Effects of Scene Detection, Scene Prediction, and Maneuver Planning Visualizations on Trust, Situation Awareness, and Cognitive Load in Highly Automated Vehicles

MARK COLLEY, Institute of Media Informatics, Ulm University, Germany
MAX RÄDLER, Institute of Media Informatics, Ulm University, Germany
JONAS GLIMMANN, Institute of Media Informatics, Ulm University, Germany
ENRICO RUKZIO, Institute of Media Informatics, Ulm University, Germany

The successful introduction of automated vehicles (AVs) depends on the user's acceptance. To gain acceptance, the intended user must trust the technology, which itself relies on an appropriate understanding. Visualizing internal processes could aid in this. For example, the functional hierarchy of autonomous vehicles distinguishes between perception, prediction, and maneuver planning. In each of these stages, visualizations including possible uncertainties (or errors) are possible. Therefore, we report the results of an online study (N=216) comparing visualizations and their combinations on these three levels using a pre-recorded real-world video with visualizations shown on a simulated augmented reality windshield. Effects on trust, cognitive load, situation awareness, and perceived safety were measured. *Situation Prediction*-related visualizations were perceived as worse than the remaining levels. Based on a negative evaluation of the visualization, the abilities of the AV were also judged worse. In general, the results indicate the presence of overtrust in AVs.

CCS Concepts: • **Human-centered computing → Empirical studies in HCI**.

Additional Key Words and Phrases: Autonomous vehicles; self-driving vehicles; semantic segmentation; intention prediction; machine learning.

ACM Reference Format:
Mark Colley, Max Rädler, Jonas Glimmann, and Enrico Rukzio. 2022. Effects of Scene Detection, Scene Prediction, and Maneuver Planning Visualizations on Trust, Situation Awareness, and Cognitive Load in Highly Automated Vehicles. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 6, 2, Article 49 (June 2022), 21 pages. https://doi.org/10.1145/3534609

1 INTRODUCTION

Automated vehicles (AVs) are expected to change mobility greatly [25]. The passenger can engage in a wide variety of non-driving-related activities [19] such as reading and working or, in fully automated driving, even sleeping [59].

However, issues of novel technology affecting personal safety such as under- or overtrust become apparent. Schöttle and Sivak found that 75% were at least slightly concerned about system failure in unexpected situations [63] and Kyriakidis et al. [47] report that the reliability of these AVs worries potential users. Without sufficient trust (i.e., undertrust), usage of this potentially life-saving technology could be scarce. Prior work investigated the effect of highlighting other vehicles under bad weather conditions [70], pedestrian intention [10],

Authors' addresses: Mark Colley, mark.colley@uni-ulm.de, Institute of Media Informatics, Ulm University, Ulm, Germany; Max Rädler, max.raedler@uni-ulm.de, Institute of Media Informatics, Ulm University, Ulm, Germany; Jonas Glimmann, jonas.glimmann@uni-ulm.de, Institute of Media Informatics, Ulm University, Ulm, Germany; Enrico Rukzio, enrico.rukzio@uni-ulm.de, Institute of Media Informatics, Ulm University, Ulm, Germany.

2474-9567/2022/6-ART49
https://doi.org/10.1145/3534609

and general object detection by displaying the result of the semantic segmentation task [11] to address these worries. In contrast, overtrust leads to over-usage under-monitoring and could result in abuse of such systems.

In modern functional hierarchies of AVs [21, 44, 66], the input of cameras, radars, lidars, maps, and GPS is used to define an environmental model (see Figure 1). For this model, the objects in the situation first have to be detected ("Situation Detection"). Subsequently, the most likely future states have to be determined ("Situation Prediction"). Finally, this can be integrated into the AV's "Maneuver Planning". Providing information to the user on any of these levels could increase and calibrate trust and enhance the technical maturity assessment of the vehicle. As there are uncertainties in each of the levels, an appropriate visualization is required. On the level of *Situation Detection*, Colley et al. [11] recently employed the output of the semantic segmentation task to visualize the AV's recognized objects. Semantic segmentation is used to gain information on objects in a scene by attributing every pixel of an image to a class (e.g., vehicle or pedestrian) and is, therefore, "an enabling factor for a wide range of applications" [17, p. 1] such as AVs. With regards to *Situation Prediction*, Colley et al. [10] evaluated the use of distinctive states of pedestrian intention finding that an Augmented Reality (AR) visualization was preferred and increased trust in the automation. Kunze et al. [45] found that especially hue conveyed uncertainty in the planned AV's trajectory. While each of these visualizations alone can calibrate trust and enhance the assessment of AVs, currently, there is no systematic comparison of these.

Therefore, we designed and conducted an online between-subjects video-based survey (N=216). We used state-of-the-art neural networks for the semantic segmentation and the pedestrian intention detection task. This enabled us to avoid potential biases in our understanding of current approaches. As there are no pure vision-based approaches to estimating the other vehicle's trajectories, we created the visualization of the other and own vehicles' trajectories manually.

Contribution Statement: (1) A framework on visualization levels based on the functional hierarchy of AVs. (2) The findings of an online study (N=216) based on a video of a real-world ride visualized using a state-of-the-art semantic segmentation model [8], state-of-the-art pedestrian intention recognition [55], and manually integrated trajectory predictions for the other road users as well as the ego trajectory. Results show that *Situation Prediction*-related visualizations were perceived worse and that the results indicate the presence of overtrust in AVs.

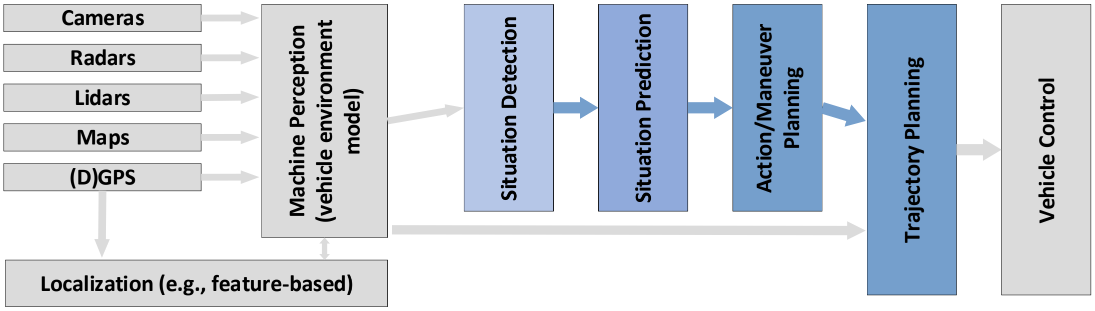


Fig. 1. Functional hierarchy of AVs based on [21].

2 IN-VEHICLE VISUALIZATIONS

This work builds on previous work on visualizations in (semi-)automated vehicles. While there are works on visualizing intentions towards pedestrians and other vulnerable road users to develop trust and avoid misunderstandings (e.g., see [9, 14–16, 20, 36]), we focus on the inward visualization.

2.1 Head-Up vs. Head-Down Displays

Head-Up Displays (HUDs) can avoid driver diversion by removing the necessity to look down because of the spatial proximity of visualizations to their intended location and via sources of available information [29]. However, challenges such as visual clutter and driver distraction could negatively impact driving performance [29]. Compared to Head-Down Displays (HDDs), HUDs increased manual driving performance measures (lateral and longitudinal control) [64]. However, current HUDs are relatively small (e.g., Volkswagen's HUD comprises a virtual screen size of 217 x 88 mm [1]). In the future, windshield Displays (WSDs) covering the entire windshield are envisioned. The goal is to show relevant content at continuous depth [31]. Nonetheless, technical challenges such as parallax effects remain. There already exist design spaces focusing on WSDs [30]. Here, Haeuslschmid et al. [30] define the categories Safety, Vehicle Monitoring, Navigation & geo information system, as well as Entertainment & Communication. Our proposed visualization fall under the category of Vehicle Monitoring with the subcategory of Supervision, however, not solely on the sensors, but also on the relevant algorithms. Regarding their design dimensions, our visualizations can be broadly categorized as passenger-focused (User dimension), Vehicle Monitoring (Context dimension), and 3D registered Augmented Reality (Visualization dimension) without interaction (Interaction dimension). The technology dimension is not applicable.

2.2 Visualized Information Types

Previous work evaluated different ways to communicate decisions, detections, destination, regulation, and navigation. Löcken et al. [50] inform the user of the AV's decisions via ambient light. In line with this work, Wilbrink et al. [69] proposed to use light strips indicating intention or perception. Lindemann et al. [49] used an AR WSD to highlight potential threats such as pedestrians. They provided a cube over moving vehicles indicating their behavior (e.g., dangerous or unusual). This resulted in higher situation awareness in low and high visibility scenarios than only having the basic elements *speed* and *navigation info*.

Calibrated trust [56] refers to a state where the user's trust is appropriate to the capabilities of the automated system. This avoids issues associated with over- and undertrust. Koo et al. [39] investigated the effect of *how* and *why* information for semi-autonomous vehicles. Explanatory information (i.e., *why* information) led to the highest trust. Providing information on *how* the vehicle behaves could lead to cognitive overload [39]. Nonetheless, combining the messages led to the safest driving behavior. Häuslschmid et al. [33] showed the vehicle's current situation interpretation via a world in miniature or a simulated chauffeur avatar. Trust was increased most by the world in miniature. Participants' opinions varied strongly about whether such a visualization was necessary. Colley et al. [10] compared the visualization of pedestrian intention in a Virtual Reality (VR) study between a tablet-based and an AR version simulating a WSD. Regarding cognitive load, the AR version was significantly higher rated. Currano et al. [18] also tested an AR HUD. They compared no HUD with a minimal and a complex one. Results were diverse and dependent on the dynamicity of the scene and the reported driving style of the participants. Therefore, the authors conclude that the HUD should be adaptive to both of these factors. Regarding the future trajectory of the ego-vehicle, Schneider et al. [62] evaluated explanations given via an AR WSD and a LED strip. They found that user experience for these first-time users increased with the explanations. However, the combination with a post-explanation via a smartphone app did not increase the user experience. Colley et al. [13] also evaluated an abstract representation of the perceived objects and showed that no futuristic visualization is necessary but that already current HUDs could provide appropriate information.

2.3 Visualization of Automation Uncertainty

Beller et al. [2] investigated driver-automation interaction. They were interested in whether conveying automation uncertainty could improve the interaction using a simple anthropomorphic symbol when system limits occurred. SA and trust were higher when uncertainty information was displayed. Another abstract uncertainty representation (bars indicating the fidelity of being able to operate) was used by Helldin et al. [34]. In line with Beller et al. [2], they found that the users took over control quicker. However, participants trusted the automation less when shown uncertainty information. Finally, Kunze et al. [45] used AR to present uncertainties of longitudinal and lateral control. Comparing visual variables, the authors found that hue especially conveys urgency. Kunze et al. [46] also argued against using the instrument cluster to visualize uncertainty information because it could increase workload. Therefore, they used a light strip as a peripheral cue and a vibrotactile seat. Hereby, users could pay more attention to the road. Recently, Colley et al. [11] investigated the effects of visualizations of the semantic segmentation task. They argue that these studies use abstract representations which do not enable the user to identify the cause for this uncertainty. They found that their simulated AR WSD did not increase trust or mental workload. However, the subjective situation awareness was higher, and users rated recognition-related attributes significantly better.

2.4 Trust in Automated Vehicles

Lee and See define trust “as the attitude that an agent will help achieve an individual’s goals in a situation characterized by uncertainty and vulnerability” [48, p. 51]. In their model of trust, there is a feedback loop with steps between automation and a user (i.e., the passenger). The information about the automation (e.g., communicated via a display), may vary. The trustor, i.e., the passenger, assimilates this information, and the process of belief-formation occurs, followed by a trust evolution. If sufficient trust is achieved, the trustee intents to rely on the automation. Otherwise, the model proposes to loop back in the *Information assimilation and Belief formation* phase. A trustee has three dimensions relevant for trust-building: (1) *Performance* (i.e., how well does the automation perform a task), (2) *process* (i.e., how convenient is the approach taken), and (3) *purpose* (i.e., usage based on the design purpose). In automated driving, the most relevant recommendation of Lee and See [48] is, therefore, to “show the process and algorithms of the automation by revealing intermediate results” [48, p. 74] to make the automation understandable.

Hoff and Bashier define trust as “a variable that often determines the willingness of human operators to rely on automation” [35, p. 407]. Their three-layered trust model includes dispositional, situational, and learned trust. *Dispositional trust* includes the personal background (e.g., culture, age, gender, and personality traits) of the trustor. *Situational trust* is divided into internal and external variability. External variability refers to changes occurring with altered automation complexity. Internal variability refers to the trustor’s mental capacity and psychological state. *Learned trust* is modelled in two layers: initially learned trust (trust based on prior knowledge about the automation) and dynamically learned trust (altered through interaction with the automation). In their model, trust is modelled as a loop with three elements. The *Dynamic Learned Trust* influences reliance on the automation, thereby influencing how the system performance is viewed. This view influences the *Dynamic Learned Trust*. The automation’s design features also influence this view. Additionally, at the initial interaction, the user begins with an initial reliance strategy.

Körber [40] bases his trust definition on Mayer et al. [53] and Lee and See [48]. According to Körber, Competence / Reliability (see the dimension Ability [53] or Performance [48]), Understandability / Predictability (see the dimension Integrity [53] or Process [48]), and Intention of Developers (see the dimension Benevolence [53] or Purpose [48]) influence trust. Additionally, Körber includes the factors Familiarity (with similar systems) and Propensity to Trust (i.e., how a trustor trusts automation in general; see Dispositional trust of Hoff and Bashir [35]).

Körber [40] incorporate two or more questions with five-point Likert scales for each category. As Körber grounds these questions in the models of Lee and See [48] and Mayer et al. [53], we chose this questionnaire for our evaluation.

Conclusion: While there exists previous work regarding increasing and calibrating trust, there currently is little work that focuses on directly visualizing the capabilities of the AV. Especially the coherence between abstraction of the information (i.e., *Situation Detection*, *Situation Prediction*, and *Maneuver Planning*) and trust is missing. Therefore, we implemented, combined, and compared visualization on these three functional levels.

3 CONCEPTS

This paper compares visualizations and their combinations regarding the functional hierarchy levels "Situation Detection", "Situation Prediction", and "Maneuver Planning" of AVs [21, 44, 66]. Therefore, we briefly introduce the design for each level. All visualizations are dependent on an AR WSD as this was shown to be superior to tablet-based versions [10, 11].

We propose to visualize objects in all views (i.e., windshield and peripheral or side windows). A user might not be familiar with the multitude of sensors built in an AV (front, rear, sides), thus, we expect that showing these detections calibrates trust. Less distracting methods of visualizing such as light bands (e.g., see Wilbrink et al. [69]) could be used when few objects have to be highlighted; however, our concept proposes a more granular possibility of highlighting all driving task-relevant traffic objects. We assume not visualizing objects could lead to the assumption that the AV did not detect them. Using a lightband is thus not feasible as multiple objects (e.g., pedestrians) might overlap when having the same angle in relation to the AV.

3.1 Situation Detection

On this level, we propose to visualize, in line with [6] and [11], the detection of pedestrians, and cyclists (both red), other vehicles (blue), and signposts (yellow) as these influence the AV's trajectory. For the semantic segmentation task, we used the state-of-the-art neural network Panoptic Deeplab by Cheng [7]. As Colley et al. [11, p. 3] stated: "Displaying semantic segmentation [...] encodes the uncertainty information for this task as only detected objects are visualized."

3.2 Situation Prediction

For the situation prediction, we again used the other vehicles as well as the pedestrians and bicyclists. For predicting pedestrians and bicyclists, we employed the state-of-the-art neural network by Mordan et al. [55]. This estimates the intention to cross the street. For the visualization, we opted for circles above the people as described by Colley et al. [10]. A yellow circle indicates that the AV did not clearly recognize the intention. Light blue was used if the intention was to stay on the sidewalk, and dark blue if the intention was to cross the street. As there are no vision-based approaches to vehicle intention and trajectory prediction, we inlaid the videos using Adobe Premiere Pro 2020. We chose to visualize intention classes instead of trajectories for pedestrians because for the AV's planning, primarily, the intention to cross is necessary. The predicted trajectory on the sidewalk, for example, is not necessary.

Trajectories were visualized as successive arrows. In addition, trajectories have a color gradient, which ranges from blue to pink. The bluer the color, the more certain the AV is about the predicted trajectory. The trajectories were developed bearing the previous path and the velocity in mind, which are also used by recent approaches using deep learning methods (e.g., Fernandez et al. [27], Ma et al. [52]).

3.3 Maneuver Planning

The information about the positions of other road users and their predicted intentions are used to determine the AV's trajectory (e.g., [65]). Therefore, we used the same visualization as for the trajectory prediction for other vehicles for the ego trajectory. This visualization was shown directly in front of the ego vehicle (see Figure 2 (4)).

4 ONLINE SIMULATION

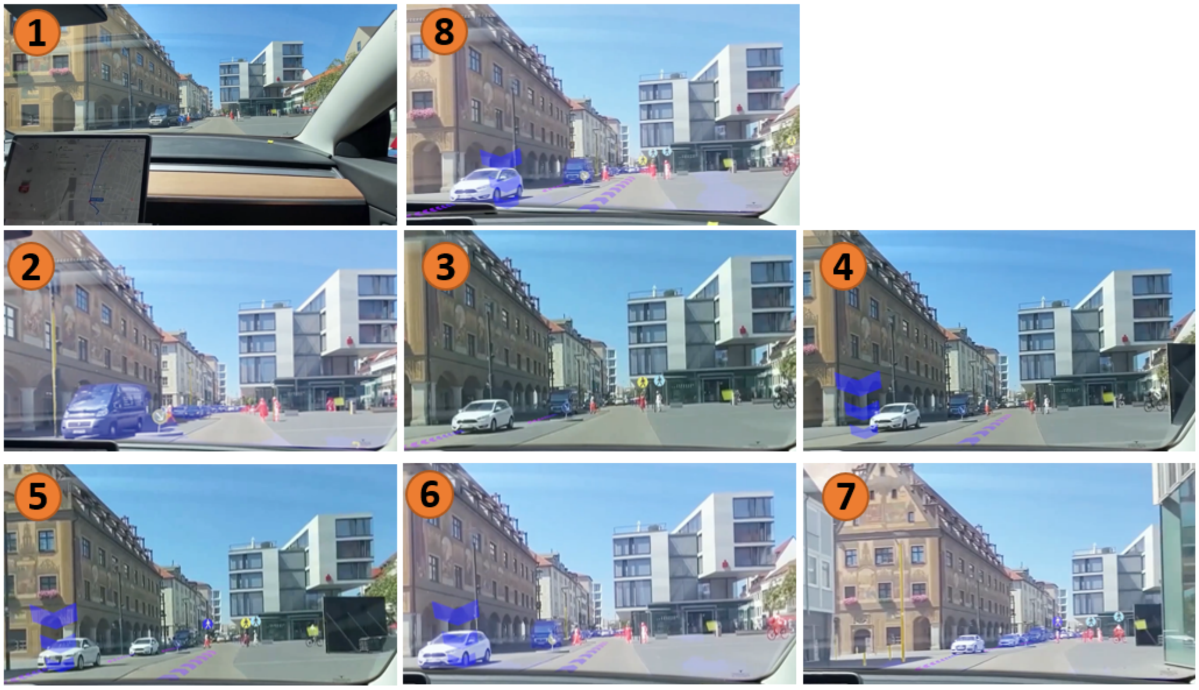


Fig. 2. Screenshots from the videos shown to participants. (1) shows the *baseline* with no visualization. (2) shows *Situation Detection*, (3) *Situation Prediction*, and (4) *Maneuver Planning*. (5) depicts *Situation Prediction and Maneuver Planning*, (6) *Situation Detection and Maneuver Planning*, and (7) *Situation Detection and Prediction*. (8) shows all combined (*Situation Detection & Prediction and Maneuver Planning*). Figures (2) to (8) were cropped for better visibility.

To evaluate the concepts, we designed and conducted a video-based online between-subject study. The following research question guided this exploratory study:

> *What impact do the "visualized objects" have on passengers in an AV in terms of (1) cognitive load, (2) trust, (3) perceived safety, (4) preference, (5) subjective situation awareness*, and (6) capability assessment?

4.1 Materials

We recorded a video in Ulm, Germany with an iPhone 11 Pro Max with 30 fps in wide-angle and Full HD (1080p) resolution for the questionnaire. We anonymized the videos (faces and license plates). In the video, the ride through a busy inner city is shown. Several people of varying age groups are walking at the side of the street. Additionally, some cross the street in front of the vehicle. At the end of the video, a parked vehicle merges into traffic. Therefore, according to Kaß et al. [38], the vehicle performs lateral and longitudinal maneuvers. The scene is rather complex due to the multitude of other traffic participants. Due to technical limitations, we had to take the video from the passenger seat.

4.2 Procedure

Every participant was randomly assigned to one of **eight** conditions. This represents a 2 x 2 x 2 design (*Situation Detection*, *Situation Prediction*, and *Maneuver Planning* all with the levels shown vs. not shown; the *independent* variables; for a description, see Section 3).

Each session started with a brief introduction, agreeing to the consent form, and a demographic questionnaire. The introduction to the capabilities of the AV was:

You will see a video of a driving session in a highly automated vehicle. The vehicle takes over lateral and longitudinal control (braking, accelerating, steering). The vehicle attempts to assess the scene and determine the intent of nearby pedestrians and cars. While watching the video, you are supposed to imagine sitting in such an automated vehicle, follow the entire journey attentively, and then assess it.

After the assigned condition, participants answered the questionnaires described below. Lastly, participants were asked for general feedback. On average, a session lasted 14 min. Participants were compensated with 1.50€. A script running in the background ensured the window was maximized. It also prevented participants from skipping or rerunning the video. This ensured equal exposure time. Additionally, we checked that at least participants used a (required) Full HD (1080p) monitor.

4.3 Measurements

We measured cognitive load using the mental workload subscale of the raw NASA-TLX [32] on a 20-point scale ("How much mental and perceptual activity was required? Was the task easy or demanding, simple or complex?"; 1=*Very Low* to 20=*Very High*) and situation awareness using the situation awareness rating technique (SART) [68]. We used the SART to assess the perceived quality of situation awareness [23]. This is a predictor of "how a person will choose to act on that situation awareness" [23, p. 86]. With high qualitative situation awareness, we would expect AV users to be less inclined to take over control with its post-automation effects [3, 54]. Therefore, automation benefits remain available. We employed the subscales *Predictability/Understandability* (*Understandability*) and *Trust* of the *Trust in Automation* questionnaire by Körber [40]. Understandability is measured using agreement on four statements ("The system state was always clear to me.", "I was able to understand why things happened."; two inverse: "The system reacts unpredictably.", "It's difficult to identify what the system will do next.") using 5-point Likert scales (1=*Strongly disagree* to 5=*Strongly agree*). To measure trust, participants indicate their agreement on the same 5-point Likert scale on two statements ("I trust the system." and "I can rely on the system."). Participants also rated their perceived safety using four 7-point semantic differentials from -3 (anxious/agitated/unsafe/timid) to +3 (relaxed/calm/safe/confident) [24]. Usability was assessed with two items closely related to the System usability Scale Brooke et al. [4] equivalent: "I think that I would like to use these visualizations frequently." and "I found the visualizations unnecessarily complex." (1=*Strongly Disagree* to 5=*Strongly Agree*). We also employed the subscales *Performance*, *Judgement*, and *Reaction* of the Situational Trust Scale for Automated Driving [37].

Participants also rated the AV's capabilities. These were assessed using self-developed single items: driving style (1=completely safe to 7=completely dangerous), object detection (three items: "The automated vehicle recognizes all pedestrians/vehicles/signposts in every situation perfectly"), prediction (two items: "The automated vehicle predicts all pedestrian intentions/vehicle paths in every scene perfectly", as well as lateral and longitudinal guidance on 7-point Likert scales (1=*Totally Disagree* to 7=*Totally Agree*).

Finally, the participant answered questions regarding expected behavior ("The automated vehicle drove as I expected at all times."; "The reasons for the automated vehicle's behavior were clear to me at all times"; "It was always clear what the automated vehicle will do next")

After all conditions, participants could provide open feedback and assessed the reasonability and necessity ("I think the visualization of the recognition of objects is reasonable/necessary)" of the visualizations using single-item ratings on 7-point Likert scales. They were also asked on 7-point Likert scales whether there was visual clutter ("I think there were too many visualizations provided.") and whether they wanted more visualization of *Situation Detection*, *Situation Prediction*, or *Maneuver Planning* ("would have liked more information regarding the automated vehicle's perception/predictions/future path").

Additionally, participants were asked to rate immersion using the *Immersion* subscale of the Technology Usage Inventory (TUI) [42].

5 RESULTS

5.1 Data Analysis

To compare the conditions, we use Kruskal-Wallis tests to account for the between-subject study design. For the factor analysis in the case of non-parametric data, the non-parametric ANOVA (NPAV) as described by Lüpsen [51] was employed. For post-hoc tests, we used Bonferroni correction. R in version 4.1.2 and RStudio in version 2021.09.0 was employed. All packages were up to date in February 2022. Effect sizes were calculated using Rosenthals's formula [61] unless stated otherwise.

We used the package *ggstatsplot* [58] in version 0.9.1. These figures include a boxplot as well as a violin plot showing the data distribution. They also include statistical details (test used, number of observations, effect size, confidence interval). Therefore, we do not rewrite these in text.

5.2 Participants

We computed the required sample size before the experiment via an a-priori power analysis using G*Power [26]. To achieve a power of .8, with an alpha level of .05, 224 participants should result in medium effect size (0.26 [28]) in a one-way ANOVA.

We recruited *N*=241 participants, however, 25 had to be excluded due to failed attention checks, leaving a final sample of *N*=216 (160 female, 52 male, 4 non-binary) via prolific.co. The participant pool was restricted to US citizens to avoid confounding effects of traffic handedness (right-hand vs. left-hand traffic) or culture [60]. Using an online participant database allowed us to circumvent biases found when recruiting mostly from a student population (e.g., see almost three-quarters of CHI publications in 2014 [5]). Participants indicated that their highest educational level was College (164), followed by High School (48) and Vocational Training (4). Regarding their employment status, 104 are employees, 73 are students at a college, 6 are at a school, 21 are self-employed, 10 are job-seeking, and 2 indicated "other". On average, participants were *M*=26.06 (*SD*=7.76) years old. All participants hold a valid driver's license for, on average, *M*=7.43 (*SD*=5.62) years. On 5-point Likert scales (*1 = Strongly Disagree — 5 = Strongly Agree*), participants showed medium interest in AVs (*M*=3.79, *SD*=1.17), believed AVs to ease their lives (*M*=3.88, *SD*=1.07), and believed AVs to become reality by 2031 (*M*=4.16, *SD*=.90). Immersion (measured using the Immersion subscale of the TUI [42]) of participants was moderate (*M*=16.13, *SD*=5.75).

5.3 Cognitive Load and Trust in Automation

A Kruskal-Wallis test found a significant difference between the conditions for the mental workload. However, pairwise comparisons using Dunn's test revealed no significant differences.

The NPAV found a significant main effect of *Situation Prediction* on mental workload (χ^2 (1)=10.78, *p*=0.001). With a prediction (*M*=12.40, *SD*=4.86), mental workload was significantly higher compared to no prediction (*M*=9.89, *SD*=5.49).

The NPAV and the Kruskal-Wallis test (*p*=0.80) found no significant effects on Understandability.

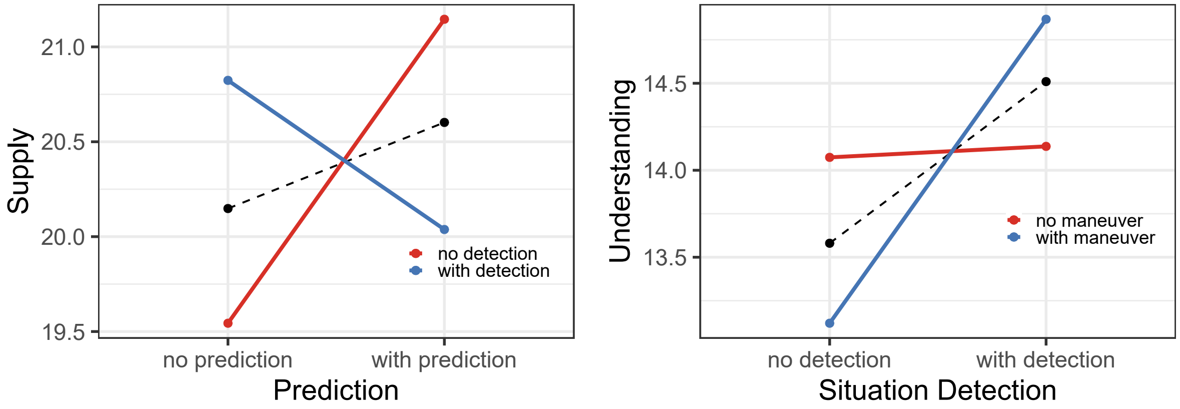
A Kruskal-Wallis test found a significant difference between the conditions for trust. However, pairwise comparisons using Dunn's test showed revealed no significant differences.

The NPAV found a significant main effect of *Maneuver Planning* on Trust (χ^2 (1)=10.12, p=0.001). With no maneuver trajectory (M=3.54, SD=.99), trust was significantly higher than with the maneuver trajectory (M=3.11, SD=.99).

5.4 Situation Awareness

Kruskal-Wallis tests showed no significant differences between the conditions for situation awareness (p=0.51) nor its subscales (Demand: p=0.30; Supply: p=0.29; Understanding: p=0.06).
A one-way ANOVA revealed that there was no significant effect on situation awareness.

The NPAV found a significant main effect of *Situation Prediction* on the subscale Demand (χ^2 (1)=6.20, p=0.013). Demand was higher with (M=15.14, SD=3.39) than without the predictions (M=13.76, SD=3.96). The NPAV found no significant main or interaction effects on the subscales Supply or Understanding.



(a) IE of *Situation Detection* × *Situation Prediction* on Supply

(b) IE of *Situation Detection* × *Maneuver Planning* on Understanding

Fig. 3. IEs on SART subscales.

The NPAV found a significant interaction effect (IE) of *Situation Detection* × *Situation Prediction* on Supply (χ^2 (1)=4.96, p=0.026; see Figure 3a). While the subjectively assessed supply rose with prediction and no detection, the supply was reduced when detection was present.

The NPAV found a significant main effect of *Situation Detection* on Understanding (χ^2 (1)=5.79, p=0.016). The NPAV found a significant IE of *Situation Detection* × *Maneuver Planning* on Understanding (χ^2 (1)=3.98, p=0.046; see Figure 3b). While Understanding remained the same with and without detection when no maneuver was displayed, with a maneuver, it was higher with than without detection.

5.5 Perceived Safety

A Kruskal-Wallis test found a significant difference between the conditions for perceived safety (p=0.04). However, pair-wise comparisons using Dunn's test revealed no significant differences.

The NPAV found a significant three-way IE of *Situation Detection* × *Situation Prediction* × *Maneuver Planning* on perceived safety (χ^2 (1)=6.09, p=0.014; see Figure 4). Without an ego trajectory, the perceived safety declined with a detection visualization when prediction visualization is present. With an ego trajectory, perceived safety

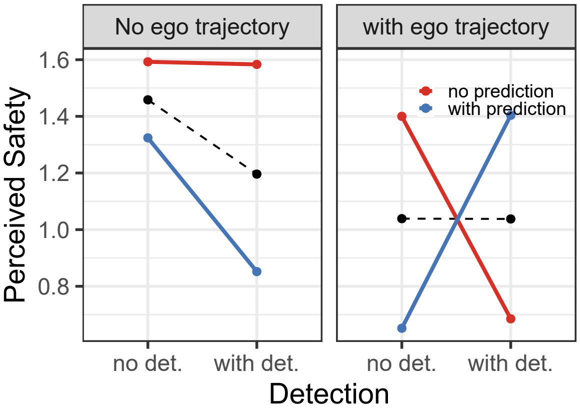
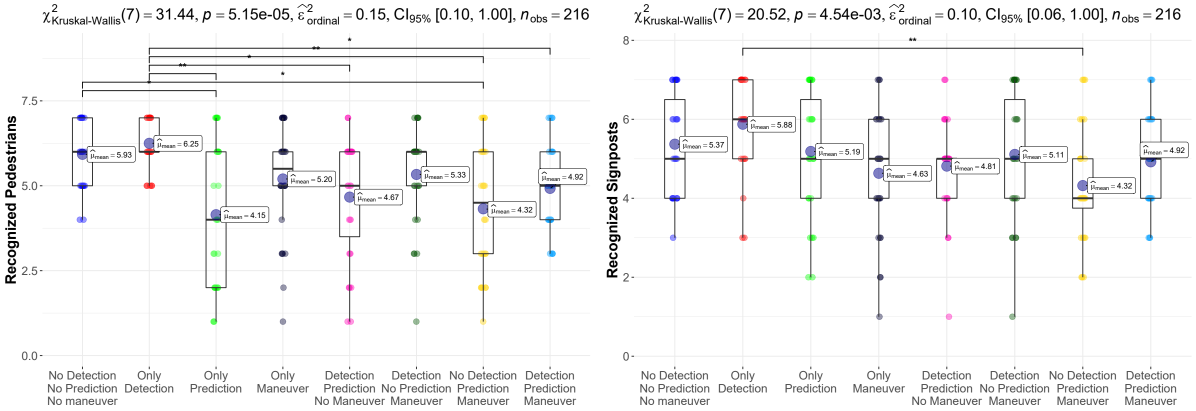


Fig. 4. IE of *Situation Detection* × *Situation Prediction* × *Maneuver Planning* on perceived safety.

declined when detection was but prediction was not visualized, and increased with prediction and detection compared to only prediction.

5.6 Recognition and Prediction Capabilities Assessment



(a) Results on recognized pedestrian assessment.

(b) Results on recognized signposts assessment.

Fig. 5. Results of Kruskal-Wallis tests.

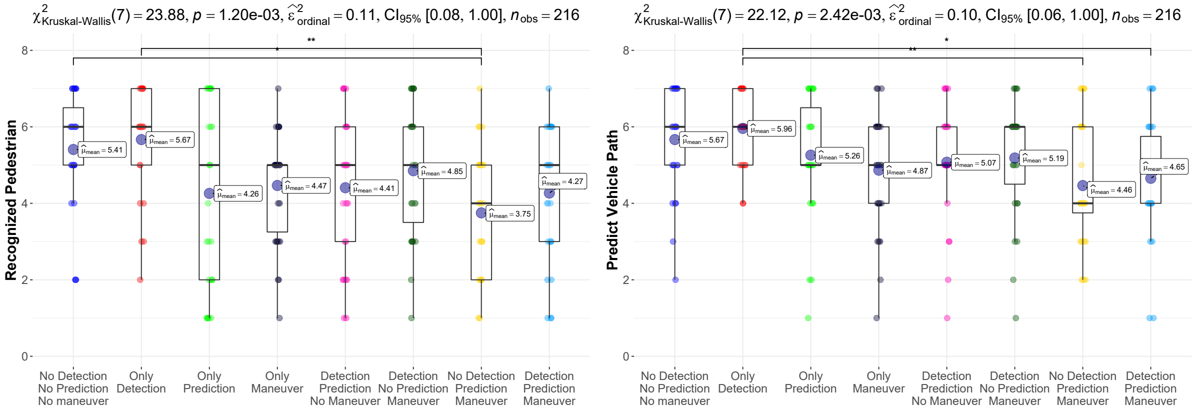
A Kruskal-Wallis test found highly significant differences between the conditions for assessing recognized pedestrians (see Figure 5a). The NPAV found a significant main effect of *Situation Prediction* on assessment of recognized pedestrians (χ^2 (1)=23.79, p<0.001). Participants believed the AV to recognize more pedestrians when the prediction was **not** visualized (M=5.65, SD=1.35) than it being shown (M=4.51, SD=1.80).

A Kruskal-Wallis test found a significant difference between the conditions for assessing recognized vehicles (p=0.01). However, pair-wise comparisons using Dunn's test revealed no significant differences.

The NPAV also found a significant main effect of *Situation Prediction* on the assessment of recognized vehicles (χ^2 (1)=12.01, p<0.001). Again, participants believed the AV to recognize more vehicles without prediction (M=5.73, SD=1.32) than with it being shown (M=5.02, SD=1.62).

A Kruskal-Wallis test found highly significant differences between the conditions for assessing recognized pedestrians (see Figure 5b).

The NPAV found a significant main effect of *Situation Prediction* (χ^2 (1)=5.55, p=0.019) and of *Maneuver Planning* (χ^2 (1)=8.44, p=0.004) on assessment of recognized signposts. In both cases, without the visualization, the scores were higher.



(a) Results on predicted pedestrians assessment.

(b) Results on predicted vehicle path assessment.

Fig. 6. Results of Kruskal-Wallis tests.

A Kruskal-Wallis test found highly significant differences between the conditions for assessing predicted pedestrians (see Figure 6a).

The NPAV found a significant main effect of *Situation Prediction* (χ^2 (1)=12.52, p<0.001) and of *Maneuver Planning* (χ^2 (1)=8.05, p=0.005) on the assessment of predicted pedestrians. In both cases, the values were higher (i.e., better) when no visualization was provided.

A Kruskal-Wallis test found highly significant differences between the conditions for assessing predicted vehicles paths (see Figure 6b).

The NPAV found a significant main effect of *Situation Prediction* (χ^2 (1)=7.91, p=0.005) and of *Maneuver Planning* (χ^2 (1)=12.77, p<0.001) on assessment of predicted vehicle paths (χ^2 (1)=7.91, p=0.005). In both cases, the values were higher (i.e., better) when no visualization was provided.

The NPAV and a Kruskal-Wallis test found no significant effects on the clarity of the next AV action.

5.7 Driving Style

Kruskal-Wallis tests found no significant differences neither for driving style (p=0.48) nor longitudinal (p=0.16) or lateral control (p=0.11).

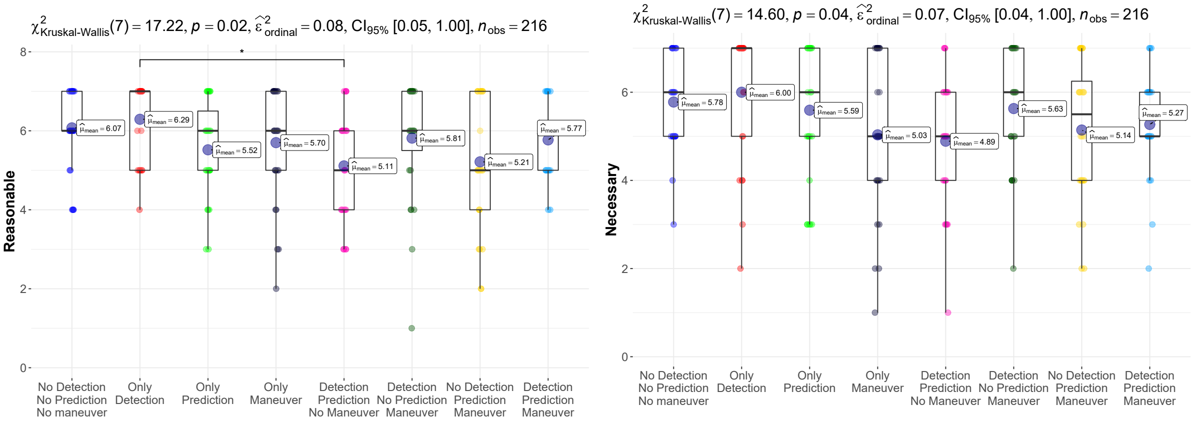
The NPAV found no significant effects on driving style and longitudinal control. The NPAV found a significant main effect of *Situation Prediction* on lateral control (χ^2 (1)=4.72, p=0.03). With no prediction, lateral control (M=5.99, SD=1.06) was rated significantly better than without (M=5.64, SD=1.27).

5.8 Performance, Judgment, Reaction

We measured the items Performance, Judgment, and Reaction of the Situational Trust Scale for Automated Driving [37]. Kruskal-Wallis tests found no significant differences neither for Performance (p=0.39) nor Judgment (p=0.83) or Reaction (p=0.29). The NPAV also found no significant effects on the assessment of AV's Judgement or Reaction.

The NPAV found a significant main effect of *Situation Prediction* on assessment of Performance (χ^2 (1)=5.01, p=0.025). With a prediction visualized, participants believed to have performed better than the AV (M=4.28, SD=1.67) than without (M=3.75, SD=1.78).

5.9 Reasonability, Necessity, and Visual Clutter



(a) Condition-wise comparisons for reasonability.

(b) Condition-wise comparisons for necessity.

Fig. 7. Condition-wise comparisons for reasonability and necessity.

A Kruskal-Wallis test found a significant difference between the conditions for the reasonability (see Figure 7a). The NPAV found a significant main effect of *Situation Prediction* on reasonability (χ^2 (1)=12.31, p<0.001). Without a prediction, the reasonability was rated higher (M=5.95, SD=1.26) than with a prediction (M=5.40, SD=1.32).

The NPAV also found a significant main effect of *Situation Prediction* on necessity (χ^2 (1)=4.59, p=0.032). With no prediction, necessity was rated higher than with prediction visualization.

A Kruskal-Wallis test found significant differences between the conditions for necessity (see Figure 7b). However, pairwise comparisons using Dunn's test revealed no significant differences.

The NPAV found a significant IE of *Situation Detection* × *Situation Prediction* on necessity (χ^2 (1)=4.44, p=0.035; see Figure 8). With a prediction, the necessity for perception was much lower.

In total, both the reasonability (see Figure 7a) and necessity (see Figure 7b) were rated high. Interesting, however, is that the *baseline* also received high values.

A Kruskal-Wallis test found highly significant differences between the conditions for the assessment of visual clutter (see Figure 9a).

The NPAV found a significant main effect of *Situation Prediction* on assessment of visual clutter (χ^2 (1)=37.48, p<0.001). With a prediction, visual clutter was rated significantly higher (M=4.16, SD=1.85) than without (M=2.53, SD=1.74).

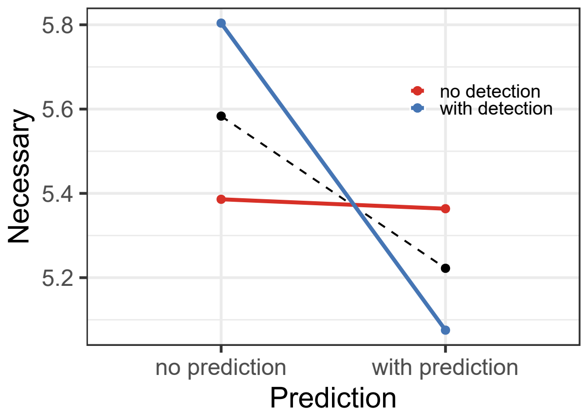
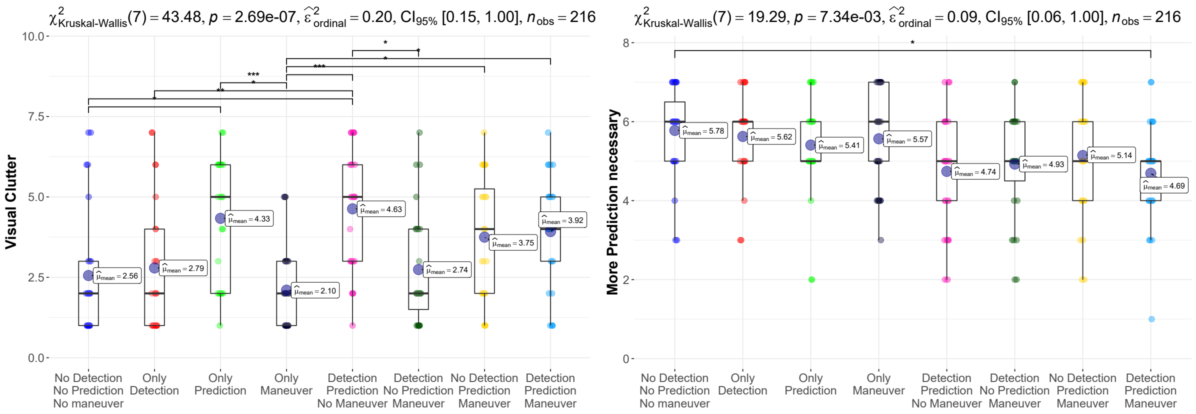


Fig. 8. IE of *Situation Detection* × *Situation Prediction* on necessity.



(a) Condition-wise comparisons for visual clutter.

(b) Condition-wise comparisons for necessity of more prediction-related information.

Fig. 9. Condition-wise comparisons for visual clutter and prediction-related information.

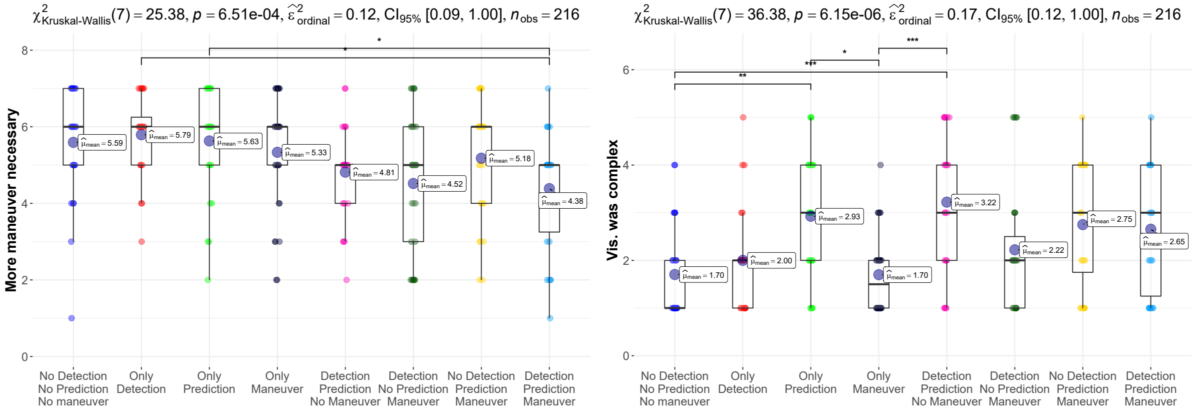
5.10 Need for Additional Information & Usage

A Kruskal-Wallis test found significant differences between the conditions for the necessity to add perception-related information (p=0.002). However, pairwise comparisons using Dunn's test revealed no significant differences.

The NPAV found a significant main effect of *Situation Detection* on need of more perception-related information (χ^2 (1)=19.85, p<0.001). With no perception related information, this need was significantly higher (M=5.28, SD=1.42) than with such visualizations (M=4.33, SD=1.60).

A Kruskal-Wallis test found significant differences between the conditions for the necessity to add prediction-related information (see Figure 9b).

The NPAV found a significant main effect of *Situation Detection* on need of more prediction-related information (χ^2 (1)=7.32, p=0.007). The NPAV also found a significant main effect of *Situation Prediction* on the need for more prediction-related information (χ^2 (1)=7.06, p=0.008). With no perception and no prediction information, the necessity was rated significantly higher.



(a) Condition-wise comparisons for the necessity of more maneuver-related visualization.

(b) Condition-wise comparisons for assessment of visualization complexity.

Fig. 10. Condition-wise comparisons for the necessity of more maneuver-related visualization and visualization complexity.

A Kruskal-Wallis test found significant differences between the conditions for the necessity to add maneuver-related information (see Figure 10a).

The NPAV found a significant main effect of *Situation Detection* on the need for more maneuver-related information (χ^2 (1)=9.54, p=0.002). Likewise, the NPAV found a significant main effect of *Maneuver Planning* on the need for more maneuver-related information (χ^2 (1)=7.28, p=0.007). With no maneuver and no prediction information, the necessity was rated significantly higher.

The NPAV and a Kruskal-Wallis test (p=0.49) found no significant effects on intended frequent usage. The mean values were moderate (between 2.96 and 3.59).

A Kruskal-Wallis test found significant differences between the conditions for visualization complexity (see Figure 10b).

The NPAV found a significant main effect of *Situation Prediction* on visualization complexity (χ^2 (1)=30.08, p<0.001). The visualization of the prediction, while not being rated very complex, was rated significantly more complex (M=2.89, SD=1.32) than not having it (M=1.90, SD=1.08).

A Kruskal-Wallis test found no significant differences between the conditions for AV behavior conformity (p=0.09).

The NPAV found a significant main effect of *Situation Prediction* on AV behavior conformity (χ^2 (1)=3.91, p=0.048). With a prediction, the conformity with expectations was rated lower (M=5.35, SD=1.27) than without (M=5.64, SD=1.26).

The NPAV found a significant IE of *Situation Detection* × *Maneuver Planning* on clarity of reasons more behavior (χ^2 (1)=5.15, p=0.023; see Figure 11). While participants believed that the reasons for the AV's behavior were clear without an ego trajectory, this belief became lesser with the ego trajectory when no detection was shown. With a detection, the belief became higher.

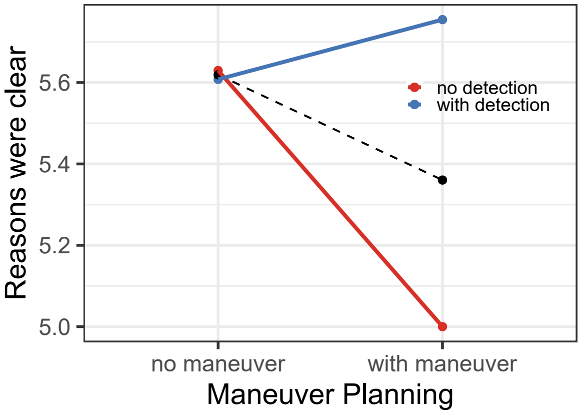


Fig. 11. IE of *Maneuver Planning* × *Situation Prediction* on clarity of reasons.

5.11 Open Feedback

Participants acknowledged that the visualizations helped them understand how an AV works. Participants mentioned that the vehicle mostly did a good job in identifying objects and their intentions but was still worried about what would happen in an unexpected situation ("like a dog running into the street or a nearby driver not paying attention"). Regarding the information level, one participant would have " like[d] more information on how the self automated car actually operates at a fundamental level". Others especially liked the red color for pedestrians as in "signaling that the car perceived them as a risk".

6 DISCUSSION

Compared to previous work (e.g., [10, 11]), our work showed mixed results of visualizing system transparency, for example, on cognitive load. Interestingly, we found that no displaying information about the *Situation Detection*, *Situation Prediction*, or *Maneuver Planning* could lead to higher assessments of the AV than displaying these (e.g., see Figure 6a. The results show, however, that visualization can increase awareness of AV capabilities. Interestingly, providing more information about the AV's internal processes did not consistently lead to higher scores neither in cognitive load, trust, perceived safety, capability assessment, or visual clutter valuations. In this light, we discuss our findings concerning trust calibration and system transparency requirements.

6.1 Calibration of User's Mental Model

For the videos, we employed state-of-the-art object detection [7] and intention determination [55]. Therefore, we were able to portray current AV's capabilities realistically. Despite not combining multiple sensor data (Lidar, radar, vision), these were, probably even better than current on-board computation possibilities as we were under no time constraint for their computation.

The imperfect recognition both for the *Situation Detection* and *Situation Prediction* directly incorporates the capabilities of the system (in contrast to anthropomorphic [2] or abstract levels [34]). While it is difficult to assess whether the mental model of potential users is calibrated, our data shows that additional information visualization on the different levels of the functional hierarchy differently impacts the users. While *Situation Detection* and *Maneuver Planning* related information had little significant impact on the scores of mental workload and situation awareness, *Situation Prediction* visualizations negatively impacted the mental workload, the Demand subscale of the SART, and capability assessment. Perceived safety (see Figure 4) was even highest either with **no** visualization or with only *Situation Detection* visualized.

Interestingly, almost no condition showed significantly higher values for any of the dependent variables compared to the *baseline* (not showing any information). In general, we see two possible explanations for this: either the reliance values (trust, perceived safety) were already calibrated based on previous knowledge or the users relied on the AV despite not knowing nor understanding its internal processes (thereby potentially showing overtrust). As participants only reported medium interest and it is, therefore, unlikely that they have sufficient knowledge about AV internals, we interpret this as a case of overreliance/overtrust in technology [43]. Therefore, we interpret our findings so that the proposed visualizations are appropriate for calibrating users' mental models. Schneider et al. [62] employed comparable visualizations at least for the *Maneuver Planning* visualizations, however, these did not incorporate uncertainty information directly. Until the technology is mature, we believe that visualizations including uncertainty information are more appropriate for users.

6.2 Hierarchy of Data Visualization

In the functional hierarchy of AVs, the later the stage, the more information is present and combined. While no previous information is necessary for the detection, the prediction relies on the detected objects, and for the maneuver, both the perception and prediction of other road users are necessary. Therefore, the later in the hierarchy, the more abstract the visualization is. Other systems that represent uncertainty in AVs use even more abstract representations (anthropomorphic [2] or abstract levels [34]). Abstract information, however, always includes a loss of information.

We could not show consistent effects of the visualization of different levels of the functional hierarchy or the combinations. The reasons for this could be manifold. For example, the visualizations could have been insufficient. Another reason could be that participants were rather forced to evaluate their own (subconscious) beliefs about other road users' intentions, especially pedestrians, with the predicted and visualized intention predictions. This assumption could explain the increased mental workload (see Section 5.3). As inferring intention is difficult [55] and some deductions were wrong, this could have led to a critical assessment of the capabilities and also to a Halo Effect [57], a cognitive bias that lets people deduce system attributes based on impressions of other attributes. This could explain the significantly lower scores in the assessment of lateral control. Nonetheless, our results for trust, for example, were consistent with previous results in similar settings [11]. Our video-based results did differ compared to the VR approach by Colley et al. [10], where only visualizing pedestrian intention lead to trust values of $M = 5.33$ ($SD = 1.68$) [10] while our trust values found were all around ≈3. Therefore, previous work would have suggested stronger effects. Thus, future work also has to account for the medium utilized for measuring, for example, trust.

In summary, we found no dependence between functional levels (i.e., increasing abstraction and complexity) and participants' ratings. However, as the reasons for this may be manifold, this should be further investigated in future work.

6.3 Visualizations as a Tool to Teach the Public

Overtrust in novel technology is a common challenge [43]. Our data support the prevalence of overtrust in AVs as numerous related subjective variables were negatively impacted by the presence of the visualizations which showed an algorithm working in real-time along with the corresponding uncertainties (flickering, etc.). The driving style of the AV was judged better without visualization of *Situation Prediction* even though the displayed driving scene was always the same (see Section 5.7). In addition, participants believed to have better performed than the AV when the *Situation Prediction* was visualized (see Section 5.8). This suggests that the participants expected a better prediction of situations based on current systems.

Therefore, we argue that providing different visualizations can be a viable tool to teach the general public about the current technological state-of-the-art. We believe that it is unlikely that a manufacturer will provide these visualizations to potential customers as these would also show potential inadequacies. However, such honest communication is necessary to calibrate trust and communicate reliable information about current possibilities. This is crucial in the case of non-perfect technology being introduced into public traffic (SAE Level [67] 3 or 4). Therefore, we envision that these visualizations could be used, for example, online to educate the public. As introduction to AVs matter [41], we argue that such a tool could be crucial to provide an unbiased picture of the technology and prevent "autonowashing" [22], that is, a misalignment between actual capabilities and presentation or marketing.

6.4 Visual Clutter & Practical Implications

The system, including the combinations (especially the *Situation Detection & Prediction and Maneuver Planning* system), introduced visual clutter. However, this did not significantly increase cognitive load or decrease trust. We argue that this visual clutter is beneficial for an initial trust-building process as we expect the salient visualizations to catch the user's attention. The proposed visualizations could be employed both for semi-automated (SAE Level 3) and automated driving (SAE Level 4 and 5). In semi-automated driving, the user can then potentially better assess situations in which the AV might not be able to perform sufficiently. In addition, the user can better assess whether a total disengagement from the driving task is appropriate in automated driving.

We were not able to identify significant negative impacts of displaying additional information. Thus, we argue that to assess the internal characteristics of the AV, visualizations of all three levels combined seem to be most appropriate. The user should then be able to turn these on or off selectively. The video employed in the study showed a first-person view from the front-row passenger side. Future work should consider whether the seating (front row vs. back row, left vs. right) alters the assessment of these visualizations.

Therefore, we consider the introduced visualizations to be a suitable tool to visually demonstrate the functionality and functional limitations of AVs, especially to novice users.

6.5 Limitations

The demographic information shows that mostly younger female participants took part in the study. Therefore, transferability to other age groups is not clear. Approximately 10% of our participants were self-employed, which is double the nationwide average in the USA in December 2021[1]. Also, the number of students is higher compared to the national average (approximately 19.6 million in fall 2019)[2]. While the required participant number was determined a priori and is sufficient, the results should be considered keeping the subjective nature of the dependent variables incorporating higher variance in mind. The relatively high number of conditions (eight) could have also prohibited the detection of underlying effects when applying the conservative Bonferroni correction. Additionally, we only focused on subjective dependent measures. While we used state-of-the-art neural networks to determine *Situation Detection* and *Situation Prediction* of pedestrian intentions, we had to manually simulate the ego and the trajectory of the other vehicles. While this was also done by others (e.g., [62]) and was done with relevant criteria (i.e., speed and previous paths) in mind, we can not infer that our simulation corresponds to real-world algorithms.

In future studies, the immersion of the setting could be enhanced by using, for example, virtual reality and simulators with higher degrees of freedom (e.g., [12]).

[1] https://www.bls.gov/news.release/pdf/empsit.pdf, see Table A8; Accessed: 17.01.2022

[2] Back-to-school statistics; Accessed: 17.01.2022

7 CONCLUSION

Overall, we showed the potential of presenting visualizations of the three functional hierarchy levels *Situation Detection*, *Situation Prediction*, and *Maneuver Planning* and their combinations to AV users. We used real-world footage and a state-of-the-art semantic segmentation model [7] as well as a model to determine pedestrian intention [55] evaluate the visualizations with *N*=216 participants. We found that especially *Situation Prediction*-related visualizations were received negatively and negatively impacted the attributed AV's capabilities. Additionally, our data support the presence of overtrust in the AV. Therefore, our work and the proposed visualizations are discussed as a possibility to educate the general public about AVs' capabilities as they directly incorporate inherent uncertainty information. Future work should focus on developing specific guidelines for these visualization concepts.

ACKNOWLEDGMENTS

The authors thank all study participants. This work was conducted within the projects 'Interaction between automated vehicles and vulnerable road users' (Intuitiver) funded by the Ministry of Science, Research and Arts of the State of Baden-Württemberg, as well as 'Semulin' and 'SituWare' funded by the Federal Ministry for Economic Affairs and Energy (BMWi).

REFERENCES

[1] Volkswagen AG. 2020. Head-up-Display. https://www.volkswagen-newsroom.com/de/head-up-display-3957. [Online; accessed: 07-AUGUST-2020].

[2] Johannes Beller, Matthias Heesen, and Mark Vollrath. 2013. Improving the Driver–Automation Interaction: An Approach Using Automation Uncertainty. *Human Factors* 55, 6 (2013), 1130–1141. https://doi.org/10.1177/0018720813482327 arXiv:https://doi.org/10.1177/0018720813482327 PMID: 24745204.

[3] S. Brandenburg and E. M. Skottke. 2014. Switching from manual to automated driving and reverse: Are drivers behaving more risky after highly automated driving?. In *17th International IEEE Conference on Intelligent Transportation Systems (ITSC).* IEEE, New York, NY, USA, 2978–2983. https://doi.org/10.1109/ITSC.2014.6958168

[4] John Brooke et al. 1996. SUS-A quick and dirty usability scale. *Usability evaluation in industry* 189, 194 (1996), 4–7.

[5] Kelly Caine. 2016. Local Standards for Sample Size at CHI. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '16).* Association for Computing Machinery, New York, NY, USA, 981–992. https://doi.org/10.1145/2858036.2858498

[6] Bike Chen, Chen Gong, and Jian Yang. 2018. Importance-aware semantic segmentation for autonomous vehicles. *IEEE Transactions on Intelligent Transportation Systems* 20, 1 (2018), 137–148. https://doi.org/10.1109/TITS.2018.2801309

[7] Bowen Cheng. 2020. Panoptic-DeepLab. https://github.com/bowenc0221/panoptic-deeplab. [Online; accessed: 28-JULY-2020].

[8] Bowen Cheng, Maxwell D. Collins, Yukun Zhu, Ting Liu, Thomas S. Huang, Hartwig Adam, and Liang-Chieh Chen. 2020. Panoptic-DeepLab: A Simple, Strong, and Fast Baseline for Bottom-Up Panoptic Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).* IEEE, New York, NY, USA, 12475–12485.

[9] Mark Colley, Jan Henry Belz, and Enrico Rukzio. 2021. Investigating the Effects of Feedback Communication of Autonomous Vehicles. In *13th International Conference on Automotive User Interfaces and Interactive Vehicular Applications.* Association for Computing Machinery, New York, NY, USA, 263–273. https://doi.org/10.1145/3409118.3475133

[10] Mark Colley, Christian Bräuner, Mirjam Lanzer, Marcel Walch, Martin Baumann, and Enrico Rukzio. 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* (Virtual Event, DC, USA) *(AutomotiveUI '20).* Association for Computing Machinery, New York, NY, USA, 181–191. https://doi.org/10.1145/3409120.3410648

[11] Mark Colley, Benjamin Eder, Jan Ole Rixen, and Enrico Rukzio. 2021. Effects of Semantic Segmentation Visualization on Trust, Situation Awareness, and Cognitive Load in Highly Automated Vehicles. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems.* Association for Computing Machinery, New York, NY, USA, Article 155, 11 pages. https://doi.org/10.1145/3411764.3445351

[12] Mark Colley, Pascal Jansen, Enrico Rukzio, and Jan Gugenheimer. 2022. SwiVR-Car-Seat: Exploring Vehicle Motion Effects on Interaction Quality in Virtual Reality Automated Driving Using a Motorized Swivel Seat. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 5, 4, Article 150 (dec 2022), 26 pages. https://doi.org/10.1145/3494968

[13] Mark Colley, Svenja Krauss, Mirjam Lanzer, and Enrico Rukzio. 2021. How Should Automated Vehicles Communicate Critical Situations? A Comparative Analysis of Visualization Concepts. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 5, 3, Article 94 (sep 2021),

23 pages. https://doi.org/10.1145/3478111

[14] Mark Colley and Rukzio Rukzio. 2020. A Design Space for External Communication of Autonomous Vehicles. In *Proceedings of the 12th International Conference on Automotive User Interfaces and Interactive Vehicular Applications (AutomotiveUI '20)*. ACM, Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3409120.3410646

[15] Mark Colley, Marcel Walch, Jan Gugenheimer, Ali Askari, and Enrico Rukzio. 2020. Towards Inclusive External Communication of Autonomous Vehicles for Pedestrians with Vision Impairments. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3313831.3376472

[16] Mark Colley, Marcel Walch, and Rukzio Rukzio. 2020. Unveiling the Lack of Scalability in Research on External Communication of Autonomous Vehicles. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, Hawaii USA) *(CHI '20)*. ACM, Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3334480.3382865

[17] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. 2016. The Cityscapes Dataset for Semantic Urban Scene Understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, New York, NY, USA, 3213–3223.

[18] Rebecca Currano, So Yeon Park, Dylan James Moore, Kent Lyons, and David Sirkin. 2021. Little Road Driving HUD: Heads-Up Display Complexity Influences Drivers' Perceptions of Automated Vehicles. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, Article 511, 15 pages. https://doi.org/10.1145/3411764.3445575

[19] Joost CF De Winter, Riender Happee, Marieke H Martens, and Neville A Stanton. 2014. Effects of adaptive cruise control and highly automated driving on workload and situation awareness: A review of the empirical evidence. *Transportation research part F: traffic psychology and behaviour* 27 (2014), 196–217.

[20] Debargha Dey, Azra Habibovic, Andreas Löcken, Philipp Wintersberger, Bastian Pfleging, Andreas Riener, Marieke Martens, and Jacques Terken. 2020. Taming the eHMI jungle: A classification taxonomy to guide, compare, and assess the design principles of automated vehicles' external human-machine interfaces. *Transportation Research Interdisciplinary Perspectives* 7 (2020), 100174.

[21] Klaus Dietmayer. 2016. Predicting of machine perception for automated driving. In *Autonomous Driving*. Springer Berlin Heidelberg, Berlin, Heidelberg, 407–424.

[22] Liza Dixon. 2020. Autonowashing: The greenwashing of vehicle automation. *Transportation research interdisciplinary perspectives* 5 (2020), 100113.

[23] Mica R Endsley, Stephen J Selcon, Thomas D Hardiman, and Darryl G Croft. 1998. A comparative analysis of SAGAT and SART for evaluations of situation awareness. In *Proceedings of the human factors and ergonomics society annual meeting*, Vol. 42. SAGE Publications Sage CA: Los Angeles, CA, SAGE Publications, Los Angeles, CA, USA, 82–86.

[24] Stefanie M. Faas, Andrea C. Kao, and Martin Baumann. 2020. A Longitudinal Video Study on Communicating Status and Intent for Self-Driving Vehicle – Pedestrian Interaction. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3313831.3376484

[25] Daniel J Fagnant and Kara Kockelman. 2015. Preparing a nation for autonomous vehicles: opportunities, barriers and policy recommendations. *Transportation Research Part A: Policy and Practice* 77 (2015), 167–181.

[26] Franz Faul, Edgar Erdfelder, Axel Buchner, and Albert-Georg Lang. 2009. Statistical power analyses using G* Power 3.1: Tests for correlation and regression analyses. *Behavior research methods* 41, 4 (2009), 1149–1160.

[27] Jaime B Fernandez, Suzanne Little, and Noel E O'Connor. 2019. A Single-Shot Approach Using an LSTM for Moving Object Path Prediction. In *2019 Ninth International Conference on Image Processing Theory, Tools and Applications (IPTA)*. IEEE, IEEE, New York, NY, USA, 1–6.

[28] David C. Funder and Daniel J. Ozer. 2019. Evaluating Effect Size in Psychological Research: Sense and Nonsense. *Advances in Methods and Practices in Psychological Science* 2, 2 (2019), 156–168. https://doi.org/10.1177/2515245919847202 arXiv:https://doi.org/10.1177/2515245919847202

[29] J. L. Gabbard, G. M. Fitch, and H. Kim. 2014. Behind the Glass: Driver Challenges and Opportunities for AR Automotive Applications. *Proc. IEEE* 102, 2 (2014), 124–136.

[30] Renate Haeuslschmid, Bastian Pfleging, and Florian Alt. 2016. A Design Space to Support the Development of Windshield Applications for the Car. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '16)*. Association for Computing Machinery, New York, NY, USA, 5076–5091. https://doi.org/10.1145/2858036.2858336

[31] Renate Haeuslschmid, Yixin Shou, John O'Donovan, Gary Burnett, and Andreas Butz. 2016. First Steps towards a View Management Concept for Large-Sized Head-up Displays with Continuous Depth. In *Proceedings of the 8th International Conference on Automotive User Interfaces and Interactive Vehicular Applications* (Ann Arbor, MI, USA) *(Automotive'UI 16)*. Association for Computing Machinery, New York, NY, USA, 1–8. https://doi.org/10.1145/3003715.3005418

[32] Sandra G Hart and Lowell E Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Advances in psychology*. Vol. 52. Elsevier, Amsterdam, The Netherlands, 139–183.

[33] Renate Häuslschmid, Max von Bülow, Bastian Pfleging, and Andreas Butz. 2017. Supporting Trust in Autonomous Driving. In *Proceedings of the 22nd International Conference on Intelligent User Interfaces* (Limassol, Cyprus) *(IUI '17)*. Association for Computing Machinery, New York, NY, USA, 319–329. https://doi.org/10.1145/3025171.3025198

[34] Tove Helldin, Göran Falkman, Maria Riveiro, and Staffan Davidsson. 2013. Presenting System Uncertainty in Automotive UIs for Supporting Trust Calibration in Autonomous Driving. In *Proceedings of the 5th International Conference on Automotive User Interfaces and Interactive Vehicular Applications* (Eindhoven, Netherlands) *(AutomotiveUI '13)*. Association for Computing Machinery, New York, NY, USA, 210–217. https://doi.org/10.1145/2516540.2516554

[35] Kevin Anthony Hoff and Masooda Bashir. 2015. Trust in automation: Integrating empirical evidence on factors that influence trust. *Human factors* 57, 3 (2015), 407–434.

[36] Kai Holländer, Mark Colley, Enrico Rukzio, and Andreas Butz. 2021. A Taxonomy of Vulnerable Road Users for HCI Based On A Systematic Literature Review. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, Article 158, 13 pages. https://doi.org/10.1145/3411764.3445480

[37] Brittany E. Holthausen, Philipp Wintersberger, Bruce N. Walker, and Andreas Riener. 2020. Situational Trust Scale for Automated Driving (STS-AD): Development and Initial Validation. In *12th International Conference on Automotive User Interfaces and Interactive Vehicular Applications* (Virtual Event, DC, USA) *(AutomotiveUI '20)*. Association for Computing Machinery, New York, NY, USA, 40–47. https://doi.org/10.1145/3409120.3410637

[38] Christina Kaß, Stefanie Schoch, Frederik Naujoks, Sebastian Hergeth, Andreas Keinath, and Alexandra Neukum. 2020. Standardized Test Procedure for External Human–Machine Interfaces of Automated Vehicles. *Information* 11, 3 (2020), 173.

[39] Jeamin Koo, Jungsuk Kwac, Wendy Ju, Martin Steinert, Larry Leifer, and Clifford Nass. 2015. Why did my car just do that? Explaining semi-autonomous driving actions to improve driver understanding, trust, and performance. *International Journal on Interactive Design and Manufacturing (IJIDeM)* 9, 4 (2015), 269–275.

[40] Moritz Körber. 2019. Theoretical Considerations and Development of a Questionnaire to Measure Trust in Automation. In *Proceedings of the 20th Congress of the International Ergonomics Association (IEA 2018)*, Sebastiano Bagnara, Riccardo Tartaglia, Sara Albolino, Thomas Alexander, and Yushi Fujita (Eds.). Springer International Publishing, Cham, 13–30.

[41] Moritz Körber, Eva Baseler, and Klaus Bengler. 2018. Introduction matters: Manipulating trust in automation and reliance in automated driving. *Applied ergonomics* 66 (2018), 18–31.

[42] Oswald Kothgassner, A Felnhofer, N Hauk, E Kastenhofer, J Gomm, and I Krysprin-Exner. 2013. Technology Usage Inventory. https://www.ffg.at/sites/default/files/allgemeine_downloads/thematische%20programme/programmdokumente/tui_manual.pdf. *Manual. Wien: ICARUS* 17, 04 (2013), 90. [Online; accessed: 05-JULY-2020].

[43] Thomas Kundinger, Philipp Wintersberger, and Andreas Riener. 2019. (Over)Trust in Automated Driving: The Sleeping Pill of Tomorrow?. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI EA '19)*. Association for Computing Machinery, New York, NY, USA, 1–6. https://doi.org/10.1145/3290607.3312869

[44] Felix Kunz, Dominik Nuss, Jürgen Wiest, Hendrik Deusch, Stephan Reuter, Franz Gritschneder, Alexander Scheel, Manuel Stübler, Martin Bach, Patrick Hatzelmann, Cornelius Wild, and Klaus Dietmayer. 2015. Autonomous driving at Ulm University: A modular, robust, and sensor-independent fusion approach. In *2015 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, IEEE, New York, NY, USA, 666–673. https://doi.org/10.1109/IVS.2015.7225761

[45] Alexander Kunze, Stephen J. Summerskill, Russell Marshall, and Ashleigh J. Filtness. 2018. Augmented Reality Displays for Communicating Uncertainty Information in Automated Driving. In *Proceedings of the 10th International Conference on Automotive User Interfaces and Interactive Vehicular Applications* (Toronto, ON, Canada) *(AutomotiveUI '18)*. Association for Computing Machinery, New York, NY, USA, 164–175. https://doi.org/10.1145/3239060.3239074

[46] Alexander Kunze, Stephen J. Summerskill, Russell Marshall, and Ashleigh J. Filtness. 2019. Conveying Uncertainties Using Peripheral Awareness Displays in the Context of Automated Driving. In *Proceedings of the 11th International Conference on Automotive User Interfaces and Interactive Vehicular Applications* (Utrecht, Netherlands) *(AutomotiveUI '19)*. Association for Computing Machinery, New York, NY, USA, 329–341. https://doi.org/10.1145/3342197.3344537

[47] Miltos Kyriakidis, Riender Happee, and Joost CF de Winter. 2015. Public opinion on automated driving: Results of an international questionnaire among 5000 respondents. *Transportation research part F: traffic psychology and behaviour* 32 (2015), 127–140.

[48] John D Lee and Katrina A See. 2004. Trust in automation: Designing for appropriate reliance. *Human factors* 46, 1 (2004), 50–80.

[49] Patrick Lindemann, Tae-Young Lee, and Gerhard Rigoll. 2018. Catch my drift: Elevating situation awareness for highly automated driving with an explanatory windshield display user interface. *Multimodal Technologies and Interaction* 2, 4 (2018), 71.

[50] Andreas Löcken, Wilko Heuten, and Susanne Boll. 2016. AutoAmbiCar: Using Ambient Light to Inform Drivers About Intentions of Their Automated Cars. In *Adjunct Proceedings of the 8th International Conference on Automotive User Interfaces and Interactive Vehicular Applications* (Ann Arbor, MI, USA) *(AutomotiveUI '16 Adjunct)*. Association for Computing Machinery, New York, NY, USA, 57–62. https://doi.org/10.1145/3004323.3004329

[51] Haiko Lüpsen. 2020. R-Funktionen zur Varianzanalyse. http://www.uni-koeln.de/~luepsen/R/. [Online; accessed 25-SEPTEMBER-2020].

[52] Yuexin Ma, Xinge Zhu, Sibo Zhang, Ruigang Yang, Wenping Wang, and Dinesh Manocha. 2019. Trafficpredict: Trajectory prediction for heterogeneous traffic-agents. , 6120–6127 pages.

[53] Roger C Mayer, James H Davis, and F David Schoorman. 1995. An integrative model of organizational trust. *Academy of management review* 20, 3 (1995), 709–734.

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

[55] Taylor Mordan, Matthieu Cord, Patrick Pérez, and Alexandre Alahi. 2021. Detecting 32 Pedestrian Attributes for Autonomous Vehicles. *IEEE Transactions on Intelligent Transportation Systems* (2021), 1–13. https://doi.org/10.1109/TITS.2021.3107587

[56] Bonnie M Muir and Neville Moray. 1996. Trust in automation. Part II. Experimental studies of trust and human intervention in a process control simulation. *Ergonomics* 39, 3 (1996), 429–460.

[57] Richard E Nisbett and Timothy D Wilson. 1977. The halo effect: evidence for unconscious alteration of judgments. *Journal of personality and social psychology* 35, 4 (1977), 250.

[58] Indrajeet Patil. 2021. Visualizations with statistical details: The 'ggstatsplot' approach. *Journal of Open Source Software* 6, 61 (2021), 3167. https://doi.org/10.21105/joss.03167

[59] Bastian Pfleging, Maurice Rang, and Nora Broy. 2016. Investigating User Needs for Non-Driving-Related Activities during Automated Driving. In *Proceedings of the 15th International Conference on Mobile and Ubiquitous Multimedia* (Rovaniemi, Finland) *(MUM '16)*. Association for Computing Machinery, New York, NY, USA, 91–99. https://doi.org/10.1145/3012709.3012735

[60] Amir Rasouli and John K Tsotsos. 2019. Autonomous vehicles that interact with pedestrians: A survey of theory and practice. *IEEE Transactions on Intelligent Transportation Systems* 21, 3 (2019), 900–918.

[61] Robert Rosenthal, Harris Cooper, and L Hedges. 1994. Parametric measures of effect size. *The handbook of research synthesis* 621, 2 (1994), 231–244.

[62] Tobias Schneider, Joana Hois, Alischa Rosenstein, Sabiha Ghellal, Dimitra Theofanou-Fülbier, and Ansgar R.S. Gerlicher. 2021. ExplAIn Yourself! Transparency for Positive UX in Autonomous Driving. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, Article 161, 12 pages. https://doi.org/10.1145/3411764.3446647

[63] Brandon Schoettle and Michael Sivak. 2014. *A survey of public opinion about autonomous and self-driving vehicles in the US, the UK, and Australia*. Technical Report. University of Michigan, Ann Arbor, Transportation Research Institute.

[64] Missie Smith, Joseph L. Gabbard, and Christian Conley. 2016. Head-Up vs. Head-Down Displays: Examining Traditional Methods of Display Assessment While Driving. In *Proceedings of the 8th International Conference on Automotive User Interfaces and Interactive Vehicular Applications* (Ann Arbor, MI, USA) *(Automotive'UI 16)*. Association for Computing Machinery, New York, NY, USA, 185–192. https://doi.org/10.1145/3003715.3005419

[65] Oliver Speidel, Maximilian Graf, Thanh Phan-Huu, and Klaus Dietmayer. 2019. Towards courteous behavior and trajectory planning for automated driving. In *2019 IEEE Intelligent Transportation Systems Conference (ITSC)*. IEEE, IEEE, New York, NY, USA, 3142–3148.

[66] Ömer Şahin Taş, Florian Kuhnt, J Marius Zöllner, and Christoph Stiller. 2016. Functional system architectures towards fully automated driving. In *2016 IEEE Intelligent vehicles symposium (IV)*. IEEE, IEEE, New York, NY, USA, 304–309.

[67] SAE Taxonomy. 2014. *Definitions for terms related to on-road motor vehicle automated driving systems*. Technical Report. Technical report, SAE International.

[68] Richard M Taylor. 2017. Situational awareness rating technique (SART): The development of a tool for aircrew systems design. In *Situational awareness*. Routledge, Abingdon, UK, 111–128.

[69] Marc Wilbrink, Anna Schieben, and Michael Oehl. 2020. Reflecting the Automated Vehicle's Perception and Intention: Light-Based Interaction Approaches for on-Board HMI in Highly Automated Vehicles. In *Proceedings of the 25th International Conference on Intelligent User Interfaces Companion* (Cagliari, Italy) *(IUI '20)*. Association for Computing Machinery, New York, NY, USA, 105–107. https://doi.org/10.1145/3379336.3381502

[70] Philipp Wintersberger, Anna-Katharina Frison, Andreas Riener, and Tamara von Sawitzky. 2019. Fostering User Acceptance and Trust in Fully Automated Vehicles: Evaluating the Potential of Augmented Reality. *PRESENCE: Virtual and Augmented Reality* 27, 1 (2019), 46–62. https://doi.org/10.1162/pres_a_00320 arXiv:https://doi.org/10.1162/pres_a_00320