Advances in Human Factors of Transportation, Vol. 148, 2024, 604–614
https://doi.org/10.54941/ahfe1005252

Interaction Concepts for Delivery Robots: Exploring Evaluation Criteria and Design via Two User-Centered Approaches

Vivian Lotz, Eva-Maria Schomakers, Chantal Lidynia, and Martina Ziefle

Human-Computer Interaction Center, RWTH Aachen University, Aachen, 52074, Germany

ABSTRACT

With the introduction of automated driving technologies onto public roads, requirements for human-vehicle interfaces have changed. Yet, little is known about what constitutes effective and enjoyable user interactions. To address this issue, we explored different concepts and modalities for interaction. Three concepts (graphical user interface, voice user interface, and wristband) were developed via two user-centered design processes. The first one was a participatory design approach (i.e., [1] co-design session, [2] initial evaluation via an online survey, and [3] redesign.) The second one was a design thinking sprint (i.e., a series of workshops: [1] emphasize, [2] define, [4] ideate, [5] and prototype.) Finally, three concepts were prototyped and comparatively evaluated. For evaluation, we used the requirements identified during the development processes. The comparative evaluation highlighted that while all three interaction concepts were viable, the graphical user interface and voice user interface showed greater promise in meeting user requirements than the wristband, which scored relatively low in aesthetic appeal, efficiency, and maturity.

Keywords: Delivery robot, User-centered design, Co-design, User experience, Technology acceptance

INTRODUCTION

The advent of automated driving technologies on public roads has transformed the requirements for human-vehicle interfaces. Additionally, the COVID-19 pandemic further accelerated the growth of the e-commerce sector, necessitating the development of contactless or minimal-contact delivery methods. This led to the increased deployment of automated delivery robots (ADRs), such as those developed by Starship, Kiwibot, and Nuro, which have undergone testing in various locations across the U.S. and U.K. (NuroInc., 2016; Starship, 2014). Initially designed for small-scale, single-order deliveries and operating on sidewalks, some undertakings are exploring autonomous solutions capable of handling larger volumes of transport (e.g., Pluta, 2017; Schomakers et al., 2022). Despite these advancements, integrating perspectives from all stakeholders impacted by automated delivery robots — logistics employees, customers, and pedestrians — remains a critical, yet often overlooked, component in their design (Mumford, 2000). Ensuring safe and smooth interactions with the new technology is essential for its successful implementation.

Research Aim

This article investigates user interaction with automated delivery robots, guided by two research goals: [1] identifying usage criteria for interfaces and [2] evaluating the alignment of co-designed interaction concepts with these criteria.

BACKGROUND

Co-Design and Design Thinking

Co-design, or participatory design, involves designers, end-users, stakeholders, and other parties collaboratively shaping solutions, aiming to democratize the design process by integrating the expertise and insights of all involved (Steen, 2013). It emphasizes integrating technology within its social context (Suchman, 2002) and, through continuous user engagement, ensures solutions meet the intended audiences' needs, preferences, and contexts, fostering ownership and satisfaction (Harrington et al., 2019). Proven effective in various fields, including human-robot interaction for specific groups such as children and older adults (e.g., Björling and Rose, 2019; Lee et al., 2017), co-design has been instrumental in developing socially functional robots.

Design thinking parallels co-design, focusing on empathy, collaboration, and iterative evaluation to solve problems creatively. It follows a structured, five-phase process: [1] understanding user needs, [2] defining problems and objectives, [3] ideating solutions, [4] prototyping, and [5] testing, with ongoing refinement based on feedback (Kelley and Kelley, 2013; Plattner et al., 2009).

Delivery Robots and Automated Micro-Vehicles

A range of research projects (e.g., Höffner, 2019; Marsden et al., 2018) explore automated delivery and freight transport vehicles. The main distinctions between the different concepts are (1) the used infrastructure (i.e., road, sidewalk, bike lane, or in-doors) and - as autonomous vehicles are not yet certified - (2) the type of human supervision (follow-approach or remote monitoring) (Baum et al., 2019).

Trials primarily feature small sidewalk robots with remote monitoring options. Only a few vehicles are already licensed and tested on public roads (e.g., NuroInc., 2016). Less common but notable are follow-approach vehicles such as DuckTrain (Schomakers et al., 2022), the PostBOT (Pluta, 2017), and smaller-sized sidewalk robots (e.g., Piaggio, 2024). Figure 1 shows some robot examples.

This article mainly focuses on a follow-me ADR with high flexibility regarding the used infrastructure and platooning options.

Figure 1: Examples of follow-ADR (Reske, 2020; Schomakers et al., 2022).

Design Space: Communication Modalities

Primary and secondary tasks in automotive interfaces are typically executed using haptic and touch inputs, with feedback provided via visual displays (Detjen et al., 2021). However, with delivery robots lacking a dedicated space for displays, replicating existing interfaces does not work. Prior research has explored more intuitive modalities such as vocal communication, action recognition, shape-changing interfaces, and gesture inputs, which are worth exploring for delivery robots (e.g., Darvish et al., 2018; Neto et al., 2019).

METHODOLOGY

Building on general preferences and barriers identified in an earlier study (Lotz et al., 2022a), we sought to identify and test interaction concepts through a user-centered design process. This process was carried out in Germany between 2021 and 2023 using qualitative and quantitative empirical methods.

Data Collection and Procedures

We conducted **co-design workshops** (N = 6) to explore general usage criteria and draft initial design concepts for Graphical User Interfaces (GUI). Subsequently, a GUI was developed and evaluated via an online survey (N = 202) (for results see Lotz et al. (2022b)). The prototype was then refined based on the survey results.

In parallel, we conducted a 12-week **design thinking sprint** with three participants to develop alternative concepts to the GUI. The sprint consisted of three workshops ([1] Emphasize & Define, [2] Define, and [3] Ideate and Prototype) in which participants consolidated learnings, interspersed with periods for independent work. Each workshop combined seminar-style teaching of methods (e.g., brainstorming methods) with practical application of these methods. During the first phase, the participants conducted a literature review, interviewed potential users (N = 7). During the second phase, insights were narrowed by defining and prioritizing the identified needs and necessary interaction steps. The final phase consisted of re-reviewing literature, brainstorming, prioritizing solutions, drafting concept explanations, and rating concepts. After drafting some solution ideas, participants selected eight concepts. The concepts were evaluated by the design team and three independent evaluators.

Lastly, three concepts were prototyped and tested via **user tests** (N = 20). Here, a between-subject design was used, i.e., participants only tested one of the concepts due to time constraints. Concepts were tested in pairs to incite discussions while testing the interaction. Post-test, experiences were shared in interviews and quantified via a questionnaire, assessing effort expectancy, performance expectancy, safety concerns, privacy concerns, social norm, attitude, prototype maturity, and questions regarding the implementation of core features.

Participation was voluntary, without providing monetary incentives. Participants were assured of privacy and provided informed consent. Further, the option to withdraw participation at any time was stressed. All studies introduced the technology, its application context, and the objective. For qualitative methods, semi-structured protocols were used. Responses in surveys were measured on 6-point Likert scales.

Table 1. Sample characteristics per study (GUI design process).

		Co-design workshop	GUI questionnaire	Design sprint interviews	User test
N		6	202	7	20
Gender	Male	3 (50 %)	63 (31 %)	3 (43 %)	3 (15 %)
	Female	3 (50 %)	139 (69 %)	4 (57 %)	17 (85 %)
Age	M	27.17	31.90	36.14	21.05
	SD	1.86	13.79	23.55	2.63
	Range	22–43	18–66	23–84	19–27
Nation-ality	German	6 (100 %)	139 (64 %)	-	-
	Turkish	-	73 (36 %)	-	-

Participants

Table 1 shows sample characteristics per study. The workshop and design sprint interviews were balanced in terms of gender. However, the questionnaire responses were slightly biased towards females, and most user test participants were female. In all studies, the average age was below that of the German population (DeStatis, 2015). Notably, the user testers were relatively young. Workshop participants included laypeople and designers.

The user test sample showed a medium to high mean affinity for technology interaction (M = 3.71, SD = 0.89), privacy disposition (M = 3.60, SD = 1.05), and intention to use delivery robots (M = 4.03, SD = 1.54).

RESULTS

Next, we describe our results, starting with the derived design criteria and prototypes. This is followed by the user test results. Qualitative data were analyzed via qualitative content analysis using deductive categorization based on user experience and technology acceptance literature (e.g., Schrepp et al., 2014; Venkatesh et al., 2012), supplemented by inductively derived categories. Analysis was done according to Kuckartz (2012) and Mayring (1994).

Quantitative data was analyzed by descriptive and inferential statistics. The significance level was set to $\alpha = .05$. Reliability was checked before analysis (Cronbach's $\alpha \geq .7$; see Table 3).

Design Criteria

Insights from the GUI co-design workshops and the criteria identified by the design sprint participants were largely congruent, with the only difference being that the co-design participants focused – due to harsher time constraints – directly on how to archive the requirements. In contrast, the design sprint team initially tried to define abstract criteria while remaining open to how to solve them via interaction design. Solutions were only defined in a subsequent step.

GUI workshop participants first and foremost emphasized the need for easy, efficient interaction, for which they employed several strategies to archive (see Table 2). Moreover, safety and ease of use were often discussed interchangeably, as aiming for intuitive interfaces can minimize distraction and enhance traffic safety. Trust and control were addressed by opting for a non-playful aesthetic and including feedback on the robot's status and intentions. Seriousness was highlighted as a relevant design criterion.

The **design sprint** team identified six main criteria to focus on during the design, listed in Table 2. The team drafted eight interaction concepts during the subsequent ideation phase, which were then rated.

Table 2. Key insights on design criteria and strategies derived from the co-design workshop and the design sprint.

GUI workshop	Design sprint emphasize phase	Evaluation criteria
Minimizing information and input options Reducing sub-menus Including redundancies, e.g., using both icons and labels; including serval ways to archive interaction goals Increasing screen size Increasing icon size Show information context-dependent (trip preparation vs. on-road) Non-playful aesthetics Robot state and intention Accessibility: color-blindness	Ease of use Efficiency Traffic safety Suitability for diverse users: Age and experience; walking versus cycling Trust calibration with appropriate levels of control and awareness Multitasking (i.e., enable carrying out secondary or primary tasks such as monitoring traffic or carrying freight)	Feasibility Reliability alignment with requirements Flexibility Trust Excitement

Designs and Prototypes

GUI design drafts predominantly featured maps showing user and robot locations to provide orientation. The interaction was tailored for different contexts (driving versus preparation), with distinct screens to reduce distraction while on the road. One group introduced voice control to facilitate non-visually demanding inputs while driving. The designs were refined, tested via an online questionnaire (Lotz et al., 2022b), and re-refined. The final (clickable) prototype used during the user tests is depicted in Figure 2.

Interaction concept drafts derived from the design sprint focused on enabling a smooth transition between interacting as a pedestrian and as a cyclist. Another aspect was the switching between different communication partners as the robots can be operated as a convoy. The best-rated concept was a wristband featuring buttons to make inputs and audio, vibration, and light robot-to-user feedback (see Figure 2).

As the co-design workshop and the design sprint team highlighted the suitability of voice control, a voice user interaction (VUI) concept was additionally drafted.

Participants were given manuals for orientation and functionality guidance. For the voice concept, a research team member simulated the robot's responses, following a predefined script to ensure the interactions were casual, friendly, and concise.

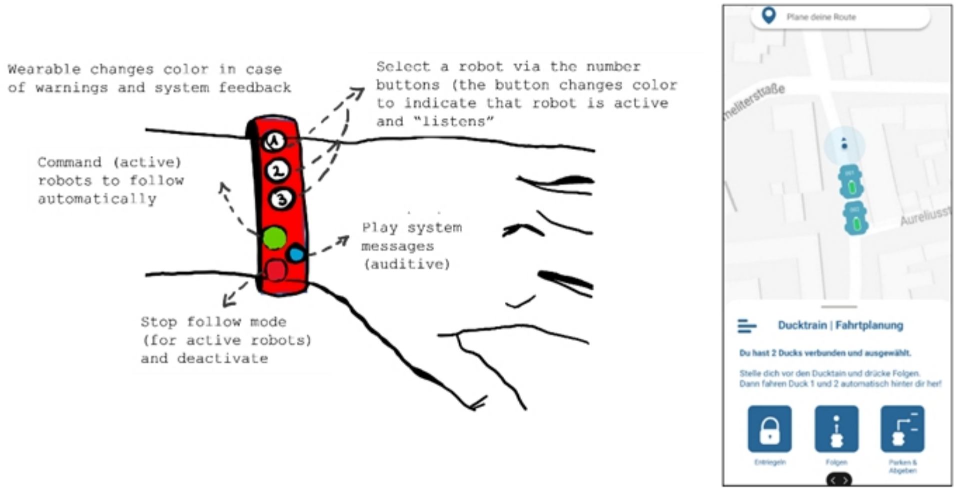

Figure 2: Illustration of the wristband (left) and GUI (right) prototypes.

Prototype Perception and Evaluation

Respondents rated the prototypes mostly positive, with an overall preference for the graphical user interface ($M_{gui} = 4.64$, $SD_{gui} = 1.41$). The voice interaction was rated close behind ($M_{vui} = 4.43$, $SD_{vui} = 0.53$), and the wristband showed the lowest usage intention ($M_{wristband} = 3.83$, $SD_{wristband} = 1.17$) (see Table 3).

Analyses of variance revealed that there was a significant difference between the prototypes regarding the constructs: "attitude toward using" ($F(2,17) = 3.668$, $p = .047$, $\eta_p^2 = .301$), "privacy concerns" ($F(2,17) = 6.406$, $p = .008$, $\eta_p^2 = .430$), "perceived maturity" ($F(2,17) = 5.982$, $p = .011$, $\eta_p^2 = .413$), and "efficiency" ($F(2,17) = 6.637$, $p = .007$, $\eta_p^2 = .438$) Post hoc comparisons using the Tukey HSD tests indicated that the attitude toward the voice interaction was significantly more positive than toward the wristband (M_{vui} = 4.95, M'$_{wristband}$ = 3.56, p =.041). Privacy concerns were more pronounced for the voice interaction than the wristband (M_{vui} = 3.14, $M_{wristband}$ = 1.33, p = .008). Further, the wristband was perceived as less ready to use than both other prototypes (M_{gui} = 3.86, M_{vui} = 3.38, $M_{wristband}$ = 1.72, $p_{gui\text{-}wristband}$ = .011, $p_{vui\text{-}wristband}$ = .48).

A closer examination of the safety concern items shows that all queried concerns scored low for the GUI (see Figure 3). In contrast, the VUI was linked to concerns regarding misinterpretations of user inputs (M = 4.29, SD = 0.95), system reliability (M = 4.14, SD = 1.16) and distractions (M = 4.14, SD = 0.90). For the wristband, reliability and misinterpretations seem not to bee too concerning for the participants, while the accident risk was evaluated similar to that of the VUI -slightly above the scales neutral point (M = 4.00, SD = 1.67).

Qualitative feedback pointed out the wristband's inefficiency and unappealing design. The GUI testers encountered challenges in switching between robots which is also reflected in the survey results (M = 3.43, SD = 0.79, see Figure 3). Participants also faced confusion about possible and necessary inputs with the VUI, even tough they rated the information amount (M = 4.57, SD = 0.79) and understandability (M = 5.00, SD = 0.82) as sufficient (see Figure 3). Lastly, test observations suggested varied user preferences for levels of control over the robots' actions: some preferred simply stating the end goal, while others detailed specific steps for the robot to follow.

Table 3. Reliability and descriptive statistics for included constructs (N = 20).

Construct	Cronbach's α	M			SD			Adapted from
		GUI	Voice	Band	GUI	Voice	Band	
Effort expectancy	0.836	4.89	4.71	3.88	0.99	0.65	0.75	(Osswald et al., 2012)
Performance expectancy	0.846	4.38	4.24	4.78	0.93	0.81	1.44	(Osswald et al., 2012)
Safety concerns	0.897	2.60	3.94	3.50	0.86	0.80	1.51	Self-developed
Social norm	0.833	3.71	3.71	3.56	1.22	0.80	1.03	(Osswald et al., 2012)
Attitude towards using	0.843	4.52	4.95	3.56	0.69	1.01	1.11	(Osswald et al., 2012)
Privacy concerns	0.868	1.93	3.14	1.33	0.93	1.21	0.41	(Belanger and Xu, 2015)
Behavioural intention interaction	0.923	4.64	4.43	3.83	1.41	0.53	1.17	(Osswald et al., 2012)
Prototype maturity	0.941	3.86	3.38	1.72	1.65	0.65	0.85	Self-developed

DISCUSSION

User-centered design approaches weer used to understand the requirements and challenges of interface design for ADRs. This article presented the multi-stage design processes, the insights on design criteria, and user test results.

Learnings from the Design Process

Guidelines and Evaluation Criteria for User-ADR Interaction Design

The evaluation of interaction concepts focused on critical criteria: ease of use, efficiency, and trustworthiness, with trust built through user control, awareness, and professional design. These factors were perceived as crucial to ensuring safety in traffic. Results hinted that user preferences for trust and control levels might vary. Other criteria included making the system accessible, adaptable to users' experience levels, and capable of supporting multitasking. These considerations are consistent with research on autonomous vehicle usability and acceptance (DIN, 2018; Madigan et al., 2016; Osswald et al., 2012).

Do the Co-Designed Interaction Concepts Meet the Evaluation Criteria?
All three concepts were evaluated positively and met most of the evaluation criteria, with each concept excelling in specific areas while also revealing distinct opportunities for refinement. The wristband, although innovative, fell short in metrics such as maturity, efficiency, and aesthetic appeal, indicating a need for redesign. The graphical user interface successfully met criteria related to effort expectations and safety perceptions, except for the challenge of switching between communication partners. While offering an intuitive means of interaction, the voice interaction needs further refinement regarding user orientation and privacy measures. Interestingly, voice interaction scored high on safety concerns, which might be explained by how safety concerns were operationalized (i.e., risk of misinterpretations, accidents, unreliability, interface-induced distractions, and negative impact on driving).

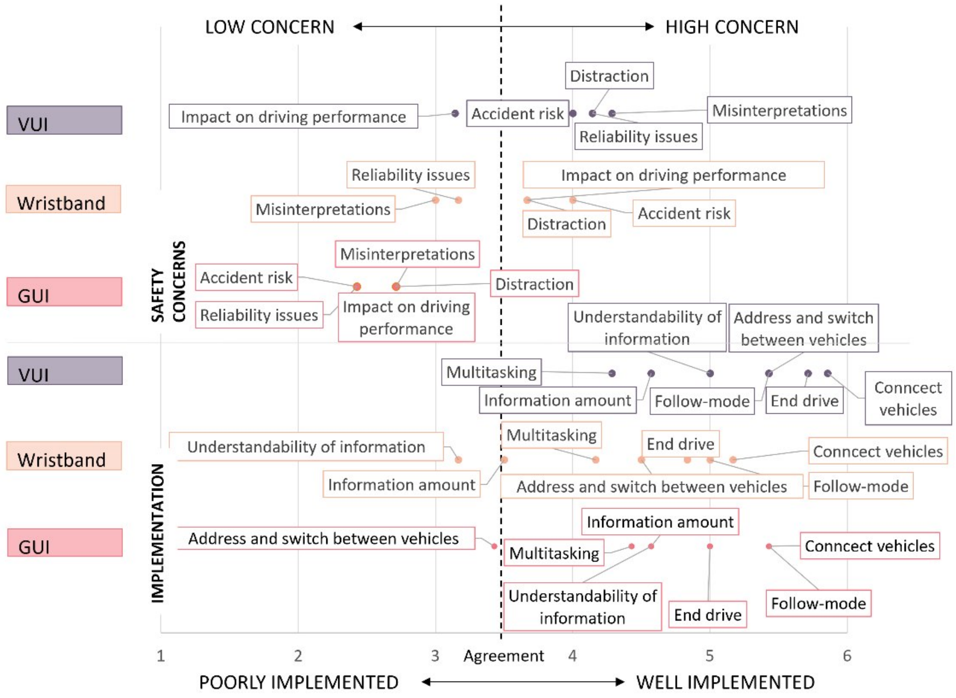

Figure 3: Evaluations of safety concerns and implementation of the three interface concepts.

Remarks on the Processes

Despite efforts to incorporate user diversity, design teams and participant groups remained largely uniform, leading to speculative concerns about the needs of diverse user groups without solid evidence. Future projects should involve a wider array of participants, including non-native speakers and older adults, to broaden the design perspective. While the co-design workshops effectively generated results, participants struggled with information overload, and some issues in group dynamics occurred. Adding follow-up workshops for a more in-depth discussion might be beneficial in light of the already extensive session that was carried out. In contrast, the design sprint offered plenty of time for ideation. However, participants defaulted to familiar concepts, highlighting a clear need for strategies that encourage more creative and less conventional thinking.

Limitations and Outlook

The chosen approach yielded valuable insights into evaluative criteria and into designing interaction concepts to meet them. While all prototypes require further refinement and iterative evaluation, initial recommendations were proposed, and common challenges were identified.

It is essential to note some constraints when interpreting the findings. The size of the co-design group was relatively small ($N = 6$), which, however, does not necessarily decrease validity (Toner, 2009) - especially because the obtained insights were later validated via an online survey and a user test.

Engaging users in the design process and testing concepts early presents clear benefits, notably ensuring that the outcomes are more closely aligned with genuine needs. Nevertheless, early testing has limitations, such as the inability to test systems under authentic conditions. Evaluating complex factors such as safety can be challenging for participants in such early stages. Moreover, questions regarding the scalability of the results remain open. Future research, with more mature prototypes and under realistic conditions, is essential to validate findings.

ACKNOWLEDGMENT

The authors would like to thank the participants. The research was funded by the Ministry of Culture and Science of the state of North Rhine-Westphalia with the funding code 321-8.03.07-127598 (project ACCESS!).

REFERENCES

Baum, L., Assmann, T., Strubelt, H., 2019. State of the art - Automated micro-vehicles for urban logistics. IFAC-Pap., 9th IFAC Conference on Manufacturing Modelling, Management and Control MIM 2019 52, 2455–2462.

Belanger, F., Xu, H., 2015. The role of information systems research in shaping the future of information privacy. Inf. Syst. J. 25, 573–578.

Björling, E., Rose, E., 2019. Participatory Research Principles in Human-Centered Design: Engaging Teens in the Co-Design of a Social Robot. Multimodal Technol. Interact. 3, 8.

Darvish, K., Wanderlingh, F., Bruno, B., Simetti, E., Mastrogiovanni, F., Casalino, G., 2018. Flexible human–robot cooperation models for assisted shop-floor tasks. Mechatronics 51, 97–114.

DeStatis, 2015. Altersdurchschnitt der Bevölkerung sank 2015 auf 44 Jahre und 3 Monate [WWW Document]. Stat. Bundesamt. URL https://www.destatis.de/DE/Presse/Pressemitteilungen/2017/06/PD17_197_12411.html (accessed 2.1.24).

Detjen, H., Faltaous, S., Pfleging, B., Geisler, S., Schneegass, S., 2021. How to Increase Automated Vehicles' Acceptance through In-Vehicle Interaction Design: A Review. Int. J. Human–Computer Interact. 37, 308–330. https://doi.org/10.1080/10447318.2020.1860517

DIN, E. I., 2018. 9241–11: Ergonomie der Mensch-System-Interaktion–Teil 11: Gebrauchstauglichkeit: Begriffe und Konzepte.

Harrington, C. N., Borgos-Rodriguez, K., Piper, A. M., 2019. Engaging Low-Income African American Older Adults in Health Discussions through Community-based Design Workshops, in: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. Presented at the CHI '19: CHI Conference on Human Factors in Computing Systems, ACM, Glasgow Scotland Uk, pp. 1–15.

Kelley, T., Kelley, D., 2013. Creative confidence: unleashing the creative potential within us all. William Collins, London.

Kuckartz, U., 2012. Qualitative Inhaltsanalyse: Methoden, Praxis, Computerunterstützung. Beltz Juventa.

Lee, H. R., Šabanović, S., Chang, W.-L., Nagata, S., Piatt, J., Bennett, C., Hakken, D., 2017. Steps Toward Participatory Design of Social Robots: Mutual Learning with Older Adults with Depression, in: Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction. Presented at the HRI'17: ACM/IEEE International Conference on Human-Robot Interaction, ACM, Vienna Austria, pp. 244–253.

Lotz, V., Schomakers, E.-M., Ziefle, M., 2022a. Users' Preferences for the Communication with Autonomous Micro-Vehicles, in: 2022 IEEE Intelligent Vehicles Symposium (IV). Presented at the 2022 IEEE Intelligent Vehicles Symposium (IV), pp. 1168–1173.

Lotz, V., Schomakers, E.-M., Ziefle, M., 2022b. Acceptance of app based interaction between automated micro vehicles (AMV) and their users. Presented at the 13th International Conference on Applied Human Factors and Ergonomics (AHFE 2022).

Madigan, R., Louw, T., Dziennus, M., Graindorge, T., Ortega, E., Graindorge, M., Merat, N., 2016. Acceptance of Automated Road Transport Systems (ARTS): An Adaptation of the UTAUT Model. Presented at the Transportation Research Procedia, pp. 2217–2226.

Mayring, P., 1994. Qualitative Inhaltsanalyse, in: Boehm, A., Mengel, A., Muhr, T. (Eds.), Texte verstehen : Konzepte, Methoden, Werkzeuge, Schriften zur Informationswissenschaft. UVK Univ.-Verl. Konstanz, Konstanz, pp. 159–175.

Mumford, E., 2000. A Socio-Technical Approach to Systems Design. Requir. Eng. 5, 125–133.

Neto, P., Simão, M., Mendes, N., Safeea, M., 2019. Gesture-based human-robot interaction for human assistance in manufacturing. Int. J. Adv. Manuf. Technol. 101, 119–135.

NuroInc., 2016. Nuro—On a mission to better everyday life through robotics. [WWW Document]. Nuro. URL https://www.nuro.ai/ (accessed 2.1.24).

Osswald, S., Wurhofer, D., Tr, S., 2012. Predicting information technology usage in the car: towards a car technology acceptance model 8.

Piaggio, 2024. Der revolutionäre Charakter von GITA und KILO [WWW Document]. piaggio. URL https://www.piaggio.com/at_DE/piaggio-world/about-us-piaggio/der-revolutionare-charakter-von-gita-und-kilo/ (accessed 2.1.24).

Plattner, H., Meinel, C., Weinberg, U., 2009. Design thinking: Innovation lernen-Ideenwelten öffnen. mi-Wirtschaftsbuch.

Pluta, W., 2017. Roboter: Wenn der Postbot dem Postboten die Post trägt - Golem.de [WWW Document]. URL https://www.golem.de/news/postbot-die-post-testet-zustellungsroboter-1710-130446.html (accessed 2.1.24).

Reske, M., 2020. UrbANT - Projekt [WWW Document]. URL https://urbant.de/de/projekt.html (accessed 2.1.24).

Schomakers, E.-M., Klatte, M., Lotz, V., Biermann, H., Kober, F., Ziefle, M., 2022. Analysis of the potential of a new concept for urban last-mile delivery: Ducktrain. Transp. Res. Interdiscip. Perspect. 14, 100579.

Schrepp, M., Hinderks, A., Thomaschewski, J., 2014. Applying the User Experience Questionnaire (UEQ) in Different Evaluation Scenarios.

Starship, 2014. About Starship: The global leader in autonomous delivery [WWW Document]. Starship Technol. Auton. Robot Deliv. URL https://www.starship.xyz/company/ (accessed 2.1.24).

Steen, M., 2013. Co-Design as a Process of Joint Inquiry and Imagination. Des. Issues 29, 16–28.

Suchman, L., 2002. Located accountabilities in technology production 14.

Toner, J., 2009. Small is not too Small: Reflections Concerning the Validity of Very Small Focus Groups (VSFGs). Qual. Soc. Work 8, 179–192. https://doi.org/10.1177/1473325009103374

Venkatesh, V., Thong, J. Y. L., Xu, X., 2012. Consumer Acceptance and Use of Information Technology: Extending the Unified Theory of Acceptance and Use of Technology. MIS Q. 36, 157–178.