Effects of Semantic Segmentation Visualization on Trust, Situation Awareness, and Cognitive Load in Highly Automated Vehicles

Mark Colley
mark.colley@uni-ulm.de
Institute of Media Informatics, Ulm University
Ulm, Germany

Benjamin Eder
benjamin.eder@uni-ulm.de
Institute of Media Informatics, Ulm University
Ulm, Germany

Jan Ole Rixen
jan.rixen@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

ABSTRACT

Autonomous vehicles could improve mobility, safety, and inclusion in traffic. While this technology seems within reach, its successful introduction depends on the intended user's acceptance. A substantial factor for this acceptance is trust in the autonomous vehicle's capabilities. Visualizing internal information processed by an autonomous vehicle could calibrate this trust by enabling the perception of the vehicle's detection capabilities (and its failures) while only inducing a low cognitive load. Additionally, the simultaneously raised situation awareness could benefit potential take-overs. We report the results of two comparative online studies on visualizing semantic segmentation information for the human user of autonomous vehicles. Effects on trust, cognitive load, and situation awareness were measured using a simulation (N=32) and state-of-the-art panoptic segmentation on a pre-recorded real-world video (N=41). Results show that the visualization using Augmented Reality increases situation awareness while remaining low cognitive load.

CCS CONCEPTS

• **Human-centered computing** → **Empirical studies in HCI**; *Interaction techniques.*

KEYWORDS

Autonomous vehicles; self-driving vehicles; semantic segmentation.

ACM Reference Format:
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 *CHI Conference on Human Factors in Computing Systems (CHI '21), May 8–13, 2021, Yokohama, Japan.* ACM, New York, NY, USA, 11 pages. **https://doi.org/10.1145/3411764.3445351**

CHI '21, May 8–13, 2021, Yokohama, Japan

ACM ISBN 978-1-4503-8096-6/21/05…$15.00
https://doi.org/10.1145/3411764.3445351

1 INTRODUCTION

With autonomous vehicles (AVs), mobility is expected to change fundamentally [14]. The passenger can engage in a wide variety of non-driving related tasks [11], such as sleeping, reading, or playing games [42]. Additionally, AVs enable higher mobility for the elderly, people with impairments, or people too young to drive. However, with the introduction of AVs, issues of novel technology affecting personal safety such as under- or overtrust emerge. Schöttle and Sivak found that 75% were at least slightly concerned about system failure in unexpected situations [45]. The reliability of AVs is also a worry of potential users [32]. With undertrust, usage of this potent technology could be scarce. Previous work investigated the effect of highlighting, for example, other vehicles under bad weather conditions [50] or pedestrian intention [9] to address these worries. In contrast, overtrust leads to over-usage and could result in abusive usage of such systems. The first prerequisite for safe driving is the detection and recognition of relevant 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" [10, p. 1] such as AVs. While the networks used for semantic segmentation are evaluated statistically on test sets, colorized pictures or videos are shown as qualitative examples [10]. Providing the information on detection and recognition to the vehicle's user could increase and calibrate the trust, improve situation awareness, and enhance the technical maturity assessment of the vehicle. As the user can directly perceive miscategorizations or highly frequent changes, this visualization directly shows when the AV is uncertain about other objects and ,therefore, calibrates trust and maturity assessment instead of trying to increase it.

The work's contributions are: (1) Results of a literature analysis on the used colorization in semantic segmentation visualization. (2) An Augmented Reality (AR) and a tablet-based semantic segmentation visualization technique, and (3) findings of two online studies based on a simulation (N=32) and a video of a real-world ride visualized using a state-of-the-art model [8]. Results show that

the AR visualization reduced cognitive load in the simulation-based study and increased situation awareness (SA).

2 IN-VEHICLE VISUALIZATIONS

In the research field of in-vehicle visualizations, Head-Up Displays (HUDs) are viewed as an approach to avoiding driver diversion. Gabbard et al. [15] highlight their advantages: no need to look down, spatial proximity, and novel sources of available information. Challenges are mainly technical. However, visual clutter and driver distraction could also negatively impact driving performance [15]. Compared to traditional Head-Down Displays (HDDs), HUDs were shown to increase performance measures (lateral and longitudinal control) [46]. Current HUDs are relatively small (e.g., Volkswagen's HUD has a virtual screen size of 217 x 88 mm [1]). Windshield Displays (WSDs) are the next step in the development of HUDs by covering the entire windshield. Finally, the goal is to show content at continuous depth [17].

While this work mainly focuses on aiding the driver of a vehicle, this technology could also be employed in AVs to calibrate user trust and improve SA. This could increase the willingness to use this novel technology. High usage of AV technology is necessary to take advantage of the potential benefits of it [21]. Previous work investigated various communication means to communicate decisions, detections, destination, regulation, and navigation. Löcken et al. inform the user of the decisions of their AV with ambient light [34]. Wilbrink et al. [49] also proposed to use light strips to indicate intention or perception. Lindemann et al. [33] used an AR WSD to highlight threats such as pedestrians and provided a cube over moving vehicles indicating their behavior (e.g., dangerous or unusual). This resulted in higher SA in low and high visibility scenarios than only having the basic elements *speed* and *navigation info*.

Calibrated trust [37] 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. [26] investigated the effect of providing different types of information (*how* and *why* information) for the actions of semi-autonomous vehicles. Explanatory information (i.e., *why* information) led to highest trust. Additionally, providing on *how* the vehicle behaves could lead to cognitive overload [26]. Still, combining both messages resulted in the safest driving behavior. Häuslschmid et al. [19] showed the vehicle's current situation interpretation. This was realized 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 is needed. Colley et al. [9] compared the visualization of pedestrian intention in a VR study between a tablet-based and an AR version. The AR version was implemented as if a WSD was already available. The AR version was significantly better rated by participants in terms of cognitive load.

The visualization of automation uncertainty was less investigated. Beller et al. [2] investigated how conveying automation uncertainty could improve driver-automation interaction. For this, they displayed a simple anthropomorphic symbol when system limits occurred. Results showed that, for takeovers, a longer time-to-collision was available. Additionally, SA and trust were higher when uncertainty information was displayed. Helldin et al. [20] also used an abstract representation of uncertainty. They support the findings that the users took over control quicker, but their participants trusted the automation less when shown uncertainty information. Kunze et al. [30] used AR to present uncertainties of longitudinal and lateral control, i.e., the ability to steer and accelerate/decelerate. This resembles the definition of Kaß et al. [24] who divide AV maneuvers into lateral (driving straight ahead, turning (left, right), and changing the lane, (left, right)) and longitudinal maneuvers (keeping a constant speed, decelerating, and accelerating). At a standstill, these maneuvers are *remaining to stand* (0 km/h), *driving forward*, and *reversing* [24]. Several visual variables were evaluated in a sorting study. Results showed that especially hue conveys urgency. In another study, Kunze et al. [31] argued that having to use the instrument cluster for visualization of uncertainty information can increase workload. Therefore, a light strip as a peripheral cue and a vibrotactile seat were added. Results showed that this enabled users to put more attention on the road. These studies have in common that an abstract representation of uncertainty is used. Several factors could be relevant to this uncertainty and are, therefore, invisible to the user.

3 SEMANTIC SEGMENTATION

Semantic segmentation of pictures describes the process of classifying every pixel of an image. In semantic segmentation, every pixel is attributed a class, for example, "vehicle" or "pedestrian". In instance segmentation, every pixel is attributed to the instance of a class. In panoptic segmentation, these two approaches are combined [8]. Bowen et al. open sourced their implementation of a panoptic segmentation on GitHub [7]. This model is trained on the Cityscapes dataset [10], a large-scale dataset providing annotations to train models for segmentation. This implementation achieves over 80% mean Intersection over Union (mIoU), therefore, "setting the new state-of-art" [8, p. 12475] in June 2020.

To evaluate how semantic segmentation is visualized, in April 2020, we queried the IEEXplore Library, the ACM Digital Library, and Google Scholar (search query: `(Semantic Segmentation) AND (Autonomous Cars)`). We limited the search to the years 01/2012-04/2020. The inclusion criteria were that the paper (1) had to be about semantic segmentation in the AV context and (2) that an example image showing the colorization had to be included. In summary, a total of **157** papers were analysed. **50** contained an example image. The PRISMA statement [36] is shown in Figure 1.

This categorization was done by the first and second author. Disagreements were resolved via discussions. We did not categorize paper for the exact RGB value of the object colorization but instead used a fixed set of colors which are shown here. The results show that there is no common visualization (see Figure 2). However, most work seems to lean towards the visualization as proposed by Cordts et al. for the *Cityscapes* dataset [10]. The lack of standardization can be explained by the goal of research on semantic segmentation: not the colorization but the accuracy of classified pixels is the relevant metric. Visualization is used for demonstration purposes and to make data human-readable and easily accessible.

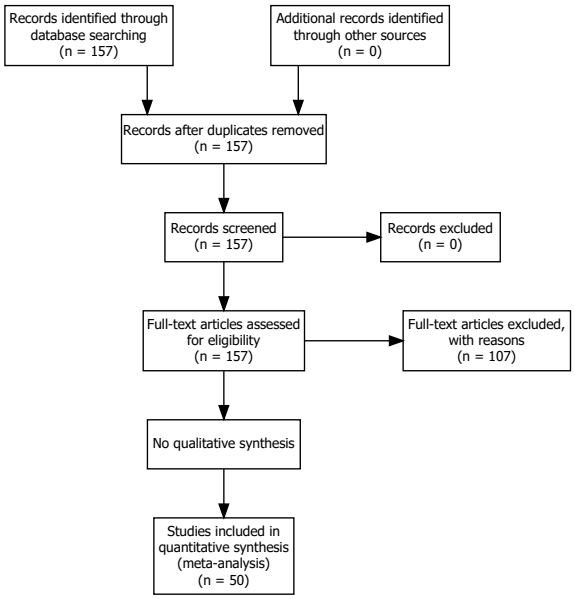


Figure 1: PRISMA [36] statement of paper selection process.

4 CONCEPT

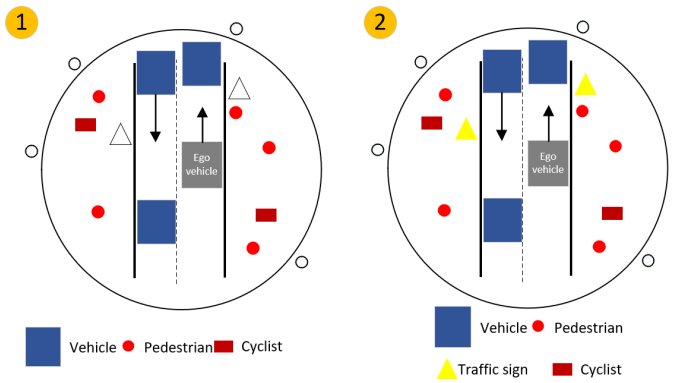


Figure 3: Schematic representation of the visualization: only dynamic (1) and dynamic+static objects (2).

We propose to visualize the results of the segmentation task to the users. Following Chen et al. [6], pedestrians, vehicles, and cyclists are the most important traffic objects for semantic segmentation. Amongst others, signposts are the second most important traffic objects. These can be grouped into *dynamic* and *static* (signposts) objects. These objects have a great impact on the trajectory and in the case of dynamic objects, a misinterpretation could be fatal. Therefore, visualizing these could calibrate trust. The visualization technique could be altered depending on the maturity of the technology: **Tablet**-based in the center stack, which is inspired by current vehicles such as Tesla or Mercedes and represents the current state-of-the-art, and **AR**-based where the semantic segmentation information is directly visualized as an overlay to the object. AR represents the ultimate goal of spatial information distribution. We propose to visualize objects in all views (i.e., windshield and peripheral or side windows), as a user might not be familiar with the multitude of sensors built in an AV (front, rear, sides), thus, showing these detections is expected to calibrate trust. Less distracting methods of visualizing such as lightbands (see [49]) 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 they were not detected. 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. Figure 3 shows this concept technology-independently. Displaying semantic segmentation information only allows for the assessment of *object detection*; however, it directly encodes the uncertainty information for this task as only detected objects are visualized. The relevant object class information (e.g., pedestrian, vehicle, cyclist) is encoded via hue. Thus, hue is not used to display uncertainty of the detection but to distinguish classes. Urgency information as proposed by Kunze et al. [30] in turn was not included as the passenger was not required or intended to interact with the shown AV. While it would be possible to colorize every detected object using the same color which varies regarding detection uncertainty, we argue that knowing the actual detected class is relevant as these lead to different assumptions and behaviors for the AV. Therefore, we used a different method to visualize uncertainty compared to Kunze et al. [30] but also use hue as it is easily distinguishable.

5 ONLINE SIMULATION PRE-STUDY

To evaluate the concepts, we designed and conducted a video-based online within-subject study. We recruited N=32 participants (9 female, 23 male) via participant recruitment mailing lists of our university and online media (Facebook, WhatsApp; ad-hoc sample). Participation was voluntary. On average, they were M=27.06 (SD=9.66) years old. All participants hold a valid driver's license on average M=3.94 (SD=1.37) years. On 5-point Likert scales (*1 = Strongly Disagree — 5 = Strongly Agree*), participants showed high interest in AVs (M=4.03, SD=1.00), believed AVs to ease their lives (M=3.78, SD=.94), but were skeptical about whether they become reality by 2030 (M=3.53, SD=.80). Immersion of participants was moderate (M=14.25, SD=4.64) using the *Immersion* subscale of the Technology Usage Inventory (TUI) [28].

The following research question guided this exploratory study:

What impact do the variables "visualization technology" and "visualized objects" have on passengers in an AV in terms of (1) affective state, (2) cognitive load, (3) trust, (4) SA, (5) preference, and (6) capability assessment?

5.1 Procedure

Every participant encountered **five** conditions, a *baseline* with no visualization of the semantic segmentation and a 2 x 2 design (*visualization technology* with two levels: tablet vs. AR and *visualized*

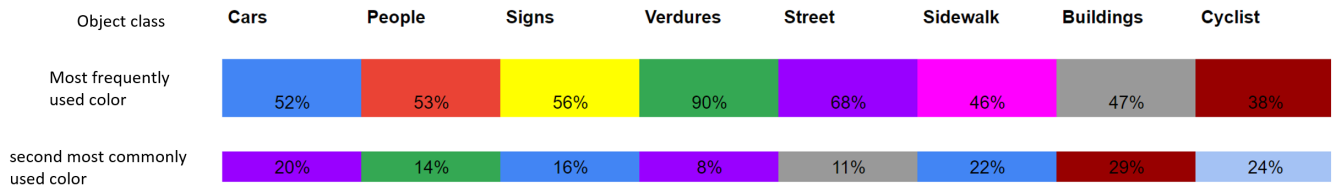


Figure 2: Color distribution of the object visualization for the evaluated publications.

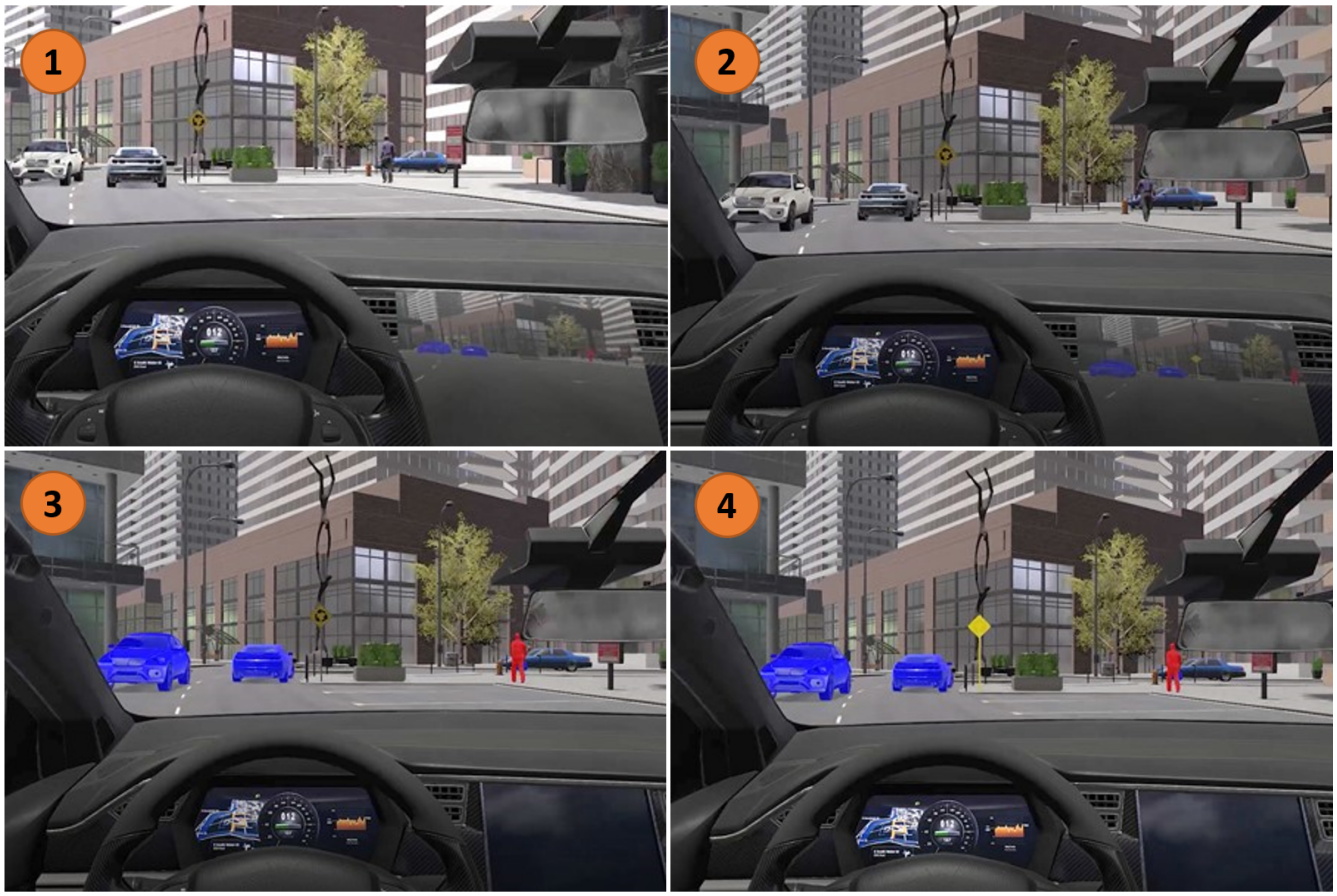


Figure 4: Interior of the simulated Tesla X with the semantic segmentation visualizations: *tablet dynamic system* (1), *tablet dynamic+static system* (2), *AR dynamic system* (3), *AR dynamic+static system* (4)

objects with two levels: dynamic vs. dynamic+static objects; the *independent* variables).

Each session started with a brief introduction, agreeing to the consent form, and a demographic questionnaire. The five conditions were then presented in counterbalanced order. The introduction to the capabilities was given as follows:

You will see various videos of a highly automated ride through a simulated environment. The vehicle takes over the lateral and longitudinal guidance. The vehicle tries to detect the objects in the scenery. This detection is presented to you in different ways. You are supposed to follow them closely and then assess them. Each video will last approximately 50 seconds.

For this, we recorded 5 videos of the simulation in Unity [48] showing the same scene but varying in semantic segmentation visualization (dynamic vs. dynamic + static; see Figure 4). The videos show a ride in a lively city with pedestrians crossing twice over crosswalks, and one pedestrian and one bicyclist crossing the street without a crosswalk. According to Kaß et al. [24], the vehicle performs lateral (i.e., driving straight ahead and turning multiple times)

as well as longitudinal (i.e., accelerate and decelerate/break) maneuvers. After each condition, participants answered the questionnaires described below. Lastly, participants were asked for general feedback. On average, a session lasted 30 min. A script running in the background ensured window maximization, that participants could not skip or rerun the video (to ensure equal exposure time), and that at least a (required) FullHD monitor was used.

5.2 Measurements

After each condition, affective state using the self-assessment manikin (SAM) [3], cognitive load using the raw NASA-TLX [18], usability with the system usability scale (SUS) [5], trust in automation using the German version [29] of the Trust in Automation scale [23], and SA using the situation awareness rating technique (SART) [47]. The SART was used to assess the perceived quality of situation awareness [13] which may be a predictor of "how a person will choose to act on that SA" [13, p. 86]. With high qualitative SA, users of AVs are expected to be less inclined to take over control with its post-automation effects [4, 35] and, therefore, the automation with its benefits can perform the driving task. Participants also rated the AV's capabilities (detection of passers-by and vehicles, recognition of signposts, longitudinal, and lateral guidance) of the system on 6-point Likert scales.

After all conditions, participants could provide open feedback, rated their preferences of the systems from highest (*ranking = 1*) to lowest (*ranking = 5*), and assessed the reasonability and necessity ("I think the visualization of the recognition of objects is reasonable/necessary)" of the semantic segmentation using single-item ratings on 7-point Likert scales.

6 RESULTS

Dependent on the data's nature, we employed a repeated measures (parametric) or a Friedman's ANOVA to find differences between the conditions. For the factor analysis in case of non-parametric data, we used *nparLD* [39]. ANOVA-type statistics are reported. For post-hoc tests, Bonferroni correction was used. Effect sizes were calculated using the formula proposed by Rosenthal [43]. For Figure 5 and Figure 6, we used the package *ggstatsplot* [41] in version 0.6.6. These figures include a boxplot as well as a violin plot showing the distribution of data points.

6.1 Cognitive Load

Cognitive load was significantly different for the concepts, $F(2, 76) = 6.72$, $p<.001$, $r=.06$. Post-hoc analyses showed that the *tablet dynamic system* ($M=7.36$, $SD=3.57$) received, compared to the *AR dynamic+static system* ($M=5.70$, $SD=3.02$; $t(31)=3.11$, *adj.* $p=.04$), significantly worse scores. The *tablet dynamic+static system* was also significantly worse rated compared to the *AR dynamic system* ($M=5.91$, $SD=3.07$; $t(31)=3.35$, *adj.* $p=.02$) and the *AR dynamic+static system* ($t(31)=3.72$, *adj.* $p<.01$). The NPVA showed a significant main effect of *visualization technology* ($F=11.28$, $df=1$, $p<.001$). Pairwise comparisons using Dunn's test revealed the difference to be significant ($p=0.003$, Z = -2.77, $r=-0.35$). Cognitive load was higher in the tablet version.

6.2 Usability

Usability was significantly different for the concepts, $F(3, 93) = 9.73$, $p<.001$, $r=.11$. Both, *tablet dynamic system* ($M=64.14$, $SD=19.27$) and *tablet dynamic+static system* ($M=62.66$, $SD=18.11$) were rated significantly worse than *AR dynamic system* ($M=73.75$, $SD=12.84$; *tablet dynamic system*: $t(31)=-3.14$, *adj.* $p=.04$; *tablet dynamic+static system*: ; $t(31)=-4.52$, *adj.* $p<.001$) and *AR dynamic+static system* ($M=77.66$, $SD=15.71$; *tablet dynamic system*: ; $t(31)=-3.75$, *adj.* $p<.01$; *tablet dynamic+static system*: ; $t(31)=-4.77$, *adj.* $p<.001$). The non-parametric variance analysis (NPVA) *nparLD* [39] showed a significant main effect of *visualization technology* ($F=18.04$, $df=1$, $p<.001$). Pairwise comparisons using Dunn's test revealed the difference to be significant ($p<0.001$, Z = 3.71, $r=0.46$). System usability was higher in the AR version.

The NPVA also showed a significant interaction effect of *visualization technology* with *visualized objects* ($F=5.75$, $df=1$, $p=.02$). The SUS score dropped in the Dynamic + Static condition for the tablet-based visualization while it rose for the AR visualization. We attribute this to the increased necessary mental mapping between the tablet and the (simulated) real world needed for the additional objects in the tablet-based system (see [9]).

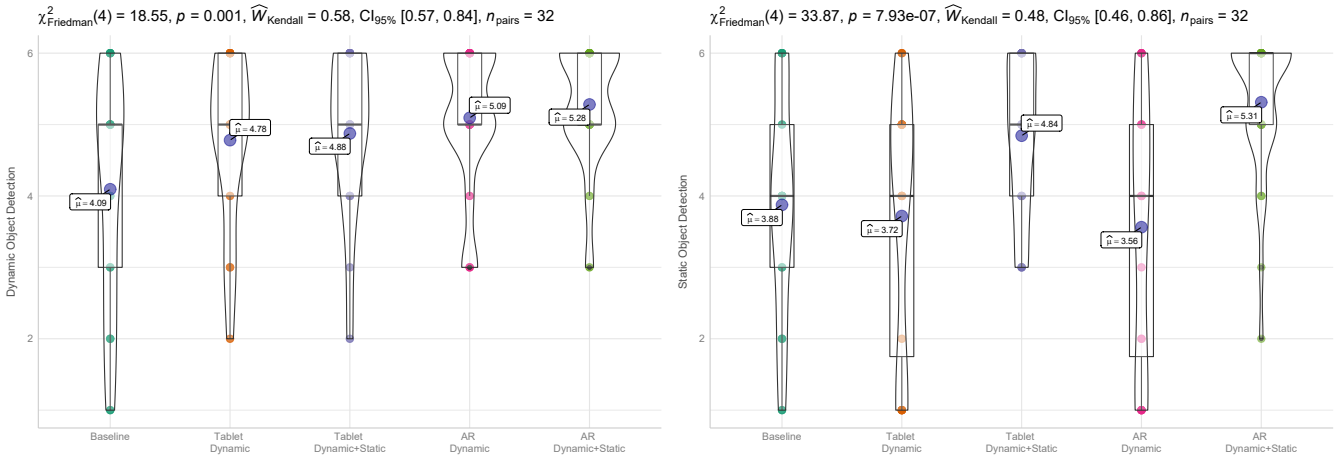
6.3 Trust in Automation & Situation Awareness

Trust was significantly different for the concepts, $F(1, 58) = 5.60$, $p<.01$, $r=.06$. The *tablet dynamic system* ($M=4.59$, $SD=1.18$; $t(31)=-3.85$, *adj.* $p<.01$) and the *tablet dynamic+static system* ($M=4.65$, $SD=1.23$; $t(31)=-3.60$, *adj.* $p=.01$) received significantly lower trust ratings than the *AR dynamic+static system* ($M=5.17$, $SD=1.13$). The *baseline* received ratings of $M=4.29$ ($SD=1.55$) and almost reached significance (*adj.* $p=0.051$) compared to the *AR dynamic+static system.*

The NPVA showed a significant main effect of *visualization technology* ($F=13.10$, $df=1$, $p<.001$) on trust. Dunn's test showed this to be significant ($p=0.015$, Z = 2.18, $r=0.27$). Trust was significantly higher when using AR.

A Friedman's ANOVA showed a significant difference in the subjective SA score (χ^2 (4)=13.95, $p<.01$). Post-hoc tests revealed that the *baseline* ($M=16.25$, $SD=6.41$) and the *tablet dynamic+static system* ($M=16.94$, $SD=5.59$) received significantly lower score than the *AR dynamic+static system* ($M=19.56$, $SD=6.26$).

The NPVA showed a significant main effect of *visualization technology* on SA ($F=4.70$, $df=1$, $p=.03$). Dunn's test showed that this almost reached significance ($p=0.0501$, Z = 1.64, $r=0.21$). Regarding the subscale Demand, the NPVA showed a significant main effect of *visualization technology* ($F=4.03$, $df=1$, $p=.04$, AR: $M=12.64$, $SD=4.21$, Tablet: $M=13.69$, $SD=4.02$).Dunn's test showed that this did not reach significance ($p=0.10$, Z = -1.26, $r=-0.16$). The NPVA also showed a significant main effect of *visualization technology* on the subscale Supply ($F=4.57$, $df=1$, $p=.03$, AR: $M=17.27$, $SD=3.55$, Tablet: $M=18.28$, $SD=4.12$). Dunn's test showed significance ($p=0.04$, Z = -1.80, $r=-0.23$). Regarding the subscale Understanding, the NPVA showed significant main effects of *visualization technology* ($F=19.81$, $df=1$, $p<.001$, AR: $M=14.41$, $SD=2.83$, Tablet: $M=12.56$, $SD=3.34$) and *visualized objects* ($F=10.79$, $df=1$, $p=.001$, dynamic+static: $M=14.00$, $SD=3.22$, dynamic: $M=12.97$, $SD=3.16$). Dunn's tests showed both



(a) Perceived ability to detect dynamic objects. **(b) Perceived ability to detect static objects.**

Figure 5: Perceived ability to detect dynamic and static objects.

differences to be significant (*visualization technology*: $p<.001$, Z = 3.15, r=0.39; *visualized objects*: p=.03, Z = -1.93, r=-0.24).

6.4 Capability Assessment, Preference, and Reasonability & Necessity

A Friedman's ANOVA showed a significant difference in the perceived ability to detect dynamic objects (χ^2 (4)=18.55, p=.001, see Figure 5a) and static objects (χ^2 (4)=33.87, $p<.001$, see Figure 5b). Post-hoc tests showed that for dynamic objects, the *baseline* was considered to detect dynamic objects significantly worse than the *AR dynamic+static system*. For static objects, post-hoc tests showed that the AV with no visualization (i.e., *baseline*), the AV with the *tablet dynamic system*, and the AV with the *AR dynamic system* were perceived to recognize static objects significantly worse than the AV with the *AR dynamic+static system*.

However, for the assessment of how far away both dynamic and static objects are detected, no significant differences were found. Participants believed dynamic to be recognized approximately 30 m and static objects approximately 35 m away. The correct answer for both was 35 m. We found no significant differences in the assessment of longitudinal or lateral control of the vehicle. This is in line with findings of Colley et al. [9] who also found no significant differences.

There was a clear ranking: both AR systems (*AR dynamic+static system*: M=2.06, SD=1.22; *AR dynamic system*: M=2.31, SD=1.23) received the best ratings, followed by both tablet systems (*tablet dynamic+static system*: M=3.34, SD=1.21; *tablet dynamic system*: M=3.38, SD=1.21). The *baseline* was rated as the least favorite (M=3.91, SD=1.38).

A Friedman's ANOVA showed a significant difference in the mean rankings (χ^2 (4)=31.12, $p<.001$). Post-hoc tests showed that, compared to the *AR dynamic+static system*, both tablet systems and the *baseline* were rated significantly worse (see Figure 6). The *baseline* was also rated significantly worse than the *AR dynamic system*.

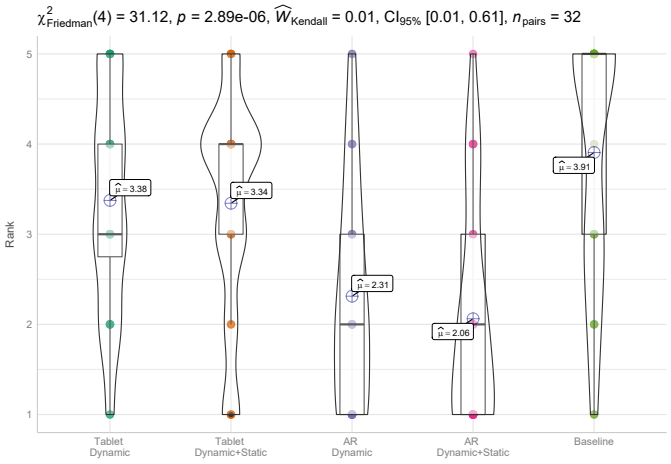


Figure 6: Ranking of the systems.

Participants rated the semantic segmentation as very reasonable (M=6.06, SD=1.19) and necessary (M=5.03, SD=1.38).

6.5 Open Feedback

Feedback about the visualization was mainly positive (20 participants). One participant highlighted "I was able to understand what the car is seeing and what not, [...]. I felt more safe and in control when visualisation was present. Without I was alert at every encounter and was hoping that the car did not crash into the approaching subject" [P32]. Participants also stated that the tablet was too distracting. Some improvement proposals were given, these included using fish-eye lenses for a higher field of view in the tablet, acoustic support for people with vision impairments, and adjustable color settings.

7 ONLINE STUDY WITH REALISTIC FOOTAGE

The first study showed that the AR system with the visualization of dynamic **and** static objects was rated best in most measurements. These results confirm the findings of Colley et al. [9] who report AR systems to increase trust and reduce cognitive load. However, this study was conducted using a Unity simulation in which all objects were perfectly colored in. Therefore, we increased external validity in a second study via videos taken in the real-world and applying the state-of-the-art model Panoptic-DeepLab [8].

For this, we used the same procedure and measurements. Again, participant recruitment mailing lists and online media (Facebook, WhatsApp; ad-hoc sample) were used for participant recruitment. There was no overlap between samples. Participation was voluntary. A *baseline* with no semantic segmentation was compared to the *AR dynamic+static system* version (see Figure 7). The video was taken in Ulm, Germany with an iPhone 11 Pro Max with 30 fps in wide-angle and FullHD resolution. We anonymized the videos (faces and license plates). The video shows a ride through the busy inner city with multiple people of varying ages crossing the street. Additionally, a parked vehicle at the roadside merges into the traffic. According to Kaß et al. [24], the vehicle performs lateral (i.e., driving straight ahead and once) as well as longitudinal (i.e., accelerate and decelerate/break) maneuvers. The scene is more complex than in the pre-study as more pedestrians are present and a vehicle merges.

Another difference is the participants' point of view: in this study, the video is taken from the passenger seat due to technical limitations.

Participants encountered **two** videos, one video without semantic segmentation information (the *baseline*; see Figure 7 (1)) and one with the beforehand applied semantic segmentation. These were presented in randomized order. Both videos had a duration of 3 min and 37 s. A session lasted approximately 20 min.

N=41 participants (21 female, 20 male) took part in the study. These were, on average, M=27.63 (SD=8.07) years old and reported low (M=2.57, SD=.68) propensity to trust [27]. 35 participants came from Germany, 2 from the USA, 1 from Colombia, 1 from Indonesia, 1 from Romania and one did not reveal the country. On 5-point Likert scales (*1 = Strongly Disagree — 5 = Strongly Agree*), participants showed high interest in AVs (M=4.51, SD=.64), believed AVs to ease their lives (M=4.24, SD=.66), but were skeptical about whether they become reality by 2030 (M=3.41, SD=1.07). The average score for Immersion was again moderate (M=14.85, SD=4.70). This Immersion score was used to assess results' reliability, which was medium in both studies.

8 RESULTS

In the following, we report the results of the statistical analysis. Descriptive and inferential statistics are reported. Depending on whether the data were normally distributed or not, we employed t-tests (Cohen's for effect size) or Wilcoxon Signed Rank tests (Rosenthal's [43] formula for effect size).

8.1 Situation Awareness

Values for the subjective assessment of SA were *baseline*: M=16.61, SD=5.99 vs. *AR dynamic+static system*: M=21.02, SD=5.70. A t-test revealed a highly significant difference regarding the subjective reports on SA ($t(40) = 3.4$, p=0.002, r=0.53; see Figure 8(a)). No significant differences were found for the Demand subscale (p=0.98). However, significant differences were found for the Understanding ($t(40) = 3.72$, $p<0.001$, r=0.58; see Figure 8(a)) and the Supply subscale ($t(40) = 2.19$, p=0.03; see Figure 8(c)). Both values were significantly higher in the *AR dynamic+static system*.

8.2 Mental Load, Trust, Usability, Reasonability & Necessity

We found no significant differences for mental load, trust, and usability assessments. The *baseline* received ratings of M=8.05 (SD=4.77), the *AR dynamic+static system* M=8.85 (SD=4.56) for mental effort. Trust was measured using the Trust in Automation scale [23]. The overall score was medium and almost equal (*baseline*: M=3.31, SD=.76 vs. *AR dynamic+static system*: M=3.32, SD=.75). Usability was also rated as medium [44] (*baseline*: M=68.11, SD=16.81 vs. *AR dynamic+static system*: M=66.89, SD=16.34). Participants believed such a visualization to be reasonable (M=5.00, SD=1.58) and necessary (M=4.56, SD=1.60).

8.3 Attribution of Capabilities

Participants rated all recognition-related attributes significantly better in the *AR dynamic+static system* condition compared to the *baseline* (see Figure 9). All effects were of moderate size. For longitudinal (*baseline*: M=5.44, SD=1.21; *AR dynamic+static system*: M=5.80, SD=.95) and lateral (*baseline*: M=5.61, SD=1.22; *AR dynamic+static system*: M=5.83, SD=1.07) control, no significant differences were found.

8.4 Open Feedback

Formulated opinions in the open feedback varied. One participant highlighted the capability of the visualization to calibrate trust:

> "The visualization made me realize, how bad the recognition still works. Without any visualization I definitely trusted the system more (over-trust)." [P16]

Others proposed some visual adjustments. Three participants did not want signposts to be visualized, one would "have preferred to only see outlines for example and try to keep the flickering to a minimum or best to none at all" [P25]. However, the flickering was intentional to convey the uncertainty of the segmentation task. Another suggestion was to leave out turning signals from the visualization.

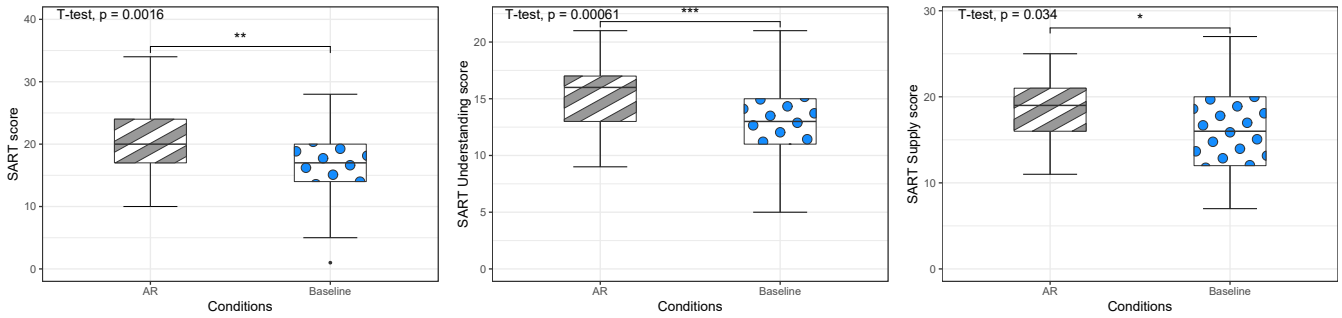
9 DISCUSSION

Overall, the *AR dynamic+static system* received higher ratings in the subjective assessment of SA and was rated as reasonable and necessary in both studies. The usage of the system did not result in any significantly lower rating regarding trust or cognitive load.

9.1 Calibration of User Expectations

In the realistic study, the model determined objects in real-time. This was not flawless and some participants mentioned the accompanying "flickering" as disturbing. However, this is seen as a tool to convey the actual recognition capabilities of an AV. While trust

Figure 7: Screenshot from the video presented to participants in the *baseline* condition (1) and the *AR dynamic+static system* semantic segmentation (2) condition.



(a) SART combined score. **(b) SART subscale understanding.** **(c) SART subscale supply.**

Figure 8: Results of the SART.

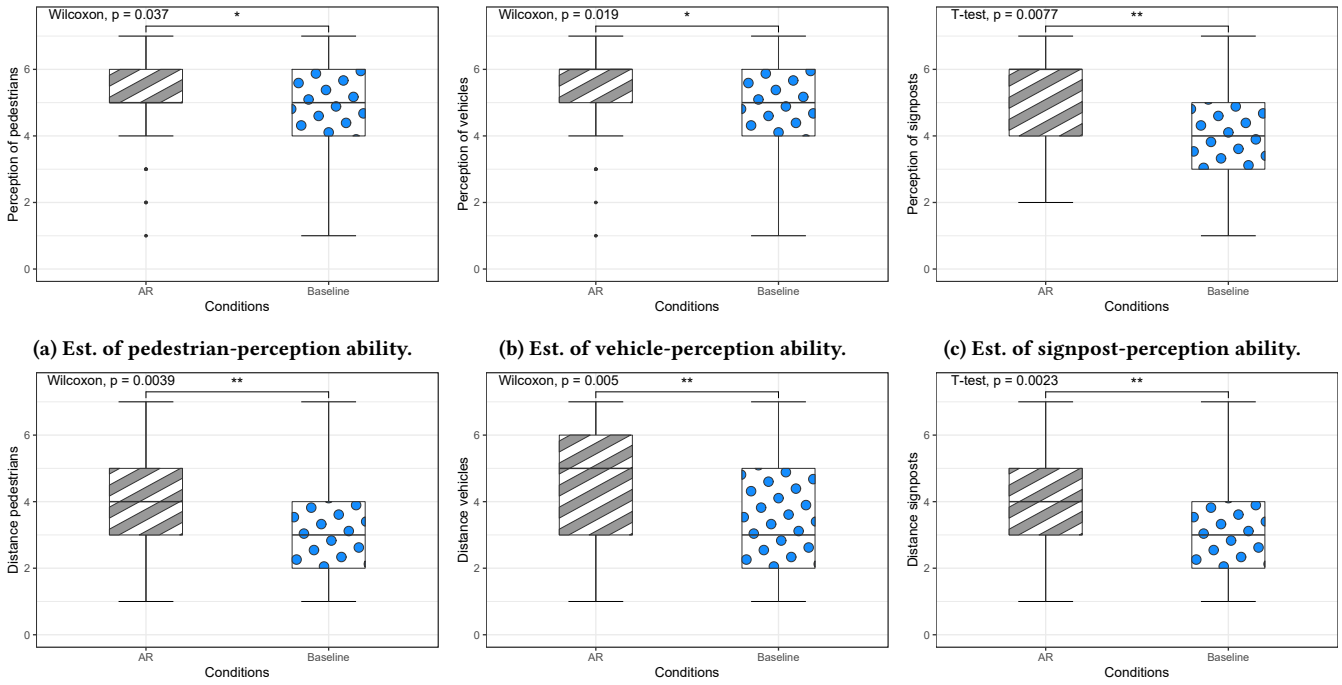
was about equal, the AV with the AR system was rated better in all perception-related questions. The Halo Effect [38], a cognitive bias that augments the perception of a system attribute based on another positive system attribute, seems to be not at work as the assessment of lateral and longitudinal control was not significantly different. As perception was rated significantly better (see Figure 9) with an AR visualization, we assume that a user of an AV will, at first, be very skeptical about the possibilities of such vehicles without a communication/visualization of capabilities. Therefore, the proposed visualization seems to be appropriate to calibrate user expectations about the vehicle's capabilities while still maintaining a moderate level of trust.

9.2 Abstract vs. Concrete Visualizations

There exist numerous works on the effect of transparent systems. These systems use abstract representations (anthropomorphic [2], abstract levels [20], or circles [9]). In the proposed concept, a concrete visualization of detected objects is used. Abstract information visualization is especially helpful if the underlying data set is large [25]. However, this abstraction inherently is accompanied by a loss in information. This is especially true for the abstract visualization of uncertainty. Uncertainty is difficult for people to understand and they, therefore, avoid it [40]. Thus, an abstraction might lead to even more difficulties. Especially in the context of automated driving, uncertainty is important at least for the introductory phase. Questions such as "Do I have to overtake?" will arise in uncertain situations. With a more concrete visualization, such questions could be more easily answered. The user can assess the capabilities of the vehicle and make an informed decision. With regards to the second study, subjective SA (see Section Situation Awareness) and capabilities (see Section Attribution of Capabilities) were rated significantly higher with the concrete visualization compared to the *baseline*. No significant differences were found for mental load (see Section Mental Load, Trust, Usability, Reasonability & Necessity). Thus, we conclude that a more abstract visualization is unnecessary as the loss in information prevents the user to gain a more profound understanding of the AV. The effects of this visualization as a potential distraction from relevant information, however, have to be taken into account. One possibility is to include a distance-based visualization. Depth prediction is already possible and could be used to enhance our proposed visualization [16].

9.3 Visual Clutter & Practical Implications

The system evaluated in the realistic study introduced a lot of visual clutter in the scene. This did not, however, significantly increase cognitive load or decrease trust. Still, some participants mentioned the need to **not** highlight lights or pillars. We agree that this is not



(a) Est. of pedestrian-perception ability. **(b) Est. of vehicle-perception ability.** **(c) Est. of signpost-perception ability.**

(d) Est. of pedestrian-perception distance. **(e) Est. of vehicle-perception distance.** **(f) Est. of signpost-perception distance.**

Figure 9: Results for the estimation (est.) of vehicle capabilities.

as relevant for the calibration of trust and SA. Additionally, participants argued that stopping and traffic lights should be excluded from the visualization as the meaning is hidden by the colorization. Again, the authors agree with these statements. The visualization could distract AV users. However, the distracting nature of our approach actually is beneficial as its salient nature nudges the driver to engage with the functionality. Based on this initial calibration, the user only then can decide to engage in non-driving task related activities and, if wanted, turn off the visualization.

These studies further solidify the benefits of using some sort of visualization of vehicle capabilities at least during the introduction of AVs. The studies further show the advantages of AR/WSD visualization. This technology is not yet available for entire windshields due to its challenges (e.g., parallax). We believe, however, that this is only a matter of time. For the second study, we used a state-of-the-art semantic segmentation model Panoptic-Deeplab [8]. While we needed approximately 0.7 s per frame on an RTX 2060, more specified hardware and future developments are likely to make this real-time capable. Therefore, our approach only requires WSD to become available. Still, the visualization should be customizable, that is, users should be able to define which object types should be visualized.

9.4 Discrepancy Between Studies

The simulation study confirmed some of the results of related work that transparent systems [12, 26] and AR systems [9, 19, 50] lead to higher trust, acceptance, and perceived safety. While in the online study the trust was not significantly different, it was still lower and almost reached significance. Also, both studies showed that an AR visualization of object recognition increased SA. For trust measurement, Hock et al. advise to "Refrain from introducing the system as flawless, if trust is of interest in the study" [22, p. 112], which we did. Still, for trust, the ratings in the study with realistic footage were almost the same. A potential explanation is that trust measurement in a simulation is not transferable to a real-world scenario. Potentially, the apparent nature of the simulation leads to a more "game-like" assessment of the situation. This is, however, not supported by the TUI scores, which were also almost the same. Another explanation is that the situations assessed were just too different and, therefore, the trust difference is higher in calm situations as seen in the simulation. The third factor could be the position of the camera. In the realistic footage, the video, due to technical requirements, had to be taken from the passenger's seat. This study cannot answer this question, however, this should be studied in the future to assess the external validity of simulator studies.

10 LIMITATIONS & FUTURE WORK

In both studies, a reasonable number of participants took part (N=32 and N=41). However, the demographic information shows that this was mostly a younger target group. It is not clear whether our findings are transferable to other age groups. Additionally, we only focused on subjective dependent measures. We targeted the simulation's results' transferability by conducting a second study with real-world footage and a state-of-the-art segmentation model. However, this transfer came with some drawbacks: the different setting of the ride and the different position of the camera. Therefore, there are several potential confounding factors and the results cannot

be directly compared. Additionally, one participant in the second study came from Indonesia with left-handed traffic. While an analysis with excluded data from the Indonesian participant found no alterations in the significance of any of the findings, this aspect could still have an impact on the perception of the visualization and should be targeted in future work. Additionally, future work should evaluate potential intention recognition [9] and connected driving information uncertainty visualization.

11 CONCLUSION

Overall, we showed the potential of presenting semantic segmentation visualization to users of AVs. First, we investigated technology-dependent visualization in a simulation study (N=32). This work further solidified that AR-based solutions are the most promising [9]. Afterward, in a second study, we increased external validity by using real-world footage and a state-of-the-art semantic segmentation model to further evaluate the *AR dynamic+static system* with N=41 participants. While not all results of the first study are supported, participants still reported subjective higher SA and assessed the perception capabilities as higher. Improvement proposals include the avoidance of coloring in the stopping and traffic lights. This work further enhances the body of knowledge on factors for a successful introduction of AVs.

ACKNOWLEDGMENTS

We thank David Dobbelstein for his support as well as all study participants. This work was conducted within the project 'Interaction between automated vehicles and vulnerable road users' (Intuitiver) funded by the Ministry of Science, Research and Arts of the State of Baden-Württemberg.

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] Margaret M Bradley and Peter J Lang. 1994. Measuring emotion: the self-assessment manikin and the semantic differential. *Journal of behavior therapy and experimental psychiatry* 25, 1 (1994), 49–59.

[4] 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**

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

[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, Christian Bräuner, Mirjam Lanzer, Walch Marcel, Martin Baumann, and Rukzio Rukzio. 2020. Effect of Visualization of Pedestrian Intention Recognition on Trust and Cognitive Load. 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/10.1145/3409120.3410648**

[10] 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.

[11] 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.

[12] Na Du, Jacob Haspiel, Qiaoning Zhang, Dawn Tilbury, Anuj K. Pradhan, X. Jessie Yang, and Lionel P. Robert. 2019. Look who's talking now: Implications of AV's explanations on driver's trust, AV preference, anxiety and mental workload. *Transportation Research Part C: Emerging Technologies* 104 (2019), 428–442. **https://doi.org/10.1016/j.trc.2019.05.025** ID: 271729.

[13] 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.

[14] 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.

[15] 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.

[16] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J. Brostow. 2019. Digging Into Self-Supervised Monocular Depth Estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, New York, NY, USA, 3828–3838.

[17] 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**

[18] 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.

[19] 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**

[20] 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**

[21] Philipp Hock, Franziska Babel, Johannes Kraus, Enrico Rukzio, and Martin Baumann. 2019. Towards Opt-Out Permission Policies to Maximize the Use 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, 101–112. **https://doi.org/10.1145/3342197.3344521**

[22] Philipp Hock, Johannes Kraus, Franziska Babel, Marcel Walch, Enrico Rukzio, and Martin Baumann. 2018. How to Design Valid Simulator Studies for Investigating User Experience in Automated Driving: Review and Hands-On Considerations. 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, 105–117. **https://doi.org/10.1145/3239060.3239066**

[23] Jiun-Yin Jian, Ann M Bisantz, and Colin G Drury. 2000. Foundations for an empirically determined scale of trust in automated systems. *International Journal of Cognitive Ergonomics* 4, 1 (2000), 53–71.

[24] 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.

[25] Tanja Keller and Sigmar-Olaf Tergan. 2005. *Visualizing Knowledge and Information: An Introduction.* Springer Berlin Heidelberg, Berlin, Heidelberg, 1–23. **https://doi.org/10.1007/11510154_1**

[26] 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.

[27] 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.

[28] 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].

[29] Johannes Kraus, David Scholz, Dina Stiegemeier, and Martin Baumann. 0. 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* 0, 0 (0), 0018720819853686. **https://doi.org/10.1177/0018720819853686** arXiv:https://doi.org/10.1177/0018720819853686 PMID: 31233695.

[30] 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**

[31] 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**

[32] 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.

[33] 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.

[34] 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**

[35] 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**

[36] David Moher, Alessandro Liberati, Jennifer Tetzlaff, and Douglas G Altman. 2009. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *Annals of internal medicine* 151, 4 (2009), 264–269.

[37] 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.

[38] 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.

[39] Kimihiro Noguchi, Yulia R Gel, Edgar Brunner, and Frank Konietschke. 2012. nparLD: an R software package for the nonparametric analysis of longitudinal data in factorial experiments. *Journal of Statistical Software* 50, 12 (2012), 23.

[40] Scott E. Page. 2008. Uncertainty, Difficulty, and Complexity. *Journal of Theoretical Politics* 20, 2 (2008), 115–149. **https://doi.org/10.1177/0951629807085815** arXiv:https://doi.org/10.1177/0951629807085815

[41] Indrajeet Patil. 2018. ggstatsplot: 'ggplot2' Based Plots with Statistical Details. **https://doi.org/10.5281/zenodo.2074621**

[42] 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**

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

[44] Jeff Sauro. 2011. Measuring usability with the system usability scale (SUS).

[45] 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.

[46] 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**

[47] 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.

[48] Unity Technologies. 2019. *Unity*. Unity Technologies. **https://unity.com/**

[49] 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**

[50] 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