Advances in Human Factors of Transportation, Vol. 148, 2024, 549–559
https://doi.org/10.54941/ahfe1005246

User Requirement Analysis for Voice and Gesture Interactions With Delivery Robots: An Interview Study

Vivian Lotz, Eva-Maria Schomakers, and Martina Ziefle

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

ABSTRACT

Urban mobility is rapidly changing. While increasing delivery volumes, traffic congestion, and a demand to reduce mobility-induced emissions challenge inner-city logistics, emerging technologies such as automated delivery robots might offer relief. Here, various interaction concepts are conceivable to ensure safe navigation and smooth communication with them. This study qualitatively examines user requirements and prevailing user perceptions of two communication modalities (voice interface and gesture interface) for interacting with delivery robots. We conducted 24 scenario-based interviews. Each interview included a part in which participants actively tried voice or gesture commands for operating the delivery robots. The practice part was intended to ensure that all participants get as realistic a sense as possible of how interactions might feel so that the discussion of requirements was not purely based on potentially flawed imaginations. Results were analyzed using qualitative content analysis and revealed universal barriers (e.g., ambiguousness of inputs) and differences regarding the two modes (e.g., privacy concerns).

Keywords: Human-machine-interaction, Interview study, Voice interaction, Gesture interaction, Delivery robots, Micro-vehicles

INTRODUCTION

Mobile robots have been a topic of research for some time. With the advancement of automation, they have been more and more discussed as a potential solution for the increased demand to transport goods in dense urban environments (Lyons & McDonald, 2023; Baum et al., 2019). Delivery robots are already used in several locations in the U.K. and U.S. (Starship, 2014).

As full automation still has limits and legal approval is still pending in most countries, some delivery robots employ a follow-me strategy (Baum et al., 2019).

Such vehicles are designed to track and automatically follow a person or object via sensors or cameras as they move. Unlike fully autonomous vehicles or robots, they require human intervention and are designed to work with a human user, e.g., a delivery person.

This study focused on a vehicle sized 1m times 2.2m developed to transport larger goods. With a maximum speed of 25 kph and the option to drive in convoys, these vehicles are designed to be used primarily on bike lanes or sidewalks with the option to move on roads. For a more comprehensive overview and illustration of the robots, see (Schomakers et al., 2022).

The prospect of using follow-me robots to address last-mile logistics challenges is promising, yet integrating new transportation modes into complex urban systems is not without risks. Vehicle automation often faces apprehension and potential rejection, making an apriori understanding of what constitutes acceptance crucial (Brell et al., 2021; Kyriakidis et al., 2015; Othman, 2021). Users need to be able to use follow-me technologies easily and safely. A critical aspect of this is understanding user preferences for interaction modalities, as how users interact with these robots influences their acceptance and overall experience.

Previous research has mainly focused on the acceptance of delivery robots quite generally (Pani et al., 2020; Yuen et al., 2022). Interaction design has hitherto been explored only sporadically, e.g. by (Dautzenberg et al., 2021)

This study focuses on voice and gesture interactions as primary modes of communication. These interaction modalities are particularly relevant in dense urban environments where users may need hands-free, visually unobtrusive, and hygienic ways to control the robots, making them more practical (Chen et al., 2017). However, in a prior study, we found that while voice control is considered desirable, there was some reservation among people regarding gesture control (Lotz et al., 2022), which makes contrasting the reasons people have for and against using them worth exploring.

Research Aim

This article explores the (1) challenges and (2) user requirements for interacting with delivery robots using speech or gestures in urban settings.

The main goal was to delve into the specific needs of each interaction mode and understand the reasons behind users' requirements and usage barriers. Given the lack of a structured framework for modality overarching interaction design features, we opted for an exploratory research method, using semi-structured interviews with prospective users to compile an overview of requirements and barriers.

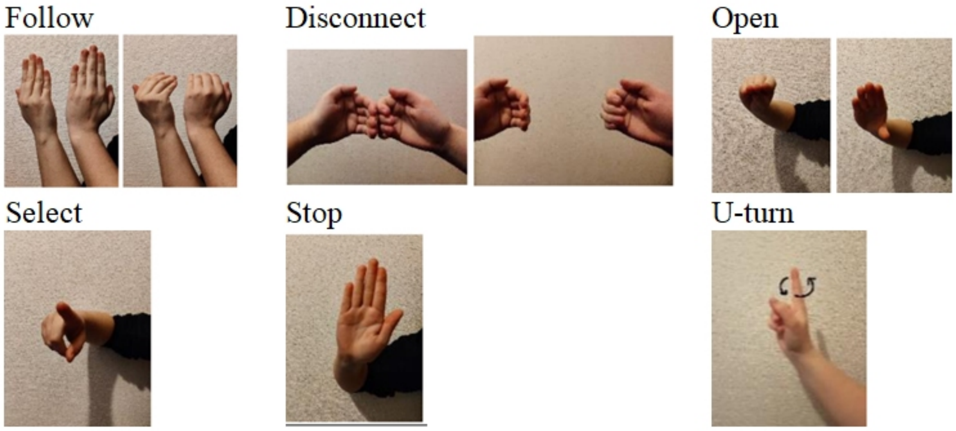

Figure 1: Overview of gesture user inputs used.

METHOD

Interview Procedure

In total, 24 semi-structured scenario-based interviews (9 focused on voice and 15 on gesture interactions) were conducted. Interviews were conducted in Germany in December 2021. Audio recordings were verbatim transcribed. Only essential parts, such as hands during the gesture interaction, were filmed to ensure participant anonymity during trial interactions.

The study employed a method mix, combining semi-structured interviews with scenario-based trial interactions (Patton, 2005).

The interview guideline was based on a literature review, pretested with five participants, and refined for clarity. Before the interview, participants were introduced to delivery robots and the study's objective. Demographic information, including gender, age, education, and experience with voice and gesture interactions, was collected via an online survey.

The interview was structured into three main parts. The first part included the introduction and a warm-up, where participants shared their prior experiences with voice or gesture interactions. The second part was slightly different between gesture- and voice-focused groups. Gesture interaction participants were asked to memorize standard commands (Figure 1). The gestures were chosen based on a literature review (Abendroth et al., 2019; ESC, 2021; Loehmann et al., 2013). This step was omitted for voice-focused interviews to capture intuitive verbal interactions with the robots. Both groups then engaged in a realistic scenario involving transporting furniture with two vehicles, completing three tasks: let two robots follow, unload two robots, and switch to interact with only one robot. The trial aimed to give participants a tangible experience of the interactions for a more informed discussion on benefits, barriers, and requirements. During the final part of the interview, participants reflected on their experiences, discussing perceived advantages, challenges, and necessary features.

Analysis

Results were analyzed using qualitative content analysis according to (Kuckartz, 2012; Mayring & Fenzl, 2019). Categories were primarily established deductively based on acceptance and user experience-related constructs (Schrepp et al., 2014; Venkatesh et al., 2012) and complemented by inductive categories. After categorization, the statements were checked for differences in requirements and barriers between modalities.

Sample

Since the robots are primarily designed for urban areas, the sample was selected with particular regard to urban dwellers. In total, 15 people participated in the gesture-focused interviews and 9 in the voice-focused interviews (M = 29.71, SD = 13.662). Both groups were primarily female (gesture-group: 11,73.3 % women; voice-group: 7,77.8 % women) and well-educated (gesture-group: 14,93.3 %; voice-group: 5,66.7 % university entrance qualification or higher).

Participants were experienced in using gestures for operating touchscreens (M = 4.13, SD = 1.356, max = 5 "regular use") or game consoles (M = 3.53, SD = 1.246) - e.g., Wii sports - but not in air gestures as used in the interviews (M = 2.27, SD = 0.884). Further, participants stated they were familiar with and actively used voice assistants in various contexts, such as service hotlines (M = 4.00, SD = 0.000) and navigation systems (M = 3.22, SD = 1.093).

RESULTS

This study explored perceived barriers and interaction requirements of two interface modalities for interaction between users and (follow-me) delivery robots. The following section presents the results of the interview study.

Perceived Barriers to Interact

When discussing which aspects might deter from using a specific mode of interaction for delivery robots, participants raised concerns about safety, reliability, lack of control, ease of use and learning, efficiency issues, and how others perceive the interaction. Most of these aspects are intertwined with each other, and often, one concern fuels another.

Safety. First and foremost, participants were concerned about traffic safety, fearing that using the mode might lead to increased accident risks while at the same time talking about the unclear liabilities if such accidents occurred when automated systems were involved. Safety concerns were heavily influenced by the perception that both interaction modalities were unfamiliar and (still) unreliable.

"You might cause accidents and hurt someone. Who would be responsible? Is it even possible to prove the system was malfunctioning?" [female, 22 years, gesture interviews].

Moreover, participants raised privacy and data protection concerns about using voice control.

"Data protection is also an issue. Your voice will be recorded, and you could do whatever with the recordings." [female, 21 years, voice interviews].

Further, participants expressed concerns that gesture and voice control could increase the risk of theft or misuse of delivery robots. They were apprehensive about the lack of secure authentication methods and transparency of communication partners, especially in gesture control. This absence of a clear, visible connection or identification between the user and the robot led to concerns that bystanders could easily disrupt the interaction and potentially hijack the robot.

"At that point, I would be worried that some stranger would find it fun to stop my robot. And then my robot listens to that person and not me." [female, 23 years, voice interviews].

Reliability. Participants doubted the technology's readiness, stressing its tendency to make errors, particularly in voice interactions. The perceived unreliability was further aggravated by external interferences like weather, noise, and traffic, which were expected to increase system errors. A primary concern was the system misinterpreting user commands, leading to unintended reactions, a problem they expected in both voice and gesture modalities, leading to decreased safety of users and other road users.

"You're not sure what to say exactly - and that leads to misunderstanding. Then that thing (the robot) does things you did not want it to do." [female, 21 years, voice interviews].

Perceived lack of control. A key issue was the ambiguous communication direction in both modes, causing users to feel powerless. This ambiguity was expected to lead to unintended system behaviors. Users anticipated problems when trying to command one specific robot among several and worried that the robot might struggle to identify its intended user and vice versa.

"If I want to take just one robot. But there's the other still nearby. Both might think they have to come with me." [female, 27 years, voice interviews].

Some respondents also said they did not like giving inputs to let the robot act autonomously – especially if the action could not be monitored directly.

"Well, it might be a little weird if I tell him to drive into the garage. You just don't have him in your sight anymore. I find that a bit weird." [female, 22 years, voice interviews].

The final barrier in this category was the robots' unpredictable behavior, leading to unreliable outcomes. Participants anticipated that the robots might behave unexpectedly, complicating effective control. This issue mainly arises because voice and gesture interactions offer a more extensive range of responses than traditional graphical user interfaces or controllers, typically leading to more predictable and consistent system behaviors.

"It is not like using a remote control where I know for certain what happens if I push a button." [male, 50 years, gesture interviews].

Ease of use and learning. Particularly for gesture control, participants stated its complexity, unfamiliarity, and lack of transparent input as perceived disadvantages. They expected learning the gesture "language" to be challenging due to its unfamiliar nature. Moreover, the range of possible inputs added confusion, leaving respondents uncertain about the available commands, the correct sequence of inputs, and their flexibility. Even though more pronounced for gesture interaction, these concerns applied to both gesture and voice modes.

"It is not that hard to learn the gestures. But you need some time to get the sequence right. You have to get a feeling for what is possible." [female, 23 years, gesture interviews].

Efficiency. Concerning efficiency, several respondents believed that voice control might be less time-efficient than graphical user interfaces. They attributed this to the perceived unreliability of voice control and the likelihood of repeating commands multiple times.

"It's undoubtedly faster to use an app. You just have to click to form a new connection. Much faster than talking to those guys (the robots)." [female, 27 years, voice interviews].

Social norm. Another mentioned barrier was the potential impact of gesture or voice control in busy environments. Some respondents were concerned about being watched by others and the possibility that their gestures, meant as commands for the robots, might be misinterpreted as communication directed at the people around them.

"I wouldn't be embarrassed. But others might misinterpret when I interact with my robot. They might find that weird or think I want them to come." [female, 33 years, gesture interviews].

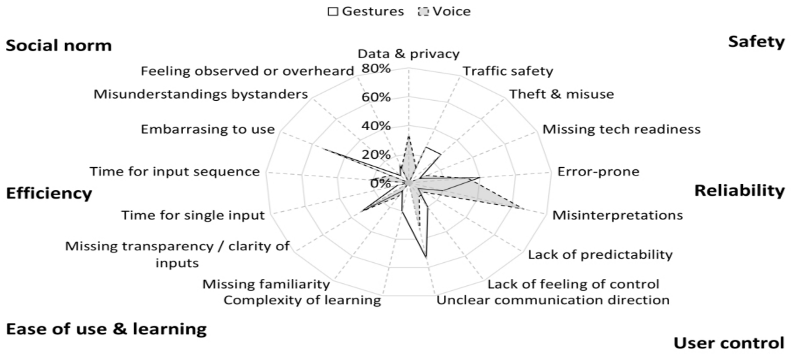

Figure 2: Overview of mentioned barriers for voice and gesture interactions in % (i.e., the share of interviews in which the barrier was mentioned).

Differences between modalities. Figure 2 illustrates the frequency of different barriers mentioned for both interaction modes. Privacy was mainly a concern for voice interactions, whereas gesture interactions brought more attention to traffic safety and theft risks. Misinterpretations and inefficiency were more prominent in voice interactions than gesture interactions, while ease of use and learning challenges were mainly associated with gesture interactions. Additionally, concerns about how others perceive the use of the interface were primarily raised in discussions about gesture-based interactions.

Requirements

In total, the participants mentioned eight requirement categories, namely: (1) Ease of use and learning, (2) risk and safety, (3) dependability and user control, (4) reliability and effectiveness, (5) accessibility and user diversity, (6) stimulation, (7) efficiency, and a (8) pleasant response behavior. The identified requirements largely align with the previously discussed barriers for both interaction modes. New elements such as accessibility, stimulation, and suitable response behavior were also emphasized.

Ease of use mainly comprised aspects of consistency within the system and with similar systems and prior experience, initial training, simplicity, clarity, transparency of provided information, and personalization.

Regarding **safety expectations**, it was stressed that the severity and likelihood of associated risks should be low, and people's data should be appropriately protected. Participants also discussed the need to define liabilities and regulations, avoid interaction-induced distractions, and include user authorization procedures.

To develop an adequate level of **trust in and control of the system**, participants wished for predictability and unambiguousness, i.e., transparent communication direction and partners, predictability of the system's next steps, and enabling user interventions anytime from anywhere. Discussed measures included appropriate system feedback and information.

Concerning **reliability and effectiveness**, participants stressed that unwanted system reactions and failures to react to inputs quickly lead to frustration. Further, the **speed** with which interaction goals are archived, inputs can be made, or the system reacts should be as quick.

Moreover, participants agreed that **accessibility** is essential when designing interactive systems. The discussion of accessibility requirements revolved around the system being reliable regardless of used language, talking habits (volume, speed, pitch, or dialect), and impairments and being understandable regardless of the user's cultural background, expertise, and age.

"Moderator: Why is accessibility important to you? P: I imagine it would be practical. Someone with a different language background should be able to talk to it in English without going through the menu." [female, 21 years, voice interviews].

Stimulation aspects were mainly mentioned by the participants in the gesture interaction interviews but rarely by the voice group.

"Moderator: Why do you like gesture interaction? P: It might be exciting. And other people around me also have something to look at" [female, 51 years, gesture interviews].

Lastly, for the aspect of the robot's **response behavior**, participants agreed that it should feel unobtrusive while still being cheerful and friendly. Another important aspect was the transparent communication of errors and system limitations so that the user could build trust.

"For example, if he does not recognize me, he could do a sad smile and apologize for that and make clear to me that it is an error on his part." [female, 23 years, gesture interviews].

DISCUSSION

This section discusses findings from the interviews, starting with the key findings. Lastly, the study's limitations are summarized.

There were six highly interlinked aspects (safety, reliability, control, ease of use, efficiency, and social norm) respondents were concerned about when thinking about using gesture or voice control to operate a delivery robot, most of which are congruent with established factors from research on user experience and technology interaction research (Schrepp et al., 2014; Venkatesh et al., 2012). The most notable additions to earlier research were the issue of in-transparency of the communication direction. The question is how the user recognizes that the system aims information at him and vice versa. Second, there is uneasiness with unmonitored autonomous robot actions. Trust has been recognized in earlier research as crucial (Choi & Ji, 2015), yet during the interviews, it wasn't directly mentioned by the participants. Instead, their stated expectations, such as reluctance to leave robots unmonitored and anticipating unreliability, subtly implied issues of trust. This suggests that mistrust, akin to concerns about traffic safety, often arises from anticipated failures to meet expectations. Thus, while not always explicitly stated, the importance of trust can be inferred from the way people talk about their concerns. It suggests that understanding the nuances of human perceptions requires looking beyond what is directly said to also consider what is implied or indicated through other means, and that trust hinges on how well it performs. It should be noted that while missing trust can hinder adoption over-trust is just as problematic as it can lead to complacency, safety issues, and a misalignment of expectations.

As expected, there were some differences regarding the perceived usage barriers between the two interaction modes. Concerns about privacy or miscommunication between the user and system and a missing overview of which user inputs are available were more pronounced for voice than gesture interactions. In contrast, gesture interaction was more heavily linked to worries about traffic safety issues, responding to inputs from bystanders, and a general uneasiness to use this novel interaction form in public spaces.

Interestingly, the data protection issue was primarily discussed for voice interactions, possibly because the issue is more salient and frequently discussed in the media in the context of voice assistants. However, whether there is a difference in the importance of privacy protection between the modalities needs to be re-evaluated using quantitative methods.

Overall, the most pressing concern for both interaction modes was whether or not such an interaction was safe (enough) for them, other road users, and the robot.

Possible countermeasures might be to define clear regulations and liabilities and communicate them to users so that if incidents happen, users know the handling and consequences - because this uncertainty seems to be the root cause for their uneasiness. Further, linking the user and the robot more tangible or visible to the user might also help. For example, including an authentification step before the interaction is possible, including cues that the robot is now listening to the user, or adding visual feedback about the coupling of the user and robot might be helpful.

Moreover, the results revealed a trade-off between the wish for complete information to convey a feeling of control and too much information leading to higher usage complexity and distractions from monitoring the traffic, which must be carefully balanced when designing interactions.

Regarding requirements, most suggestions aimed to mitigate the previously outlined perceived barriers. Additional aspects included ensuring accessibility and stimulation and designing the response behavior to be friendly while remaining unobtrusive.

Strengths and Limitations

Identifying relevant challenges and requirements builds a solid basis for deepening and elaborating on acceptance determinants. Further, the present study takes a first step towards uncovering the rationale behind peoples' preferences and highlights potential differences between the perception of and requirements for interaction modes. The obtained insights enable the focus of subsequent quantitative validations and experimental analyses on relevant aspects.

While the results are valuable, it should be noted that scalability is still underexplored (Colley et al., 2020). In the present study, the experience took place in a controlled laboratory environment, excluding bystanders or other vehicles. In future work, the proposed concepts should be evaluated in a realistic environment to gain further insights into how they integrate into the broader traffic landscape. Another aspect for future work is to explore the trade-off between enhanced system transparency and explainability of automation by providing corresponding information and keeping user distractions low by limiting the displayed information to essential information.

The sample was skewed towards young female participants. Therefore, perceptions of older and male people should be considered in future studies. Lastly, the study was conducted with German participants, which limits generalizability. While the overarching challenges are assumed to be universal (e.g., ensuring safety), there are international differences in traffic law, behavior, technology acceptance, and technical approval processes (Güliz Uğur, 2017; Özkan et al., 2006).

PRACTICAL IMPLICATIONS AND CONCLUSION

Although most discussed findings align with the existing literature on human-machine interaction, comparing modes provides new insights into aspects that hamper or vice versa drive the willingness to engage with interactive systems.

From the obtained insights, it is clear that both modalities are still met with some apprehension, and successful implementation requires resolving technology-inherent issues such as faulty input detection and finding solutions to system-related issues such as the unclear state of liabilities. Central to all endeavors should be to ensure the system's safety and design to increase its robustness against faulty user inputs. Safety was linked to nearly all other mentioned requirements and barriers during the interviews. As such, it should be included as a central variable in modeling the acceptance of such systems.

Apart from the overarching issues that apply to most novel technologies, specific issues to work on are (1) how to convey the participants and the direction of communication. Further, some thought has to be put into (2) how the user can learn and become proficient with the range of available commands and develop an accurate mental model of the automated system's capabilities and limits. Initial training can help, as might the use of supplementary visual cues.

ACKNOWLEDGMENT

The authors would like to thank all participants for their valuable input. This research was funded by the Ministry of Culture and Science of the state of North Rhine-Westphalia (321-8.03.07-127598) and the Federal Ministry for the Environment, Nature Conservation, Nuclear Safety, and Consumer Protection (67KI31073).

REFERENCES

Abendroth, B., Müller, A., Zöller, I. & Ateu, S., 2019. *Gestensteuerung intuitiv gestalten - Eine Betrachtung ausgewählter Bedienfunktionen im Fahrzeug*. [Online] Available at: https://gfa2019.gesellschaft-fuer-arbeitswissenschaft.de/inhalt/C.3.1.pdf [accessed: 12 10 2022].

Baum, L., Assmann, T. & Strubelt, H., 2019. State of the art-Automated micro-vehicles for urban logistics. *IFAC-PapersOnLine,* 52(13), pp. 2455–2462.

Brell, T., Philipsen, R., Biermann, H. & Ziefle, M., 2021. Social Acceptance of Autonomous Driving--The Importance of Public Discourse and Citizen Participation. In: *Smart Transportation.* s.l.: CRC Press, pp. 1–17.

Chen, L.-C., Cheng, Y.-M., Chu, P.-Y. & Sandnes, F. E., 2017. Identifying the usability factors of mid-air hand gestures for 3D virtual model manipulation. *Universal Access in Human--Computer Interaction. Designing Novel Interactions: 11th International Conference, UAHCI 2017, Held as Part of HCI International 2017, Vancouver, BC, Canada, July 9–14, 2017, Proceedings, Part II 11*, pp. 393–402.

Choi, J. K. & Ji, Y. G., 2015. Investigating the importance of trust on adopting an autonomous vehicle. *International Journal of Human-Computer Interaction,* 31(10), pp. 692–702.

Colley, M., Walch, M. & Rukzio, E., 2020. Unveiling the lack of scalability in research on external communication of autonomous vehicles. *Extended abstracts of the 2020 chi conference on human factors in computing systems*, pp. 1–9.

Dautzenberg, P., Voß, G. M. I., Ladwig, S. & Rosenthal-von der Pütten, A., 2021. Investigation of different communication strategies for a delivery robot: the positive effects of humanlike communication styles. *30th IEEE International Conference on Robot \& Human Interactive Communication (RO-MAN)*, pp. 356–361.

ESC, E. S. L. C., 2021. *Spread the sign..* [Online] Available at: https://www.spreadthesign.com/de [accessed: 12 10 2022].

Güliz Uğur, N., 2017. Cultural Differences and Technology Acceptance: A Comparative Study. *Journal of Media Critiques,* Band 11, pp. 123–132.

Kuckartz, U., 2012. *Qualitative inhaltsanalyse.* s.l.: Beltz Juventa.

Kyriakidis, M., Happee, R. & de Winter, J., 2015. Public opinion on automated driving: Results of an international questionnaire among 5000 respondents. *Transportation research part F: traffic psychology and behaviour,* Band 32, pp. 127–140.

Loehmann, S., Knobel, M., Lamara, M. & Butz, A., 2013. Culturally independent gestures for in-car interactions}. *Human-Computer Interaction--INTERACT 2013: 14th IFIP TC 13 International Conference, Cape Town, South Africa, September 2-6, 2013, Proceedings, Part III 14*, pp. 538–545.

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

Lyons, T. & McDonald, N., 2023. Last-Mile Strategies for Urban Freight Delivery: A Systematic Review. *Transportation Research Record,* 2677(1), pp. 1141–1156.

Mayring, P. & Fenzl, T., 2019. Handbuch Methoden der empirischen Sozialforschung. In: s.l.: Springer, pp. 633–648.

Othman, K., 2021. Public acceptance and perception of autonomous vehicles: A comprehensive review. *AI and Ethics,* 1(3), pp. 355–387.

Özkan, T. et al., 2006. Cross-cultural differences in driving behaviours: A comparison of six countries. *Transportation research part F: traffic psychology and behaviour,* 9(3), pp. 227–242.

Pani, A., Mishra, S., Golias, M. & Figliozzi, M., 2020. Evaluating public acceptance of autonomous delivery robots during COVID-19 pandemic. *Transportation research part D: Transport and environment,* Band 89, p. 102600.

Patton, M. Q., 2005. Qualitative research. *Encyclopedia of statistics in behavioral science.*

Schomakers, E.-M.et al., 2022. Analysis of the potential of a new concept for urban last-mile delivery: Ducktrain. *Transportation Research Interdisciplinary Perspectives,* Issue 14, p. 100579.

Schrepp, M., Hinderks, A. & Thomaschewski, J., 2014. Applying the user experience questionnaire (UEQ) in different evaluation scenarios. *Design, User Experience, and Usability. Theories, Methods, and Tools for Designing the User Experience: Third International Conference, DUXU 2014, Held as Part of HCI International 2014, Heraklion, Crete, Greece, June 22–27, 2014, Proceedings, Part I 3*, pp. 383–392.

Starship, 2014. *Starship: The global leader in autonomous delivery.* [Online] Available at: https://www.starship.xyz/company/ [accessed: 02 08 2023].

Venkatesh, V., Thong, J. Y. & Xu, X., 2012. Consumer acceptance and use of information technology: Extending the unified theory of acceptance and use of technology. *MIS quarterly,* pp. 157–178.

Yuen, K. F., Cai, L., Lim, Y. G. & Wang, X., 2022. Consumer acceptance of autonomous delivery robots for last-mile delivery: Technological and health perspectives. *Frontiers in Psychology,* Band 13, p. 953370.