group (control, training) were conducted to analyse Hypothesis 1, which states that the subjectively experienced situational awareness and communication improve in the training group compared to the control group from baseline to post. The requirements were checked and outliers were removed ($SD > 2.5$). SA was normally distributed for all groups, as assessed by the Kolmogorov-Smirnov test ($p > .05$). Communication was only partially normally distributed. Since visual inspection suggests a normal distribution for all groups and the procedure is robust to violations of this requirement, the procedure was continued. Regarding the ANOVA for SA, there was a statistically significant effect of time ($F(1, 93) = 8.46$, $p = .005$, partial $\eta^2 = .08$.). There was no statistically significant interaction between time and group ($F(1, 93) = 2.62$, $p = .109$, partial $\eta^2 = .03$). For communication there was no statistically significant effect of time ($F(1, 93) = 2.35$, $p = .128$, partial $\eta^2 = .03$) and no statistically significant interaction effect between time and group ($F(1, 93) = 0.00$, $p = .968$, partial $\eta^2 = .00$).

The graphical representations of the effects showed that the trend of the data for SA developed in the direction we hypothesised (see Figure 2). For communication data of both control and training group run parallel (see Figure 3).

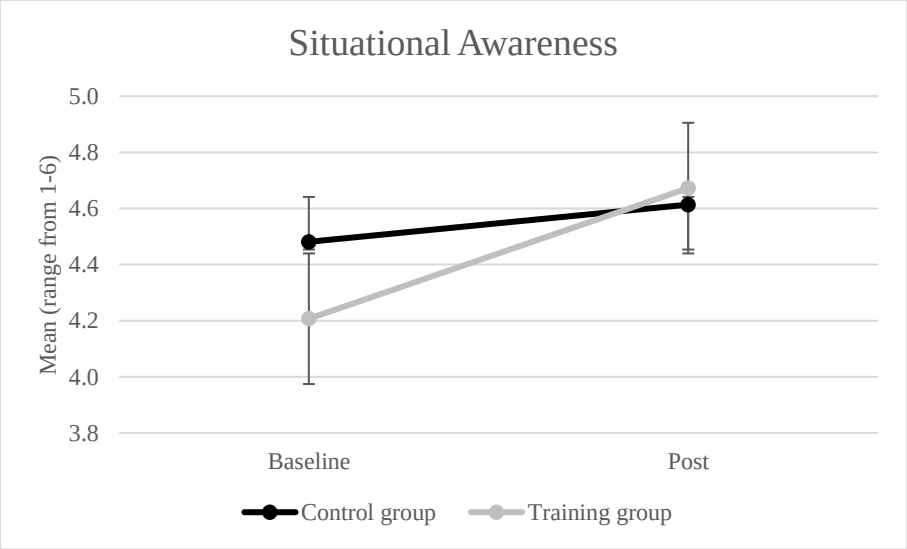


Figure 2. Mean subjective situation awareness in control group and training group before and after training. Error bars reflect Standard Error.

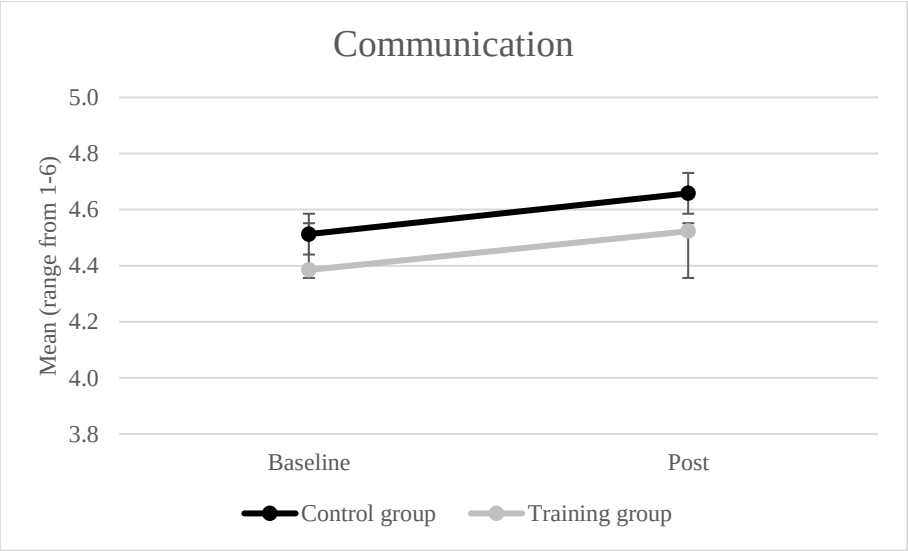


Figure 3. Mean subjective communication in control group and training group before and after training. Error bars reflect Standard Error.

Summing up, regarding Hypothesis 1, there is no interaction effect between time and group for SA as well as for communication. Power analysis showed that with the given sample size an effect size of .34 could be found to be significant. It can therefore be concluded that with the present sample size in this ongoing research project, no final assessment can be made for Hypothesis 1.

Observed behavioural changes through the training

During the operational exercises, the trainees' behaviour in terms of communication and situational awareness was observed and evaluated. The ratings of the two independent raters were averaged so that each sub team received one score for each exercise. Due to the small number of teams (n = 13) surveyed so far, the observational data were only analysed descriptively. In relation to Research Question 1, which looked at the development of participants' communication and situational awareness assessed by external raters, it could be seen that there were no differences between the control and training groups for the teams as a whole, as the scores improved slightly in both groups. However, when looking at the data separately for inner (i.e. attack squad) and outer (i.e. water squad, machinist, and possibly hose squad) team, it could be seen that the training group showed increases in communication and situational awareness in the outer team from baseline to post, while the control group showed constant ratings. These increases could not be confirmed to the same extent for the inner team of the training group. Furthermore, the inner team of the control group also showed an increase in communication. For an overview see Table 1.

Table 1. Means (SD) of communication and SA.

	Com BL	Com Post	SA BL	SA Post
Control group	2.27 (0.31)	2.46 (0.47)	2.25 (0.35)	2.43 (0.42)
Inner team	2.22 (0.37)	2.61 (0.46)	2.19 (0.39)	2.54 (0.40)
Outer team	2.31 (0.26)	2.31 (0.47)	2.31 (0.31)	2.31 (0.43)
Training group	2.48 (0.44)	2.79 (0.33)	2.23 (0.36)	2.75 (0.39)
Inner team	2.75 (0.44)	2.75 (0.40)	2.33 (0.45)	2.79 (0.44)
Outer team	2.21 (0.25)	2.83 (0.30)	2.13 (0.25)	2.71 (0.39)

Note. Means, range from 1 = never to 4 = constantly, Com=communication, BL=baseline, SA=situational awareness

Summing up, regarding Research Question 1, it can be stated that there are no differences between the groups as a whole. When looking at the inner team, there is an increase in communication in the control group but not in the training group. SA increases in both groups in the inner team. In the outer team, there is an increase in communication and situational awareness in the training group which cannot be seen in the control group.

Evaluation of the training

Regarding Research Question 2, which looked at how participants rated the team training in terms of the outcome dimensions and the training design, the evaluation immediately after the training showed that participants rated the training positively in relation to both the different outcome variables and the training design. All mean scores were in the upper third of the scale (range from 1 to 5). The scores for the outcome dimensions ranged between $M = 3.82$ and $M = 4.58$ (SD between 0.43 and 0.82) and for the training design between $M = 3.59$ and $M = 4.44$ (SD between 0.49 and 0.71). An overview is shown in Figures 4 and 5.

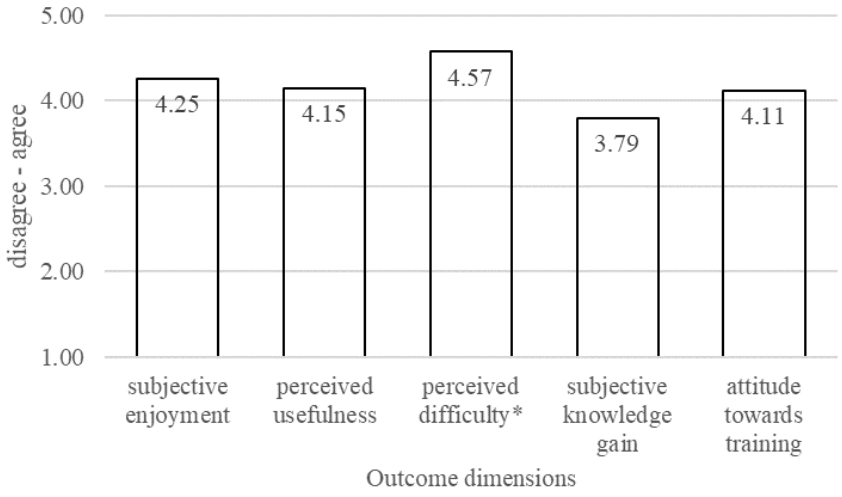


Figure 4. Means of the Outcome dimensions. Perceived difficulty was reverse coded so that high scores indicate low difficulty.

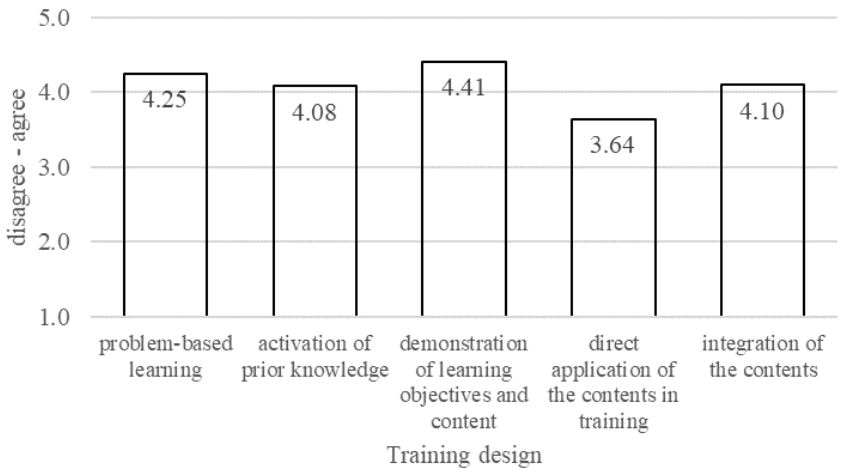


Figure 5. Means of the variables of Training design.

Summing up, regarding Research Question 2, it can be shown that the participants enjoyed the team training in regard to subjective enjoyment. They also perceived it as useful and not too difficult. Furthermore, the participants reported a gain in knowledge regarding the contents of the team training and have a positive attitude towards it. These results can be associated with learning success and positive attitudes towards teamwork skills (Ritzmann et al, 2014).

Regarding the training design, the participants confirmed that the design principles that promote learning according to Merrill (2002) were fulfilled. Meeting these requirements for the training design supports learning and the transfer of what has been learned into everyday work (Ritzmann et al., 2014).

Discussion

The present study aims to develop and evaluate a non-technical skills training specifically adapted to the needs and working requirements of trainees in the fire service. To date, to the authors' knowledge, there is no non-technical skills team training like that and accordingly no scientific literature evaluating a non-technical skills team training for trainees in the fire service. Therefore, a qualitative and quantitative requirements analysis was conducted which was then used to develop learning objectives for the team training. Based on those, the team training was developed, containing the modules communication, (shared) situational awareness & shared mental model, cooperation/support, decision making, and leadership. The new team training was evaluated with the help of operational exercises using pre-post-control group design with a questionnaire and external observations as measurements.

Regarding Hypothesis 1, which stated that situational awareness and communication improve in the training group compared to the control group from baseline to post, no significant results could be shown. The expected training effects can be visualised as a trend regarding SA, which increases slightly more in the training group compared to the control group from baseline to post. No effects could be demonstrated for

communication. The fact that we have not yet been able to demonstrate positive changes in the non-technical skills, as other studies from other sectors (medicine and aviation) have (Hagemann et al., 2017; Salas et al., 1999a), may be due to the fact that the presented study has not yet been completed. This means that the control group is currently substantially larger than the training group. In addition, the overall sample size is not yet sufficient to find even medium effect sizes. A trend could be seen in the SA data, but the control group values were substantially higher from the beginning. This could be due to the fact that the participants are newcomers to the profession who cannot yet reliably assess their own performance, as they have not yet encountered real challenges and missions. It is also possible that the training group evaluates itself more critically than the control group, because only the training group has been made aware of the issues. Briggs et al. (2015) described that SA is critical for performance in HRTs and should be addressed in trainings. By improving SA, mistakes and subsequently critical situations or accidents can be avoided and firefighting operations can be performed successfully (Endsley, 1999b; Wilke, 2006). Improving communication can have positive effects on the number of accidents (Molesworth & Estival, 2015) and safety at the workplace (Flin et al., 2008).

The results of Research Question 1 show that the non-technical skills communication and situational awareness, assessed by raters, tend to be partially better in the training group compared to the control group after the team training. The improvement of communication through team training is also consistent with findings in the literature (e.g. Armour Forse et al., 2011; Salas, 1999a). The tendency to show better communication and situational awareness in the training group compared to the control group only occurs when looking at the outer team. The division into an inner (i.e. attack squad) and an outer (i.e. water squad, machinist, and possibly hose squad) team is a peculiarity in firefighting groups. In the present observational data, only the outer team in the training group shows an improvement between before and after the training compared to the control group. During an operational exercise, the demands on the inner team are noticeably higher than on the outer team. The inner team stays mainly inside a training house in the operational exercises of the present study. There it is confronted with external stressors such as darkness, smoke, and noise. These have an aggravating effect on teamwork. The outer team works mainly in daylight and with less noise. Under these conditions, it might be much easier for the outer team to try out and successfully implement the learnings from team training in terms of communication and situational awareness. This means that inner teams might need more opportunities for exercising in order to be able to apply the newly learned skills. In addition, it could be useful to create more opportunities for trainees in the fire service to practice teamwork skills under challenging conditions. As there is also an increase in situational awareness and communication from baseline to post in the control group, these results must be viewed with caution. It is possible that the effects also occur without training.

Descriptive data show that the training was positively evaluated by the trainees (Research Question 2), which can be interpreted as a positive result for the effectiveness of the training in terms of positive attitudes towards teamwork, knowledge gain and transfer of the training contents (Hagemann & Kluge, 2013; Ritzmann et al., 2014).

Strengths & limitations

The training is evaluated through field research. This means that the study does not take place in a simulator under completely standardised laboratory conditions with students as test subjects, but that the training is evaluated with the real target group during real operational exercises. Of course, there are difficulties with this procedure, such as cancellations or postponements due to Corona-related restrictions or weather-related influences. However, the advantages are that the evaluation is extremely practice-oriented and the training procedures are optimally adapted to the fire department. The use of field research also improves external validity, as participants are in their familiar environment (classrooms, training building), which facilitates the transfer of the newly learned skills into their everyday (working) life. A disadvantage for external validity is the possible effect of observation during pre-measurement. Participants might be influenced to be more attentive during team training than trainees would be who do not participate in a study. They know that they will be observed again during an operational exercise after the team training and want to perform well there.

Internal validity is somewhat limited, as the participants are not randomly assigned to a group, but the starting date of their apprenticeship determines the assignment. It is also not tested, e.g. through a knowledge test, how attentive each participant was during team training and how much training content has stuck accordingly. But with the modules being conducted in attendance and as interactively as possible, this should be the best way to keep attention as high as possible.

In the control group presented here, the evaluation of self-assessed situational awareness might show a ceiling effect, as they have a higher score in the baseline than the training group. This could prevent a significant effect of time being found in the control group as was found in the training group. However, the difference between control and training group at baseline is not significant.

A key advantage of the study is the mixed-methods design. Self-reports and external observation ratings were used so that two interdependent perspectives were considered. Furthermore, two well-trained external observers rated interdependently direct at the scene as well as retrospectively using videos.

Future perspectives and practical implications

In the further course of the study, more participants will be surveyed and there will be a second post measurement point six months after the first post measurement. So further relationships and effects, as for example a long-term behavioural change in the training group, can be evaluated.

It may be necessary for people working in the inner team to have more than one practice opportunity in order to successfully apply the content of the team training. Therefore, future studies should provide participants with several practice opportunities where they observe what changes occur within the inner team.

The team training developed is designed to be very practice-oriented and should be used as a standard in the apprenticeship in the fire service. The application in other contexts, e.g., in the volunteer fire department, is also desirable and will be advanced in the further course of the project.

Conclusion

This study shows mixed findings on whether the team training specifically developed for fire service trainees helps improve the quality of non-technical skills. The evaluation of the team training by the participants shows first indications of the effectiveness of the training. The further course of the project with proceeding data collection will show whether the initial trends can be verified.

References

Armour Forse, R., Bramble, J.D., & McQuillan, R. (2011). Team training can improve operating room performance. *Surgery, 150*, 771-778. https://doi.org/10.1016/j.surg.2011.07.076

Badke-Schaub, P. (2012). Handeln in Gruppen [Acting in groups]. In P. Badke-Schaub, G. Hofinger, and K. Lauche (Eds.), *Human Factors* (pp. 121-139). Berlin, Heidelberg: Springer.

Baker, D.P., Day, R., & Salas, E. (2006). Teamwork as an essential component of high-reliability organizations. *Health Services Research, 41*, 1576–1598. https://doi.org/10.1111/j.1475-6773.2006.00566.x

Bohmann, F.O., Guenther, J., Gruber, K., Manser, T., Steinmetz, H., & Pfeilschifter, W. (2021). Simulation-based training improves patient safety climate in acute stroke care (STREAM). *Neurological Research and Practice, 3*(1), 37. https://doi.org/10.1186/s42466-021-00132-1

Branlat, M., Fern, L., Voshell, M., & Trent, S. (2009). Understanding coordination challenges in urban firefighting: A study of critical incident reports. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (pp. 284-288). https://doi.org/10.1518/107118109x12524441080740

Briggs, A., Raja, A.S., Joyce, M.F., Yule, S.J., Jiang, W., Lipsitz, S.R., & Havens, J.M. (2015). The role of nontechnical skills in simulated trauma resuscitation. *Journal of Surgical Education, 72*, 732-739. https://doi.org/10.1016/j.jsurg.2015.01.020

Cannon-Bowers, J.A. & Bowers, C.A. (2011). Team Development and Functioning. In *APA Handbook of Industrial and Organizational Psychology* (pp. 597–650).

Cannon-Bowers, J.A. & Salas, E. (1998). Team Performance and Training in Complex Environments: Recent Findings from Applied Research. *Current Directions in Psychological Science, 7*, 83–87. https://doi.org/10.1111/1467-8721.ep10773005

Cannon-Bowers, J.A., Salas, E., & Converse, S.A. (1993). Shared mental models in expert team decision making. In N.J. Castellan (Ed.), *Individual and group decision making: Current issues* (pp. 221-246). Hillsdale, NJ, USA: Lawrence Erlbaum Associates.

Cannon-Bowers, J.A., Tannenbaum, S.I., Salas, E., & Volpe, C.E. (1995). Defining Competencies and Establishing Team Training Requirements. In R.A. Guzzo

and E. Salas (Eds.), *Team effectiveness and decision making in organizations* (pp. 333-380). San Francisco, CA, USA: Jossey-Bass.

Cohen, J. (1988). Statistical power analysis for the behavioral sciences. Hillsdale, NJ: L. Erlbaum Associates.

Endsley, M.R. (1995a). Toward a Theory of Situation Awareness in Dynamic Systems. *Human Factors, 37*, 32-64.

Endsley, M.R. (1995b). A taxonomy of situation awareness errors. *Human Factors in Aviation Operations, 3*, 287-292.

Fire Near Miss - Lessons learned become lessons applied (2016). Opposing Hose Lines Hamper Crews. Available online: http://www.firefighternearmiss.com/Reports?id=7447 (accessed on 11 April 2022).

Flin, R.H., O'Connor, P., & Crichton, M.T. (2008). *Safety at the sharp end: A guide to non-technical skills*. Aldershot, UK: Ashgate.

Flin, R.H., O'Connor, P., & Mearns, K. (2002). Crew resource management: improving team work in high reliability industries. *Team Performance Management: An International Journal, 8*, 68-78. https://doi.org/10.1108/13527590210433366

Gaba, D.M., Howard, S.K., Fish, K.J., Smith, B.E., & Sowb, Y.A. (2001). Simulation-Based Training in Anesthesia Crisis Resource Management (ACRM): A Decade of Experience. *Simulation and Gaming, 32*(2), 175–193. https://doi.org/10.1177/104687810103200206

Hagemann, V. (2011). *Trainingsentwicklung für High Responsibility Teams [Training development for high responsibility teams]*. Lengerich: Papst Verlag.

Hagemann, V., Herbstreit, F., Kehren, C., Chittamadathil, J., Wolfertz, S., Dirkmann, D., Kluge, A., & Peters, J. (2017). Does teaching non-technical skills to medical students improve those skills and simulated patient outcome? *International journal of medical education, 8*, 101-113. https://doi.org/10.5116/ijme.58c1.9f0d

Hagemann, V. & Kluge, A. (2013). The Effects of a Scientifically Based Team Resource Management Intervention for Fire Service Teams. *International Journal of Human Factors and Ergonomics, 2*, 196-220. https://doi.org/10.1504/IJHFE.2013.057617

Hagemann, V., Kluge, A., & Kehren, C. (2015). Evaluation of Crew Resource Management Interventions for Doctors-on-Call. In D. de Waard, J. Sauer, S. Röttger, A. Kluge, D. Manzey, C. Weikert, A. Toffetti, R. Wiczorek, K. Brookhuis, and H. Hoonhout (Eds.), *Proceedings of the Human Factors and Ergonomics Society Europe Chapter 2014 Annual Conference* (pp. 237-253). ISSN 2333-4959 (online).

Hunt, G.J.F. & Callaghan, K.S.N. (2008). Comparative issues in aviation and surgical crew resource management: (1) are we too solution focused? *ANZ Journal of Surgery, 78*, 690–693. https://doi.org/10.1111/j.1445-2197.2008.04619.x

Kluge, A., Sauer, J., Schüler, K., & Burkolter, D. (2009). Designing training for process control simulators: a review of empirical findings and current practices. *Theoretical Issues in Ergonomics Science, 10*, 489-509. https://doi.org/10.1080/14639220902982192

Landis, R.T. & Koch, G.G. (1977). The Measurement of Observer Agreement for Categorical Data. *Biometrics*, *33*, 159-174.

Leape, L.L., Lawthers, A.G., Brennan, T.A., & Johnson, W.G. (1993). Preventing Medical Injury. *QRB - Quality Review Bulletin*, *19*, 144-149. https://doi.org/10.1016/S0097-5990(16)30608-X

Marks, M.A., Mathieu, J.E., & Zaccaro, S.J. (2001). A temporally based framework and taxonomy of team processes. *Academy of Management Review*, *26*, 356-376.

McIntyre, R.M. & Salas, E. (1995). Measuring and managing for team performance: Lessons from complex environments. In R.A. Guzzo and E. Salas (Eds.), *Frontiers of industrial and organizational psychology. Team effectiveness and decision making in organizations* (pp. 9-45). Jossey-Bass.

Merrill, M.D. (2002). First principles of instruction. *Educational Technology Research and Development*, *50*, 43-59.

Molesworth, B.R. & Estival, D. (2015). Miscommunication in general aviation: The influence of external factors on communication errors. *Safety Science*, *73*, 73-79. https://doi.org/10.1016/j.ssci.2014.11.004

Morgan, B.B., Salas, E., & Glickman, A.S. (1993). An analysis of team evolution and maturation. *Journal of General Psychology, 120*, 277-291.

Nieva, V.F., Fleishman, E.A., & Rieck, A. (1985). *Team Dimensions: Their Identity, Their Measurement and Their Relationships*. Fort Belvoir, VA. https://doi.org/10.21236/ada149662

Omodei, M.M., McLennan, J., & Reynolds, C. (2005). *Identifying the Causes of Unsafe Firefighting Decisions: A Human Factors Interview Protocol* (Bushfire CRC Project D2. 3 Safety in Decision Making and Behaviour, Tech. Rep. No 1). Melbourne, Australia: Australian Bushfire Cooperative Research Centre.

Paige, J.T., Kozmenko, V., Yang, T., Paragi Gururaja, R., Hilton, C.W., Cohn, I., & Chauvin, S.W. (2009). High-fidelity, simulation-based, interdisciplinary operating room team training at the point of care. *Surgery, 145*, 138-146. https://doi.org/10.1016/j.surg.2008.09.010

Rosen, M.A., DiazGranados, D., Dietz, A.S., Benishek, L.E., Thompson, D., Pronovost, P.J., & Weaver, S.J. (2018). Teamwork in healthcare: Key discoveries enabling safer, high-quality care. *American Psychologist, 73*(4), 433–450. DOI: 10.1037/amp0000298

Ritzmann, S., Hagemann, V., & Kluge, A. (2014). The Training Evaluation Inventory (TEI) - Evaluation of Training Design and Measurement of Training Outcomes for Predicting Training Success. *Vocations and Learning*, *7*, 41-73. https://doi.org/10.1007/s12186-013-9106-4

Salas, E., Fowlkes, J.E., Stout, R.J., Milanovich, D.M., & Prince, C. (1999a). Does CRM training improve teamwork skills in the cockpit? Two evaluation studies. *Human Factors*, *41*, 326-343.

Salas, E., Prince, C., Bowers, C.A., Stout, R.J., Oser, R.L., & Cannon-Bowers, J.A. (1999b). A methodology for enhancing crew resource management training. *Human Factors*, *41*, 161–172

Salas, E., Sims, D.E., & Burke, C.S. (2005). Is there A "big five" in teamwork? *Small Group Research*, *36*, 555-599. https://doi.org/10.1177/1046496405277134

Salas, E., Wilson, K.A., Burke, C.S., Wightman, D.C., & Howse, W.R. (2006). Crew Resource Management Training Research, Practice, and Lessons Learned. *Reviews of Human Factors and Ergonomics, 2*, 35-73. https://doi.org/10.1177/1557234X0600200103

Smith-Jentsch, K.A., Johnston, J H., & Payne, S.C. (1998). Measuring Team-Related Expertise in Complex Environments. In J.A. Cannon-Bowers and E. Salas (Eds.), *Making Decisions Under Stress* (pp. 61-87). Washington: American Psychological Association.

Sundstrom, E., Meuse, K.P. de, & Futrell, D. (1990). Work teams: Applications and effectiveness. *American Psychologist, 45*, 120–133. https://doi.org/10.1037/0003-066X.45.2.120

Unfallkommission „Tübingen“ [Accident Commission "Tübingen"] (2006). *Bericht zum Einsatz „Tübingen - Reutlinger Straße 34/1“ [Report on the operation "Tübingen - Reutlinger Straße 34/1"]*. Baden-Württemberg: Landesbranddirektor Innenministerium [Baden-Württemberg: State Fire Director Ministry of the Interior]. Available online: https://www.atemschutzunfaelle.de/download/Unfaelle/u20051217-tuebingen-bericht-unfallkommission.pdf (accessed on 8 April 2022).

Weick, K.E. & Sutcliffe, K.M. (2003) *Das Unerwartete managen [Managing the Unexpected]*. Stuttgart: Klett-Cotta.

Wilke, J.P. (2006). *Fordern und Fördern - Führungspraxis für Feuerwehrleute [Challenge and encourage - leadership practice for firefighters]*. Stuttgart: Kohlhammer.

Critical decision making with a highly automated UAV – a case study

Nicolas Maille
ONERA – The French Aerospace Lab
France

Abstract

In transportation and aerospace, more automation and autonomy are continuously added to systems. The ability of human operators to effectively monitor and interact with these systems, poses significant challenges. This research focuses on critical decisions that largely rely on the system capabilities but need to be validated and made under the responsibility of the operator. In the context of unmanned combat air vehicles (UCAV), the experiment focuses on how the communication strategy of semi-autonomous systems modifies the operators' understanding of the situation and the final decision. The study has been conducted in an immersive simulator with a 30 minutes ecological military scenario where the operator had to manage a full mission, including an unplanned missile firing decision. The experiment included the use of physiological measures related to electrodermal and cardiac activities. The paper reports the results of the decision-making performances and the analyses of the physiological parameters. It appears that the communication strategy has an impact on the situation awareness of the operator, the decision taken, and the evolution of the physiological parameters.

Introduction

The development of highly automated vehicles, from autonomous cars for civilian applications to Unmanned Combat Air Vehicle (UCAV) for military operations is profoundly changing the way people interact with these systems. Although the word "autonomy" implies that systems will be able to perform actions on their own, in real-world applications, these autonomous systems must still cooperate with humans who may be responsible for effectively monitoring the behaviour of systems, directing them when needed, or acting as teammates and collaborating on decision-making. The ability of human operators to oversee and manage these systems appropriately when needed is a major challenge. Endsley (2017) wrote "*an automation conundrum exists in which as more autonomy is added to a system, and its reliability and robustness increase, the lower the situation awareness of human operators and the less likely that they will be able to take over manual control when needed*". Questions about autonomous driving and how humans adapt to taking control of these vehicles are currently under investigation (Morgan et al., 2016; Eriksson & Stanton, 2017; Morgan

In D. de Waard, S.H. Fairclough, K.A. Brookhuis, D. Manzey, L. Onnasch, A. Naumann, R. Wiczorek, F. Di Nocera, S. Röttger, and A. Toffetti (Eds.) (2022). Proceedings of the Human Factors and Ergonomics Society Europe Chapter 2022 Annual Conference. ISSN 2333-4959 (online). Available from http://hfes-europe.org

et al., 2018). For example, recent studies investigate the effects of takeover signal lead time or modality on automated vehicle takeover performance (Huang & Pitts, 2022).

In the case of drones, the main issue is no longer taking manual control of the vehicle, but rather being responsible for monitoring the mission, assessing the overall situation in relation to the mission objectives, and collaborating in decision-making (Barnes & Evans, 2016). From an operational perspective, humans may soon act as managers 'on' or 'over' the decision-making loop, rather than in it (Mayer, 2015). As a result, a new context is emerging, characterized by humans managing a machine-driven decision loop. In the military domain, a strong requirement is the ability to operate in more contested air environments, which implies reduced data transmission, automated on-board analysis of raw intelligence data, and greater autonomy when navigating or tracking targets (Mayer, 2015). Increasing the level of autonomy allows for more irregular monitoring by the operator and raises the issue of "neglect time" (i.e., the length of time the system can operate autonomously before reporting back to the human) and "interaction time" (i.e., the period of time during which the system and the human communicate and define the next actions) (Olsen & Goodrich, 2003). Stress is high during the neglect period because of the uncertainty of what the system is doing and whether it will actually return a communication. Stress is also high in the interaction period because communication must be fast, efficient, and accurate (Hancock & Szalma, 2008). These communication constraints amplify the difficulties for the operator who does not have access to the continuous evolution of the situation but who may have to enter the decision loop at any moment and cooperate in highly critical and urgent decisions. While the operator has only a limited amount of time to weigh, verify and gather all critical information, such an interaction process can create a high workload and stress. It is worth pausing to reflect on the operator's ability to contribute effectively and take responsibility for the final decision.

This study focuses on the effects of the timing of communications on the decision making process. Human computer interaction studies addressing human-drone interaction generally indicate that greater transparency about the drone's behaviour helps the operator to monitor the mission (Mirri et al., 2019). Nevertheless, when communications must be sparse for operational reasons, the effect of communication choices on the operator's supervision and decision is little studied. The main contribution of this work is to evaluate in an ecological simulation how the communication strategy implemented by a UCAV impacts the final decision made by the operator and can change the stress, workload and situation awareness of this frontline operator. The use case involves a UCAV operating in a hostile environment in which a trade-off must be made between stealth (for survivability) and communications with the ground station. This work is based on two main hypotheses:

H1: An early communication strategy reduces the stress and workload of the decision-making process.

H2: An early communication strategy helps the operator to make the right decisions.

This experiment is part of a larger research project, but the results presented in this paper focus on one decision to be made during the mission. Further hypotheses on the

evaluation of global awareness and workload at the end of the mission have been defined but are not presented here.

Material and method

Participants

The study was approved by University of Aix-Marseille Ethics Committee (ref. Prop. 2018-24-05-001) and involved twenty students and junior research scientists (5 females, 15 males), aged between 20 and 38 years old (M=27.4; SD=4.9). There were all civilian employees of an aerospace laboratory. All subjects volunteered to take part in the study and gave their full informed consent before taking part in the experiment.

Task

The military operational context used to create an ecological task was based on an armed reconnaissance mission, which is one of the typical missions envisaged for future UCAVs. The main objective of the mission is to obtain detailed information on enemy activity in a given area, with a contested air space (Fig. 1). Even if the mission is not dedicated to attacking predetermined targets, the aircraft must be armed and capable of identifying threats and conducting air strikes on targets of opportunity. The mission usually involves medium-range infiltration into a contested environment, reconnaissance of the area and exfiltration from enemy territory.

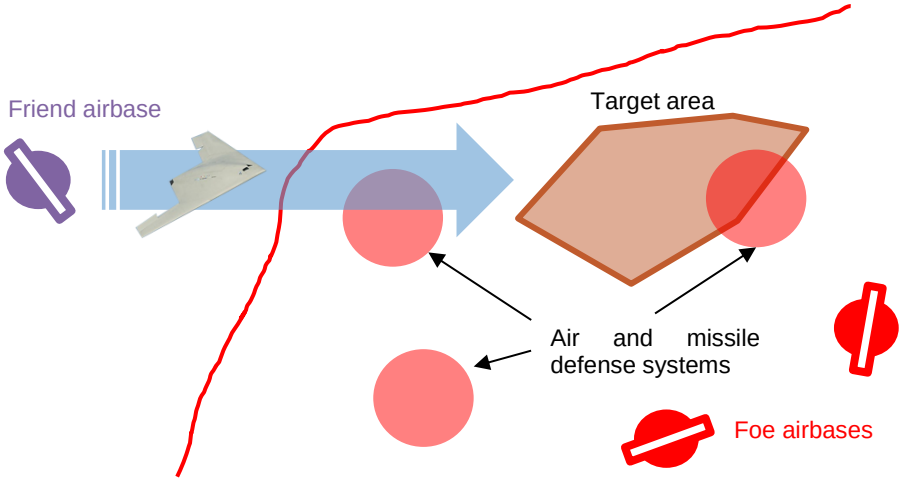


Figure 1. Armed reconnaissance mission. The red line indicates the border of the contested air space.

In this study, the simulated UCAV was equipped with terrain following capacities, air-ground missiles and highly automated features that gave it the ability to autonomously identify threats and adapt the mission if necessary. To increase survivability, communications in foe territory were severely restricted, but the UCAV took pictures of targets and threats and sent them back to the operator when a

communication point was reached. The use of the weapon to strike targets of opportunity had to comply with the rules of engagement and was under the operator's responsibility. A set of rules, in line with those used in military battlefields, was defined for this experiment. Briefly, the use of weapons was limited to enemy military targets with hostile intent and permission to fire was given to the crew only if no collateral damage was expected. Otherwise, permission to engage the target was given to a higher hierarchical level and the operator had to request permission to fire.

The participant had to monitor a complete 30 minutes mission. The initial flight plan to achieve the mission objectives was already inserted in the flight management system. This plan contained three targets to be observed by the UCAV. Pictures of these targets were to be taken and sent back to the ground station when communication between the UCAV and the ground station was allowed. Specifically, the flight plan contained sections where the UCAV communicated continuously with the ground station, updating its position, sending all available data (images) and other sections where only intermittent communication was allowed at predefined communication waypoints. Nevertheless, the initial flight plan could be modified by the UCAV if threats were encountered during the mission. Depending on the operational situation, these threats could be avoided (the UCAV moved away from the threat) or engaged (the flight plan was modified to create a missile firing opportunity on this new target). In the second case, the operator was responsible for the final firing decision, which had to be in accordance with the rules of engagement. In addition, throughout the firing window, a continuous communication channel was maintained between the UCAV and the ground station, so that all available data to support decision making was displayed to the operator. On the ground station's touch screen interface, the operator had the option to initiate or deny the attack. A radio communication system allowed direct communication with headquarters (in this case the experimenter) if approval was required by the rules of engagement before executing the action.

During the mission, two threats were identified by the UCAV and led to changes in the flight plan. Only the first, which was a missile firing opportunity, required a decision by the operator and is considered in this article. The operator had to collaborate with the UCAV and decide whether or not the target should be engaged. The process involved three distinct parts:

- Perception: extracting useful elements from the environment to understand the actual situation at that time and place on the battlefield and evaluate possible collateral damages.
- Deduction: Select the relevant rules of engagement and decide what to do. Depending on the situation, the operator could deduce that engagement was not allowed, or that engagement was only possible with the approval of higher headquarters, or that she/he could take responsibility for the missile firing.
- Action: abort the attack, contact headquarters, or validate the attack.

All participants had the same scenario and the right decision was to contact headquarters.

Experimental conditions

This study employed a 2 between-subject design (N=20). Two communication strategies (C1 and C2) were used.

C1: Early communication strategy. In this condition, the UCAV inserted new communication waypoint whenever new threats were encountered. The operator was therefore immediately informed of the change in the flight plan and knew that a firing opportunity was possible. However, the operator did not have the necessary information at that time to make the decision. This data was only provided at the beginning of the firing window.

C2: Late communication strategy. In this condition, the flight plan change was made without being communicated to the operator. The change was only sent to the operator when the firing window was started.

Thus, condition C1 favoured the transmission of new data to the operator while the second favoured the survivability of the UCAV.

Experimental device

Figure 2. UCAV flight simulator (left) and its dedicated touchscreen user interface (right).

An immersive UCAV simulator (Fig. 2) was used to run the scenario with a dedicated touchscreen interface allowing (1) the monitoring of the UCAV trajectory and the visualization of the new the flight plan when modified by the UCAV (right part of Fig. 2); (2) the visualization of the target's images sent by the UCAV (central part of Fig 2), and (3) the visualization of a continuous stream of full-motion video during the firing decision process (also on the central part of Fig 2, instead of the image management interface).

BioPac (MP150) was used to collect physiological data. Electrodermal activity (EDA) and cardiac activity (ECG) were recorded. Both raw signals were acquired at a sampling rate of 1250 Hz. The Biopac ECG100C amplifier used a band-pass filter of 35Hz and 0.5Hz. The Biopac EDA100C amplifier used with a low pass filter set at 10Hz. As both hand were used for the experiment, we used the recommended foot sites (right foot) for EDA recording (Boucsein et al., 2012).

Procedure

Once the ECG and EDA physiological acquisition systems were set up, the participants were briefed on the operational context of the mission, including the rules of engagement, and installed in the simulator. A presentation of the user interface was given and then the participants performed four training scenarios to familiarize themselves with the UCAV monitoring and the firing decision-making process. If they wished to continue with the other phases of the trial, participants signed a consent form. They then performed the 30 minutes scenario before fulfilling NASA-TLX (Hart & Staveland, 1988) (workload) and QUASA (McGuinness 2004) (situation awareness) questionnaires. Participants received a full verbal debrief. The experiment lasted approximately 2 hours.

Data analysis

A factorial independent measure design was employed. The independent variable was the experimental condition with two levels: early and late communication strategy. So 10 subjects (M=27.5; SD=5.5) managed the UCAV with an early communication strategy (C1) while the 10 others (M=27.3; SD=4.5) managed the UCAV with a late communication strategy (C2). Both electrodermal and cardiac activities were analysed thanks to the AcqKnowledge 4.1 © software.

Each participant's raw ECG data was processed using the built-in "Detect and Classify Heartbeats" function to estimate the R-wave peaks (minimum BPM 30; maximum BPM 240; R wave threshold 50% Max R peak level). A visual inspection was used to remove unreliable R peaks and related R-R intervals before calculating the mean R-R interval for the one-minute baseline and for the firing window. The metric used is then the difference, for each participant, between the mean R-R interval for the firing window and the mean R-R interval for the baseline. This normalization allows for comparison between subjects.

Each participant's raw EDA signal was visually inspected to remove parts with noisy data (foot movement) and SRC were identified thanks to the AcqKnowledge "Locate SRCs" function. For each participant, the first 500 seconds of the experiment (before the threat was detected) are used as a baseline (extract mean and SD) for a participant z-score transformation. Then, the normalized EDA allows for comparison between subjects. Two measures are used, as shown in figure 3.

- **Measure 1:** The difference between the EDA value prior to threat identification (mean value for the 60 seconds prior to detection of the new threat and communication of the flight plan change to the operator) and the higher EDA value obtained during the firing decision window.
- **Measure 2:** The largest amplitude of the skin conductance response (SCR) during the firing decision window. It measures the phasic change in electrical conductivity of the skin related to the drone request for the firing decision.

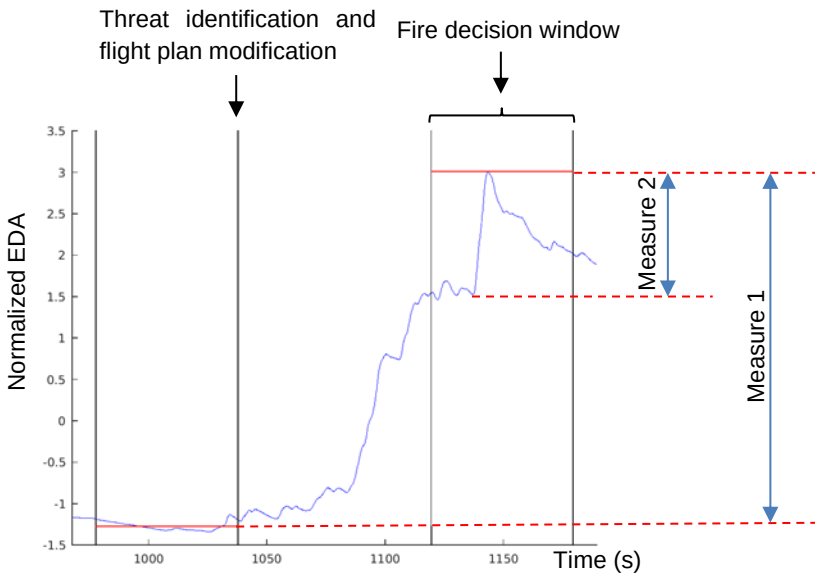


Figure 3. Measures used to characterize the galvanic skin response induced by the decision-making process.

Results

All statistical tests reported are two-tailed with alpha levels of .05. Effect sizes were determined using Cohen's d with ≥.2, ≥.5 and ≥.8 indicating small, medium and large effect sizes. In figures, the central rectangle spans the first quartile to the third quartile, the black segment inside the rectangle shows the median, the diamond gives the mean, error bars represent the "inner fence" and unfilled circles outliers.

Testing Hypothesis 1

The first hypothesis is that the early communication strategy should reduce the stress and workload associated with the decision-making process, compared to the late communication strategy. Changes in stress level and workload are assessed by physiological parameters: heart rate and skin conductance. In both experimental condition, the identification by the UCAV of a new threat and the request to the operator to validate or not an attack on this new target should increase the level of stress and workload during the firing window. In both conditions an increase in heart rate and skin conductance is expected. This first hypothesis will be confirmed if these modifications are shaped by the experimental conditions.

As already stated, the analysis of the heart rate is made on the bases of R-R intervals expressed in seconds. In both conditions, a reduction of the R-R interval is observed (Fig. 4) and corresponds to an increase of the heart rate in beat per minute. Nevertheless, the reduction of the R-R interval compared to baseline is significant only for condition C2 (Paired t-test; C1: t = 1.325, df = 9, p-value = 0.218, Cohen's d= 0.256; C2: t = 3.368, df = 9, p-value = 0.008, Cohen's d= 0.574). So, a significant

heart rate increases (in beat per minute, compare to baseline) is observed only for condition 2. The comparison of the reduction of R-R interval between both conditions is not significant (Two Sample t-test; t = 0.940, df = 18, p-value = 0.360, Cohen's *d* =0.420).

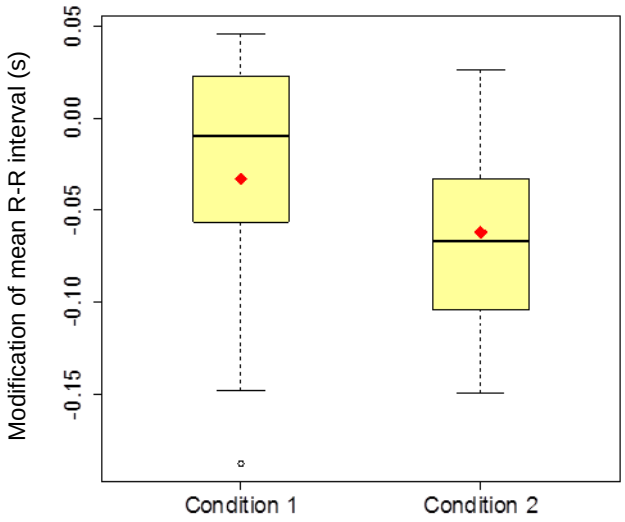


Figure 4. Differences between the mean R-R interval during the fire decision windows and the mean R-R interval during the minute that precedes the detection of the threat, by condition.

To conclude, there is no clear increase in heart rate during the decision making process in the first condition (C1) whereas this increase is significant with the late communication strategy (C2). Although this result is not confirmed by the direct comparison between both conditions, it is therefore likely that with a larger number of subjects, the difference between the two conditions should become significant.

For the electrodermal activity, the statistical analysis of measure 1 indicates that in both conditions an increase in skin conductance is observed (Paired t-test; C1: t = -5.920, df = 9, p-value = 0.000, Cohen's *d* = 2.135; C2: t = -4.300, df = 9, p-value = 0.002, Cohen's *d* = 1.372). Nevertheless, no significant difference appears between the two experimental conditions (Two sample t-test; t = -0.566, df = 18, p-value = 0.578, Cohen's *d* = 0.253). Thus, the detection of the new threat, and the opportunity to attack it, implies an activation of the sympathetic branch of the autonomic system from the operator. Moreover, this physiological response seems to be equivalent, whatever the communication strategy of the UCAV.

When measurement 2 is analysed, the skin response related to the drone request during the firing window is slightly different for the two experimental conditions, although the predefined statistical level is not reached (Two sample t-test; t = -1.554, df = 18, p-value = 0.138, Cohen's *d* = 0.695). It is likely that with some additional participants these differences become significant as the Fligner-Policello test (U* = 2.465, p-value = 0.015) considers that the difference is significant. The overall increase in skin

conductance can therefore be considered steeper in the second condition and more concentrated during the decision-making process.

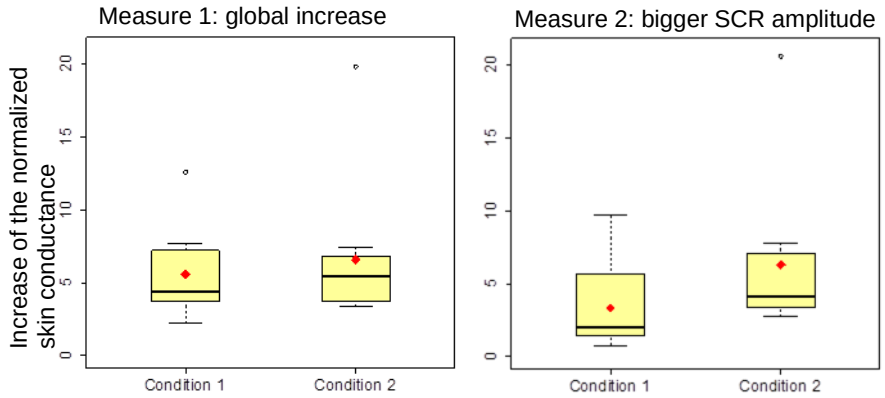


Figure 5. Modification, by measure and condition, of the normalized skin conductance related to the threat detection and fire decision-making process.

In conclusion, the physiological parameters indicate that the early communication strategy slightly reduces stress levels and workload, with a smaller increase in heart rate and an increase in skin conductance that spread out over time.

Testing hypothesis 2

The second hypothesis is that the early communication strategy should improve the operators' performance in making a firing decision. Two indicators are considered here, the validity of the response and the reaction time. In this experiment, a complete detection and identification of all elements in the scene, as well as an accurate application of the rules of engagement, should lead the operator to request a firing clearance from headquarters. Of the three possible actions (fire, request permission, abort), the first is a clear violation of the rules of engagement, the second is the expected one and the third is sub-optimal. Table 1 shows the results of the experiment.

Table 1. Summary of operators' decisions, by condition

	Condition 1	Condition 2
Fire	1	6
Request permission	8	4
Abort	1	0

A χ^2 test with this contingency table indicates that differences between the two conditions are just over the predefined threshold (X-squared = 5.905, df = 2, p-value = 0.052). It is very likely that with some additional participants these differences become significant. From an operational point of view, a majority of operators violated the rule of engagement in experimental condition 2, which is simply not acceptable.

Let us now look at the response time, calculated as the time between the request for a shot by the UCAV and the moment when the operator starts an action (fire, request permission, or abort). The t test (t = -1.409, df = 18, p-value = 0.176, Cohen's d = 0.630) indicates that the difference is not significant. From an operational point of view, the average response time is 4.5 seconds shorter in condition 1 (Fig. 6), which can be a real advantage in case of hostile enemy reaction.

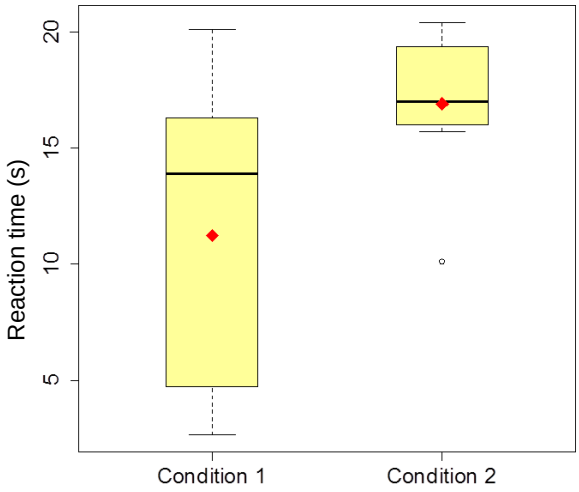


Figure 5. Reaction time by condition

Thus, these results broadly support the early communication strategy (C1) where better decisions are made in a shorter time. Nevertheless, other operators should be taken into consideration to confirm the robustness of these results.

Discussion

This study focuses on issues related to supervision and critical cooperative decision making involving an operator and a highly automated system. With the increasing level of automation and autonomy and the diffusion of more and more sophisticated objects in everyday life, it is of interest to better understand and assess their impact on human decisions. A plausible military scenario was chosen to outline a situation where the operator has to take, under strong time pressure, the responsibility of a critical decision that has been largely prepared by the partially autonomous system. Implemented in an immersive UCAV simulator, an ecological experiment was conducted to better understand the impact of the UCAV's communication strategy on the decision-making process and the overall monitoring of the automated system. This article only reports the results related to critical decision making.

Firstly, the physiological parameters recorded indicate that the decision-making process is always accompanied by an increase in electrodermal response and heart rate, reflecting an increase in stress level and workload. However, when the autonomous system warns the operator that a major decision has to be made soon, the increase in stress level (as captured by the electrodermal response) is spread over the

available time. Heart rate analysis indicates that the overall increase in stress and workload appears to be slightly lower in this condition.

Secondly, the communication strategy has a real impact on the final decision and far fewer wrong decisions are made when the operator knows a few minutes before that he will have to assess a situation and make a decision. With the early communication strategy, the operator has a better perception of the elements of the battlefield and makes more accurate use of the rules of engagement. In addition, the reaction time is shorter.

It is also found that the warning process modifies the operator's level of alertness and vigilance. Thus, although the time windows in which the operator can acquire the required information and construct his decision are exactly the same, the operator is more effective when he/she has been prepared to act. An interesting observation is also that, in the late communication strategy, the wrong decision was always to fire (rather than to abort the attack), as if the strong time pressure pushed the operator to follow the system's decision. Such a result needs to be studied with other experiments, but it is consistent with the notion of complacency towards automation that has already been studied in the aeronautical field. Finally, analyses of QUASA, NASA-TLX and debriefing data (not presented in this paper) indicate that there is no real difference in the overall mission. Users do not report that the late communication strategy is more uncomfortable and do not realise that their decisions were not in line with the rules of engagement.

Further studies are now needed to reinforce these results but also to determine how to reduce complacency towards automation. One perspective is to work on how the autonomous system can better 'explain' to the end user what is relevant to the decision.

References

Barnes, M.J., & Evans A.W. (2016). Soldier-Robot Teams in Future Battlefields: An Overview. In F. Jentsch, and M. Barnes (Eds.), *Human-Robot Interactions in Future Military Operations* (pp. 9-30), New York: Routledge.

Boucsein, W., Fowles, D.C., Grimnes, S., Ben-Shakhar, G., Roth, W.T., Dawson, M.E., & Filion, D.L. (2012). Publication Recommendations for Electrodermal Measurements. *Psychophysiology*, *49*, 1017–1034.

Endsley, M.R. (2017). From Here to Autonomy: Lessons Learned From Human–Automation Research. *Human Factors*, *59*, 5–27.

Eriksson, A., & Neville A.S. (2017). Takeover Time in Highly Automated Vehicles: Noncritical Transitions to and From Manual Control. *Human Factors, 59,* 689–705.

Hancock, P.A., & Szalma, J.L. (2008). Stress and Performance. In P. Hancock and J. Szalma (Eds.), *Performance under Stress* (pp. 1-18), Aldershot: Ashgate.

Hart, S.G., & Staveland, L.E. (1988). Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. *Advances in Psychology, 52,* 139-183.

Huang, G., & Pitts, B.J. (2022). Takeover Requests for Automated Driving: The Effects of Signal Direction, Lead Time, and Modality on Takeover Performance. *Accident Analysis & Prevention, 165,* 106534.

Mayer, M. (2015). The New Killer Drones: Understanding the Strategic Implications of next-Generation Unmanned Combat Aerial Vehicles. *International Affairs 91*, 765–780.

McGuinness, B. (2004). Quantitative Analysis of Situational Awareness (QUASA): Applying Signal Detection Theory to True/False Probes and Self-Ratings. In ICCRTS. Copenhagen, Dk.

Mirri, S., Prandi, C., & Salomoni, P. (2019). Human-Drone Interaction: State of the Art, Open Issues and Challenges. In *MAGESys'19 Proceedings* (pp. 43–48). Beijing, China: ACM.

Morgan, P.L., Alford, C., & Parkhurst G. (2016). *Handover Issues in Autonomous Driving: A Literature Review* (Project Report). Bristol, UK : University of the West of England.

Morgan, P.L., Alford, C., Williams, C., Parkhurst, G., & Pipe, T.. (2018). Manual Takeover and Handover of a Simulated Fully Autonomous Vehicle Within Urban and Extra-Urban Settings. In N. Stanton (Eds.) *Advances in Human Aspects of Transportation* (pp. 760–771). Cham: Springer.

Olsen, D.R., & Goodrich, M.A. (2003). Metrics for Evaluating Human-Robot Interactions. In PerMIS'03 Worksop Proceedings. Gaithersburg, MD, USA.