Team At Your Service: Can Multiple Conversational Agents Increase Functional Specificity in Automated Driving?

DINARA TALYPOVA, University of Applied Sciences Hagenberg, Austria
PHILIPP WINTERSBERGER, University of Applied Sciences Hagenberg, Austria

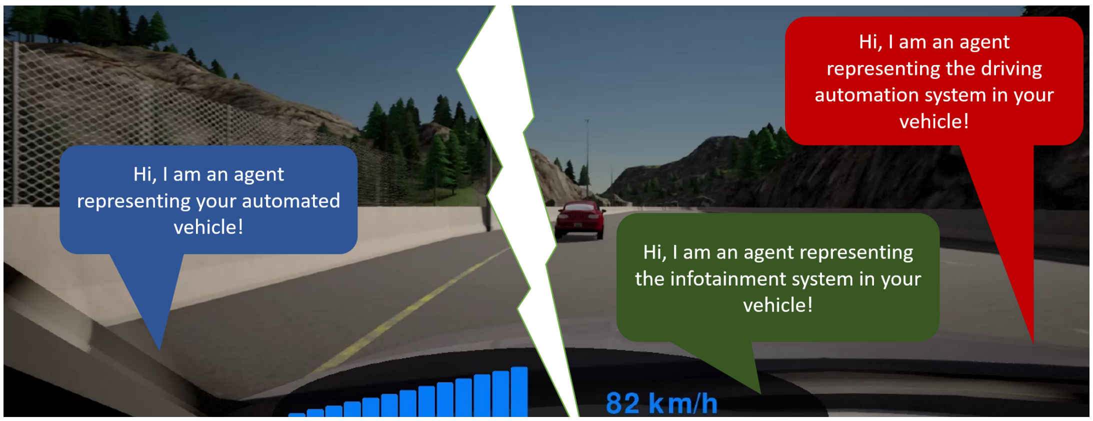

Fig. 1. We investigated how good participants are able to calibrate their trust in different in-vehicle sub-systems, where conversational agents represented either the entire car (left) or individual sub-systems (right).

Only a few works so far have addressed functional specificity for trust formation. Previous research indicated that users could hardly distinguish between sub-systems and integrate all system perceptions into a global assessment. Thus, we conducted a user study where participants had to supervise an automated vehicle (AV) equipped with conversational agent(s). In two conditions, the vehicle was represented either by two agents that portrayed the driving automation and the infotainment system or by a single agent that embodied both systems. We hypothesized that a clear differentiation between sub-systems could allow drivers to better calibrate their trust. However, our results show quite the opposite. Correlation analyses suggest that participants' functional specificity was high, and they based their situational and general trust ratings mainly on the perception of the driving system. We conclude that functional specificity can be supported with clear communication about the presence of different subsystems in AV.

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

Additional Key Words and Phrases: Trust in Automation, Automated Vehicles, Driving Simulator Study, Trust Calibration, SAE level 2 driving

ACM Reference Format:
Dinara Talypova and Philipp Wintersberger. 2023. Team At Your Service: Can Multiple Conversational Agents Increase Functional Specificity in Automated Driving?. In *TRAIT @CHI'23: Workshop on Trust and Reliance in AI-Assisted Tasks, April 22–28, 2023, Hamburg, GER*. ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Overtrust is a critical issue in the interaction between humans and AI-driven systems. After the deadly accident of a Tesla driver, the National Transport Safety Board (NTSB) concluded that *"the driver was playing a mobile game while using Autopilot before the crash, and investigators also determined he was overly confident in Autopilot's capabilities"* [14]. Especially monitoring over longer time periods is challenging even for *"highly motivated human beings"* (i.e., an "irony of automation" [2]). It was argued that mitigating such issues requires drivers to calibrate their trust so that their expectations of and trust in the driving automation system fits its actual capabilities [4]. Consequently, a wide range of studies has addressed trust calibration for automated vehicles (AVs) [6, 7, 12, 13]. A significant issue preventing a successful calibration of trust may lie in users' perception of an AV as a single entity, rather than a collection of sub-systems. An experiment by Frison et al. [3] has demonstrated that drivers can hardly calibrate their trust in a level 2 driving automation system (DAS) since they are influenced by other sub-systems in the vehicle, where both the performance and the visual design (aesthetics) of an in-vehicle infotainment system (IVIS) influenced their trust levels. These so-called "Halo-effects" negatively influence drivers' functional specificity (the degree to which users can distinguish between sub-systems), which is an important factor in the trust calibration model by Lee and See [10].

In this paper (which is a short version of our study published at ACM AutomotiveUI'22 [17]), we address the issue of functional specificity for trust calibration in AVs. We conducted a driving simulator study where participants had to cooperate (i.e., monitor and intervene) with an AV whose system makes sporadic errors (missing to detect a slower lead vehicle or red light) while being moderately distracted (reading texts on a smartphone and interacting with conversational agents in the car). We aimed at (1) validating the assumption that drivers have low functional specificity and (2) finding out if functional specificity increases when different sub-systems (the driving automation system and the IVIS) are portrayed with different conversational agents.

2 TRUST AND CONVERSATIONAL AGENTS IN AUTOMATED VEHICLES

Trust in automation can be defined as *"the attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability"* [10], and users can either distrust, overtrust, or appropriately trust a system. Appropriate, or "calibrated", trust means that a user's subjective trust levels match the system's performance/capabilities [10]. However, calibration is not the only important trust dimension in the widely cited paper by Lee and See [10], which also discusses the factors "resolution" and (temporal/functional) "specificity". Resolution describes the mapping of performance changes to trust, where low resolution means that large performance changes would only slightly influence trust levels, whereas high resolution indicates a presumably precise mapping of trust and performance. Specificity is similar to resolution. Temporal specificity describes how long it takes until performance changes become visible in trust (i.e., low temporal specificity leads to a delayed trust update, high describes a situation where performance changes are reflected in trust immediately). Functional specificity, in turn, describes the degree to which users can calibrate their trust to different sub-systems – while high functional specificity describes a situation where users can successfully calibrate towards performance variations of multiple systems, low functional specificity indicates that users can hardly distinguish between sub-systems and integrate all system perceptions to a more global assessment. According to [20], only a small number of works addressing trust calibration consider and discuss the additional dimensions of resolution and specificity. One experiment that explicitly addressed the dimension of functional specificity was conducted by Frison et al. [3]. They have confronted study participants with performance variations of a level 2 DAS and an IVIS system and different user interface designs (aesthetics). Their results suggest that users integrate all system perceptions into the trust formation process and can hardly distinguish between sub-systems. Consequently, they concluded that drivers in AVs suffer low functional specificity.

Conversation with both agents and other people is expected to grow increasingly popular as vehicles of the future become more highly automated [8]. That said, spoken interactions with conversational agents have only recently been studied in an automated vehicle context. A Wizard of Oz study found that participants in a fully automated pod vehicle placed more trust in a highly anthropomorphized conversational agent which engages in small talk than in a command-based agent, while both engendered more trust than a touchscreen interface with the same features [9]. This finding built on similar research in which participants in a simulator study reported higher confidence in the system when interacting with a conversational agent as compared to a touchscreen [1]. To prevent Halo-effects, a multi-agent approach may be fruitful, allowing designers to align communicative style with communication function. Indeed a recent study investigated the feasibility of multiple conversational agents in a driving simulator, sparking increased interest in the approach [11].

3 METHOD AND RESEARCH QUESTIONS

Consequently, we investigated if functional specificity is low in AVs by explicitly assessing the sub-systems with corresponding items. Such an approach could help to obtain more transparent results regarding what participants base their trust on. If the results of [3] can be confirmed, solutions will be needed to help users distinguish between multiple in-vehicle systems. As one prototypical solution, we finally assess if functional specificity increases when in-vehicle systems are portrayed with different user interfaces (in particular, in the form of two independent conversational agents). Therefore, we ask:

- **RQ1:** Can it be confirmed that functional specificity is low in automated driving?
- **RQ2:** Can functional specificity be increased by representing different sub-systems as independent agents?

We conducted a driving simulator study, where participants interacted with a simulated SAE level 2 driving automation as well as an in-vehicle infotainment system. Both systems provided a conversational interface to communicate with the driver, where we distinguished between two conditions. In the first condition (*Single Agent*), a single conversational agent represented all the functions in the vehicle (i.e., DAS and IVIS). In the second condition (*Multiple Agents*), two different conversational agents (with different voices and sentence structures, see below) represented the DAS and the IVIS, respectively.

3.1 Conversational Agent Design

We recorded three different agents using Google WavNet [15]. A difficulty here was that we wanted to guarantee that a potential difference in trust formation was solely attributed to the sub-systems and not to the voices themselves. For example, research has shown that users may rate the trustworthiness of agents differently when using a male vs. a female voice [16]. Although different genders would easily allow participants to distinguish between the sub-systems, gender stereotypes could influence the results [18]. Hence, we chose to only vary pitch and sentence structure and utilized the Wavelet-C demo voice with three different pitches. The default neutral pitch was used to represent the whole vehicle in the first condition. For the second condition, the lower-pitched voice was used to represent the DAS, and the higher-pitched voice to represent the IVIS. To not suffer additional biases emerging from a representation of the agents, we did not include avatars or portray them visually in any way. For the IVIS, we recorded interactions where the agent asks the driver if it should turn on the music, set a reminder for a calendar appointment, or read out a just

Fig. 2. Study setup: participants had to supervise a level 2 DAS while communicating with WoZ-controlled conversational agents. In the *single agent* condition, we used the default pitch and a more neutral wording, while in the *multiple agents* condition, the higher-pitched IVIS agent delivered non-driving related information and the lower-pitched DAS agent communicated supervision reminders more strict.

received text message. For the DAS, we recorded different supervision reminders to stay attentive and monitor the road environment. As expressed, in condition *Single Agent*, all these interactions used the neutral voice, while in condition *Multiple Agents*, the IVIS interactions used the higher-pitched voice, and the DAS agent used the lower-pitched voice. During the drive, the interactions were controlled by an experimenter in a Wizard-of-Oz (WoZ) manner by triggering the pre-recorded voice lines in particular sections of the route. Further, we slightly modified the sentence structure of the supervision reminders in a way that the DAS agent would communicate in a more strict manner than the neutral agent in the *single agent* condition (see Figure 2).

3.2 Driving Simulation and Study Setup

We set up a simple driving simulator using a Logitech G29 steering wheel with pedals and a large display (see Figure 2). We adapted the scenario as described by Frison et al. [3], where an AV drives on the left lane of a highway at 100km/h and encountered 12 slower (70km/h) lead vehicles. In most of these situations, the lead vehicle was detected by the driving automation system, which slowed down and waited until the lead vehicle swerved to the right before accelerating again. However, in 2 of the 12 situations (83% reliability), the lead vehicle was not detected, creating the need for manual intervention. In a second part of the route, the vehicle left the highway and entered a small town with 5 intersections. Here, the vehicle would normally detect red lights and stop. Similarly, in one of the 5 intersections, the red light was not detected, and again a manual intervention was necessary. To intervene and clear the situation, participants had to manually apply the brake pedal (no steering was necessary for the entire experiment). To put the participants into additional tension and make monitoring more difficult, we requested them to read a news article on their smartphone during the drive, which they had to summarize in a few sentences after the trip was completed.

4 USER STUDY

We conducted a user study with the just described setup, where participants experienced both conditions *Single Agent* and *Multiple Agents* in a within-subjects design.

4.1 Participants and Procedure

N=25 participants (17 male, 8 female) aged between 20 and 31 (M=23.28) years completed the experiment. First, participants received a short briefing where we explained the different tasks (i.e., monitoring and intervening if necessary, interacting with the conversational agents), completed the pre-test survey and demographics, and performed a short test drive to get familiar with the scenario. Then, they performed two drives, one in condition *Single Agent* and one in condition *Multiple Agents* in counterbalanced order. After each condition, they completed a set of trust scales.

4.2 Measurements

We collected subjective ratings (all assessed on a 7-point Likert scale from "strongly disagree" to "strongly agree") for participants' trust (driving performance measurements are described and evaluated in [17]). After each condition, we administered the Situational Trust Scale for Automated Driving [4] as well as the Trust in Automation Scale [5]. Additionally, we included single-item ratings for trust and mistrust in the individual sub-systems as well as the overall AV (i.e., "I trust/mistrust the infotainment system/driving system/overall system").

4.3 Results

All scales showed acceptable reliability (Cronbach's alpha $> .6$), so we calculated scale values by averaging the individual items of the respective scales and treated the results as continuous data for our analyses (see Table 1). Regarding the single-item questions for trust/distrust in the individual sub-systems as well as the overall system and the Trust in Automation scale, we subtracted the distrust score from the trust score, which in some cases (i.e., distrust higher than trust) has led to negative values. The evaluation was performed using IBM SPSS V27. We conducted correlation analyses using the Pearson coefficient, and comparisons between conditions were carried out with non-parametric Wilcoxon tests or t-tests, depending on the data following a normal distribution.

To investigate our research questions, we conducted multiple correlation analyses. Looking at the trust ratings for the individual sub-systems shows interesting results: In both experimental conditions, participants' trust in the IVIS did not correlate with any of the other trust ratings, i.e., trust in the overall system, STS-AD, or the Automation Trust Scale. In contrast, their trust in the driving automation sub-system significantly correlated with all subsequent trust ratings in both conditions; overall trust (*Single Agent*: $r = .83, p < .001$; *Multiple Agents*: $r = .81, p < .001$), STS-AD (*Single Agent*: $r = .64, p < .001$; *Multiple Agents*: $r = .88, p < .001$), and the Automation Trust scale (*Single Agent*: $r = .79, p < .001$; *Multiple Agents*: $r = .85, p < .001$). Further, the ratings for the two sub-systems "trust in IVIS" and "trust in the DAS" did not show any significant association. This means that the participants did not "integrate" the trust ratings of the individual sub-systems towards their overall ratings but based their situational and general trust assessment mainly on the perception of the DAS (see Figure 3).

To evaluate potential differences between the two experimental conditions of the individual sub-systems being represented with only one *Single Agent* or *Multiple Agents*, we conducted Wilcoxon signed rank tests. Neither the individual ratings (trust in IVIS: $Z = -.83, p = .41$, distrust in IVIS: $Z = -1.89, p = .06$; trust in DAS: $Z = -1.42, p = .16$, distrust in DAS: $Z = -.42, p = .67$; overall trust: $Z = -1.90, p = .06$, overall distrust: $Z = -.30, p = .76$) nor the combined

Table 1. Descriptive statistics of the assessed self-rating scales. Negative values emerge by subtraction of the distrust compontent from the trust component (i.e., participants' distrust was higher than their trust ratings).

Measurement	**Single Agent M (SD, Med)**	**Multiple Agents M (SD, Med)**
Trust in IVIS	2.32 (3.25, 4.00)	3.16 (2.54, 4.00)
Trust in DAS	-1.16 (2.89, -1.00)	-.60 (3.64, 0)
Trust in overall System -	.36 (2.71, 1.00)	.72 (3.08, 1.00)
STS-AD	3.32 (1.17, 2.83)	3.69 (1.36, 3.83)
Jian et al. Trust	3.76 (1.21, 4.14)	3.79 (1.27, 4.00)
Jian et al. Distrust	3.26 (1.20, 2.80)	3.15 (1.39, 2.80)
Jian et al. Overall Trust	.50 (2.12, 1.06)	.64 (2.49, 1.29)

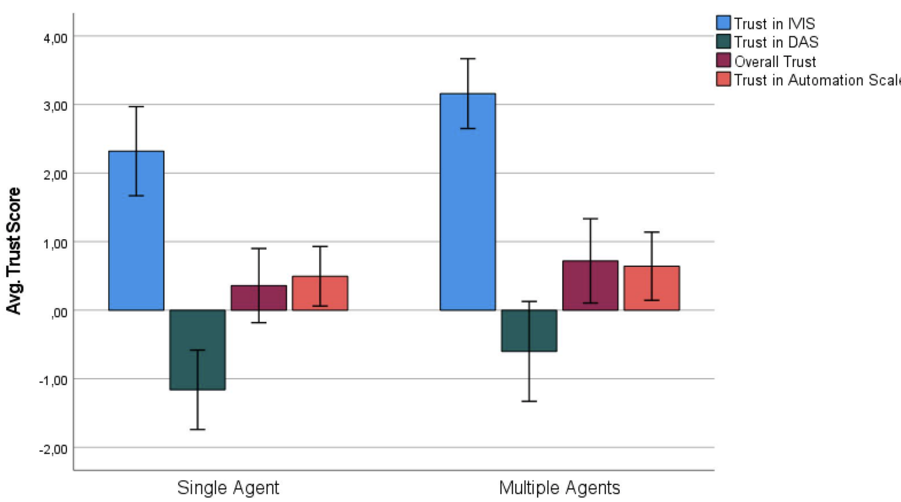

Fig. 3. Results of the individual scales (with standard error bars). The overall/general trust ratings were mainly influenced by the subjective trust in the driving automation system, indicating high functional specificity, while there were no differences regarding the AV being portrayed with only one (left) or multiple agents (right).

(trust-distrust) scores did significantly differ between the conditions (trust in IVIS: $Z = -1.62, p = .11$; trust in DAS: $Z = -1.16, p = .25$; overall trust: $Z = -1.06, p = .29$). The same result was indicated for the validated, standardized multi-item ratings. Neither the situational trust scale for automated driving (STS-AD: $Z = -1.56, p = .12$) nor the Trust in Automation Scale ($Z = -.16, p = .87$) showed significant differences between the two conditions.

Additionally, we looked at how many times our participants failed to intervene when the driving automation system showed erroneous behavior. We must report a high number of missed interventions - in both conditions, we recorded multiple crashes into a lead vehicle (*Single Agent*: 8 of 50 = 16%; *Multiple Agents*: 10 of 50 = 20%), and similarly, many missed to manually stop the vehicle at the red traffic light (*Single Agent*: 9 of 25 = 36%; *Multiple Agents*: 6 of 25 = 24%). There was no significant difference between the two experimental conditions.

5 DISCUSSION

Overall, we were surprised by the results of this experiment, as it contradicted our initial expectations, in particular regarding participants' functional specificity. Frison et al. [3] have claimed that functional specificity is low in automated vehicles since users integrate a great variety of system perceptions (such as the performance of other sub-systems but also interface aesthetics) into their trust levels. We cannot confirm the existence of such "Halo-effects" - In the presented experiment, the perception of the IVIS seemed not to influence participants' subjective trust levels of the overall system, neither from the perspectives of situational (as measured by the STS-AD) or general (as measured by the Trust in Automation Scale) trust. However, Frison et al. [3] directly modified system performance while the reliability of the sub-systems (DAS and IVIS) remained stable in our study. Another difference to [3] is that they assessed the combined stream of perceptions into a single trust score, while we explicitly assessed the sub-systems with corresponding items. Also, we explained in the participant introduction that *"the vehicle has a so-called driving automation system, which is responsible for driving, navigation, and safety, and an infotainment system, which is responsible for music, heating, as well as other connectivity and smart functions"*. Potentially, explaining the existence of and asking about different sub-systems could have primed our participants. However, this can be considered a positive result since it would be easy for vehicle designers to communicate this to drivers. Nevertheless, regarding our research questions, we cannot confirm that functional specificity is low in automated driving (**RQ1**).

Given that functional specificity was high in both experimental conditions, it is not too surprising that we could not demonstrate any beneficial effect of portraying different sub-systems with individual conversational agents. Potentially, we were too conservative in our designs, as we only modulated the pitch of the voices and the sentence structure. Other design variations such as voices with different genders or pictures/animations of visual avatars for the sub-systems could have led to different results but would also increase the probability that differences emerge from these designs rather than the mere perception of one vs. multiple in-vehicle systems. Still, our study cannot confirm that functional specificity increases by representing different sub-systems with different conversational agents (**RQ2**).

One significant problem was still present in the results: Even high functional specificity did not prevent a significant number of study participants from crashing into a lead vehicle or missing a red traffic light. We must report failure rates between 16 and 35%, which well fits the results by Victor et al. [19], where one-third of drivers crashed with an obstacle in a level 2 test track experiment, independent of system introduction or supervision reminders.

We spent quite some time discussing potential methodological issues that could have led to the obtained results. As discussed, the system description outlined that the vehicle is equipped with two systems (the DAS and the IVIS), and we did also ask about participants' trust in the individual sub-systems individually – both factors could have primed our participants. Further, we portrayed the different conversational agents only with different voice pitches and slightly adjusted sentence structure. Future experiments should include additional factors, i.e., system performance and agent portrayal.

6 CONCLUSION

In this short paper, we have presented a user study in a static driving simulator that addressed trust formation in level 2 automated vehicles. Previous work has indicated that drivers who are confronted with multiple in-vehicle systems show low functional specificity (i.e., can hardly distinguish between sub-systems when building up their trust). In our experiment, we investigated the influence of dispositional trust and different in-vehicle systems (infotainment and a driving automation system, portrayed by either a single conversational agent representing the whole vehicle or as two independent agents representing the corresponding sub-system) on participants' situational and general trust levels. Our results suggest that functional specificity was high among our participants. This is supported by correlation analyses, which show that situational and general trust in the AV was mainly driven by the perception of the driving automation system in the car but not the infotainment system. Further, we did not see any differences in trust ratings and crash rates depending on whether the vehicle was portrayed with one or multiple agents. In all conditions, about a third of participants failed at least once in their monitoring duties and either crashed with a lead vehicle or missed a red light. We conclude that it could be sufficient to inform drivers about different in-vehicle sub-systems to support trust calibration. However, it remains unclear how the problems associated with supervisory control can be resolved.

ACKNOWLEDGMENTS

We want to thank Anton Torggler and Justin Edwards for their contributions to the original publication summarized in this workshop paper.

REFERENCES

[1] Vicki Antrobus, David R. Large, and Gary E. Burnett. 2018. Trust me – I'm AutoCAB': Using natural language interfaces to improve the trust and acceptance of level 4/5 autonomous vehicles. Hague, Netherlands. https://nottingham-repository.worktribe.com/output/1218942

[2] Lisanne Bainbridge. 1983. Ironies of automation. In *Analysis, design and evaluation of man–machine systems*. Elsevier, 129–135.

[3] Anna-Katharina Frison, Philipp Wintersberger, Andreas Riener, Clemens Schartmüller, Linda Ng Boyle, Erika Miller, and Klemens Weigl. 2019. In UX we trust: Investigation of aesthetics and usability of driver-vehicle interfaces and their impact on the perception of automated driving. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–13.

[4] 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*. 40–47.

[5] 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.

[6] 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.

[7] Johannes Kraus, David Scholz, Dina Stiegemeier, and Martin Baumann. 2020. The more you know: trust dynamics and calibration in highly automated driving and the effects of take-overs, system malfunction, and system transparency. *Human factors* 62, 5 (2020), 718–736.

[8] Andrew L Kun, Susanne Boll, and Albrecht Schmidt. 2016. Shifting gears: User interfaces in the age of autonomous driving. *IEEE Pervasive Computing* 15, 1 (2016), 32–38.

[9] David R Large, Kyle Harrington, Gary Burnett, Jacob Luton, Peter Thomas, and Pete Bennett. 2019. To please in a pod: employing an anthropomorphic agent-interlocutor to enhance trust and user experience in an autonomous, self-driving vehicle. In *Proceedings of the 11th international conference on automotive user interfaces and interactive vehicular applications*. 49–59.

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

[11] Seul Chan Lee, Seona Jeong, Manhua Wang, Philipp Hock, Martin Baumann, and Myounghoon Jeon. 2021. "To Go or Not To Go? That is the Question": When In-Vehicle Agents Argue with Each Other. In *13th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*. Association for Computing Machinery, New York, NY, USA, 223–224. https://doi.org/10.1145/3473682.3481876

[12] Mengyao Li, Brittany E Holthausen, Rachel E Stuck, and Bruce N Walker. 2019. No risk no trust: Investigating perceived risk in highly automated driving. In *Proceedings of the 11th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*. 177–185.

[13] Brittany E Noah, Philipp Wintersberger, Alexander G Mirnig, Shailie Thakkar, Fei Yan, Thomas M Gable, Johannes Kraus, and Roderick McCall. 2017. First workshop on trust in the age of automated driving. In *Proceedings of the 9th international conference on automotive user interfaces and interactive vehicular applications adjunct*. 15–21.

[14] Sean O'Kane. 2020. Tesla Autopilot, distracted driving to blame in deadly 2018 crash. The Verge. Retrieved March 11, 2020 from https://www.theverge.com/2020/2/25/21153320/tesla-autopilot-walter-huang-death-ntsb-probable-cause.

[15] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. 2016. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499* (2016).

[16] Suzanne Tolmeijer, Naim Zierau, Andreas Janson, Jalil Sebastian Wahdatehagh, Jan Marco Marco Leimeister, and Abraham Bernstein. 2021. Female by Default?–Exploring the Effect of Voice Assistant Gender and Pitch on Trait and Trust Attribution. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–7.

[17] Anton Torggler, Justin Edwards, and Philipp Wintersberger. 2022. Beyond the Halo: Investigation of Trust and Functional Specificity in Automated Driving with Conversational Agents. In *Proceedings of the 14th International Conference on Automotive User Interfaces and Interactive Vehicular*

Applications. 195–203.

[18] UNESCO. 2019. *I'd blush if I could – closing gender divides in digital skills through education.* Technical Report. UNESCO.

[19] Trent W Victor, Emma Tivesten, Pär Gustavsson, Joel Johansson, Fredrik Sangberg, and Mikael Ljung Aust. 2018. Automation expectation mismatch: Incorrect prediction despite eyes on threat and hands on wheel. *Human factors* 60, 8 (2018), 1095–1116.

[20] Philipp Wintersberger. 2020. Automated Driving: Towards Trustworthy and Safe Human-Machine Cooperation. (2020).