

Potential of Socially Assistive Robotics for an Application within the Field of Active and Assisted Living

DISSERTATION

submitted in partial fulfilment of the requirements for the degree of

Doktor der Technischen Wissenschaften

by

Dipl.-Ing. Mag. Franz-Lothar Werner

Registration Number: 0126476

to the Faculty of Informatics

at the Vienna University of Technology

Advisor: Ao.Univ.Prof. Dipl.-Ing. Dr. Wolfgang L. Zagler

The dissertation has been reviewed by:

Assoc. Prof. Raymond H. Cuijpers, PhD

Asst. Prof. Özge Subaşı, PhD

Vienna, 30th April, 2020

Franz-Lothar Werner

Technische Universität Wien

A-1040 Wien • Karlsplatz 13 • Tel. +43-1-58801-0 • www.tuwien.at

this page intentionally left blank

Erklärung zur Verfassung der Arbeit

Hiermit erkläre ich, dass ich diese Arbeit selbstständig verfasst habe, dass ich die verwendeten Quellen und Hilfsmittel vollständig angegeben habe und, dass ich die Stellen der Arbeit, einschließlich Tabellen, Karten und Abbildungen, die anderen Werken oder dem Internet im Wortlaut oder dem Sinn nach entnommen sind, auf jeden Fall unter Angabe der Quelle als Entlehnung kenntlich gemacht habe.

Vienna, 30th April, 2020

Franz-Lothar Werner

this page intentionally left blank

Abstract

The not-too-distant future will see an increasing demand for elder care and a shortage of professional and informal caregivers. Ageing society would benefit from technological systems that are capable of supporting older people in a way that allows them to stay independent for longer in their own home. Socially assistive robots (SARs) [Feil-Seifer2005] could become an aid for older users as they introduce unique support prospects. They provide multimodal communication channels and thereby allow for intuitive interaction. Their physical embodiment was found to influence the user's acceptance, adoption and perception and their ubiquitous mobility allows them to cover a large number of use cases that otherwise would have to be carried out by a variety of dedicated systems.

Still, socially assistive robotics is a burgeoning field of science with yet unclear potential in terms of end-user acceptance, uptake by users and impacts on the quality of care and daily life. Solutions developed so far are mostly in a state of research and general technological readiness is low due to their inherent technical complexity. Current methods for development and evaluation of such solutions are rather vague and often not replicable, limiting the potentials of integration and spreading of SARs. This dissertation aims to enhance current methodologies for user-centred evaluation and to give indications of the potential of such systems in supporting older adults.

Within the scope of this dissertation, a series of three prototypes of socially assistive robotic solutions that support older users and their caregivers was evaluated. A reiterating participative evaluation process was applied to investigate performance, acceptance and impact factors as well as methods to evaluate SARs. Each demonstrator was evaluated together with end-users and domain experts in user studies within laboratory and living-lab settings.

Integrative methods for the user-centred evaluation of SARs in real-life-like settings were developed, evaluated and validated by means of user studies with the developed prototype systems. Results of the conducted user studies are given regarding the systems' dependability, acceptance, applicability, motivational abilities and potential impacts for use by the target groups of older users, secondary end-users such as relatives, carers, care experts and therapists, and tertiary users such as managers of care centres. Design guidelines for future research clearly stating reusable methods and strategies to develop and evaluate assistive robotics are given, including

recommendations for successful human-robot interaction within applied domains of assistive technologies.

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.



Kurzfassung

In nicht allzu ferner Zukunft ist auf Grund der demographischen Entwicklung in Österreich und Europa ein steigender Bedarf an Altenpflege und ein Mangel an professionellen und informellen Pflegekräften zu erwarten. Die alternde Gesellschaft würde von Systemen profitieren, die in der Lage sind, ältere Menschen so zu unterstützen, dass diese länger unabhängig in ihrem eigenen Zuhause leben können. Socially Assistive Robots (SAR) [Feil-Seifer2005] könnten zu einem Hilfsmittel für ältere Nutzerinnen und Nutzer werden, da sie einzigartige Unterstützungsperspektiven bieten. Sie stellen multimodale Kommunikationskanäle zur Verfügung und ermöglichen so eine intuitive Interaktion. Ihre physische Verkörperung hat Einfluss auf die Akzeptanz und Wahrnehmung durch Nutzerinnen bzw. Nutzer und ihre Mobilität ermöglicht es verschiedenste Anwendungsfällen abzudecken, die sonst von einer Vielzahl von dedizierten Systemen durchgeführt werden müssten.

Dennoch ist die sozial-assistive Robotik ein junges Wissenschaftsgebiet mit noch unklaren Potenzialen in Bezug auf Endnutzerakzeptanz, Auswirkungen auf die Qualität der Pflege und des täglichen Lebens. Bisher entwickelte Lösungen befinden sich meist im Forschungsstadium und die technologische Reife ist aufgrund der inhärenten technischen Komplexität gering. Aktuelle Methoden zur Entwicklung und Bewertung solcher Lösungen sind oft vage und nicht reproduzierbar, was den wissenschaftlichen Wert der vorgelegten Evidenz einschränkt. Diese Arbeit zielt darauf ab, aktuelle Methoden zur nutzerzentrierten Evaluation zu verbessern und gibt Hinweise auf die Potenziale solcher Systeme zur Unterstützung älterer Erwachsener.

Im Rahmen dieser Arbeit wurde eine Serie von drei Prototypen von sozial-assistiven Roboterlösungen zur Unterstützung älterer Nutzer und Nutzerinnen, sowie ihrer Betreuerinnen und Betreuer evaluiert. Ein sich wiederholender partizipativer Evaluationsprozess wurde angewandt, um Performanz-, Akzeptanz- und Wirkungsfaktoren, sowie Methoden zur Evaluierung von sozial unterstützenden Robotern zu untersuchen. Jeder Prototyp wurde zusammen mit Endnutzerinnen und -nutzern sowie Fachexpertinnen und -experten in Nutzerstudien in Labor- und Wohnlaborsituationen evaluiert.

Methoden zur nutzerzentrierten Evaluierung von sozial-assistiven Robotern in realitätsnahen Umgebungen wurden entwickelt, evaluiert und durch Nutzerstudien mit den entwickelten Prototypensystemen validiert. Die Ergebnisse dieser Nutzerstudien

werden hinsichtlich der Zuverlässigkeit, Akzeptanz, Anwendbarkeit, Motivationsfähigkeit und möglichen Auswirkungen auf die Zielgruppen der älteren Nutzerinnen und Nutzer, der sekundären Endnutzerinnen und -nutzer, wie z.B. Angehörigen, Betreuerinnen und Betreuern, Pflegeexpertinnen und -experten und Therapeutinnen und Therapeuten und der tertiären Nutzerinnen und Nutzer wie z.B. Führungskräften von Pflegezentren dargestellt. Es werden Gestaltungsrichtlinien für die zukünftige Forschung vorgestellt, die wiederverwendbare Methoden und Strategien zur Entwicklung und Bewertung von assistiver Robotik nennen und Empfehlungen für zielführende Mensch-Roboter-Interaktion gegeben.

Acknowledgements

This thesis was written over a period of nearly seven years. During this long time I had the privilege to meet and work with many great people in international or national collaborations. I would like to use this opportunity to express my gratitude to the people who supported me during my work on this dissertation or otherwise made an impact on the work presented.

I would like to send very warm thanks to the many colleagues who worked with me, mostly over the duration of several years within research projects that allowed me to pursue the undertaken research. I would like to thank my colleagues at raltec for the great time we had at the institute and their support of my work. In particular Johannes Oberzaucher for his vision, great team-skills and in-dept discussions within the projects KSERA and PhysicAAL, Walter Hlauschek for the experienced, profound and unstressed lead of our institute which strengthened our backs to concentrate on our research topics, Daniela Krainer for her tireless dedication and her expertise in therapy within the project PhysicAAL and, of course, my wife Katharina Werner who worked with me day and night on the projects.

I would like to thank my supervisor Wolfgang Zagler for inspiring and supporting me to undertake research in the field of assistive technologies, for strongly influencing and shaping the research-field as it is now, as well as for his direct support and feedback on this thesis.

Many thanks also to my colleagues at our partner institutions such as Robert Trapp and Sabine Payr from the Austrian Research Institute for Artificial Intelligence for sharing their highly valued expertise within our project PotenziAAL, Elena Torta, Marco Bazzani, John Lemberger, Hadas Lewy and all others from the project KSERA for generating a friendly and productive atmosphere within the team.

I would also like to thank the colleagues at the Institute for Design & Assessment of Technology at the TU Wien for providing me a location and environment to work at. In particular Geraldine Fitzpatrick for the possibility to get integrated at the institute, Naemi Luckner, Michael Urbanek and Florian Güldenpfennig for their encouraging words on pursuing the dissertation, our break-chats and our continuing work in the field.

I am particularly thankful for all the support I received from end-users, care personnel and care institutions who voluntarily agreed to take part in the studies, invest their time and resources into the idea of assistive solutions for their daily care and work, despite the often large gap between care needs and the capabilities of the presented technological prototypes.

Special thanks belong to my family, firstly to my wife Katharina who supported me through all the years and did everything possible to provide me the enormous amount of time needed to pursue the research and finalize this dissertation. Thank you (and excuse me for taking so much time) to my daughters Johanna und Theresa for doing without me, in particular during the final writing phase.

Finally I would like to thank my parents for always supporting me, financing me and believing into my academic career.

Declarations

This dissertation was created in the period from 2013 to 2020 aside of three research projects and builds upon and includes knowledge gained within project-based funded research; the author's role within the respective projects is disclosed and the projects' input to the dissertation is given in the following paragraphs.

The EU-funded project "KSERA" mainly aimed at exploring the capabilities of socially assistive robots (SARs) for older people and developed a proof-of-concept solution. Our institute was partner within the project and responsible for the design and evaluation of the prototype and to implement a user-centred design (UCD) approach within a living-lab setting. I partly led the institute's part of the project towards the end of the project and was responsible for the "Integration and Evaluation" work-package during this phase. During the whole project, I largely contributed to the planning of the evaluation phase and the design of the methodology for the assessment of SARs. I was responsible not only for the conduction of trials but also partly for the integration of technical components to fit the trial sites' needs, and I conducted all Austrian trials. I was further responsible for the performance analysis and parts of the acceptance and impacts analysis. Within this dissertation, the methodology for evaluation developed during this project is described and taken as a starting point for the development of the presented SSUT method. The main results of my work within this project are included in condensed form in the results section of this dissertation.

The nationally funded research project "PhysicAAL" aimed at the development of a proof-of-concept of a SAR solution for the support of physical training. I was co-author of the research proposal and thereby shaped the research and its goals from the beginning to fit this dissertation. I was responsible for the design of the prototype from the start of the research and the overall research-project for the second half of the project including development activities, the planning and conduction of the evaluation phase and the analysis of results. The main contributions of other members of the team were parts of the prototype development, organizing the user involvement and research on the background and state-of-the-art regarding physiotherapy. The results of this project are included in the result section of this dissertation.

The main aim of the nationally funded study "PotenziAAL" was to review the current state-of-the-art of assistive robotics and discuss its potential. I was responsible for the research proposal and the setting of research goals together with a partnering institute.

I conducted several small-scale studies that led to main contributions. In particular, I conducted an analysis of the state-of-the-art of assistive robotics and current R&D, an analysis of user requirements and needs in relation to SARs, an analysis of the acceptance of assistive robotics and technical potentials and limitations as well as the analysis of methodological issues in current research. The state-of-the-art section within this dissertation profited from the results of this project because the generated overview could be used to frame the scope. Further, the section summary and discussion of results profited from the better overview and knowledge gained regarding the general issues and potentials of robotics.

Results gained within the above-mentioned projects are partly included in this dissertation if the main contribution to the result was given by the author.

To give credit to the various inputs of colleagues and the fact that ideas often arise out of interaction with others, within this dissertation I generally prefer to use the word “we” when talking about achievements and results.

List of Published Papers

The following list details the author's research papers that were written alongside the dissertation and published as journal articles, conference proceedings or, in one case, as a book chapter.

1. F. Werner, "A survey on current practices in user evaluation of companion robots," in *Evaluation Methods Standardization for Human-Robot Interaction*, C. Jost, B. Le Pvdic, T. Belpaeme, B. Cindy, D. Chrysostomou, N. Crook, M. Grandgeorge, and N. Mirnig, Eds. Springer, to be published.
2. K. Werner, F. Werner, "Assessing User Needs and Requirements for Assistive Robots at Home", in *Proceedings of the AAATE – Assistive Technology: Building Bridges*, Budapest, 2015, pp. 174-179.
3. S. Payr, F. Werner, and K. Werner, "AAL Robotics: State of the field and challenges," in *Proceedings of the eHealth2015 – Health Informatics Meets eHealth*, Vienna, 2015, pp. 117–124.
4. S. Payr, F. Werner, and K. Werner, "Potential of Robotics for Ambient Assisted Living," Study published by the Austrian Research Promotion Agency, Vienna, 2015.
5. E. Torta, F. Werner, D. O. Johnson, J. F. Juola, R. H. Cuijpers, M. Bazzani, J. Oberzaucher, J. Lemberger, H. Lewy, and J. Bregman, "Evaluation of a Small Socially Assistive Humanoid Robot in Intelligent Homes for the Care of the Elderly," *Journal of Intelligent and Robotic Systems*, vol. 76, no. 1, pp. 57–71, 2014.
6. D. Krainer, F. Werner, and J. Oberzaucher, "Performance of a socially assistive robot as trainer for physical exercises for older people," in *Proceedings of the 7th German AAL-Congress*, Berlin, 2014.
7. F. Werner, D. Krainer, J. Oberzaucher, and K. Werner, "Evaluation of the Acceptance of a Social Assistive Robot for Physical Training Support Together with Older Users and Domain Experts," in *Proceedings of the AAATE – Assistive Technology Research Series, Assistive Technology: From Research to Practice*, vol. 33, P. Encarnacao, L. Azevedo, G. J. Gelderblom, A. Newell, and N.-E. Mathiassen, Eds. IOS Press, 2013, pp. 137–142.

8. F. Werner, K. Werner, and J. Oberzaucher, "Evaluation of the acceptance of a socially assistive robot by older users within the project KSERA," in *Proceedings of the 6th German AAL Congress – Lebensqualität im Wandel von Demografie und Technik*, Berlin, 2013.
9. F. Werner and D. Krainer, "A Socially Assistive Robot to Support Physical Training of Older People – An End User Acceptance Study," in *Proceedings of the 5th international Conference on Social Robotics (ICSR 2013)*, 2013, pp. 562–563.
10. E. Torta, J. Oberzaucher, F. Werner, R. H. Cuijpers, and J. F. Juola, "Attitudes Towards Socially Assistive Robots in Intelligent Homes: Results From Laboratory Studies and Field Trials", *Journal of Human-Robot Interaction*, vol. 1, no. 2, pp. 76–99, 2013.
11. K. Werner, J. Oberzaucher, and F. Werner, "Evaluation of Human Robot Interaction Factors of a Socially Assistive Robot Together with Older People," in *Proceedings of the Sixth International Conference on Complex, Intelligent, and Software Intensive Systems*, 2012, pp. 455–460.
12. F. Werner, J. Oberzaucher, and K. Werner, "Real-life evaluation of a socially assistive robot," *Gerontechnology – International journal on the fundamental aspects of technology to serve the ageing society*, vol. 11, no. 2, p. 382, 2012.

Parts of the papers listed above were reworked into chapters from this dissertation:

- **User-centred evaluation methods (chapter 2.2.2):** parts of paper 4 were updated and integrated into this chapter and at the time of writing, an updated version is prepared for publication as a book chapter for Springer (paper 1).
- **Performance and acceptance results of the first prototype (chapters 4.3.1 and 4.3.2):** the presented results are summarized from the author's work in papers 12, 10 and the project report [KSERA2012].
- **Evaluation results of the second prototype (chapter 4.5):** the presented results are summaries from the author's work in papers 5, 8 and the project report [KSERA2012a].

Content

ABSTRACT.....	V
KURZFASSUNG.....	VII
ACKNOWLEDGEMENTS.....	IX
DECLARATIONS.....	XI
LIST OF PUBLISHED PAPERS.....	XIII
1 INTRODUCTION	1
1.1 MOTIVATION AND BACKGROUND.....	2
1.1.1 <i>Motivation No. 1: Demographic, economic and societal challenges</i>	2
1.1.2 <i>Motivation No. 2: Robotics as an age-old dream</i>	4
1.1.3 <i>Motivation No. 3: Promising and challenging intersection of AAL and robotics</i>	6
1.2 PROBLEM STATEMENT.....	7
1.3 GOALS AND RESEARCH QUESTIONS.....	10
1.4 CONTRIBUTION.....	12
1.5 METHODOLOGICAL APPROACH.....	14
1.6 STRUCTURE OF THIS DISSERTATION.....	16
2 STATE OF THE ART.....	19
2.1 (SOCIAL) ASSISTIVE ROBOTICS.....	19
2.1.1 <i>Typical functionalities of SARs</i>	26
2.1.2 <i>Uptake of SARs in current research & commercialization</i>	28
2.2 USER-CENTRED EVALUATION METHODS.....	30
2.2.1 <i>Methodology of the literature survey</i>	30
2.2.2 <i>Results of the literature survey</i>	32
3 INITIAL EVALUATION FRAMEWORK.....	47
3.1 USER-CENTRED DESIGN AS BASE FOR THE EVALUATION FRAMEWORK.....	47
3.2 THE LIVING LAB APPROACH.....	49
3.3 EVALUATION OF ROBOTIC SOLUTIONS – A DEFINITION.....	52
3.4 PRINCIPLES OF A SAR EVALUATION WITHIN A LIVING LAB ENVIRONMENT.....	54
3.5 DEVELOPMENT OF THE USER RESEARCH METHODOLOGY.....	56
3.5.1 <i>Problem description</i>	56
3.5.2 <i>Goals and requirements</i>	57
3.5.3 <i>Development sequence</i>	58
3.6 EVALUATION DOMAINS AND FACTORS.....	60
3.6.1 <i>The performance domain</i>	61
3.6.2 <i>The acceptance domain</i>	63
3.6.3 <i>The impacts domain</i>	66

3.7	THE USER GROUPS	67
3.8	THE TEST SITES.....	68
3.9	KEY ORGANIZATIONAL METHODS.....	71
3.9.1	<i>Trial Plan</i>	71
3.9.2	<i>Test plan</i>	71
3.9.3	<i>Installation and support plan</i>	72
3.9.4	<i>Global study design – time flow</i>	73
3.10	KEY USER RESEARCH METHODS.....	74
3.10.1	<i>System pre-tests</i>	74
3.10.2	<i>Short-term scenario-based user trials (SSUT)</i>	74
3.10.3	<i>Group discussions</i>	79
3.11	KEY METHODS USED FOR DATA ANALYSIS.....	80
4	AN ASSISTIVE ROBOT TO SUPPORT VULNERABLE OLDER USERS.....	85
4.1	BASIC CONCEPT AND IDEA.....	85
4.1.1	<i>User needs, use cases and scenarios</i>	85
4.1.2	<i>The prototype system</i>	86
4.2	IMPLEMENTATION OF THE EVALUATION FRAMEWORK FOR THE FIRST PROTOTYPE (E1).....	88
4.2.1	<i>Evaluation goals</i>	88
4.2.2	<i>Evaluation model</i>	88
4.2.3	<i>Evaluation methodology</i>	89
4.2.4	<i>User group</i>	93
4.2.5	<i>Test setting</i>	93
4.2.6	<i>Evaluation procedure and test flow</i>	93
4.2.7	<i>Analysis of the gathered data</i>	95
4.3	EVALUATION RESULTS OF THE FIRST PROTOTYPE (E1).....	95
4.3.1	<i>Summary of performance results</i>	95
4.3.2	<i>Summary of acceptance results</i>	98
4.3.3	<i>Applicability of the solution and acceptance from the view of care experts (secondary users)</i>	101
4.3.4	<i>Summary of results regarding prospective impacts</i>	103
4.3.5	<i>Recommendations for further development</i>	104
4.3.6	<i>Contribution to HRI design based on evaluation results</i>	105
4.3.7	<i>Lessons learned regarding the methodological approach</i>	107
4.4	IMPLEMENTATION OF THE EVALUATION FRAMEWORK FOR THE SECOND PROTOTYPE	110
4.4.1	<i>Evaluation goals</i>	110
4.4.2	<i>Evaluation model</i>	111
4.4.3	<i>Evaluation methodology</i>	112
4.4.4	<i>User group</i>	118

4.4.5	<i>Test setting</i>	118
4.4.6	<i>Evaluation procedure and test flow</i>	118
4.5	SUMMARY OF EVALUATION RESULTS OF THE SECOND PROTOTYPE (E2)	120
4.5.1	<i>Performance results</i>	120
4.5.2	<i>Acceptance results</i>	126
4.5.3	<i>Impacts and added value of the presented solution</i>	131
4.6	SUMMARY AND DISCUSSION OF EVALUATION RESULTS	133
4.6.1	<i>Performance, acceptance and added values</i>	133
4.6.2	<i>Lessons learned regarding the methodological approach</i>	135
4.7	HEURISTICS FOR FURTHER DESIGN AND DEVELOPMENT (DESIGN PRINCIPLES).....	137
5	CONCEPT OF A SAR FOR PHYSIOTHERAPY AND REFINEMENT OF THE	
	METHODOLOGICAL FRAMEWORK	141
5.1	BASIC CONCEPT AND IDEA.....	141
5.1.1	<i>Background regarding physical therapy</i>	142
5.1.2	<i>Design principles</i>	144
5.1.3	<i>The prototype system</i>	144
5.2	EVALUATION MODEL.....	145
5.3	EVALUATION FACTORS AND RESEARCH QUESTIONS.....	147
5.3.1	<i>Performance</i>	147
5.3.2	<i>Acceptance</i>	148
5.3.3	<i>Impacts and added values</i>	149
5.4	EVALUATION METHODOLOGY	150
5.4.1	<i>Methodology of the laboratory validation</i>	151
5.4.2	<i>Methodology of pre-trials</i>	151
5.4.3	<i>Methodology of the main user trials</i>	153
6	RESULTS OF THE EVALUATION OF AN ASSISTIVE ROBOT TO SUPPORT THE PHYSICAL	
	THERAPY OF OLDER USERS (E3)	161
6.1	RESULTS FROM PRE-TRIALS.....	161
6.2	EVALUATION RESULTS OF THE MAIN EVALUATION PHASE	167
6.2.1	<i>Performance results, technical performance and usability</i>	167
6.2.2	<i>Acceptance results</i>	171
6.2.3	<i>Impacts and added values</i>	193
6.2.4	<i>Discussion of results with secondary users</i>	197
6.3	SUMMARY AND DISCUSSION OF EVALUATION RESULTS	200
6.3.1	<i>Summary of usability results</i>	200
6.3.2	<i>Summary of acceptance results</i>	202
6.3.3	<i>Summary of impacts and added values</i>	205

6.3.4	<i>Summary of results from discussion with secondary users and AAL experts.....</i>	206
7	SUMMARY AND DISCUSSION.....	207
7.1	SUMMARY OF EVALUATION RESULTS.....	207
7.1.1	<i>Summary of performance results.....</i>	207
7.1.2	<i>Summary of acceptance results.....</i>	213
7.1.3	<i>Summary of prospective impacts.....</i>	221
7.2	SUMMARY OF EVALUATION METHODS.....	225
7.3	LIMITATIONS OF THE PRESENTED RESEARCH	232
7.4	SUMMARY OF ETHICAL, SOCIAL AND LEGAL ASPECTS OF ASSISTIVE ROBOTICS	237
8	CONCLUSIONS AND PROPOSED FUTURE STEPS.....	243
	REFERENCES	247
	ANNEX.....	264
A1	ATTITUDE TOWARDS TECHNOLOGY QUESTIONNAIRE – USAGE OF TECHNICAL DEVICES IN DAILY LIFE.....	265
A2	EXAMPLE OF AN INTERACTION FLOW USED WITHIN E3	267
A3	EXAMPLES OF A CUSTOM DEVELOPED QUESTIONNAIRE FOR SATISFACTION ASSESSMENT WITHIN E1.....	268
A4	EXAMPLE OF A CUSTOMIZED QUESTIONNAIRE TO ASSESS MOTIVATIONAL CAPABILITIES OF THE ROBOTIC SYSTEM	271
A5	EXAMPLES OF QUESTIONNAIRES USED WITHIN EVALUATION PHASE 3.....	273
A6	INVITATION TO THE FINAL WORKSHOP WITHIN E3.....	290
A7	EXAMPLE OF INFORMED CONSENT DOCUMENT USED WITHIN E3	293

List of Figures

FIGURE 1: OVERVIEW ON MAIN CONTRIBUTIONS AND THEIR LOCATION WITHIN THIS DISSERTATION.	14
FIGURE 2: EVALUATION PHASES E1 TO E3 AND USER-RESEARCH METHODS FOR RESULT GATHERING DURING THE COURSE OF THIS DISSERTATION.	15
FIGURE 3: STRUCTURE OF THE DISSERTATION AROUND THE MAIN CONTRIBUTIONS.	17
FIGURE 4: INVESTMENT INTO ASSISTIVE ROBOTICS RESEARCH, MODIFIED FROM [PAYR2015].	28
FIGURE 5: DISTRIBUTION OF ROBOT SUBTYPES AMONG CURRENT RESEARCH ACTIVITIES, MODIFIED FROM [PAYR2015].	29
FIGURE 6: USER-CENTRED DESIGN PROCESS, ADAPTED FROM [HCIINTERNATIONAL2011].	48
FIGURE 7: OVERVIEW ON THE DEVELOPMENT OF AN EVALUATION METHODOLOGY	59
FIGURE 8: BASIC HOLISTIC EVALUATION MODEL.	61
FIGURE 9: TAM MODEL – DAVIS 1989.	63
FIGURE 10: UTAUT MODEL.	64
FIGURE 11: ALMERE MODEL BY HEERINK ET AL. [HEERINK2010].	64
FIGURE 12: EXAMPLE OF A SETUP WITHIN THE “LIVING LAB SCHWECHAT”, PREVIOUSLY PUBLISHED BY THE SAME AUTHOR IN [TORTA2014].	69
FIGURE 13: IMPRESSIONS FROM THE TEST ENVIRONMENT.	69
FIGURE 14: USER INTERFACE AT THE TECHNICAL CONTROL ROOM SHOWING TWO CAMERA VIEWS (LEFT SIDE) AND TECHNICAL DATA ON THE LOCATION OF ROBOT AND USER AS WELL AS TECHNICAL OUTPUT DATA (RIGHT SIDE).	70
FIGURE 15: TYPICAL EVALUATION TIME FLOW.	73
FIGURE 16: PRINCIPLE METHOD FOR RESULT GENERATION.	75
FIGURE 17: OVERVIEW ON THE USED METHODS FOR DATA GATHERING, ANALYSIS AND FUSION.	83
FIGURE 18: OVERVIEW OF THE FIRST (E1) AND SECOND (E2) PROTOTYPE, ADAPTED FROM [WERNER2013] ...	87
FIGURE 19: EVALUATION DOMAINS AND METHODS USED FOR THE EVALUATION OF THE FIRST PROTOTYPE.	89
FIGURE 20: ACCEPTANCE FACTORS REGARDING THE PERCEPTION OF THE SAR, ADAPTED FROM [WERNER2012], (A) ANTHROPOMORPHISM, (B) ANIMACY, (C) LIKEABILITY, (D) PERCEIVED INTELLIGENCE, (E) PERCEIVED SAFETY.	99
FIGURE 21: EXAMPLE OF AN "INTERACTION BLOCK". THE COLOURED BOXES ARE PLACEHOLDERS FOR INTERACTION CAPABILITIES; IN OUR CASE TOP-DOWN: VOICE, GESTURES AND MIMICS.	106
FIGURE 22: EXAMPLE FRAGMENT OF AN INTERACTION FLOW.	107
FIGURE 23: EVALUATION DOMAINS AND METHODS USED FOR THE EVALUATION OF THE SECOND PROTOTYPE. .	112
FIGURE 24: RELATIONS OF CONSTRUCTS WITHIN THE ALMERE MODEL [HEERINK2010]. GREYED-OUT CONSTRUCTS WERE NOT USED DURING THE TRIALS. FIGURE ORIGINALLY PUBLISHED BY THE SAME AUTHOR IN [TORTA2014].	114
FIGURE 25: COMPARISON OF PERFORMANCE RESULTS BETWEEN E1 (LIGHT GREY) AND E2 (DARK GREY),	

ADAPTED FROM [WERNER2013], THE CHART PRESENTS EXACT MEASUREMENTS, HENCE NO ERROR BARS ARE GIVEN.	121
FIGURE 26: QUANTITATIVE RESULTS TO THE CUSTOM-DEVELOPED QUESTIONNAIRE REGARDING THE PERCEIVED EASE OF USE. ERROR BARS INDICATE THE STANDARD DEVIATION, DATA FROM ALL ITERATIONS OF E2 EXCLUDING PARTICIPANTS THAT PERCEIVED TECHNICAL MALFUNCTIONS (SCORE FROM 1 TO 5, 1 = "NOT AT ALL, 5 = "TOTALLY").	125
FIGURE 27: PERCEPTION OF THE ROBOT FOR E1 (LEFT) AND E2 FIRST ITERATION (MIDDLE) AND E2 SECOND ITERATION (WIZARD OF OZ, RIGHT), PREVIOUSLY PUBLISHED IN [KSERA2012A], ALL DATA, INCLUDING TECHNICAL ISSUES.	126
FIGURE 28: MEAN VALUES OF THE CONSTRUCTS WITHIN INTENTION TO USE DURING THE FIRST AND SECOND ITERATION OF E2. THE CHART WAS UPDATED FROM A PUBLISHED VERSION BY THE SAME AUTHOR IN [TORTA2014].....	128
FIGURE 29: EMOTIONAL FEELING RIGHT AFTER THE TEST ON A SCALE FROM 1 = NOT AT ALL TO 7 = VERY MUCH, ERROR BARS INDICATE THE STANDARD DEVIATION.	130
FIGURE 30: OVERVIEW OF THE 3RD PROTOTYPE, ADAPTED FROM [WERNER2013B].	145
FIGURE 31: EVALUATION MODEL FOR THE EVALUATION OF A SAR SYSTEM FOR PHYSIOTHERAPY SUPPORT WITHIN E3.	146
FIGURE 32: EVALUATION MATRIX DETAILING THE USED EVALUATION METHODS AND HOW THEY INTERLINK WITH EVALUATION DOMAINS.	147
FIGURE 33: FULL ALMERE MODEL WITH ITS CONSTRUCTS AND THEIR LINKS TO PREDICT THE USE OF A SYSTEM.	148
FIGURE 34: IMPRESSION OF THE GYMNASIUM USED FOR TESTING.	152
FIGURE 35: TOP VIEW OF THE TEST ENVIRONMENT AT THE SCHWECHAT SENIOR-CITIZEN CENTRE. SAR SYSTEM COMPONENTS AND TEST COMPONENTS ARE SHOWN.	154
FIGURE 36: USER INTERFACE FOR THE EXPERIMENTERS AS USED DURING E3.	155
FIGURE 37: SCREENSHOT OF TRAINING VIDEO USED FOR COMPARISON (SOURCE: YOUTUBE.COM).	157
FIGURE 38: SCREENSHOT OF THE USED TRAINING PROGRAMME (EA SPORTS ACTIVE ON THE NINTENDO WII), SOURCE: HTTP://WWW.TECHNOLOGYTELL.COM.	158
FIGURE 39: RESULTS OF MOTIVATION AND TRAINING SUPPORT (N = 14).	161
FIGURE 40: RESULTS OF USEFULNESS (Q5-Q9) OF THE TRAINING SYSTEM (N = 14).	162
FIGURE 41: RESULTS OF GENERAL IMPRESSION (Q11-Q13) OF THE TRAINING SYSTEM (N = 14).	163
FIGURE 42: RESULTS OF THE CHARACTERISTICS OF THE ROBOT TRAINER (Q14-Q21), (N = 11).	165
FIGURE 43: COMPARATIVE ANALYSIS OF DIFFERENT WAYS OF TRAINING SUPPORT (Q22, Q24, Q25), (N 12).	166
FIGURE 44: USABILITY RESULTS REGARDING TRAINING EXPLANATION AND FEEDBACK.	168
FIGURE 45: GODSPEED CONSTRUCTS FOR E3: (A) ANTHROPOMORPHISM, (B) ANIMACY, (C) LIKABILITY, (D) PERCEIVED INTELLIGENCE, (E) PERCEIVED SAFETY.	172

FIGURE 46: GODSPEED COMPARISON WITH THE FIRST ITERATION OF E2 (LEFT) AND E3 (RIGHT). (A) ANTHROPOMORPHISM, (B) ANIMACY, (C) LIKABILITY, (D) PERCEIVED INTELLIGENCE, (E) PERCEIVED SAFETY.	176
FIGURE 47: GODSPEED CONSTRUCT MEANS FOR COMPARISON WITH OTHER ROBOTIC SYSTEMS (FROM LEFT TO RIGHT: E3, E2, JAMES ROBOT, MAGABOT)- (A) ANTHROPOMORPHISM, (B) ANIMACY, (C) LIKABILITY, (D) PERCEIVED INTELLIGENCE, (E) PERCEIVED SAFETY – ERROR BARS INDICATE ONE STANDARD DEVIATION.	178
FIGURE 48: NAO ROBOT USED FOR E1 TO E3 (LEFT), JAMES ROBOT (MIDDLE) [FOSTER2012] AND MAGABOT (RIGHT) FOR COMPARISON.	178
FIGURE 49: MEAN SCORES OF THE ALMERE MODEL CONSTRUCTS, ERROR INDICATORS SHOW THE STANDARD ERROR OF THE MEAN.	180
FIGURE 50: COMPARISON OF MEAN SCORES OF SELECTED CONSTRUCTS WITH SIMILAR ASSISTIVE ROBOTIC PROJECTS, ERROR BARS INDICATE THE STANDARD ERROR OF THE MEAN. E3: DARK GREY, E2: GREY, MIKA TRIALS: LIGHT GREY.	187
FIGURE 51: MIKA ROBOT, IMAGE TAKEN FROM [XU2012].	188
FIGURE 52: RESULTS OF USEFULNESS (Q5-Q9) OF THE TRAINING SYSTEM (N = 12).	188
FIGURE 53: SCORE FOR FULFILMENT OF USER EXPECTATIONS GIVEN IN ABSOLUTE NUMBERS.	189
FIGURE 54: RESULTS OF THE POST-TEST FEELING QUESTIONNAIRE (N = 12).	191
FIGURE 55: RESULTS OF MOTIVATION AND TRAINING SUPPORT (N = 12).	192
FIGURE 56: COMPARATIVE ANALYSIS OF QUANTITATIVE RESULTS OF THE DIFFERENT TRAINING SUPPORT OPTIONS.	194
FIGURE 57: PLANNED USAGE OF THE COMPARISON SYSTEMS.	196

List of Tables

TABLE 1: EXAMPLES OF THE HISTORY OF ROBOTICS.	5
TABLE 2: EXAMPLES OF ASSISTIVE ROBOTS ADAPTED FROM [PAYR2015].	19
TABLE 3. TECHNOLOGY-READINESS LEVELS AS PROPOSED BY NASA [NASA2015].	32
TABLE 4. PART OF EVIDENCE TABLE FOR LABORATORY TRIALS OF THE INTEGRATED PROTOTYPE.	34
TABLE 5. EVIDENCE TABLE FOR USER TRIALS OF THE INTEGRATED PROTOTYPE (PART 1).	37
TABLE 6. EVIDENCE TABLE FOR USER TRIALS OF THE INTEGRATED PROTOTYPE (PART 2).	38
TABLE 7. EVIDENCE TABLE FOR FIELD TRIALS IN REAL ENVIRONMENTS.	40
TABLE 8. EVIDENCE TABLE FOR FIELD TRIALS IN REAL ENVIRONMENTS (PART 2).	41
TABLE 9: CONDENSED RESULTS REGARDING THE USED SCENARIOS FROM WORKSHOPS WITH PRIMARY AND SECONDARY USERS.	132
TABLE 10: OVERVIEW OF EVALUATION FACTORS USED DURING THE EVALUATION OF THE INTENTION TO USE THE SYSTEM, BASED ON [HEERINK2010].	148
TABLE 11: METHODS USED AND EVALUATION DOMAINS STUDIED WITHIN THE THREE PHASES OF E3.	151

TABLE 12: DESCRIPTIVE STATISTICS OF GODSPEED CONSTRUCTS.	171
---	-----

TABLE 13: DESCRIPTIVE STATISTICS OF THE ALMERE MODEL.	179
--	-----

Acronyms and Abbreviations

AAL – Active and Assisted Living

ANOVA – ANalysis Of Variances

ANX, ATT, FC, ITU, PAD, PENJ, PEOU, PS, SI, SP, TTS, TRUST – Acceptance factors of the Heerink-Model

CEIT – Central European Institute of Technology

COPD – Chronic Obstructive Pulmonary Disease

E1, E2, E3 – Evaluation phases one to three

HCI – Human-Computer Interaction

HRI – Human-Robot Interaction

KSERA – European Union funded research project “Knowledgeable Service Robots for Aging”

LL – Living Lab

MANOVA – Multivariate ANalysis Of Variations

MSPSS – Multidimensional Scale of Perceived Social Support

NASA – National Aeronautics and Space Administration

PANAS – Positive and Negative Affect Schedule

PhysicAAL – Austrian nationally funded research project to evaluate the applicability of socially assistive robotics for training of older people at home

PotenziAAL – Austrian nationally funded study that undertook an analysis of robotics in the field of Active and Assisted Living

PT1, PT2, PT3 – Prototype one to prototype three

RQ – Research question

ReMIND – European Union funded research project “Robotic Epartner for Multitarget INnovative activation of people with Dementia”

SAR – Socially Assistive Robot

SSUT – Short-term Scenario-based User Trials

SUMI – Standard Usability Measurement Inventory

TAM – Technology Acceptance Model

TRL – Technology Readiness Level

UCD – User-Centred Design

UTAUT – Unified Theory of Acceptance And Use of Technology

1 Introduction

Within the last few years, research efforts in the field of assistive robotics have strongly increased both within the Active and Assisted Living (AAL) community but also in the Human-Robot Interaction (HRI) and robotics research communities, as is illustrated by the number of prototypes developed, for example, in the “Domeo”,¹ “Florence”,² “KSERA”,³ “Companionable”,⁴ and “ALIAS” projects.⁵ These and other projects have shaped the AAL robotics domain and over 15 research projects on the European level alone are currently running to target the development of robotics to support older users at home or in care facilities.

Although various types of robots are imaginable to support older users performing activities of daily living at home, currently the AAL robotics community predominantly researches one particular type, the multi-purpose companion robot (see also section 2.1) as shown by the large number of scientific projects undertaken with these particular systems. They are popular in research as they have the potential to target a wide range of individual user needs and provide an anthropomorphic multi-modal user interface that could be of interest for non-tech-savvy seniors. To predict the long-term impacts of SARs remains one of the most prominent challenges in the area.

A variety of user research methods can be applied to evaluate the developed prototypes of companion robots from multiple perspectives such as the technical performance in a real-life setting, the usability of prototypes, the acceptance among users and the impact on care and the users’ lives. Although the evaluation aims between existing research projects are to some extent similar, the evaluation methods implemented vary strongly both in quality and quantity. Evaluation methods currently in use are mostly derived from the field of human-computer interaction, a discipline that aims to understand how human users experience computer-based applications and tools [Bødker2015]. However, the interaction between humans and physically embodied and socially

¹ <http://www.aal-domeo.org>

² <http://www.florence-project.eu>

³ <http://ksera.ieis.tue.nl>

⁴ <http://www.companionable.net>

⁵ <http://www.aal-europe.eu/projects/alias/>

situated machines is inherently different and needs specific consideration [Dautenhahn2007a]. Several researchers found that the methodology used to evaluate assistive robots in current research suffers limitations, is rather vague and often not replicable, which limits the outcome of undertaken studies [Papadopoulos2019, Bemelmans2012]. This dissertation challenges this limitation by offering a toolkit of methods and methodologies that integrate a human-centred view into the applied area of SARs in AAL.

This work takes a “human-centred view” on HRI [Dautenhahn2007a] and develops new methods and methodologies, in particular for the evaluation of HRI. The developed methods are applied in research studies with three prototypes and central results both from the methodological development and the studies are presented.

1.1 Motivation and background

1.1.1 Motivation No. 1: Demographic, economic and societal challenges

A demographic change is currently under way in most developed countries and particularly in Europe, including Austria. The ratio and absolute number of older people is currently rising and often predicted to rise even more sharply over the coming decades as the baby-boomer generation hits retirement age, for example by WHO [WHO2015a]. In particular, the number of old and very old people is going to increase, whereas the number of children is declining because Europe’s fertility rates are far below sustainability [Eurostat2019]. As a result of these demographic effects, a study found that the ratio of German people at working age will by 2030 already drop from currently 61% to 54% of the total population [HWWI2015], reducing the dependency ratio between carers and care-takers and posing an increasing risk for the country’s economy. Another detailed investigation undertaken in three federal German states concludes that employment bottlenecks for nurses and social workers are already evident and will be exacerbated in the future because of the named demographic effects and predicts challenging work conditions for social workers [Fuchs2013].

In addition to these demographic trends, chronic diseases such as high blood pressure, diabetes and chronic obstructive pulmonary disease (COPD) are also on the rise as well as the number of multimorbidities among older people, aggravating care needs and costs [Potenziaal2015].

From an Austrian viewpoint, it was found that allowances for nursing care have already more than doubled (even without considering inflation) within the 20 years between 1997 and 2017 showing that the effects are already taking place and solutions are needed sooner rather than later [BMASGK2018].

Another issue aggravating challenges from demographic changes are the effects of societal changes. Lower family cohesion, lower rates of marriage, higher rates of divorce [Eurostat2017] and a population agglomeration in big cities lead to higher numbers of older people living alone, outside of a stable local network of relatives and friends who could provide needed informal care [Potenziaal2015]. This is of particular importance as informal care in Europe is of high economic value, making up a significant part of the total care provided [VandenBerg2004].

In this landscape, the factual motivation of this dissertation is to contribute to the care of older members of our society via technologies that are sustainably developed to support the amount of care needed for the predicted numbers of people in formal and informal care.

Active and Assisted Living as a technical contribution

These aforementioned demographic challenges are obvious and have been well known for many years already and it seems clear that solutions have to be searched for in several domains including politics, social sciences and health sciences, including gerontology. However, given technical advances, in particular regarding embedded computing, wearable, pervasive and ubiquitous systems and the wide availability of Internet access, the chances of a contribution to tackle demographic issues by technological means have risen.

Because of this opportunity, about 15 years ago assistive technologies started to receive interest from research and politics and a new research field was established, the so called “Ambient Assisted Living” (now “Active Assisted Living”), short AAL, focusing on the support of older users. Conferences such as the “German AAL Kongress” were initiated in 2008,⁶ and international conferences such as the AAATE,⁷ and the ICCHP started to include AAL tracks.⁸ A European initiative called the “AAL Joint Programme”

⁶ <http://www.aal-kongress.de>

⁷ <http://www.aaate.net>

⁸ <http://www.icchp.org>

was established in 2008 for funding international research activities (now called “Active and Assisted Living programme”).⁹

The main goal of AAL is to support the daily life of users in need, in particular older users, by means of technical methods, concepts, systems or services in order to prolong the timespan of independent living and enhance the quality of life [Georgieff2008].

Because AAL technologies typically interact closely with older and vulnerable user groups, and to conquer the risk of technically driven developments running aside of users’ true needs, AAL research promotes participatory design, development and evaluation strategies as a basic concept and involves relevant user groups right from the beginning of research projects. This is because a paradigm AAL technology should assist users and be able to adapt to the specific needs of targeted user groups instead of forcing users to learn new technologies and user-interfaces [AALAZ].

This dissertation is placed in the area of AAL, an applied domain of technology development. The second motivation of this dissertation is to contribute to sustainable development of AAL technologies by questioning the existing methods and methodologies in their fit to the applied development domain of SARs.

1.1.2 Motivation No. 2: Robotics as an age-old dream

To build an artificial counterpart of a human being that acts as a servant is an age-old dream of mankind. The earliest known manifestation of this idea is the “Golem” – a mystical hefty creature created magically out of clay to physically serve its master. This fantasy, which dates back about 2,000 years, was documented already in the Jewish Talmud and is still a common theme picked up in science fiction.

In the early 19th century, the very same idea was fuelled again by new technological possibilities. Mechanical automates were designed and also built that closely resemble humans and seemingly come to life [Hoffmann1819].

Duffy [Duffy2003] calls the idea of building a fully anthropomorphic synthetic human “the ultimate quest of many roboticists” and seems to have a point as in fact, when the “iCub” project started in 2004,¹⁰ clearly one of the key motivations was to build an artificial machine that resembles a human for its own sake. One of the core developers even stated the idea was created out of an inner urge to create a human counterpart





⁹ <http://www.aal-europe.eu>

¹⁰ <http://www.icub.org>

[Schanze2010]. Because it was already well known that the system would not comply with the high functional expectations triggered by the high-tech humanoid design, the form of a small child was chosen.

Table 1 gives examples of early and recent manifestations of the idea to build artificial humans.

Table 1: Examples of the history of robotics.

			
<p>Prague Golem reproduction¹¹</p>	<p>Jaquet-Droz automata¹²</p>	<p>iCub robot platform¹³</p>	<p>“Pepper” robot¹⁴</p>

Humanoid robots are perfectly suited to realizing the dream of an artificial human-like servant. Operto et al. argues that the motivation to develop robots that resemble humans certainly does not spring only from rational, engineering or utilitarian reasons but also from psycho-anthropological motivations [Operto2008].

Because of this psychological effect leading to an intrinsic motivation, substantial efforts are already being undertaken today to realise anthropomorphic robots that could support us in various ways, just like humans can. Large restrictions in terms of affordability and technical feasibility still have to be solved but the technological advance is proceeding quickly and given current dedication to the topic of robotics worldwide, it seems just a matter of time until we can cope with them. Indeed, a product recently came onto the market that could be a start to turning this vision into reality. The robot “Pepper” by the *Softbank* cooperation is the newest advancement of the

¹¹ <https://en.wikipedia.org/wiki/Golem#/media/File:Prague-golem-reproduction.jpg>

¹² http://www.forensicgenealogy.info/contest_412_results.html

¹³ <http://juxi.net/projects/iCub/>

¹⁴ <https://www.softbankrobotics.com/emea/en/pepper>

developments of the French robotic company Aldebaran and because of its high-volume of production units, comes already at a price of around €8,000 – an affordable price for private customers [Payr2015].¹⁵

It seems almost certain that one day, likely in the not-too-distant future, we will deal with anthropomorphic robots that operate in our homes and we will take substantial amounts of our time to interact with them. Because of their anthropomorphism these systems will suit well to tasks that ask for human-like abilities and given current research directions, one of these tasks will be to help care for older people.

Since SARs will presumably be part of our daily living, we have to ensure that already now, during the phases of early research and development, we consider all implications – social, ethical, legal or technical – that could arise from an introduction of such systems into our lives. **As a contribution, one major goal of this dissertation is to find out if and how such systems are able to support older users and provide experiences of possible implications of their use in real life.**

1.1.3 Motivation No. 3: Promising and challenging intersection of AAL and robotics

Robotics seems like a good fit for active and assisted living because of its inherent unique features. Aside of being able to solve physical tasks such as lifting and carrying patients, anthropomorphic robots in particular could be the one-size-fits-all solution that assists in the multiplicity of daily care tasks.

In addition to a technical tool that supports older users during daily life, robots can also be used as multi-modal, easy-to-use interfaces for a target group that is typically not well trained in using computers to help the introduction of technical services. The social presence inherent to physical objects was also shown to have positive influences on the adoption of the technology by the target groups. (See also chapter 2.1 for further details about the potential benefits of assistive robotics.)

From another perspective, AAL contributes to robotics as it gives additional meaning to the undertaken research and provides interesting applied-research questions that help to further develop robotics to finally evolve with solutions that are capable of performing in complex real-life environments in close proximity to its (vulnerable) users.

¹⁵ <http://www.softbank.jp/en/>

Of course, when combining two research fields, not only are the benefits combined but also the difficulties.

Well known issues include the lack of technical reliability of existing prototypes that currently hinders an effortless integration into real-life scenarios, in particular of complex human-like robotic solutions, as well as limitations regarding the marketability of complex and rather expensive solutions, especially within the as yet barely developed AAL market.

Additional to these, the use of technical solutions and, in particular, of robots for the vulnerable user-group of older people and people with complex disabilities such as dementia rightly raises many concerns from legal, ethical and social viewpoints, such as personal safety, the risk of social isolation due to (social) contacts with machines instead of humans, potential privacy violations due to the autonomy of robots, potential issues arising from social robots acting as companions instead of tools and others, as described elsewhere by the same author [Payr2015].

Until now, it remains mostly unknown how the introduced social and technical challenges can be seamlessly integrated. These briefly outlined issues concerning the combination of AAL and robotics make the research within this field very interesting and motivates researchers from different disciplines such as social sciences, ethics, psychology, medicine and business to join the discussion on the feasibility of robotic solutions for the care of older people.

The last motivation of this dissertation is to introduce an integrated set of design recommendations for applied researchers of SARs in the AAL domain. The implications aim to integrate technical and social challenges, and introduce case studies on how they were negotiated throughout three different projects.

1.2 Problem statement

Socially Assistive Robots (SAR) were first defined in 2005 by Feil-Seifer et al. [Feil-Seifer2005] and are hence still a very young research field with many open issues and basic questions to be answered. Two main problem domains can be identified: a) it is at present mostly unclear to what extent SAR solutions would be accepted by the target groups and what impacts can be expected in real life and b) how can we measure the various performance factors (both technical and social) given that our research methods have mostly not been developed for the evaluation of a technology that is not only

technologically challenging but also has obvious psychological, emotional and social influences?

a) Unclear benefits of SARs due to lack of real-life studies

Few user studies in real-life settings with a SAR have so far been undertaken due to the many difficulties involved. The necessary inclusion of vulnerable user groups makes trials of such technologies time consuming, and the technical reliability of current prototypes often does not allow for field-testing without risking the safety of users. Hence, it is mostly lab-based experiments on single aspects of the design or performance of SARs that have been undertaken and published. These bottom-up experiments give valuable insights into many detailed aspects. However, it is unclear whether all these small insights will add up without interfering with each other and if they will provide a clear big picture which will give us the chance to develop SARs that are able to generate the impacts on our care systems that we hope for.

In addition to such bottom-up experiments, it seems necessary to undertake larger top-down studies with end-users in real-life settings that can provide us with a better holistic view on three problem domains:

1. **Performance**: To what extent is current cutting-edge technology able to provide what users want?
2. **Acceptance**: To what extent would the involved user groups accept a SAR system and its integration into their daily life and work? Are current SAR approaches usable by the target groups? How should we design future robotic solutions in order to increase acceptance and uptake? Which factors and parameters influence acceptance?
3. **Impacts**: What impacts can be expected from SAR systems? Can they improve the efficiency and quality of care? Can they contribute to a higher quality of life of users? Given that SAR systems are mostly complex and costly, could they achieve impacts justifying the costs, hence could solutions be cost-efficient?

As the three problem domains are linked with each other – performance influences acceptance, which influences impacts and vice versa – it is necessary to analyse these domains holistically.

b) Unclear methods to evaluate SARs in real-life settings

Methods to evaluate SARs are scarce and mostly not replicable due to their qualitative nature. Already in 2007, Dautenhahn [Dautenhahn2007a] criticized that methods, tools and methodologies that can advance our understanding of HRI and allow replication of results have yet to be found in future research. Not much has changed since then. Bemelmans et al. [Bemelmans2012] found, five years later again, that currently used research methods are rather vague and not replicable, which limits the value of the presented evidence. A new scientific culture needs to be established that is able to confirm or refute findings.

Currently used methods are derived from other fields such as sociology and psychology and although they have been used for many years in HCI contexts, knowledge on how to use and adapt them for the field of assistive robotics is limited. Typically applied research methods such as heuristics, observations and focus groups were not developed with the intention to design or assess the performance of SARs. They need to be adapted to fit the specific characteristics and challenges of robots such as their physical autonomy, the capability for human-like communication and the social presence inherent to physical objects that influence the effects of these systems. To give an example, Weiss et al. found that the applicability of HCI methods for user-experience evaluation in HRI is unclear because user experience “might be heavily influenced by the individual’s general attitude and the overall societal opinion” on robotics [Weiss2009].

Several authors mention low reproducibility as an issue. This is, on the one hand, ascribable to the qualitative nature of the undertaken research itself which does not acknowledge reproducibility as a necessity; on the other hand, it would certainly be beneficial to be able to validate findings and the oft-used mixed-research models (composed of qualitative and quantitative methods that validate each other) would also allow this to be done if there weren’t other issues. One main issue is again the lack of methodologies that are accepted as quasi-standard by the community. For that reason, most researchers are currently forced to develop own methods such as questionnaires that fit their particular research questions or adapt existing methods to the particular field of robotics. The process of describing the used methods in rigorous detail to let others reproduce them would go beyond the scope of result papers, which is why this is often not undertaken. To further aggravate the issue, different research groups use different robotic platforms, thereby limiting the reproducibility as the influence of the

platforms' particularities on the user experience and impacts remains unknown and cannot be differentiated from other influences.

The inclusion of a vulnerable user group (seniors with disabilities or health restrictions that make them potential profiteers of assistive technologies) puts restrictions on fitting research methods, as this user group can be harmed by practices and procedures of research such as being involved into studies for long durations. The participation of this user group during the course of a research project as intended by the participatory design approach is limited due to the likelihood of dropouts because of the worsening health conditions of participants. Furthermore, this group is very heterogeneous with strong inter-individual differences; most studies presented so far have failed to frame the target group (see also Chap. 2).

Methods are needed that take care of the inevitable low technical robustness of current technical prototypes by simultaneously allowing a hands-on experience, as the functionality of and acceptance towards SARs can only be grasped during direct interaction, not by mere observation, due to the psychological effects of perceived sociability.

1.3 Goals and research questions

Ad a) Unclear benefits of SARs due to lack of real-life studies

The main aim of this dissertation was to gain insights into the technical performance, user acceptance and potential impacts of socially assistive robotics by means of an iterative evaluation of recent prototypes together with relevant target groups. To what extent the developed prototype solutions comply with real-users' needs, their values and requirements was part of this goal.

Within the course of two research projects, it was hoped that the development of proofs of concept would allow for insights into the technological applicability under real-life conditions. It was planned to assess the usability and user experience as they are crucial for acceptance; likewise potential impacts on health factors, factors that influence the quality of life of users and factors influencing care efficiency and care quality would be measured.

Research topics aligned with these goals were:

RQ1: To what extent are current SAR systems applicable under real-life conditions from a technological perspective?

- a. To what extent is current SAR technology able to satisfy relevant user needs?
- b. Which flaws and challenges need to be solved on a technological base in order to allow an acceptable human-robot interaction?

RQ2: To what extent do both older users and their carers accept socially assistive robotic solutions for the support of older users at home?

- a. How do acceptance rates compare between robotic solutions and technological but non-robotic solutions?
- b. Which behaviour of SARs is socially accepted?
- c. How can solutions be integrated into the daily life of users and the daily work of carers?

RQ3: Do SAR robots have beneficial effects for the support of older people at home and, if so, which ones?

- a. To what extent is it possible to motivate users by means of a SAR solution and how does this compare to similar technological but non-robotic solutions?
- b. Can the developed SAR systems be therapeutically effective?
- c. What impacts on the quality and efficiency of care can be expected?
- d. Can solutions be cost-effective?

Ad b) Unclear methods to evaluate SARs in real-life settings

As the methodology on how to gain insights into the main questions of assistive robotics suffers limitations, a second goal of this dissertation was the development and adaption of research methods to support the evaluation of SARs. In particular, methods for the evaluation of how such devices can be evaluated in real-life contexts together with the target groups of older users, their carers, therapists and relatives were still missing and needed to be developed to gain insights on the effects of current solutions. But not only were the methods missing, but crucially a framework composed of the various existing methods that facilitates a holistic evaluation, which would seem necessary to understand SARs due to the strong interlinks between evaluation factors.

Research topics aligned with these goals were:

RQ_M1: Which current research methods can be used to assess a SAR?

RQ_M2: Which methods can be used and how can they be adapted to allow an evaluation of a SAR in settings as close to real life as possible?

RQ_M3: Which methods can be used and how can they be used to safely involve vulnerable older users (patients) and let them experience interaction with a SAR?

RQ_M4: How can existing methods be synthesized together to form a reusable evaluation framework that facilitates a holistic evaluation?

1.4 Contribution

Ad a) Unclear benefits of SARs due to lack of real-life studies

Insights into the currently achievable technical performances, typical technical issues of prototypes, typical issues regarding the feasibility of integration into users' homes and their lives as well as into care processes are given and discussed in chapters 4.3.1 (E1), 4.6.1 (E2), 6.2.1 (E3) and summed up in section 6.3.1. As a proof of concept of the applicability of current SAR solutions, the development of three prototypes was supported; the prototypes could be successfully implemented into trial setups and usage scenarios.

The main focus of the conducted user trials was laid on investigating aspects of acceptance, including usability and user experience. User-acceptance factors such as perceived usefulness, perceived ease of use, perceived enjoyment of interaction, trust in and anxiety towards the systems, social factors such as the social presence and perceived sociability of developed solutions, and external social influences were investigated in detail and are presented in sections 4.3.2 (E1), 4.6.1 (E2), 6.3.2 (E3) and were fused, summed up and discussed in section 7.1.2.

Impacts were evaluated using qualitative methods to gain insights into potential benefits of SARs in future use scenarios, including potential impacts on the quality and efficiency of care such as the therapeutic effectiveness of solutions, impacts on the quality of life of users, and impacts on care costs were estimated in the sections 4.3.4 (E1), 4.6.1 (E2), 6.2.3 (E3) and again were summed up and discussed in section 7.1.3.

Ethical, social and legal aspects of the implementation of SARs into real settings were collected within the undertaken studies and are discussed in section 7.4.

Ad b) Unclear methods to evaluate SAR in real-life settings

User-centred evaluation methods were reviewed and analysed, and conclusions on current open issues and how to avoid them are given in chapter 2.2 “User-centred evaluation methods”.

Within chapter 3, as a contribution to current research, methods for a holistic top-down analysis of SARs are proposed to supplement the bottom-up experiments. In particular, contributions to the development of a methodology for the short-term user-centred evaluation of SARs within a living-lab environment together with vulnerable older users are given. The development of the method is explained, and potential benefits over existing methods for lab-based and real-life evaluations are presented. How current methods can be ideally combined with or integrated into the new methodological framework is discussed. Technological and user-based prerequisites for using the method are given.

Current methodologies for the evaluation of technology were assessed and insights into the practical applicability of these methodologies are given as well as lessons learned from the execution of user trials within sections 4.3.7 (from evaluation phase 1), 4.6.2 (from evaluation phase 2) and 7.2 (fused from all evaluation phases).

A design method called “interaction flows” was developed to support the design of human-robot interaction specifically for use with socially assistive robotics that are capable of multi-modal interaction interfaces. This method is presented in section 4.3.6.

A set of design principles was developed based on the result of presented user trials that can be used in future research to guide the design of socially assistive robotics. The design principles are presented in section 4.7.

The following Figure 1 provides an overview on the contributions achieved within this work.

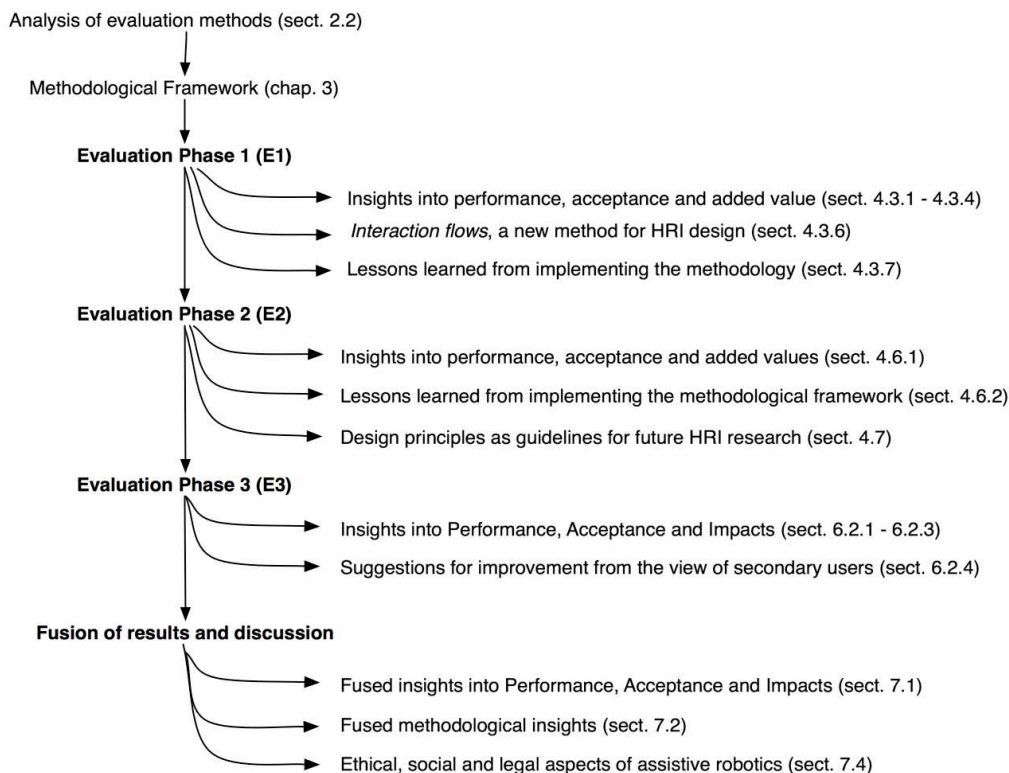


Figure 1: Overview on main contributions and their location within this dissertation.

1.5 Methodological approach

The methodology developed is based on the user-centred evaluation phase of the user-centred design process, which is often used to design, develop and evaluate assistive technology and is described within the ISO 9241 [ISO9241_2019]. The process was altered to fit the needs of assistive robotics, vulnerable target groups, a living-lab approach, and time and budget constraints of scientific research projects. Evaluation methods were derived from human-computer interaction (HCI), psychology and sociology, and adapted with recent knowledge on the specifics of SARs, in particular regarding the evaluation of the SARs capabilities and influences on humans. The methodology of this dissertation is composed of a mix of qualitative and quantitative user-research methods to gather information on the research questions with an aim to combine and compare them to gain and validate insights from different perspectives.

The methodology for the evaluation of SARs was initially developed based on own experiences from AAL research and a literature review together with experts from HCI and psychology. The methodology was implemented for the first prototype of the

European research project “KSERA” (Knowledgeable Service Robots for Aging), refined for the second prototype of “KSERA”, and later adapted towards another robotic solution within the “PhysicAAL” national scientific project. The iterative development of the methodology on the use and potential issues of the evaluation methods is described in detail in this dissertation. By means of these iterative steps, experiences on the use and potential issues of the evaluation methods could be gained and are described in this dissertation, providing a guideline on how to use the methodology in similar settings.

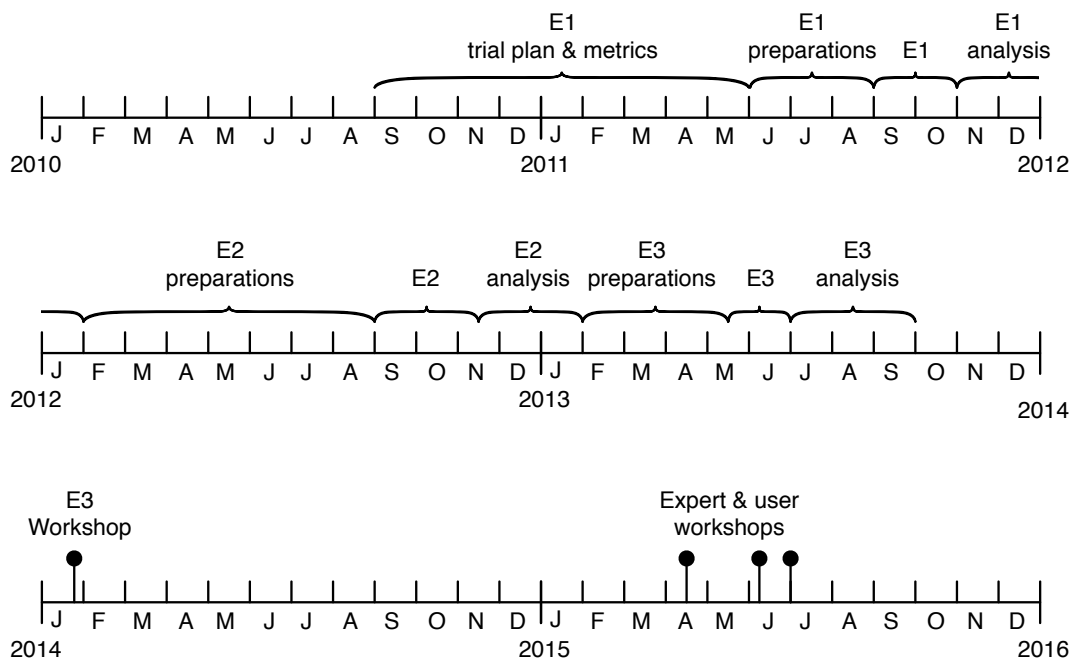


Figure 2: Evaluation phases E1 to E3 and user-research methods for result gathering during the course of this dissertation.

The presented figure 2 details the time flow during and between the evaluation phases E1 to E3.

- *E1...3 preparations*: Development of a detailed test plan including the detailed methods
- *E1...3*: Phase including several user-research methods such as Workshops, SSUT with primary, secondary and tertiary users
- *E1...3 analysis*: Result analysis and discussion with experts or user groups
- *Expert and user workshops* were conducted to reflect on the results from a top-down view and set them into the wider scope of related research.

1.6 Structure of this dissertation

Chapter 1 gives an introduction into the field of assisted technologies and details the motivation behind using high-level technology such as assistive robots as a contribution to conquer the demographic challenges in Europe. The open research problems are stated and the overall methodology used to answer the derived research questions for this dissertation is given. Finally the main contributions to the research field are carved out.

Chapter 2 presents the state-of-the-art of the research on socially assistive robotics, including the typical current technical limitations of prototypes and the limitations of current methods for development and evaluation. In particular, the state-of-the-art of methodologies used for the evaluation of robotic prototypes regarding technical performance, user experience and impacts on care as well as quality of life aspects is presented.

Chapter 3 presents a methodological concept for the evaluation of prototypes of the respective SAR solution. The evaluation phases are detailed and the evaluation methods are described in detail.

Chapter 4 describes the implementation of the evaluation concept from chapter 3 for the evaluation of two prototypes of a SAR for COPD patients. The results of the evaluation are presented and summarized, including conclusions for further refinement of research methods for the evaluation of SARs.

Chapter 5 describes the basic concept and ideas behind another research prototype of a SAR with the goal of particularly assisting healthy older adults at home or in a care centre by conducting prescribed physical training. The evaluation methods are adapted for the partly-new research aims and described, taking into account the results presented in Chapter 4, and the implementation and validation of this methodology within a set of living-lab trials with different user groups is detailed.

Chapter 6 details the user evaluation regarding the technical performance, user experience and potential impacts on the quality of life with the described methodology from chapter 5. In conclusion, the results of the user evaluation are presented and summarized to provide insights into the value of SARs for applications within the field of AAL.

Chapter 7 sets the gathered results into the context of current research and discusses the use of SARs within AAL in general as well as the limitations of the research provided and reflects critically on the presented results. The final evaluation framework is presented as a guide for future SAR evaluations.

Chapter 8 concludes the dissertation and suggests future steps to evolve the research field.

The following Figure 3 provides a visual overview of the structure of this dissertation.

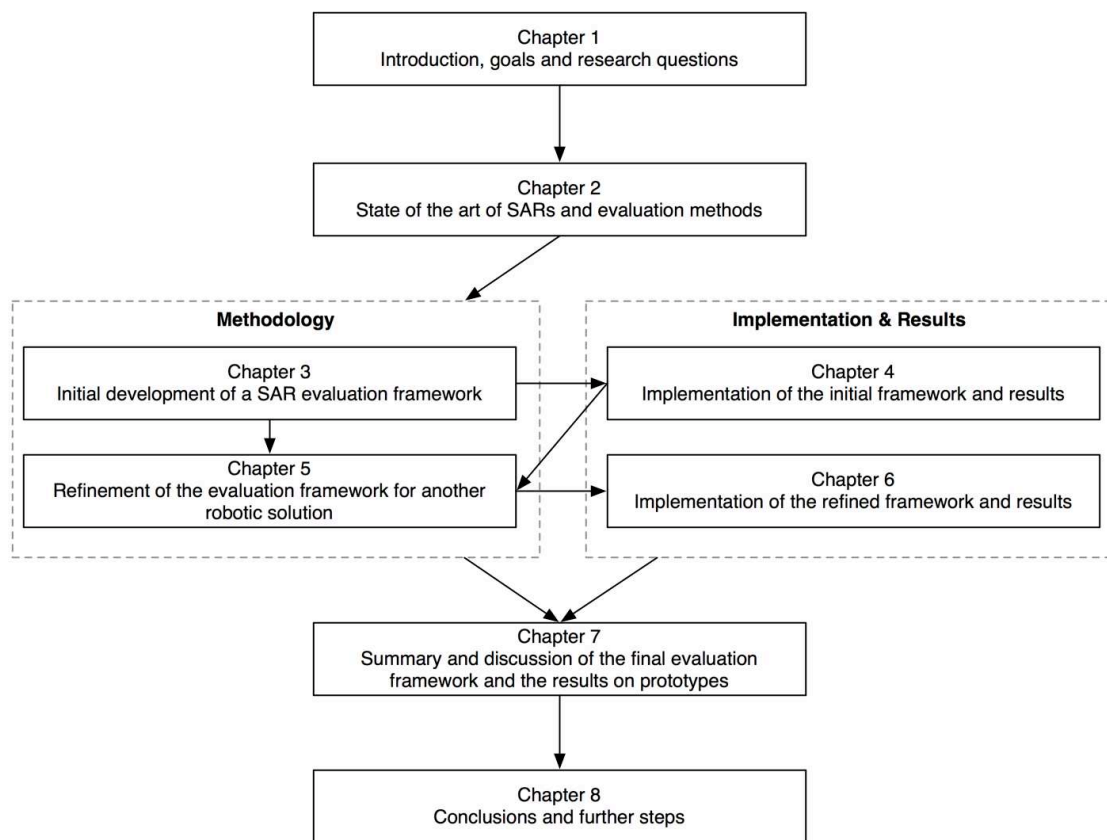


Figure 3: Structure of the dissertation around the main contributions.

this page intentionally left blank

2 State of the art







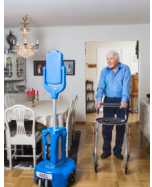



2.1 (Socially) Assistive Robotics

Assistive Robotics is a field of research that covers many different robotic solutions that share the common goal to support users with disabilities [Feil-Seifer2005]. Solutions typically help users to perform tasks by means of three main characteristics of assistive robots and their derivations:

1. The capability to provide physical support
2. The social presence inherent to physical objects and robots in particular
3. The capability to perform multi-modal human-robot interaction

As a subset of assistive robots, Feil-Seifer and Mataric defined in 2005 the term “***socially assistive robot***” (SAR) which they refer to as robots that do not provide the first feature (the physical support) but provide assistance by their multi-modal interaction capability, enhanced by the effects of their physical appearance (social presence) [Feil-Seifer2005].

Table 2: Examples of assistive robots adapted from [Payr2015].

				
a) Robotic Mobility Aid “Friend II”	b) Fetch & Carry Support “Botlr”	c) Robotic Manip. Aid “Asibot”	d) Rehabilitation Robot “Auto Ambulator”	e) Household Robot “Roomba”
				
f) Personal Care Robot “Bestic”	g) Telepresence Robot “Giraff”	h) Companion Robot “Hector”	i) Emotional Robot “Paro”	j) Entertainment Robot “Ifbot”

Examples of robots that belong to the group of SARs are (compare also the framed Table 2 g-j):

Telepresence robots, such as the “Giraffe” Robot [Coradeschi2013], typically are composed of a tablet-like interface on a movable base. Two key functionalities can be provided by such systems. Social contacts between the older users and informal and formal caregivers are facilitated by means of an easy-to-use video-conferencing system. Further, such robots can be tele-operated from a remote location by e.g. a doctor or caring relative in order to observe the user’s current situation even in circumstances where the user itself is not able to initiate interaction. This is supposed to support the safety of users living alone at home, however studies on the impacts are yet largely missing and it is still unclear whether users see the additional benefits and accept the inevitable privacy violations.

Companion robots, such as the prototype “Hector”, which was developed within the “Companionable” research project [Schroter2014], are discriminated from other categories mainly by their capability to perform a variety of different tasks. According to an early definition by Kerstin Dautenhahn:

“... a companion robot is a robot that (i) makes itself ‘useful’, i.e. is able to carry out a variety of tasks in order to assist humans, e.g. in a domestic home environment, and (ii) behaves socially, i.e. possesses social skills in order to be able to interact with people in a socially acceptable manner.” [Dautenhahn2007a]

Companion robots typically possess skills that enable them to interact with users, allowing them to serve and assist them in many different scenarios, as compared to e.g. telepresence robots that might also involve social interaction but are mostly limited to their functionality as remote presence and tele-operation devices.

Emotional robots like the “Paro” robot [Wada2007] are derived from the known beneficial effects of pet-therapy and typically resemble domestic animals such as cats or dogs. The systems are capable of a limited set of typical pet behaviours to make them appear more realistic, natural and life-like. Such robots are currently mostly used in care facilities and support the work of carers by enhancing the communication with patients and the quality of life of older users by providing a feeling of social interaction. Positive effects could be shown in the therapy of dementia patients where the systems were found to be useful in opening up the social conversation between therapists and users and to have an effect on the emotion regulation of patients [Shibata2011].

Entertainment robots with the main aim of providing amusement and game-like interaction such as the “Ifbot” [Plasticpals2009], which was used in hospitals to entertain patients by interaction and memory games, are used mainly in conjunction with interactive-training games such as serious games for brain training. Only a limited number of systems were developed primarily for the target group of older users, however the entertainment functionality is a typical part of other robot types, such as companion robots.

One motivation of the research on *socially* assistive robots is to narrow the research field by stripping it from the technically challenging and expensive physical-support abilities to allow in-depth studies of particular aspects, such as on socially meaningful and acceptable interaction. Insights gained within these focused studies can later either be integrated into robotic systems of other categories or physical-support capabilities could be added to current SAR designs, as is currently already done in particular with companion robots that in some cases include means such as an attached robotic arm for physical support or a tray to deliver goods.

Ad 1. The capability to provide physical support

Although explicitly not a feature of SARs, it is worth mentioning briefly the potentials that reside in the field of robots that can actually perform physical support for older users. The potential application areas are huge and can be categorized as follows (compare also Table 2 a-f).

Robotic mobility aids are systems that directly support the transport of users. An example is the robot “FRIEND” (“Functional robot arm with user-friendly interface for disabled people”) [UniB2005], which is a semiautonomous robotic wheelchair that supports users with disabilities and older users and is depicted in Table 2. Other examples include robotic walking frames such as the robot “Guido” [Rentschler2008]) or even autonomous cars such as the Google Car [Markoff2010]).

Fetch and carry robots are solutions that support older users mainly by delivering needed goods, such as systems that autonomously detect and bring small items within the own home or care facility. Examples include the robot “Relay” [Savioke2015], which is depicted in Table 2, and could be used to deliver smaller goods to seniors within sheltered housing. This category also comprises robots that carry shopping bags for the user such as the robot “Budgee” [FER2014] and systems that take care of disposing of rubbish such as the “DustCart” [Ferri2011].

Robotic manipulation aids are currently early prototypes that are designed to enhance the dexterity or strength of arms and hands. As an example, “ASIBOT” [Victores2014] (see also Table 2) is a manipulator with five degrees of freedom that is mounted on a wheelchair due to its weight of 10kg. It supports disabled users in daily tasks involving the need for fine manipulation accuracy such as eating, tooth cleaning and shaving.

Rehabilitation robots are systems typically implemented at hospitals or care centres that support patients and therapists in conducting neuro-rehabilitation. A common example is the exoskeleton “Auto Ambulator” [Colombo2000] that provides support by guiding the patients’ gait movements along with constant audio-visual feedback.

Household robots are dedicated systems that support household tasks at home such as floor and window cleaning, lawn-mowing, or gutter cleaning. Despite such robots being developed mostly for a general target group, older and disabled people could also benefit from reduced cleaning needs, which also makes such devices assistive robots. A well-known example is the disc-shaped floor cleaning robot “Roomba” [Forlizzi2007].

Personal care robots are systems that were developed as dedicated systems to serve particular tasks of personal hygiene and care such as robotic toilets, robotic baths and robots for feeding support. As example is the semi-autonomous eating aid for people with disabilities “Bestic” [Villarreal2011], which is shown in Table 2, and which can be best described as a small robotic arm holding a spoon that can be used to lift food from a plate to the mouth of the users, thereby facilitating the user’s autonomy whilst eating.

Ad 2. The social presence inherent to physical objects and robots in particular

As humans, we tend to treat objects and devices as social beings [Reeves1998]. We apply social rules and heuristics from our knowledge of social interaction to the domain of inanimate devices and machines. For example, people are polite to computers, talk to systems knowing they do not understand, and refuse to discard old, malfunctioning devices because they have established a bond to them. Reeves et al. also found that people “like computers with personalities similar to their own, find masculine-sounding computers extroverted, driven and intelligent while they judge feminine-sounding computers knowledgeable about love and relationships”.

Lee et al. defined social presence as “a psychological state in which virtual actors (para-authentic or artificial) are experienced as actual social actors in either sensory or nonsensory ways” [Lee2004]. Given that virtual actors trigger this “psychological state” the hypothesis was raised that real artificial actors such as robots also cause similar,

likely even stronger, effects. This hypothesis was found to be correct in many circumstances by several authors.

Wei et al. [Lee2006] compared tangible (physically present) with non-tangible (not physically present) social agents. The physical embodiment of a social agent was found to positively contribute to several measurements including the general evaluation of the agent, the social attraction of the agent and the human-agent interaction. The same authors also found that the loneliness of users positively correlates with the level of perceived social presence, suggesting a potential use in a therapeutic context with isolated patients.

Driven by the same hypothesis, Powers et al. [Powers2007] compared a computer agent with a humanoid robot in user studies in the three scenarios a) a computer agent on a screen, b) a remote robot projected on a screen, c) a physically present robot. The findings suggest that robots have higher social impacts than virtual agents, in particular because they were found to be more engaging. Further, a physically present robot has advantages over the projected control condition in terms of helpfulness, usefulness and effectiveness as a communicator. The same authors also found some negative effects from the physical presence, as users seemed to be distracted by the appearance of the robot and performed worse in remembering key points of the communication in a recall test.

A number of research studies have been undertaken and have investigated the effects of social presence for robotic agents in more detail. Jung and Lee [Jung2004] found in 2004 that the physical embodiment of social robots enhances the feeling of social presence and that this effect causes a more positive evaluation of social robots. Heerink et al. [Heerink2008] validated this finding in 2008 and could show that *social presence* positively correlates with *perceived enjoyment* during usage, which again leads to a higher *intention to use* according to his acceptance models. Similar results were also found in non-robotic studies of social presence e.g. within a computer-mediated conferencing environment where authors suggest that social presence is a very strong predictor of satisfaction [Gunawardena1997].

Within a study using the “iCat” robot [VanBreemen2005] as a conversational partner in a chess game, Leite et al. [Leite2009] found an influence of usage time on social presence, as several social presence factors (attentional allocation, perceived affective and behavioural interdependence) declined after a usage duration of five weeks. The

authors reasoned that the social abilities the robot was capable of were not enough to create and maintain the perception of social presence in the long term, suggesting that robots might appear believable and intelligent at first sight but need to be designed in a more intelligent way to allow users to benefit over a longer term.

More recently research has been directed towards the use of SARs in eldercare institutions or for use at home by older users. Regarding the social presence of SARs in eldercare scenarios, Heerink et al. found that the social presence positively correlates with the expressiveness of the robot's social behaviour [Heerink2010a], meaning that a more extroverted and expressive robot is perceived as more socially present, resulting again in higher acceptance ratings.

Ad 3. The capability to perform multi-modal human-robot interaction

The capability of robots to perform communication-supporting movements in addition to auditory and optical in- and outputs enriches the possibilities for interaction with humans and raises the potential of human-like communication. In particular, in the case of anthropomorphic robotic systems, it seems possible to realize a HRI that resembles a social human-human interaction by including the use of social cues such as co-speech gestures, simulated mimics, eye contact and body postures in addition to social voice communication.

Such social abilities for robots, and in particular for companion robots, are not only beneficial to enhance the communication and the respective functionality of the robot but are even necessary for robots that interact with humans in daily activities. As a contrast, let us consider a robot moving among people without social navigation blocking people on their way, which certainly would soon cause frustration, possibly leading even to the denial of such technology. Going even further, Kerstin Dautenhahn [Dautenhahn2007a] argues that social skills are not only a nice add-on to enhance HRI but are essential to understand and develop the cognitive skills of robots and thereby advance artificial intelligence (AI), as the social intelligence is “a key ingredient of human intelligence”. So, if we want to further explore the promising field of embodied AI, it is necessary to research social abilities of robots which are a key part of HRI.

Several authors have investigated the role of social abilities for HRI. Within an experiment comparing different levels of social expressiveness, it was found that there is a positive correlation between the level of social abilities shown by the robot and the social presence, which again enhanced the perceived enjoyment of the robot during

interaction [Heerink2008]. This finding is backed up by another experiment in a particular setting using a physical-exercises scenario with older users and with a SAR as trainer. The users strongly preferred a social variant of a robot that gave verbal praise, displayed continuity behaviours such as referencing past experiences with the user, showed humour and referred to the user by name over a robot that only gave functional feedback as needed [Fasola2012].

Several studies examined the role of non-verbal social cues within HRI. It was found that non-verbal cues such as gestures and social gaze behaviour improve the users' compliance with a robot and have a positive influence on the persuasiveness during conversation [Chidambaram2012], deictic gestures ease building a common ground in communication [Brooks2006] and even idle motions that are used during interactive breaks and do not serve a functional purpose seemingly make the robot more human-like, alive and empathic [Cuijpers2015].

Not only do non-verbal social cues enhance the users' attitude towards the robot, but they can also increase the functional performance of systems, as was shown by Van Dijk et al. [VanDijk2013]. They showed within an experiment with older users that co-speech gestures aid verbal communication as users performed better in a recall task on a message provided by a robot that used action-depicting gestures compared to a robot that did not.

Interaction also happens during navigation such as when approaching or passing by people. Several authors researched the influence of social navigation in environments with humans. Kerstin Dautenhahn [Dautenhahn2007a] investigated the differences between socially ignorant and socially interactive navigation in a laboratory context with users and found that most users disliked a robot moving behind them, blocking their path or driving on a collision course, in particular when within their personal space (3m). Proxemics were studied by several authors and it was found that users prefer similar interaction distances as in human-human interaction [Walters2009], [Takayama2009] but that the ideal conversation distance may also depend on the size of the robot, as a small robot resulted in a preference for higher distance (because users won't have to bend over it so much) [Torta2013b].

By means of embodied multi-modal interaction, properties of human-like personality can be simulated as well. Personality can be simulated by several controllable factors such as facial expression, voice (speed, pitch and volume), movement patterns, speed of

movements, text and approach path (compare also [Panek2015]). The simulated personality of a robot was found to have an influence on the acceptance of the robot. Tapus et al. [Tapus2008] gave evidence that users of an assistive robot preferred a simulated personality that matched their own and that a matching robot even produced a positive effect on users' task performance.

It seems important to keep in mind that social cues can have a positive effect only if they are used in a way that seems natural to our understanding of social communication. Under certain conditions, it could be shown that unnatural usage of social cues can negatively influence the enjoyment of HRI, as was argued by Torta et al. [Torta2012] based on an experiment in which a robot looking at a person from the side was considered less pleasant as people felt it was impolite.

2.1.1 Typical functionalities of SARs

SARs can provide a vast range of assistive functionalities that are comparable with those of today's smartphones, augmented with the ability of social interaction and the effects of social presence. Due to their lack of physical support, SARs are typically used in specific applications and combinations thereof such as:

a) reminders

Typically, reminders are used in systems that support older users, particularly users with mild cognitive impairments, to structure their day and alert them in case of appointments. For example, within the "Pearl" project [Pollack2002] a robot was built that, among other features, helped older users with schedule planning and reminded them about their regular daily activities. The "Domeo" robot developed within the AAL joint programme also included, among other features, a calendar and agenda system.¹⁶

b) motivators

Several studies suggest that robotic solutions can be convincing motivators due to their physical nature and inherent social presence [Kidd2008][Fasola2012][Saerbeck2010]. Examples of robotic prototypes that facilitate this capability are the "Autonom" robot, developed by the Massachusetts Institute of Technology's (MIT) media lab, which was used to study the motivating effects of robotics supporting users keeping their diet programme [Kidd2008]; the robot "Bandit" developed by the group around Maja Mataric at the University of Southern California (USC), which was used to study the

¹⁶ <http://www.aal-domeo.org>

motivational influence within a training programme including physical exercises [Fasola2012]; and the “iCat” social robot which was used to study the motivational influence of different behavioural roles when teaching school children [Saerbeck2010].

c) entertainers

Assistance can be provided by entertainment alone. This was found in an early study by Tamura et al. who evaluated the effects of the robotic dog “Aibo”, which was used in a scenario to entertain older users within a care home [Tamura2004]. Broadbent et al. undertook another experiment with the conversational robot “Ifbot” within a care setting and found that users were initially excited about the robot but lost interest after one month [Broadbent2009]. Entertainment functionality is since included in most companion robots for older people as a secondary feature. For example, the “Hobbit” project developed a robotic assistant for older users at home that among other features also includes entertainment functionality such as looking at photos, listening to music or playing games [Fischinger2014].

d) communicators

The facilitation of communication with family members, friends or care staff is a common target of socially assistive robotics for older users. Typical examples of robots that assist by providing communication functionalities are telepresence systems such as the robot “Giraff” [Coradeschi2013]. Additional communication features are mostly also present in multi-functional companion robots such as in the “Care-O-Bot” [Amirabdollahian2013] or the “Nao” robot used within the KSERA project [Werner2013].

e) information centres

SARs can be used as a simple-to-use information centre, in particular for computer-sceptic users. This functionality was explored in a range of research projects where SARs were typically used to inform users about their local environment or schedule, the weather conditions or the news [Werner2013]. The same concept is also implemented in the commercial robot “Pepper” which is currently used to inform customers of *Nestlé* stores in Japan [Nestle2014].

f) emotional support

The robotic seal “Paro” by *AIST* is the most-studied robot providing emotional support. The support is given by means of a simulated pet therapy and was shown to have

positive effects on the quality of life of older adults with cognitive impairment [Sabanovic2013]. The same principle has also been lately used in other robotic pets such as the commercial robotic cat “JustoCat” [Gustafsson2015].

2.1.2 Uptake of SARs in current research & commercialization

Although assistive robotics for the support of older users and carers is a relatively new research field, it has been picked up very well within the research community, as can be shown by recent research activities. The number of research projects have increased steadily over the last number of years and also the funding from the European Commission has increased over time, as is shown in Figure 4 that details both investment in euro and the overall number of research projects at European level. The decline in the year 2014 is likely not due to a loss of interest in the field, as it can be explained by a gap between the “Horizon 2020” and “Framework Programme 7” funding schemes. The underlying data for this graph also tell us that the average European research project lasts for three years and consumes €1 million of funding per year.

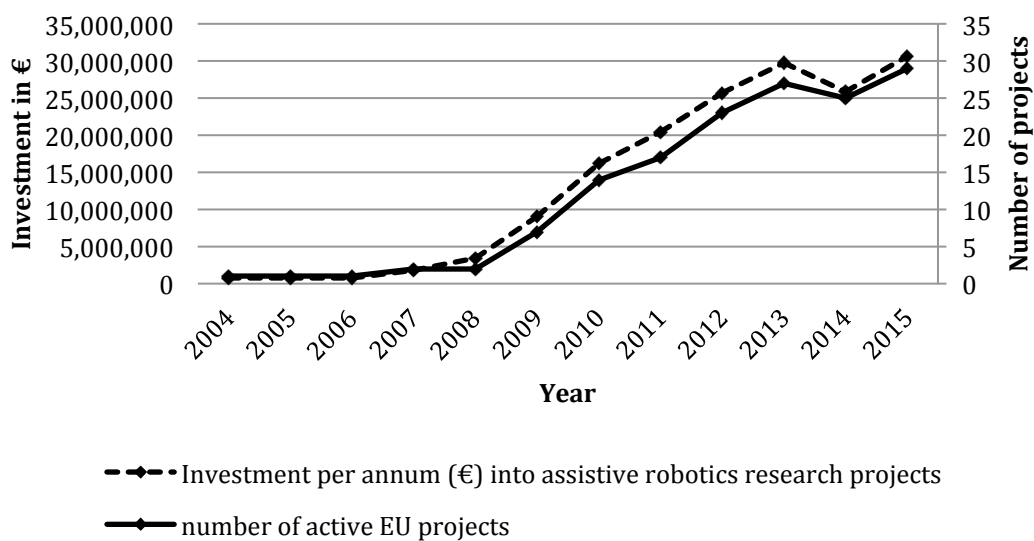


Figure 4: Investment into assistive robotics research, modified from [Payr2015].

An analysis of the goals of research projects over the past nine years shows that interestingly, the majority of research projects and funding is dedicated to the subfield of SARs; see also Figure 5.

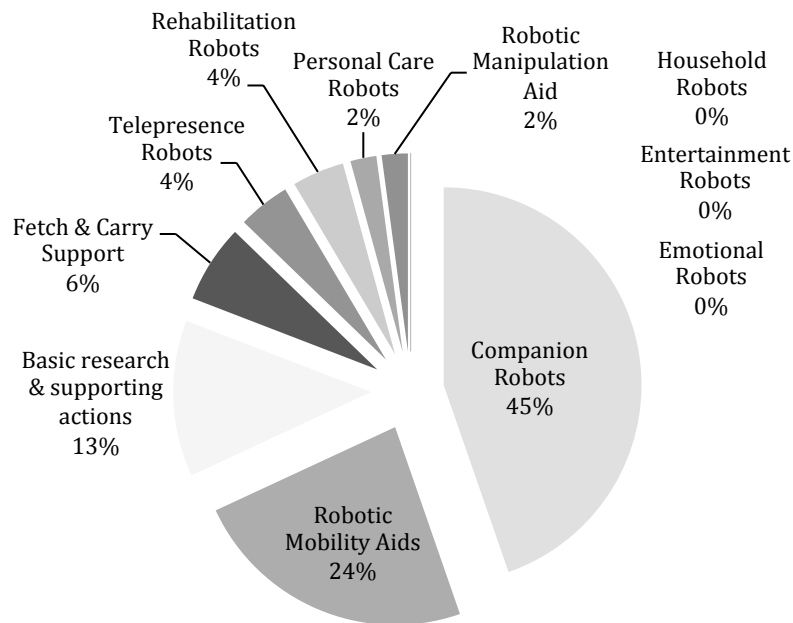


Figure 5: Distribution of robot subtypes among current research activities, modified from [Payr2015].

Given this data, companion robots (mostly SARs) qualify for 45% of the overall funding within European robotics projects, stating the high research interest into this particular field.

Since research projects increasingly step away from building own robotic solutions and rather develop new software for existing robotic platforms, a market for such solutions has built up. The most frequently used SAR solution on sale today is the robot “Nao” from *Aldebaran* robotics, of which 5,000 units were in use by educational and research institutions in 70 countries in 2014, according to the inventors [Aldebaran2014].

According to the World Robotics Survey [WRSR2015] issued in September 2015, robotic products that support older people and people with disabilities are already now a commercial success and have quickly gaining relevance, with 4,416 units sold in 2014 as compared to only 699 units in 2013.

2.2 User-centred evaluation methods

2.2.1 Methodology of the literature survey

A narrative review was conducted to gather the state-of-the-art (as of the time of writing in Q3/2018) regarding methods used to develop and in particular to evaluate SARs. This narrative review of primary literature adheres to the methodology and considerations presented by Green et al. [Green2006].

Several steps were used to acquire and select literature as sources for this review:

As a first step, literature was searched for in *Google Scholar* by using the following list of keywords: “assistive robot* method”, “assistive robot* evaluation”, “robot* evaluation method” and “user research robot*”. Interestingly only a few papers could be found detailing information on the evaluation of companion robots. As the current development of companion robots depends on large resources and long-time spans for reasons of technical complexity and the need of experts from several research domains such as technology, sociology and healthcare, the focus of the literature search was laid on the results of larger European projects which are also able to provide the necessary resources.

Projects funded by the European commission within the framework programmes and the Horizon2020 programme are listed in the “Cordis” database.^{17,18} Projects funded by the “AAL Joint Programme” (AAL-JP) are listed on the AAL-JP website.¹⁹ The Cordis database was searched for the terms “robot elderly”, “robot senior”, and “robot older”, which gave 71 results of which 23 were projects developing robots to assist older users. The AAL-JP website hosts the abstracts of 249 AAL projects which were hand searched for projects aiming to develop or evaluate a robotic solution. Seven AAL-JP projects with this aim were found, giving altogether 30 relevant projects within the field of assistive robotics for older users.

Since the evaluation usually takes place towards the end of the project, projects that are currently running and end later than Q2 2018 are highly unlikely to already have published evaluation results and were excluded from further analysis. Of the remaining 39 projects, 15 were excluded as they did not develop companion robots but other

¹⁷ http://ec.europa.eu/research/fp7/index_en.cfm

¹⁸ http://cordis.europa.eu/home_en.html

¹⁹ <http://www.aal-europe.eu>

assistive robots for older users such as exoskeletons, rehabilitation robots or a pedestrian assistant. One project (KSERA) was excluded to avoid a possible review bias since the author participated in the evaluation. Finally, 24 projects were selected for detailed analysis. The earliest project ended in 2010, the latest project ended in April 2018.

For the remaining 24 research projects, publications were searched for a) on the project's website, b) by directly contacting responsible investigators and c) by searching through publications of institutions undertaking the user trials within the project.

49 publications (43 from the project-based search, six from the general search of databases) could be identified to contain relevant information on user evaluation of robotic technologies for this review. Publications in the languages English and German with a publication date later than 2007 were considered.

Assessing the quality of studies

Papers were selected based on the quality and detail provided. Literature that did not provide basic information on the evaluation methods including the evaluation aims, trial setup, participating users and methods used for results generation were omitted. Based on these criteria, out of 49 publications, 20 key publications from 10 projects were selected for the detailed analysis within this review. For the remaining 14 projects, the information found on the evaluation procedures and methods was either too scarce or only the planned evaluation procedures rather than the actual trial results were reported.

Data extraction

Data about the evaluation aims, evaluation setup, participating user groups and used methods and metrics was extracted from literature and inserted into evidence tables for further analysis. Evidence tables are reported within the chapter.

Categorization of data

To provide a comprehensive overview on current methodologies and flows of research within robotic projects, data was structured along common themes. As methodologies used are depending on the aims and the technology readiness of the technical probes, it was decided to structure the data along the typical workflow within European projects, which again is linked to the technological advancement of the research prototypes over the evaluation phases.

The NASA (National Aeronautics and Space Administration) invented a model of technology readiness that can be used to categorize the gathered data as the methodologies used for evaluating technical prototypes depend upon the technological readiness. The model was later adopted and used within the “Multi-Annual Roadmap” project [euRobotics2016] to describe the future aims of robotics research. See also Table 1 for an overview on the levels of technology readiness. The highlighted items in the table show the technology-readiness levels (TRLs) achieved within the reviewed projects.

Table 3. Technology-readiness levels as proposed by NASA [Nasa2015].

TRL	Description
1	Basic principles observed and reported
2	Technology concept and/or application formulated
3	Analytical and experimental critical function and/or characteristic proof of concept
4	Component and/or bread board validation in laboratory environment
5	Component and/or bread board validation in relevant environment
6	System/subsystem model or prototype demonstration in a relevant environment
7	System prototype demonstration in an operational environment
8	Actual system completed and qualified through test and demonstration
9	Actual system proven through successful mission operations
10	Commercial

2.2.2 Results of the literature survey

The following discussion is structured along the presented TRLs. The used evaluation aims, methods, involved user groups and test settings are reported for each category as they vary between the different categories based on the achieved TRLs.

2.2.2.1 Laboratory trials of the integrated prototype (TRL-4)

The goals of this phase are to verify the correct functionalities of all system parts in conjunction with each other and to guarantee the sufficient reliability and stability of the prototype to allow for later evaluation phases involving the targeted user groups.

One example is given by Merten et al, who reports laboratory trials of the “Companionable” robot regarding the mechanical design of the drive system, the

mechanical framework of the robot, the system architecture including the communication networks and implemented software functionalities.²⁰ Furthermore, the safety concept was reviewed in cooperation with an independent testing laboratory. The usability of the system's interactive components was validated regarding ergonomic standards [Merten2012].

Methods used within this phase include:

a) Integration tests such as checklist type tests to validate the correct functionality of all integrated technical modules. Ad-hoc lists are used that define single test cases [UWE2010]. Integration tests typically take place within a laboratory setting or within a setting mimicking a real-life environment such as a living lab [UWE2013], [Merten2012].

b) Usability evaluation by experts who walk through the concept description and mark all positive and negative aspects they think affect the user experience are undertaken in [Schröter2014]. Heuristic evaluation, as proposed by Nielsen et al., is a specific method undertaken within this phase by HCI experts from within the project to validate the system's usability prior to conducting user trials [UWE2013], [Nielsen1990].

c) System pre-tests are conducted at homes of project members and project-related users such as grandparents of researchers that are easy to recruit and rather tolerant regarding the probable lack of functionality and usability [Pigini2012], [Pigini2013].

Checklist type functional tests are conducted similar to those within integration tests, with the exception of a setup within a real environment [Pigini2013]. As the prototype in this stage is typically not yet stable, the Wizard of Oz technique [Green2004] is of strategic importance to simulate functionalities not yet fully integrated or not yet working smoothly enough but needed in order to test other functions depending on it [Pigini2013].

d) Integration tests are conducted to gather information on potential issues regarding the integration of the robotic platform and surrounding technologies such as smart-home equipment into a real environment, in case the plan is to perform field trials at users' homes with this prototype, such as described in Pérez et al. [Pérez2014].

²⁰ www.companionable.net

Table 4. Part of evidence table for laboratory trials of the integrated prototype.

Reference	[Merten2012], [Schröter2014]	[Pigini2013][Pigini2012]	[Pérez2014]
Project name	Companionable ²¹	SRS ²²	Accompany ²³
Robot type	Companion	Companion	Companion
Robotic Platform	Scietos G3	CareOBot	CareOBot
Aims	Verification of technical specification. Validation of usability for user trials.	Measure technical performance, usability and acceptance of the prototype to generate feedback for improvement.	Get a first exploration on how to deploy a robot at a trial site. Get general opinions on the robotic use cases from potential users. Usability evaluation.
Setup	Laboratory setup.	Whole-system pre-test in real-home of project-affine users (grandparents of researcher). Functional test for a duration of 1.5 days.	The used robot was placed inside the activities room of a sheltered-housing facility.
Users	<i>unknown</i>	Two older users aged 80 and 81.	10 older users from an elderly activities facility
Methods	Functional tests of all technical systems. Safety evaluation by German TÜV.	Evaluation list for technical performance measurements of system components. Semi-structured interviews with participants implementing Wizard of Oz.	Technology probe, Interview, Observation.

2.2.2.2 Short-term user trials of the integrated prototype within realistic environments (TRL-6)

A wide array of research questions is targeted by implementing user trials of the integrated prototype outside of the field and laboratory, within controlled but still realistic settings. Focus in most projects was found to be laid on usability evaluation and evaluation of acceptance and social aspects resulting from the anthropomorphic characteristics of the robots used.

²¹ <https://www.tu-ilmenau.de/neurob/projects/finished-projects/companionable/>

²² <http://srs-project.eu>

²³ https://cordis.europa.eu/project/rcn/100743_en.html

Workshops, focus groups and group discussions

Within focus groups, group discussions or questionnaires, scenarios that provide show-cases of typical assistive functionalities are shown to groups of primary, secondary or tertiary users and user feedback is gathered [Cesta2012a][UWE2013]. The scenarios might be shown by the actual prototype or by videos of recordings of the actual prototype. In the first case too, a try-out-session was included to give participants a deeper understanding of the system's capabilities and behaviour.

The aims of using this method are to gain early input on advantages and disadvantages of the demonstrated functionalities or suggestions for improvements from a diverse user group. The method has the advantage of providing input from several participants and experts from different fields within one test session, which makes it cost-efficient as compared to short-term single-user trials.

Short-term scenario-based user trials under controlled conditions

Short-term scenario-based user trials within a setup mimicking a real user's home or a living lab is the most commonly used method implemented to evaluate assistive companions [Kosman2013], [Lucia2013], [Ihsen2013], [Fischinger2014].

Within this method, individual users are typically invited for the duration of about two hours. After an explanation of the goals and an informed consent procedure, measurements are undertaken followed by a block of pre-defined scenario-based interaction with the robotic solution in which the developed usage scenarios are demonstrated one after the other. Sometimes the scenarios are embedded in larger user stories, to give the participants an impression of how they could use the system in real life. Either final interviews or questionnaires or both conclude the test session.

This method is used to cover a wide variety of research questions such as those related to technical performance or reliability, usability, acceptance and perceived value, which were also taken up by most authors. In one case, impact measurements regarding the user's autonomy and perceived safety were undertaken [Lucia2013]. In another case, Fischinger et al. reported seeking information on the perceived value and willingness to pay which is similar to a concept that was already mentioned by Coradeschi et al. who measure the use-worthiness, which reflects whether people think this technology might be worth a try [Fischinger2014], [Coradeschi2013].

Typically, primary older users were the core group of participants. The number of users varies strongly between the studies, ranging from four [Ihsen2013] to 49 [Fishinger2014] but is generally low and hence qualitative methods are mainly used such as interviews, thinking aloud and observations during the scenario execution. [Kosman2013] used experience-sampling cards with single closed questions to assess the user impression about a scenario directly after conduction.

Other authors conducted interviews and questionnaires prior and after showing scenarios. Typically, customized questionnaires were used to specifically target the evaluation aims [Kosman2013], [Lucia2013], [Fischinger2014], which indicates a lack of standardized or well-accepted questionnaires. Lucia et al. facilitated the “AttrakDiff” questionnaire [Hassenzahl2003], which is composed of 28 items to evaluate factors of usability and user experience. The questionnaire can be used in lab as well as field studies. Within the AttrakDiff questionnaire, hedonistic and pragmatic dimensions of the user experience are studied by means of semantic differentials.

Most authors additionally involved informal caregivers as a secondary-user group, firstly to gain their views on the questions of research and, in particular, in the case of tele-care or communication functionalities which need a counterpart for communication to evaluate these specific functionalities (both from the side of the client and carer) in which case the evaluation was done with teams of participants consisting of one primary and one secondary user [Kosman2013], [Pigini2013].

Longer user trials under controlled conditions

Schröter et al. report of trials conducted in a living-lab situation in which users were invited to stay for a duration longer than the two hours typical within short-term trials. The authors clearly tried to go as close to field trials as possible without leaving the controlled environment necessary to safely conduct trials. Users stayed for two consecutive days but slept at their own homes [Schröter2014]. In contradiction to the short-term trials described above, in this case the developed usage scenarios are embedded into the users’ daily routine, providing a more realistic experience for the participants as well as also including possible repetitive or annoying situations. Only primary users were used in the described evaluation and the aims were comparable to the short-term scenario-based interactions as described above.

Table 5. Evidence table for user trials of the integrated prototype (part 1).

Reference	[Merten2012], [Schröter 2013], [Schröter2014]	[Cesta2012], [Cesta2012a]	[Kosman2013], [Melenhorst2013]
Project name	Companionable ²⁴	ExCite ²⁵	Florence ²⁶
Robot type	Companion	Tele-Presence (Companion)	Companion
Robotic Platform	Scietos G3	Giraff	Florence (developed within the project)
Aims	Validate “interaction” between robot and smart-home. Evaluation of usability and acceptance in real life.	Assess users’ reaction towards the adoption of the robotic system. Assess willingness to adopt the robotic solution, possible domains of application, advantages and disadvantages and suggestions for improvements.	Technical performance of the prototype. Usability evaluation to give recommendations for future prototypes. Gather overall impression of the users.
Setup	6x 2-day trials within an environment mimicking a real-user’s flat.	Workshop with a group of participants. Interviews with older users.	Short-term demos of scenarios in a living-lab setting mimicking a real-user’s flat.
Users	6 older users with mild cognitive impairments.	10 older adults 44 health-workers (26f, 18m) from different disciplines.	5 primary older users (4m, 1f, 68-86y), 5 informal carers, 2 tertiary users (professional tele-care support staff).
Methods	Semi-structured interviews, observations, diary, ad-hoc questionnaires.	Workshop with health-workers: presentation, try-out-session, focus group and final ad-hoc questionnaire. Interviews with older adults: (video) presentation of the robot, interview and qualitative analysis thereof.	Pre-test interview. Experience sampling cards (tailored closed question questionnaire). Post-test interview. Observations during the tests.

²⁴ <https://www.tu-ilmenau.de/neurob/projects/finished-projects/companionable/>

²⁵ <http://www.aal-europe.eu/projects/excite/>

²⁶ https://cordis.europa.eu/project/rcn/93917_en.html

Table 6. Evidence table for user trials of the integrated prototype (part 2).

Reference	[Pigini2012] 2013][Pigini2012]	[Rehrl2012] [Ihsen2013]	[Fischinger2014], [Pripfl2016]
Project name	SRS ²⁷	ALIAS ²⁸	Hobbit ²⁹
Robot type	Companion	Companion	Companion
Robotic Platform	CareOBot	Scietos A5	Hobbit (developed within project)
Aims	Evaluation of technical effectiveness, Impact on autonomy and safety, usability, acceptability / intention to adopt.	Evaluation of usability, user friendliness, system performance.	Usability of multimodal interaction possibilities, acceptance of the robot, perceived value with respect to affordability and willingness to pay.
Setup	Scenario-based test sessions with users in teams consisting of an elderly user together with an informal caregiver and / or remote operator within a test site.	Scenario-based individual-user try-out sessions. Two main trial iterations with users with 1 year in between to allow for technical modifications.	Short-term scenario-based individual trials at 3 similar test-sites in simulated real homes (living labs) decorated as living rooms.
Users	16 elderly users, 1 young disabled man. 12 informal caregivers (relatives). 5 professional operators (tertiary users from a 24hour call centre).	4 primary users (2f, 2m) 2 care givers.	49 primary users aged 70+ with typical age impairments.
Methods	Evaluation check-list for technical performance. Interactive think-aloud with moderators. Ad hoc developed questionnaires. Attrackdiff questionnaire. Focus group on safety, ethical and privacy issues after the test session.	Task-oriented test methods taking users' behaviour and comments into account. Observation during the conduction of trial scenarios. Analysis of user comments.	Wizard of Oz, Ad hoc developed questionnaires for usability, acceptance and affordability.

²⁷ <http://srs-project.eu>

²⁸ <http://www.aal-europe.eu/projects/alias/>

²⁹ <http://hobbit.acin.tuwien.ac.at>

2.2.2.3 Field trials in real environments (TRL-7)

Field trials have been undertaken by projects in more recent years and in particular by using either product-grade off-the-shelf robotic systems or functionally minimal robotic solutions e.g. with restricted ability to interact (compare also [Leite13]). Mucchiani et al. were able to use a technically advanced robotic system that was initially developed for the commercial setting of goods delivery to hotel guests [Mucchiani2017].

The goals and research questions of most projects were to gain information on the impact of a robot on care and the impacts on health and quality of life of the targeted user groups. Nevertheless, aspects of all research goals of earlier phases were also included such as measurements of social aspects, usability measurements and measurements regarding the technical performance within a real-life setting.

Typically, a within-subject design was chosen for field trials with respect to the inter-individual differences of older users and users with disabilities. Questionnaires, (semi-structured) interviews and medical measurements were used as repeated measurements prior, during and after the integration of the robot into the users' homes or care facilities to gain information on the impact of such systems on the users. User diaries and technical data-logging were the most often used methods to gain continuous information about the user experience over time and the technical performance of the systems. (See also evidence in Table 5 and Table 6).

Heylen et al. reported of a technique to use video logging within the homes of users by using cameras that would only activate when the participant pressed a button, in order to take account of privacy needs [Heylen2012].

Authors that report of standard questionnaires name the "Standard Usability Measurement Inventory" (SUMI) [Coleman1993], which was initially developed to evaluate the usability of software, and the "System Usability Scale" (SUS) [Brooke1996], to measure aspects of usability. The "Positive and Negative Affect Schedule" (PANAS) [Terraciano2003], the "Short Form Health Survey" (SF12) [Ware1996], the "Geriatric Depression Scale" [Yesavage1982] and the "UCLA Loneliness Scale" [Russel1980] were used to gain insights on the impact on quality of life and health by the introduced systems. The "Multidimensional Scale of Perceived Social Support" (MSPSS) [Zimet1988] was used to measure the impact of the system on the subjective feeling for social support which influences depression and anxiety symptomatology and hence is also a factor for quality of life.

To assess the acceptance and factors of usability of the systems, the “Almere Model” [Heerink2010] and the “Godspeed questionnaire” [Bartneck2008] were implemented as these were both specifically developed to assess acceptance factors of social robots as companions.

Table 7. Evidence table for field trials in real environments.

Reference	[Cesta2012a]	[Pérez2014]	[Heylen2012]
Project name	ExCite ³⁰	Accompany ³¹	SERA ³²
Robot type	Tele-presence (Companion)	Companion	Companion
Robotic Platform	Giraff	Developed within the project, similar appearance as CareOBot3	Nabaztag ³³
Aims	Monitor robots usage over time, measure impact on users’ health and quality of life.	Evaluate a) perceptions and attitudes towards the robot, b) impact on daily routines, c) impact on physical and psychological health.	Study HRI aspects such as attitudes towards the robot and their change over time, interaction of participants with the device.
Setup	Field trials for a duration of 3-12 months in users’ homes.	Field trial for 3 weeks at the participants’ home (in the living room).	Field trial for a duration of approx. 10 days each.
Users	Users consist of pairs of primary older users (have the system at home) and secondary formal or informal caregivers (teleoperate the system).	1 older user, male, 74years, living alone at home, technically experienced.	6 healthy primary users (aged 50+).
Methods	Repeated measurements prior, during and after integration of the robot. Evaluation with carers: ad-hoc questionnaires, SUMI questionnaire [Coleman1993], Temple Presence Inventory [Lombard2009], PANAS [Terraciano2003], structured interviews, diary.	Pre-/post interview, daily diary. Objective methods: frequency and duration of use, performance score of a health exercise, heart rate. Godspeed questionnaire [Bartneck2008]. Almere model [Heerink2010].	Analysis of video recordings. Semi-structured interviews before, during and after the test. Diary to note interesting aspects during the test duration.

³⁰ <http://www.aal-europe.eu/projects/excite/>

³¹ https://cordis.europa.eu/project/rcn/100743_en.html

³² https://cordis.europa.eu/project/rcn/89259_en.html

³³ <http://www.nabaztag.fr>

	Additionally, for evaluation with primary users: UCLA Loneliness Scale [Russel1980] SF12 [Ware1996], MSPSS [Zimet1988] Geriatric Depression Scale [Yesavage1982] Almere model [Heerink2010].	Source Credibility Scale [McCroskey1999] to measure trust in the technical system. Personal Opinion Survey (POS) [McCraty1998] to measure impact on stress.	
--	--	--	--

Table 8. Evidence table for field trials in real environments (part 2).

Reference	[Radio2017a]	[Mucchiani2017]	[Vroon2015]
Project name	Radio ³⁴	none	Teresa ³⁵
Robot type	Companion	Companion	Tele-presence (Companion)
Robotic Platform	Developed within the project.	Savioke ³⁶	Developed within the project, based on the Giraff platform
Aims	Evaluation of the usability for primary users (older people).	Understand efficacy of HRI and enhance future robot versions.	Investigate user acceptance and experience.
Setup	The users were involved for 5 days (2 days for deployment of the system at the participants' homes, three days of actual pilot study).	Field trial for 1 week (4 users) or 2 days (12 users).	Deployment of the robotic system during 4 sessions of a weekly activity to groups of participants. The robot was controlled using a Wizard of Oz approach.
Users	2 users were recruited from beneficiaries of a home-care service and from volunteers of a social care activities network. Users were excluded from the trials if unable to operate the robotic system.	16 older users living in supported apartment living.	Older users within a nursing home who were already part of a coffee and quiz activity.
Methods	Users complete a set of assessments over the course of 3 days. Day 1 is used primarily for a pre-assessment and training of the usage of the robotic system. On day 2, the system is used and the user	Immediately after each interaction (e.g. the robot delivered water, or the robot guided the users through the building) a post-interaction survey, including a questionnaire based on the Almere Model [Heerink2010], was	Qualitative approach observation and retrospective video analysis, group discussion as well as a final semi-structured interview with older residents and unstructured meetings with care staff.

³⁴ <http://radio-project.eu>

³⁵ <https://teresaproject.eu>

³⁶ <http://www.savioke.com>

	<p>experiences different scenarios such as “pill intake”, “bed transfer”, “chair transfer”, “meal preparation” over the course of the day. On day 3, usability satisfaction and quality of life questionnaires are filled out within an in-depth interview for qualitative analysis.</p> <p>Standardized assessments used include: Long-Term Care Facilities Form [Kim2015], SUS [Brooke1996], the Psychological Impact of Assistive Device Scale (PIADS) [Jutai2002].</p>	<p>conducted. Further, an observation was undertaken and project-specific parameters were noted down.</p>	
--	--	---	--

2.2.2.4 General considerations

Evaluation aims

Three main evaluation aims could be identified across the literature.

Most reviewed studies had the main goal of developing an assistive companion for older users and use study results to provide insights on how to further improve the developed companion in the future, relative to the current solution (see e.g. [Merten2012], [Kosman2013], [Fischinger2014]).

Another main goal was to show that the developed prototype has an impact on users’ care, health and/or quality of life. In this case, the implemented evaluation methods were selected to evaluate or prove impacts resulting in the necessity of long-term interactions [Cesta2012a], [Pérez2014].

A third major goal was to push the state-of-the-art in a particular research field such as HRI. In that case, evaluation was used to gain insights on the use of robotic companions in general, rather than to validate a particular development [UWE2013], [Pérez2014].

User groups

User groups were typically split into two to three sub-groups with different interests, often named “primary”, “secondary” and “tertiary” users.

In all reviewed studies, primary users were the group of older users. Different aspects were taken to qualify as primary users; in most cases healthy older users were included based on their age, such as in [Fischinger2014] or [Kosman213]. Secondary users were

often included and referred to as informal and formal carers [Rherl2012], [Cesta2012a]. Tertiary users such as technical-support staff and professional tele-operators were included in some trials [Cesta2012a], [Pigini2013].

2.2.2.5 Methodological challenges

This section presents several methodological challenges that were brought up by the authors of the reviewed literature or submerged during the review process.

Low technological readiness of the used prototypes

Several studies such as [Pigini2013] and [Schröter2014] reported technical malfunctions influencing the user evaluation. Mainly the robustness and reliability of complex technical components such as the speech recognition software were, due to their prototype state and nature, considered insufficient to undertake user trials. Issues were noted by the users and thus negatively influenced the measured user acceptance. Pigini et al. reports that in certain evaluation studies, up to 70% of the use cases they wanted to demonstrate to users showed technical issues. In particular, in uncontrolled real-life settings the technical systems lacked robustness. Pripfl et al. report of the core functionality of the robotic system being fully operational for only about 18% of the time within the conducted field trials [Pripfl2016].

Other authors found that in addition, functionalities that were technically robust showed severe issues regarding their potential for integration into real-life settings, as the used robotic system was unable to navigate around glass-objects which therefore had to be covered with sheets before operating the system [UWE2013], [Pigini2013].

It can be expected that low performance rates negatively influence study results as Heylen et al. found that a poorly designed robot frustrated people and hence biased acceptance results [Heylen2012].

In other cases, the trial methodology had to be altered to compensate for a lack of robustness. Vroon et al. changed their initial plan of conducting field trials of a three-week duration for the sole reason that they were not able to log-in their robot into the test site's Wi-Fi network [Vroon2015].

Difficulties in conducting user trials with the group of older users

Older users are a heterogeneous group with high inter-individual differences. Most reviewed projects used the chronological age as inclusion criteria to select participants, assuming that this would result in a homogenous user group. This cannot always be

expected, as Britt Östlund also argues: “... chronological age is not a sufficient measure for older people’s life situation” [Östlund2015].

The health status of participants might lead to a high number of study dropouts, in particular in case of trials running over long durations. Rehr et al. had to modify their initially planned trial methodology because the poor health status of study participants did not allow their further involvement within the trials [Rehr2012].

Within the “Teresa” research project, the trial setup was altered after researchers found users within a nursing home were incapable of filling in a questionnaire and seemed scared to participate in a formal experiment as they feared to be “not good enough” for the project and hence hesitated to sign an informed consent form [Vroon2015]. Within the “Radio” project, out of an initially planned group of ten users in a real-life evaluation, only two users were finally recruited for the trials; furthermore, only three days of trial duration were planned [Radio2017].

Lack of accepted methodologies

Ganster et al. states that the research field of assistive robotics is within an “exploratory” state in which qualitative research methods and subjective measurements are dominant [Ganster2010]. For that reason, hardly any standardized measurement instruments could be used within the reviewed projects. According to Feil-Seifer et al., it would be necessary to generate a means to compare robotic systems to each other, even if they are designed for different tasks, to establish benchmarks for effective and ethical designs of SARs [Feil-Seifer2007].

Issues regarding long-term field trials

Only one of the so far reviewed field trials (Pripfl et al. 2016) has reached the minimum duration of two months which is necessary to gain information on acceptance without the bias of the participants’ initial excitement [Broekens2009]. In more recent years, an increasing number of projects and studies tried to undertake real-life field trials. However, as the presented results suggest, almost all of the presented studies faced severe methodological problems in conducting the trials, leading mostly to a steep decrease in study participants and/or a methodological shift towards a more qualitative approach (compare also [Radio2017a, Raadio2017b, Pripfl2016]).

Heylen et al. also remind us that real-life trials are not necessarily superior to short-term trials in realistic settings, for two main reasons [Heylen2012]. Firstly, trials at

users' homes cannot eliminate experimental biases such as socially accepted answers. Secondly, although trials are conducted in real-users' homes, the character of an experiment can still be evident to the users and users also behave differently during interaction phases if they just have the research project in mind. The situation in real-life trials is therefore not comparable with real use of a purchased system.

Limitations of this review

The literature review is limited to sources from funded projects at a European level. In particular, no national or overseas publications were considered.

Within this dissertation, because of a lack of information present in peer-reviewed sources, project reports, namely public deliverables of European projects from the EU-FP7 and AAL-JP programmes, were also analysed. The scientific quality of information presented in public deliverables is not validated as they are commonly not peer-reviewed. It could be argued that deliverables are a necessary work to convince reviewers of funding organizations and might hence be rather positively phrased. However, the author believes this is not the case for the reviewed descriptions of the methodology used.

The methodologies used strongly depend on the research aim, which varies within the literature presented and does not always fit well to the chosen categorization of TRLs. In that way, the categorization is limited, but the author still thinks that the presented overview is helpful to other researchers as it can be used to find potentially fitting methods for future studies.

2.2.2.6 Conclusion

An overview on current practices and current methodologies used for the user evaluation of companion robots has been given here and has included current typical research aims, research methods, test setups and user groups.

Additionally, a discussion of methodological points was presented – in particular regarding the selection of methods, which is partly caused by a lack of available standardized methodologies and evaluation frameworks as well as a low technological readiness of used prototypes and its consequences. Given the complex technology used in robotic systems, it currently seems clear that technical issues will also be present in most evaluation phases in future studies. This is to be taken into account by the user researchers who have to ensure a system which seems to work perfectly to the user in

order not to bias the evaluation results, in particular those concerning acceptance and user experience.

All reviewed projects that tried to perform real-life field trials with robotic prototypes reported severe issues in trial execution. The lesson learned seems to be that only product-grade robotic platforms should be used within real-life trials.

Out of 39 researched projects in the field of assistive robotics, only 24 did not belong to the field of companion robotics. Hence it can be stated that the research field is currently focused on this particular type of robots, although the scientific community seems to have already taken counter measures as later projects have focused less on this particular type of robots.

The method of searching for literature based on relevant scientific projects in this area resulted in a considerably larger literature base as compared with a classic search of databases since the proper selection of keywords (both by authors who link their publications to certain keywords and the reviewer who searches for them) does not play a crucial role.

3 Initial evaluation framework

The evaluation framework described in this dissertation was developed over the course of six years and was validated within two research projects on three sets of prototypes. This chapter describes the initial definition of the framework, which consists of research domains, evaluation factors and methods to target them holistically. This basic framework was implemented and empirically tested within two scientific studies with a SAR for COPD patients. The first two evaluations (E1 and E2) are described in chapter 4. The framework was later refined to fit the needs of another research prototype for physiotherapy. The results are presented in chapter 5. The evaluation framework is finalized by providing information about its potentials and issues in chapter 6 and chapter 7.

The initial evaluation framework was derived from the concept of user-centred design (UCD), the principles behind living-lab research and development, recommendations of HCI experts, literature on evaluation methods – in particular usability evaluation and HRI evaluation – as well as own experiences with the implementation of the UCD framework in AAL projects and by the comparison of standards for the larger and more complex clinical trials.

3.1 User-centred design as base for the evaluation framework

Within the field of AAL, UCD is established as quasi-standard for the design, development and evaluation of assistive technologies to ensure that user needs and requirements are in focus and thereby minimize the risk of technologically driven developments. This process is accepted to an extent that funding agencies, such as the Austrian “Research Promotion Agency” within their “Benefit program”,³⁷ make its adherence a requirement for a positive assessment of research-proposals. The key principle of the UCD approach is to optimize a prototype or product in a way that fits the targeted users’ needs instead of forcing users to adopt a technology by adapting themselves.

Given that the research goal on SARs within the field of AAL is to support user groups that are commonly described as having little affinity with new technologies, and covering the users’ needs is considered by the author as more important than

³⁷ The Austrian „Research Promotion Agency“ can be found at <https://www.ffg.at>. The „Benefit program“ is a specific funding scheme and can be found at: <https://www.ffg.at/benefit>

technological development, it makes sense to use this process as a base for SAR development despite it being not specifically designed for the unique aspects and particular requirements of SARs, as mentioned in chapter 2.1. This is acknowledged also by other authors such as Green et al. [Green2000] who have applied the UCD approach to the development of mobile robots.

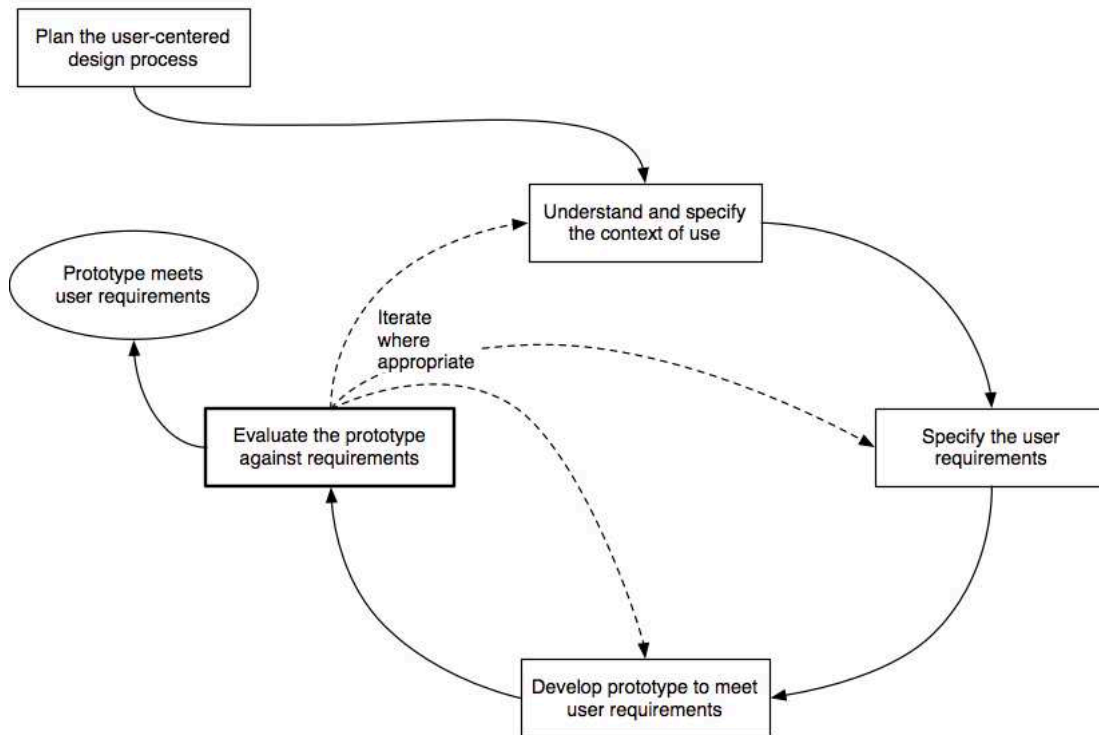


Figure 6: User-centred design process, adapted from [HCIinternational2011].

The UCD process is iterative as it applies the results of an evaluation phase to inform the design of a follow-up prototype until the user requirements are met according to the used evaluation metrics. Depending on the evaluation results, the UCD process can be iterated partially or as a whole.

UCD covers the whole design, development and evaluation of a technology. Within this dissertation, only the aspects of evaluation were investigated, whereas the other phases of the process and methods linked to those, such as methods used mainly for the purposes of design or development, are considered out of scope. Three cycles of the UCD process were used to derive the conclusions of this dissertation.

3.2 The living lab approach

Living labs (LL) can be viewed as an environment, a methodology, an innovation process or a combination of these [Kareborn2009]. The initial focus of LLs was to allow testing of new technologies together with users in real-world contexts. Since technology design and evaluation in practice play hand-in-hand, the concept was opened up to include concepts from participatory design. According to the most recent definition of ENOLL (European Network of Living Labs)³⁸ an LL is “a user-centred, open innovation ecosystem based on a systematic user co-creation approach integrating research and innovation processes in real-life communities and settings” [ENOLL2016].

Since LLs put the user at the centre of innovation, they share the goals of user-centred design, -development and -evaluation and fit well with the UCD approach to develop ICT solutions or concepts that fit the specific needs of particular target groups such as older users.

Key components of a LL are (modified from Kåreborn et al. [Kareborn2009]):

- LL participants (users) who are willing to take part in technology development. Users are composed of all subjects that might in the future use a particular technology including end-users, secondary users that support end-users and tertiary users that support the technology, its marketing or secondary users.
- The LL environment in which users interact with the technology and reflect on the planned usage scenarios.
- LL technologies and infrastructure that allows stakeholders within the LL to create new ideas and solutions.
- LL methodologies that are used for the innovation process, for example, research methods for the design or evaluation of new technology or organizational methods supporting research and innovation.
- LL companies, research institutions and all other partners that work together within the LL, sharing their knowledge and experiences.

The Living Lab Schwechat

The “Living Lab Schwechat” was initiated in 2006 in order to support the research on active and assisted living and in particular of assistive technologies that support older

³⁸ <http://www.openlivinglabs.eu>

users within Schwechat. The initiating partners were the city of Schwechat (with its senior citizen centre) and the non-profit research company *CEIT Raltec*.

The strategic goals of the city were to enhance the residents' quality of life (in particular those of senior residents), to support research for a good cause and to attract new (high-tech) companies and new residents. The goals of the research company *CEIT Raltec* were to conduct research on possible future assistive technologies, develop these together with partnering companies and to develop new methodologies for user research that were mainly needed to enable the research work within the LL context. *CEIT Raltec* focused on new technologies and their application potential for older people and society. Consequently, the research domains of smart homes, smart wearables and user research methods including ethics in user research were established. One of *CEIT Raltec's* main goals was to create technologies that allow users to stay in their own home for longer and more independently, and thereby reduce the overall cost of care and increase the quality of life of seniors.

The research work undertaken within the "Living Lab Schwechat", in which most of the trials within this dissertation were conducted, was based on the following key principles:

Cooperation: the Living Lab Schwechat" at its core is a network between a research organisation, user organisations and companies. This network is essential, as these three components are mandatory in order to perform multi-disciplinary research that targets the market and the users' demands.

The concept of a LL was chosen because the partnership with the local senior citizen centre should enable easy access to the target groups, not limited to seniors living there, but also care personnel, therapeutic experts and care management. Although it was planned to involve additional business partners for technology development and marketing, it proved to be difficult to recruit them as the field of active and assisted living is a difficult market to target and hence, only few potential business partners were available. Instead, companies were involved via funded research projects to ensure the viability of targeted solutions for a later commercialization.

Although the concept of the "Living Lab Schwechat" saw the whole city as part of the LL, only a few partners were actively involved. Aside of the initiating partners of *CEIT Raltec* and the city of Schwechat, during this dissertation the LL was composed of:

- An advisory board of senior citizens called “Seniorenbeirat”, which represents the interests of all seniors in Schwechat. This advisory board was often used as a multiplier during the recruiting phase of projects and helped to get access to participants from the primary target group.
- A cooperation platform that consisted of members of the senior citizen centre, mobile care institutions such as the local Red Cross, and *CEIT Raltec*.
- Small and medium enterprises were loosely involved via a local business-network called “Wirtschaftsplattform Schwechat”.

Realism: As one of the core principles of LL research and development (R&D) is that it is undertaken in a real-life context, including realistic settings, with real users and in real environments.

Realism was brought into the LL by generating ideas either directly by the user partners or at least validating technology-driven ideas from the beginning by user partners. Further, the design and evaluation methods employed were user focused and studies took place either at the users’ own premises or within two rooms of the senior citizen centre that were reserved for LL purposes, and hence in a setting that is realistic for systems that target an institutional care or sheltered housing setting. Most results presented in this dissertation were gathered within these rooms (see also Figure 12).

Focused on users: To avoid technology-driven R&D processes and developments that miss the users’ and markets’ demands, the focus has to stay on the users’ needs. Users need to be motivated to get involved in the design, development and evaluation processes and thereby enhance the acceptability of developed solutions. The LL approach also sees participants not just in the narrow sense as being an object for R&D activities, but it conceives human beings, citizens and the civic society as the source of innovation [Eriksson2005].

Within the “Living Lab Schwechat”, the user focus was guaranteed by ensuring users and stakeholders with a sound knowledge about the targeted user groups (such as companies for mobile nursing services and the senior citizen centre) were already involved in the idea-generation process and during all design and evaluation phases. Additionally, the methodological approach applied was derived from the principles of user-centred design, as described earlier.

State-of-the-art methodologies: It is important to constantly update the implemented methodologies either with own knowledge, which is gained during conducting studies, with literature or from knowledge-transfer from scientific partners.

A methodology for user involvement within the “Living Lab Schwechat” was developed during the course of many research projects empirically based on the UCD process and the LL ideology. A strategic partnership with the TU Wien Institute for Design and Assessment of Technology helped to develop the methods and stay informed about the state-of-the-art. In particular, ethical background and know-how on ethically sound user involvement was brought in by this partner.

3.3 Evaluation of robotic solutions – a definition

In the context of the UCD framework, an evaluation is a method to drive and refine the design of a prototype or product. Within HCI, it is mostly used to evaluate the perceived utility and value of software and in particular user interfaces such as websites. According to UCD, evaluation activities should be conducted together with users, beginning in the early phases of development. Early activities include evaluation of the concept and prototype design with methods such as mock-ups, cognitive walk-throughs or screen sketches.

The most common type of evaluation is the “usability evaluation” [Nielsen1994], which is used to assess whether a system applies to the Standard “ISO (International Organization for Standardization) 9241-11:1998”. Such usability evaluation is also a central part of the evaluation of a SAR but due to the complex nature of SARs, not sufficient to qualify the design of a SAR. Instead an evaluation of the user experience including social, psychological and impact factors is essential as those have a clear influence on the user’s willingness to adopt SAR solutions.

Evaluations are similar to experiments and hence often confused, despite certain important differences (modified from Helen Purchase [Purchase2012]):

- Evaluations are exploratory tests that aim to show that an idea works in practice (proof of concept). The outcome of an evaluation is a list of recommendations for future designs (formative evaluation) or an approval of a concept (summative evaluation).
- Evaluations are more exploratory and less constrained than experiments. For example, tasks given to the user are mostly designed to let the user perceive how

a system works without the need to measure the performance of a system. As a result, tasks in evaluations are longer and more complex compared to concise short tasks within an experiment that typically also have a clear correct answer.

- Evaluations do not depend on alternative conditions. Typical results are participants' opinions on a certain prototype or the analysis of user behaviour during interaction with a system.
- Results of evaluations are often qualitative in nature as compared to experiments, which are typically quantitative.

Despite these obvious differences between evaluations and experiments, the evaluations undertaken within this dissertation were prepared and conducted in the same scientific rigor and manner as scientific experiments, including the use of predefined research questions, pre-/post-test designs, statistical analysis of quantitative results and, in the cases of integrated experiments, also regarding the use of alternative conditions.

Given the large investment of resources to perform an evaluation with real users in a real-life context, we also inserted formal comparative experiments into the evaluation process in order to gain a higher validity of results. Hence, this methodological framework for the evaluation of SARs is an intentional combination of evaluations and experiments that can be used to study particular aspects in detail.

Within this dissertation, an *evaluation* shares three goals that are defined as:

- 1) The engineering goal: a validation of a prototype against a set of predefined metrics, leading to a proof of concept.
- 2) The design goal to inspire users and researchers: a method to gain information on how to improve future designs in the sense of a formative evaluation.
- 3) The social science goal of understanding users in real-life settings: a method to gain results on the effects of technology in the sense of a summative evaluation without the direct goal to implement the results into the next generation of prototypes but to enhance the general knowledge on potential impacts of SAR technology.

3.4 Principles of a SAR evaluation within a Living Lab environment

According to Baxter et al., the following basic principles need to be adhered to when conducting user evaluations [Baxter2015].

The participants of the evaluation should be representative of the target group and not be recruited based on the ease of recruitment considering cost, time and resource efficiency (convenience sampling). Here, a LL that involves user-partners can be helpful as these user partners can support the selection of potential trial participants based on the population's characteristics under the view of care experts (quota sampling). Also given that user partners are already taking part in the study, a large set of potential users exists to choose from. Still, it is clearly difficult to recruit users that are critical of the solution to be tested because those users, due to their criticism towards technology, are often not willing to take part in evaluations of technologies. Hence it is important to take the technological affinity of the user group into consideration when evaluating the results to diminish a potential acceptance bias.

Within an evaluation, use cases and scenarios are presented to users to let them understand the technological solutions. Often tasks are given to the users that let them explore the system. These tasks presented to the participants should not be comprised of the tasks that the system can easily handle, or is good at, but of tasks that are most relevant to the user. Now in practice this proves to be difficult as SARs, due to their multi-purpose capabilities, could serve users in a very wide array of tasks. Additionally, the relevance to the user depends on the user-specific preferences and needs. From a technological point of view, it does not make sense to develop usage scenarios which are known to be technically not feasible. Hence, tasks to be evaluated should be comprised of all developed tasks, which in the first place were selected from a list of tasks based on their technical feasibility and their potential to solve the users' needs.

The language and nonverbal cues used by researchers should be neutral and not lead or guide the participants. Given the extremely high efforts needed to prepare and conduct evaluations with SARs, this point is of very high importance and needs to be ensured by a detailed preparation including training of the researchers, pre-defined workflows and protocols that ensure the necessary amount of scientific rigor. Within the trials conducted during this dissertation, training sessions with colleagues or students were conducted to train the researchers. Guidelines were also used that partly gave the exact text to be spoken by the researchers to avoid researcher biases between trials and also

between trial sites. Additionally, the researchers that conducted the user evaluations did not take part in the development to avoid an unintentional positive influence on the participant.

The analysis needs to be based on the actual data, not on what the researchers interpret from it. The easiest way to realize this basic principle is that people who were not present during the evaluation undertake the analysis. However, during trials conducted within this dissertation, it was found that this principle does not hold universally. In several occasions, users provided data that contradicted their beliefs or intentions. This became evident during trials, for example, when they gave their intention vocally but crossed a different field on a questionnaire. In such cases, individual solutions had to be found between the participating researchers that ranged from omitting this question during the analysis to interpreting the answer in the sense as intended by the participant. As another measure to avoid this issue within this dissertation, the qualitative data is presented in direct quotes prior to analysis.

Further ethical considerations play a strong role when designing a SAR evaluation. On one side, good practices in user involvement have to be adhered to. This includes the timely information of the user prior to the trials within a meeting with the researchers and later, right before the trials, by clarifying the main goals of the trials and the participants' role within the evaluation. It also includes the appropriate language and interaction with the participant and choosing a test setting that makes the participant feel comfortable during the trials. Despite the short-term nature of LL trials, exit scenarios were developed due to the social impacts of experiences with SARs that may lead to a form of social bonding. Informed consent documents had to be signed by the participant and in particular, the use of recording technology had to be clear to the participant. Explaining how the data was gathered within the trials and used during and after the research project is another key point within this document.

Given the involvement of vulnerable user groups, ethical considerations also have to be undertaken regarding the way seniors and patients can be involved in user evaluations and whether they can be involved at all. Inviting such user groups to a LL setting can pose a real challenge to them as they have to structure their day to make this interview fit into their time plan. The journey to the LL alone might be challenging to them, let alone the approximately two hours of testing and the way back. For this reason, patients with higher grades of COPD could not be involved within the LL evaluations in Schwechat as medical support could not be arranged and the risk introduced to the

persons' lives was considered too high. This, of course, limits the representativeness of the test group, which has to be considered when analysing the results, and limits the value of gained results. Aside of this dissertation, the developed methodology was used with COPD patients within trials in Tel Aviv, Israel within a care facility that housed COPD patients and was hence also able to care for them during such trials.

The LL environment provides a realistic setting to give users the feeling of the prototype's performance under real-life conditions. At the same time, the LL is a stable and relatively secure place to operate prototypes because the environmental conditions can be controlled. This allows keeping light conditions, placement of furniture and sound conditions static and gives the users a consistent experience. Furthermore, the performance of the prototype can be increased as not only the environmental conditions but also the room-setup are known beforehand. Hence, the LL approach is well suited to provide a consistent user experience and gather the users' opinions and ideas on the prototypes and how they could be enhanced. On the counter side, the approach is not well suited to estimate the functional performance of the prototype because of the limited ecological validity. The LL is devoid of distractions such as kids making noise, pets or other people who would influence the performance of the system. Typically, within the LL, preconditions that are unrealistic for a user's home such as high-speed reliable internet connection, furniture that is well suited and chosen for the test, and room layouts that facilitate testing are used. Also, the static environmental conditions that allow the demonstration of early prototypes bias the results of performance testing positively. Such a LL is well suited to demonstrate realistic test scenarios and allows users to experience the prototype in a realistic setting, but not to test the prototype's readiness for later real-life integration.

3.5 Development of the user research methodology

Within this dissertation, user research methods had to be developed or adapted from existing frameworks as few methods could be found to evaluate SARs on their performance, acceptance and holistic impacts and none are widely accepted within the community.

3.5.1 Problem description

General issues of assessing the suitability of SAR prototypes to serve older users were already described within the problem statement in chapter 1.2. In summary, in order to let users give their opinion on a SAR concept, it is necessary to provide them with a

hands-on experience with a real prototype because the mere look of a SAR does not reveal its functionality and the social impacts become visible only when demonstrating the real artefact. This hands-on experience has to take place in a real-life context to enable users to imagine the intended real-use at home. Since user feedback is required early in the design and development process, evaluations have to take place with early prototypes that cannot show a product-grade stability and safety. Further, the evaluation has to involve participants that are vulnerable due to their age and age-related deficiencies including chronic diseases. The two latter points in combination negatively impact the viability of real-life trials at users' homes.

3.5.2 Goals and requirements

The top-level goals of method development within this dissertation are to adapt and integrate existing methods, in particular from the research fields of HCI, sociology and HRI, and thereby generate methodologies that fit for the evaluation of SARs in real-life contexts. The methods were selected based on the research questions and goals described in chapter 1.3. Additionally, the following accompanying requirements were defined:

1. The methods should allow a holistic evaluation of performance, acceptance and potential impacts of SAR technologies as those dimensions are strongly interlinked.
2. The methods developed should stand-up to any technical stability issues of the prototype to allow the validation of early developments and enhance the reproducibility of trials.
3. The methods should allow integrating subjective and objective data using qualitative and quantitative research methods within a mixed research model to cross-validate findings between different methods and data modalities and thereby enhance the scientific evidence of results.
4. The methods should allow the continuous control over the robot during its operation for safety reasons.
5. The methods should facilitate long-term investigations to gain information on possible biasing effects of initial excitement, which are known to be powerful influencing factors due to the novel nature of the introduced technology [Heerink2009].

6. The methods should be well defined and described to contribute to the general aim of creating a set of methodologies and tools that can be used to allow other researchers to replicate and validate their user studies [Dautenhahn2007a]

3.5.3 Development sequence

In order to develop the methodology of a SAR-user evaluation, several prerequisites have to be met. First the research problem needs to be defined and the research goals need to be clear and agreed on between research partners. Evaluation domains can be designed once the aims are clear. Within this dissertation, the evaluation domains were always comprised of technical performance, acceptance and impacts of the implementation of the respective solution. Detailed research questions have to be developed to guide the method selection. In the case of evaluations with incorporated experiments, testable hypotheses also need to be described that serve as predictions that can be tested against. The accompanying requirements need to be clear to be able to select the right methods from a list of potentially applicable methods. This list is generated based on an analysis of the state-of-the-art and by involving experts on the research topic. The key methods are selected based on the aforementioned requirements (which implies the knowledge of the prototype to be tested) and the list of researched potentially fitting research methods. Finally, the evaluation methodology can be designed and includes a detailed flow of events during the test and methods to analyse the results. Figure 7 summarizes this step-by-step guide within a flow chart.

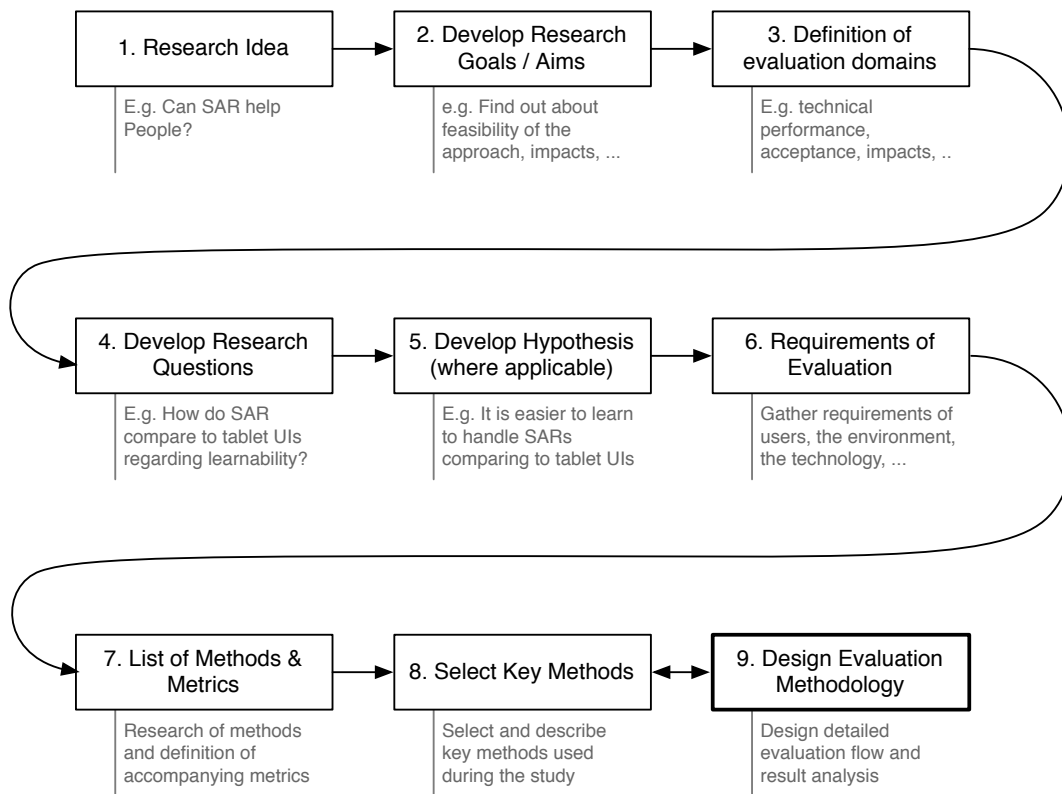


Figure 7: Overview on the development of an evaluation methodology.

It is important to mention that this flow of events can only be successful if the system to be tested is at least known in detail by step eight, to select the appropriate methods. Steps one to eight can be rather generic for SAR systems, with similar evaluation aims and can hence be well prepared in advance. Whereas the first four steps consider mainly WHAT is to be tested, steps five to nine target the question HOW the evaluation should be undertaken.

Within this dissertation, the described steps were conducted and are detailed in the following chapters:

- Step 1. The research ideas behind the conducted research are described in chapter 1.1 – Motivation and background
- Step 2. Research goals and aims are described in chapter 1.3 – Goals and research questions.
- Step 3. The evaluation domains are described in chapter 1.3 – Goals and research questions and further specified in chapter 3.6 – Evaluation domains and factors.

- Step 4. The developed research questions are broadly introduced in chapter 1.3 – Goals and research questions, and further detailed during the presentation of results in chapters 4.3 and 6.
- Step 5. Hypotheses are specific to the individual implementations of the research methodology and are presented in the result chapters.
- Step 6. Requirements of the evaluation are presented in chapter 3.5.2 – Goals and requirements.
- Step 7. A list of methods is presented in the state-of-the-art section 2.2.
- Step 8. Selected key methods that were used during the presented evaluations are detailed in chapter 3.10 – Key user research methods.
- Step 9. The evaluation methodology is generally described within the evaluation framework in chapter 3.6 and further detailed and updated in the chapters describing the implementation of the methodology (chapter 4.2 and chapter 5.4).

3.6 Evaluation domains and factors

A holistic model for the evaluation of SARs with older people in a close to real-life context was developed. The model is composed of three major evaluation domains (performance, acceptance, impacts) that can be split into several subdomains and influencing factors (see Figure 8). The model also considers three main user groups (primary-, secondary- and tertiary users) that are detailed in section 3.7 and describes a set of qualitative and quantitative methods for conducting the study and to analyse the results, as detailed in section 3.10.

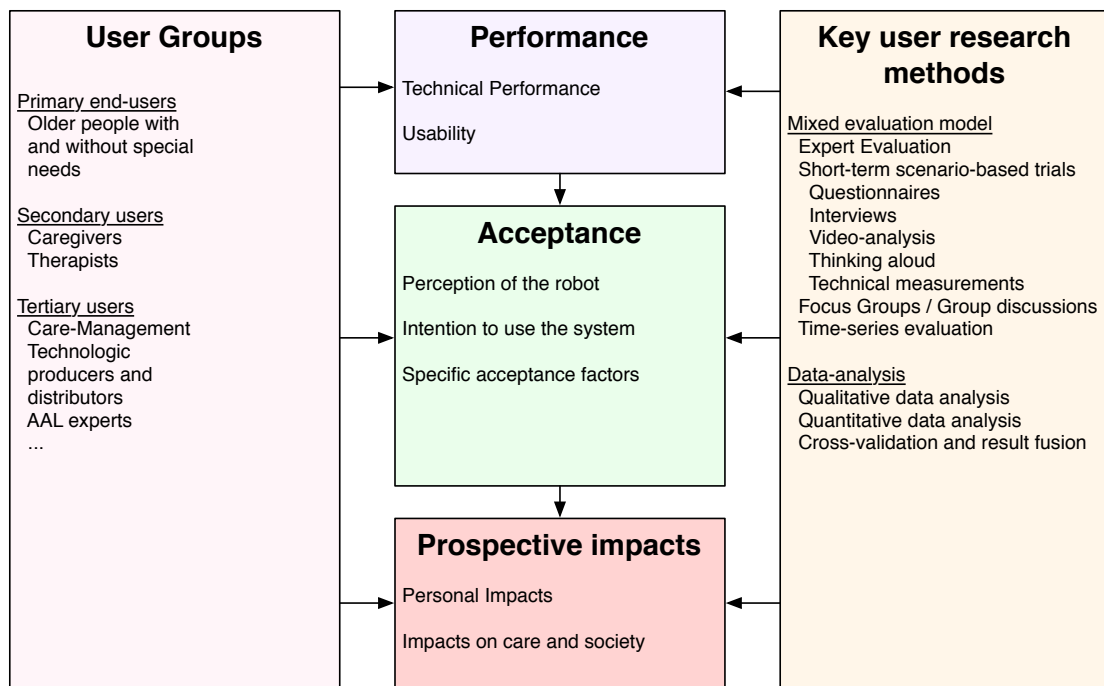


Figure 8: Basic holistic evaluation model.

3.6.1 The performance domain

This domain considers the technical performance of the used prototype and usability factors. A positive evaluation of this domain can be seen as a pre-requisite for the assessment of the following domains as acceptance is strongly influenced negatively by a lack of performance and impacts can neither be generated nor measured with a non-performing system.

The following factors are considered within the technical performance subdomain:

- Functionality of the prototype

The functionality according to the specifications ensures that the users' needs are targeted in the way the system was designed. Missing or wrong functionality influences the perceived usefulness and thereby the acceptance of the system, or too many features might overwhelm users and thereby reduce the usability of the system. The functionality is measured against the specification of the prototype. A checklist-based method is used for measurement where the delivered functionality of the prototype is compared against the specification.

- Stability of the prototype

Stability refers to the error rate of the system within a real-life situation. In contrast to the functionality, it does not measure whether a function is present and working, but whether this function is performing within a real-life context and over time. It is typically measured by means of several iterations of black-box testing. The number of iterations needed depends on the tested functionality, as more critical functionalities need to be tested more intensively. Another aspect of this factor is the technical performance of the system over time. The stability strongly influences the usability and thereby the user experience and overall acceptance of the prototype.

The following factors were considered for the usability-subdomain

- Learnability

Learnability refers to the effort and the duration spent to learn how to interact with the system. Given that the target group has a high prevalence of memory disorders, good learnability is crucial to enhance the usability, acceptance and thereby the likelihood of adoption of the system in the future. Learnability is evaluated by measuring the task performance during the first time of use in comparison with later use.

- Effectiveness

Effectiveness refers to the question as to whether the system solves a real user need and whether users can use it to effectively handle a problem. This item can be assessed by judging the outcome of users' interactions with the prototype [Frokjar2000].

- Efficiency

Efficiency is based on the effectiveness and takes into account the time and resources needed to achieve the desired outcomes. A typical quantitative indicator would be the time needed to complete a task [Frokjar2000].

- Flexibility

Flexibility describes the optionality of different usage strategies to achieve the same task. For example, SARs can facilitate a variety of communication channels; the same outcome could therefore be achieved by different input modalities. Such flexibility is also needed to allow users with deficiencies to operate the system as they might not be able to understand all the presented modalities of the SAR system.

3.6.2 The acceptance domain

“Acceptance” is an umbrella term for factors that influence the user’s willingness to use a system in the future. Several acceptance models have been developed so far, ranging from general and generic models that can be used for any kind of technology, such as the UTAUT (unified theory of acceptance and use of technology) or the TAM (technology acceptance model), to specific models such as Heerink’s Almere model for HRI [Heerink2010].

The most common tool for estimating the intended future use and the related usage behaviour is the TAM, which was first introduced by Davis et al. [Davis1989]. It is based on the measurement of perceived usefulness (PU) and perceived ease of use (PEOU). Figure 9 gives an overview of the construct interrelations.

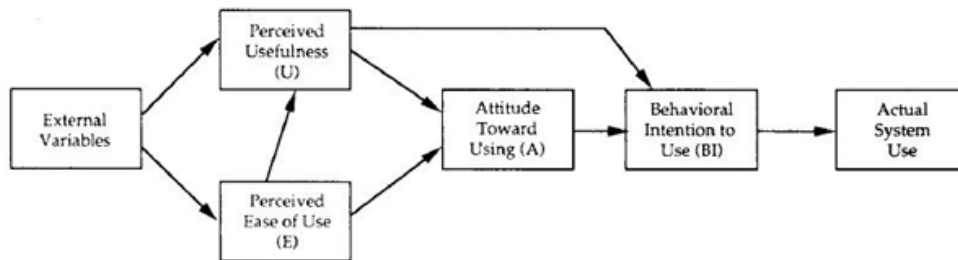


Figure 9: TAM model [Davis1989].

TAM2, TAM3 and the UTAUT modell (Venkatesh 2000, 2003 and 2008) are consistent enhancements of the TAM model and also introduce social-influencing factors and moderating factors like age, gender and experience. The inter-dependencies are shown in Figure 10. The TAM as well as the UTAUT models have been used in related projects [Heerink2010] for evaluating the acceptance of robotic solutions.

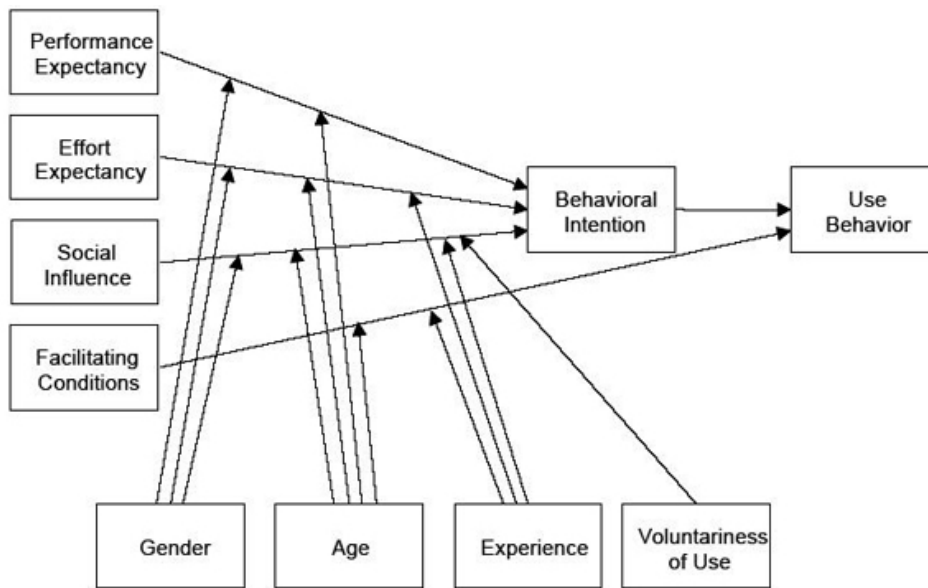


Figure 10: UTAUT model [Heerink2008].

A HRI-specific model which was developed on the basis of TAM and UTAUT is the “ALMERE” model by Heerink [Heerink2010]. Its purpose is to evaluate the acceptance of a robotic solution together with older people. It combines several constructs based on TAM, TAM2 and the UTAUT models and adds relevant constructs related to the needs of the target group of older adults and vulnerable users. In this dissertation, the evaluation concerning several acceptance-relevant factors and an intended future use were based on this model. The interrelations of the constructs are shown in Figure 11.

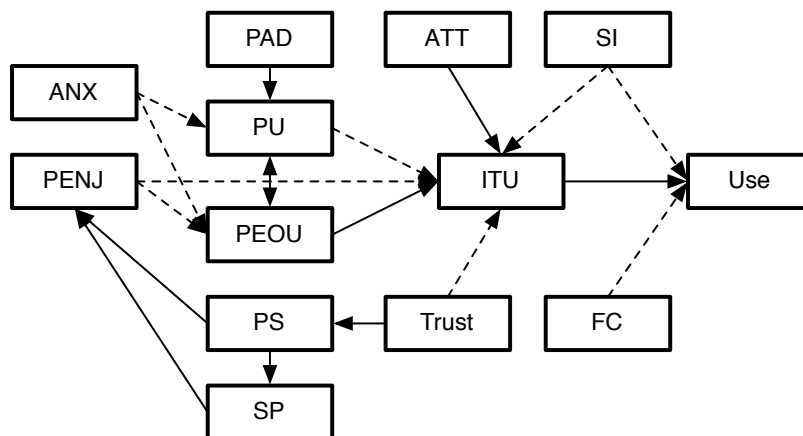


Figure 11: Almere Model by Heerink et al. [Heerink2010].

The following factors were derived from the presented acceptance models and considered for the acceptance domain:

Perception of the robot

The perception of the robot as a social entity by the participants was considered to be most relevant in addition to classical acceptance models. The perception of the robot can be assessed using the Godspeed questionnaire [Bartneck2008]. The Godspeed questionnaire was proposed by Christoph Bartneck as a measurement instrument to evaluate the users' perception of robots. The robot is the only visible component of the system for interaction aside of the comparative touch-screen and hence has a strong influence on the perception and thereby acceptance of the whole system and approach. The questionnaire is free and generally used by HRI researchers, which makes it a useful tool to compare results with other similar approaches. The questionnaire ranks the user's opinion by using semantic differential scales composed of opposing statements such as "humanlike – machinelike", with five possible grades between them.

The users' perception of the robot is assessed based on the following five specific constructs:

- Anthropomorphism

Anthropomorphism gives information about the compliance to human-like characteristics. This construct is affected by the visual appearance of the used robot and by the functionality shown within the implemented scenarios. This construct gives information about the achievement in following the human-human model for interaction.

- Animacy

Animacy rates to what extent the robot appears to be a living, organic and interactive individual rather than an inanimate machine.

- Likability

Likability shows to what extent the robot appears to be likable, kind, pleasant and nice.

- Perceived intelligence

Perceived intelligence rates the competence, intelligence, responsibility and sensibility of the robot's appearance and actions. Results showed that extrovert robots and robots perceived as intelligent result in a higher intention to use the solution.

- Perceived safety

Perceived safety refers to whether the user feels safe in using the system or feels rather anxious or surprised when interacting with the system.

Specific acceptance factors

These acceptance factors are specific to the respective prototypes and depend on their functionality. The idea behind these factors is to not only evaluate the system as a whole but to also take a closer look at acceptance factors related to those functionalities. Users might accept a particular functionality well at the same time disliking another one. By means of these specific factors, the evolvement of the set of functionalities of the SAR should be driven.

3.6.3 The impacts domain

Personal impacts

- Quality of life

This factor measures whether the tested system has an influence on the quality of life (QoL) of users (primary and secondary). Typical methods are questionnaires on the QoL such as the QoLBREF [WHO1996] which needs to be applied within a pre-/post-design with substantial exposure to the tested systems. Other less time consuming methods include qualitative interviews of participants; an estimation can also be given by deduction. (If a system solves an issue in one's life and is accepted and used, it should also enhance the QoL in this particular aspect).

- Work conditions

This factor measures the impact of the tested solution on the work conditions of secondary users (care personnel). One of the main ideas of assistive robotics is to enhance the work conditions of carers. Here, assistive robotics promises to reduce the amount of dull or dirty work and thereby prolong the time available for personal care with patients. This factor also measures the burden reduced or introduced by assistive robotic systems.

Societal impacts

- Care and the care system

Factors related to care and the care system rate the quality of care provided by secondary users and the cost of care in comparison with the cost of the technological

solution provided. Factors depend on the specific solution and may include the number of days spent in formal care institutions versus at-home care, impacts on the employment of carers, availability of the developed solutions to financially poor user groups and questions of sustainability and resources. A recent study provided a detailed set of factors to be assessed within this category in [EvAAUation2017].

3.7 The user groups

The UCD approach foresees the inclusion of the most relevant user groups as stakeholders of the future product. These user groups can be split based on their engagement with the robotic system and how they profit from its use. Within this dissertation, three user groups were defined:

1. Primary users

Primary users are people that profit directly from the use of the technology and are mainly addressed during the creation process. Within this dissertation, primary users are older people with certain special needs. The term “older users” is not clearly defined in general and often refers to the age of the users (such as 65 years and older). Within this dissertation, inclusion and exclusion criteria based on age, life situation, physical and cognitive abilities, and technological affinity were used to define the primary target group. As the criteria differ between the conducted studies, these are reported in detail within the study-specific methodology sections.

2. Secondary users

Secondary users are those who again profit directly from the use of the system and have regular contact with the system but support the primary users in their activities with the system. In all studies within this dissertation, secondary users were caregivers and therapists including informal caregivers such as relatives, friends or neighbours, formal caregivers such as social workers in home care and carers within an institutional setting, and therapists such as physiotherapists or occupational therapists.

3. Tertiary users

Tertiary users include all groups that profit indirectly but do not interact directly with the system. This user group is very diverse and includes:

- Producers and distributors of system parts
- Management of formal care institutions in which the system is used

- Social insurance institutions
- Politics who might profit in case the system has a wider impact on the social system

Given the diversity within this group, it becomes clear that not all sub-groups can be involved within a research project. Within this dissertation, participants of the management of formal care institutions were included as they were considered to be most relevant given the premature technological state of the tested solutions.

3.8 The test sites

All evaluations were undertaken within the “Living Lab Schwechat”. Most user participation including the short-term scenario-based user trials took place within a room that was provided by the Schwechat senior citizen centre and is depicted in Figure 12. This room, with simple furniture and a rectangular shape, was turned into a test environment by integrating technical components of the test system such as the robot, components that facilitate testing such as cameras and furniture that enhances the impression of a living room.

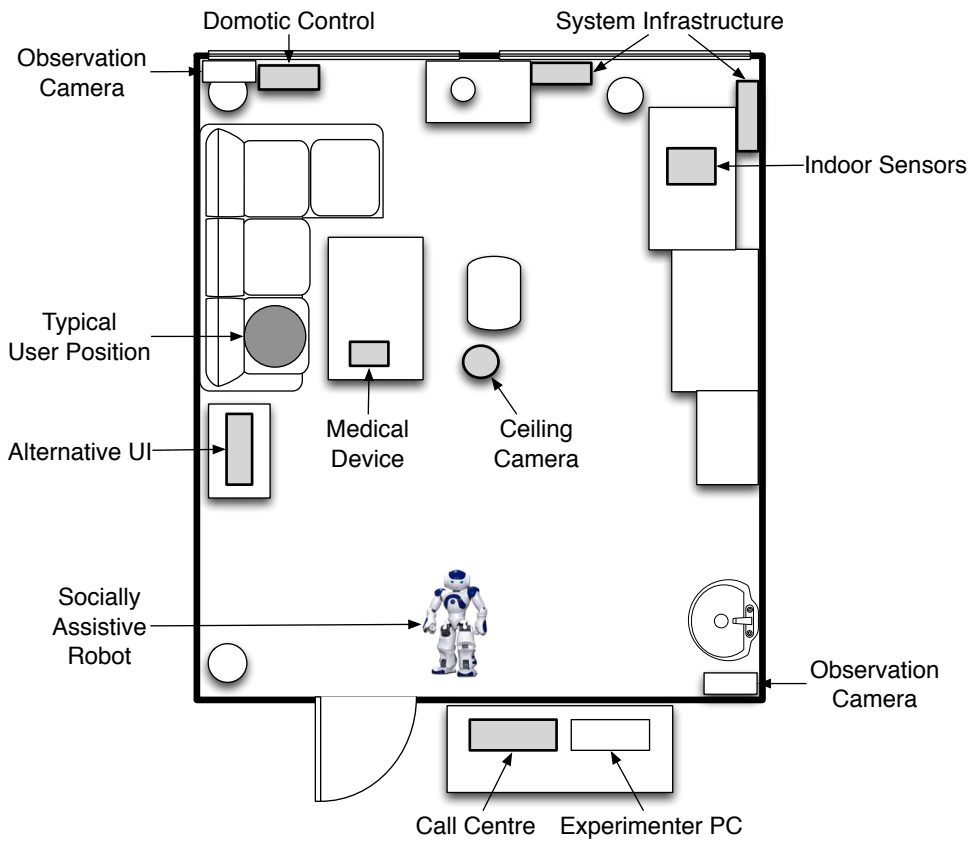


Figure 12: Example of a setup within the “Living Lab Schwechat”, previously published by the same author in [Torta2014].

The technical system components were hidden from the user in order to generate the feeling of sitting inside a living room, except from the SAR, which was prominently displayed (see Figure 13).



Figure 13: Impressions from the test environment.

Outside the room and hidden from the evaluation participant, a technical control room for two experimenters was set up who controlled the technical systems and were able to follow the scenarios inside the room visually and audibly. Figure 14 depicts a screenshot

of the user interface on the authors PC. On the left side two camera views are depicted to observe the situation. On the right side of the image, technical data is displayed. This includes another camera view from top down with augmented real-time feedback of the estimated location of the robot (green square), and the user (red square). Additionally a steering interface and technical output (black console) was used to control and monitor the system. All trials of E1 and E2 were recorded using screen recording of this interface.

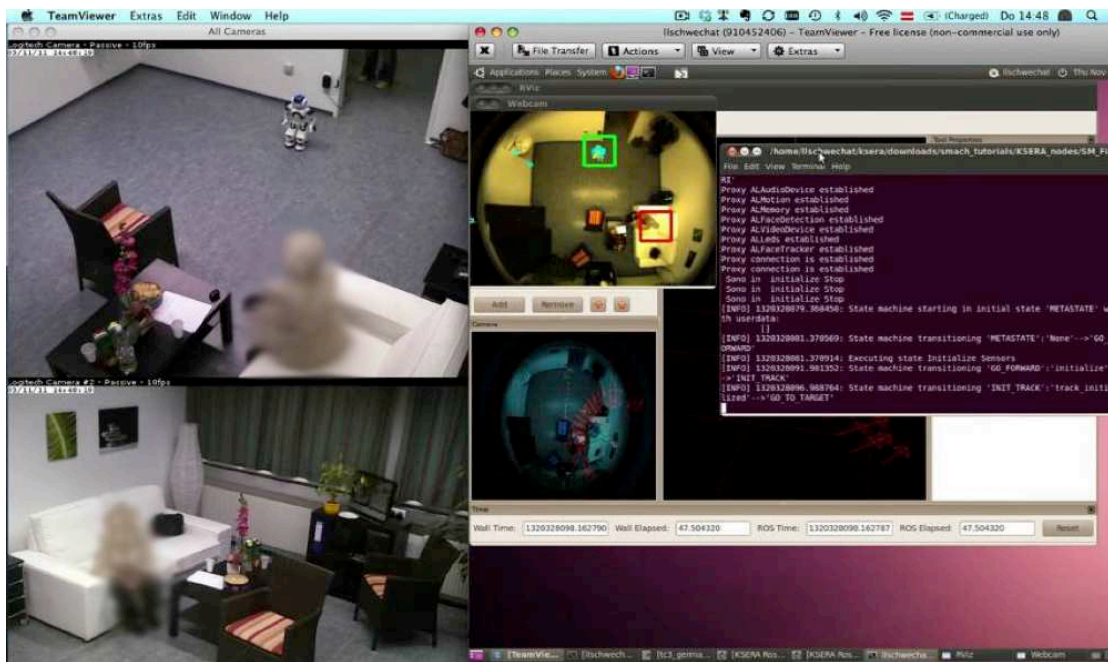


Figure 14: user interface at the technical control room showing two camera views (left side) and technical data on the location of robot and user as well as technical output data (right side).

The environmental conditions within the room were controlled as far as possible to increase the performance of technical subcomponents. The illumination was kept static by closing the curtains and relying on the room lighting; the noise from outside was kept minimal by closing the door. Furthermore, the furniture layout was arranged in a way that guided the user to stay at specific positions (such as the sofa or a chair) and to allow the robot to navigate the room, avoiding known issues such as collisions with certain furniture. This was undertaken to ensure the intended user experience, which was weighted higher than the need to assess the technical performance under realistic conditions.

In addition to the described room, the Schwechat senior citizen centre also hosts a gymnasium which was used within a pre-test to demonstrate the system to a larger group of users. See section 5.4.2 for a description of this specific test setup.

3.9 Key organizational methods

3.9.1 Trial Plan

The trial plan provides an initial framework for conducting the evaluation and is an important concept paper that allows the involved researchers to work towards the same goals. It is prepared by a panel of experts within the research project. It defines the general aims and specifies the overall objectives of the evaluation, deriving detailed research questions and hypotheses. It includes the methodology used to measure and analyse data.

It defines how the users are involved before, during and after the evaluation and gives recommendations on good practices in user involvement to allow similar procedures across trial sites. It defines which test environments are relevant and how they have to be described. As a main point, it defines the flow of events during the evaluation and gives a course timeline including milestones containing the dates the prototypes need to be ready for testing with users in order not to delay the evaluation phase.

The trial plan includes the metrics used to assess the prototype. It specifies all measurements and indicators used and how they should be applied. The trial plan includes:

- a. Detailed evaluation aims, research questions and hypothesis
- b. A short description of the technical prototype
- c. A course time plan
- d. General guidelines for the test environment
- e. Risk assessment
- f. Ethical, trust and privacy aspects

3.9.2 Test plan

The test plan is derived from the trial plan and can be seen as an instantiation of the trial plan given a concrete test setting and specific test object (technical prototype). Experts in user research and research practitioners prepare the document to ensure a practical applicability. It is important to use both concepts of the trial and test plans because of the timing inherent to the UCD. The test plan can only be written after the functionality

of the prototype is known in detail, which is typically the case after the prototype is finished which is only weeks before the trials have to start – making it impossible to spend the time necessary for good practice in research design. Hence the trial plan, which is created in parallel to the development phase and knowing only the specification of the prototype, is used to research all possibilities of appropriate methods and preselect the ones that fit best. Within the test plan, these methods only have to be further selected and specified for the current test setting and prototype, which saves time.

The test plan gives information about the detailed test setting and time plan for the conduction of the single experiments within the evaluation. It describes all experiments in detail and gives a detailed flow of experiments including all responsibilities of participating researchers. The test plan defines the exact communication protocol for communication with the test participants to ensure comparable experiments.

The test plan includes:

- a. Detailed description of the test setting
- b. Detailed description of test methods
- c. Detailed flow of tests including responsibilities of each researcher
- d. Detailed test cases
- e. Communication with the participants
- f. Detailed time plan as discussed with partners (adapted from trial plan)

3.9.3 Installation and support plan

The installation and support plan is used to verify the same setup of the trial sites with an integrated prototype across different test sites and gives the researchers a tool to take with them to the test as a memory aid on how to conduct the trials.

- a. Detailed information on how to install the system, integrate it at a test site and configure it, including a workflow of installation (order to install components, order to boot-up components)
- b. Description of necessary support actions in order to keep the system correctly performing during the conduction of trials (e.g. backup of trial data, replacement of batteries, routine checks on system functionality)
- c. Risk analysis and when to abort a test for technical reasons
- d. Contains detailed installation procedure and checklists that ensure the intended user experience (e.g. no blinking lights, sounds)

- e. Goal: Ensures comparability between trial sites

3.9.4 Global study design – time flow

The global study design is composed of a mix of timed evaluation phases and experiments that ensure the applicability of the prototypes for the evaluation, perform the evaluation and gather the results. Despite being different steps in the UCD, the development and evaluation phases are in practice strongly interleaved since evaluation activities are already used in parallel to development to verify the performance of the prototype for the user trials.

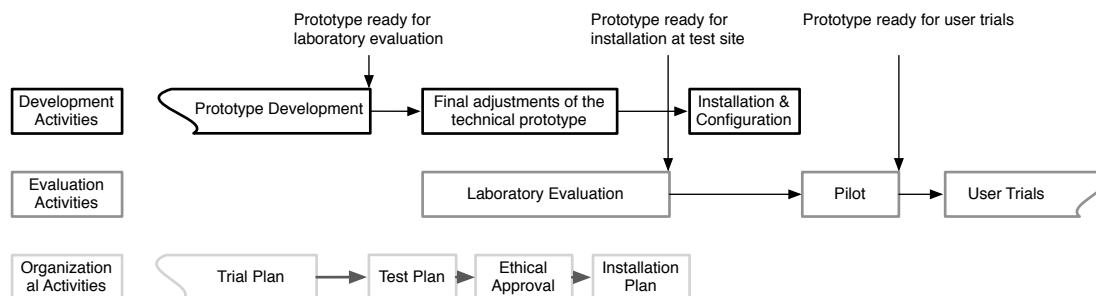


Figure 15: Typical evaluation time flow.

Figure 15 details the time flow of an evaluation, showing parallel activities for development, organization and the evaluation itself. As an important project-management tool, milestones are used to define the maturity of the technical prototype. Three maturity levels are used to define the technical functionality and robustness of the prototype and thereby ensure the applicability during the user trials.

1. Prototype ready for the laboratory evaluation

As soon as the technical prototype is ready for the laboratory evaluation, the test plan can be derived from the prepared trial plan. The first evaluation phase in the laboratory can be started. This phase is used to verify the prototype, mainly the functional and performance aspects. In case functional elements can only be evaluated with users present, technology-affine users (such as students or project members) are invited to participate in the laboratory evaluation.

2. Prototype ready for installation at the test site

Only if the prototype passed all test criteria in the laboratory, can it be considered functionally ready to be tested with real users. The system is ready to be installed at the

trial site and the installation plan can be finalized. After installation of the prototype, the system is again functionally verified by project members and together with at least one user of the target group to gain information on potential target-group specific issues.

3. Prototype ready for user trials

If the prototype also passed the pilot trials, it can be considered ready for the user trials.

3.10 Key user research methods

This chapter section gives an overview on the most-used research methods that fit the developed framework of research domains. Details such as how the methods were implemented and data was analysed are specific to the individual evaluation phases and are reported in the respective sections.

3.10.1 System pre-tests

Prior to conducting user trials, the technical system has to be evaluated towards its readiness for user trials regarding the intended functionality, performance (stability, reliability, usability) and safety. This evaluation can be conducted at the same test site as the later user trials (see chapter 3.8 – “The test sites”) and by inviting test users such as students or project members. The technical performance (stability, reliability and safety) is evaluated by means of black-box testing using pre-defined test cases. The usability is evaluated based on Nielsen’s Heuristics [Nielsen1990]. Results of the performance evaluation are communicated to the development team and fixes are implemented whenever possible. In case a fix may seem unrealistic due to the technical complexity or time needed, a Wizard of Oz approach [Kelley1984] can be taken.

3.10.2 Short-term scenario-based user trials (SSUT)

This method is considered the main method behind the undertaken evaluations of SARs. The method was derived by combining the LL approach with usability evaluation according to the user-centred design and augmenting it with the need for a holistic evaluation method due to the multi-modal concept of SARs and their known social influence on humans.

This method was developed while keeping in mind the requirements and goals described in chapter 3.5 – “Development of the user research methodology”. The basic goals were to evaluate performance, acceptance and impact factors by actively involving

older users and providing them with a hands-on experience with a running system and within a realistic context of use.

The core idea behind this method is to bring the user into a controlled real-life context in which a technology can be demonstrated and explored whilst result-generating methods can be implemented safely at the same time.

This controlled real-life context is generated by:

- a) Locating the test environment within the intended future application location, e.g. within this dissertation, in a user's home, a sheltered housing facility or a care centre.
- b) Furnishing the test environment to resemble the targeted environment for later application, e.g. within this dissertation, a user's flat or living room. This includes in particular the hiding of any test equipment that would not be present in a real-life evaluation.
- c) Providing the user with a temporal and situational real-life context by describing a user story that lets the user imagine the current time, setting and situation such as in a simple case: "Imagine you are sitting at your breakfast table, just finished eating and would like to call a friend."

Putting the user into a real-life context and into a realistic scenario empowers the test participant to imagine the real use of the demonstrated technology and hence to provide more accurate feedback as when compared to laboratory trials or experiments [Kareborn2009].

Once the real-life context was generated, the technology can be demonstrated to alter the usual flow of events that the participant knows. This way, the user becomes able to reflect on the benefits or drawbacks that this alternative brings to her or his life.

The reactions of the user to this alteration can be measured subjectively by asking about the user's opinion but also more objectively by observing the user and measuring reactions directly.

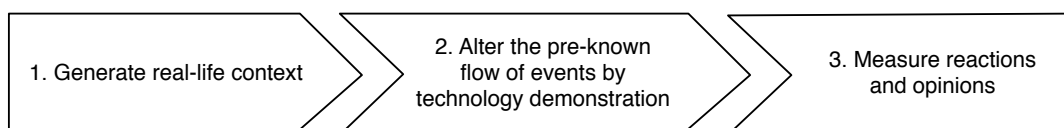


Figure 16: Principle method for result generation.

As such, this method serves as a container for more specific result-generating methods and enables different test designs, including the use of more experimental designs such as pre-/post-measurements and comparative test designs.

To generate the results before, during and after demonstrating the technology, various methods can be implemented; the following were used within this dissertation and are found to fit the principal approach:

Questionnaires

Questionnaires can be used to gain quantitative and qualitative results; in particular after the demonstration of the prototype. It seems important to capture the users' opinions directly after the demonstration of the prototype to avoid influencing the user and tapping the short-term memory. Questionnaires can also be used before the test to gather either general information about the user; e.g. such as the general technology acceptance, or as a base for comparison within a pre-/post-experimental setup.

Interviews

Qualitative interviews such as semi-structured interviews [Smith1995] can be used to gain qualitative results and generate new ideas and insights into possibly new acceptance and impact factors. Interviews can be conducted after technology interaction by an interviewer that enters the test environment (and thereby ends the phase of technology demonstration) or beforehand, to gain information about the expectations towards using the system.

Thinking aloud

The concurrent thinking-aloud method [Kuusela2000] is used to obtain feedback on the technology and its use directly during the interaction. The method in particular is useful as it gives the feedback during the context of use and thereby clarifies why the feedback is given. Additionally, the feedback is more accurate due to users not having to reflect on earlier experience but providing their opinion from their working memory. Before the technology demonstration, the user is asked to verbalize thoughts and findings during the demonstration. These thoughts are recorded and transcribed and can be analysed qualitatively.

Observation

Observation [Holzinger2005] can be used to obtain more objective results during the interaction of human and robot. By means of this method, results on the users'

behaviour and the technical performance of the system can be generated. To reduce the intrusiveness of the method, the user should not notice the observation (e.g. the researcher watches through a semi-transparent glass or via camera). However, due to ethical reasons, the participant has to be informed about video taking within the informed-consent document. In the sense of preserving the real-life context, cameras should not be easy to spot and should be well integrated into the room.

Within this dissertation, participants were always aware that they were being observed and also recorded, as this was part of the informed consent and due to ethical requirements, but care was taken that the observers always stayed in the background to blend-in with the setting. In most parts of the conducted SSUT trials, one of the experimenters was observing the scene from outside the room via camera or through the open entrance door.

Technical measurements

Technical measurements can be undertaken prior, after and during the demonstration to gather information about the technical performance of the used system within a close to real-life setting. The incorporation of technical-performance assessment into the user trials is important as several functionalities of SARs can only be assessed during interaction with a real user (e.g. face recognition, voice recognition, navigation algorithms). Technical measurements can either be automatically generated (e.g. logging) or made by researchers during observation of the scene. For that latter method, checklist-type sheets have to be prepared that guide the researchers in their assessment as the assessments have to be undertaken live and typically quickly. Alternatively, technical measurements can be performed retrospectively by video analysis if logging output of the system is recorded as well.

3.10.2.1 Positive aspects of the methodology

The SSUT method is targeted for situations where field trials would be technically not feasible, too expensive or time consuming. Additionally, field-trials do not necessarily provide better results as they still suffer from certain biases such as the Hawthorn effect, as found by Heylen et al. [Heylen2012], and are less flexible in terms of integration of methods – such as observation which is hard to realize over the long-term in private premises due to ethical constraints.

The described methodology tries to take the best of the two worlds: field trials and laboratory assessments and can be put in between them. The real-life context derived from field trials enables participants to anticipate a later real use and thereby strengthens the ecological validity of the method. The presented method comes close to the ideal to conduct trials in the same setting in which the technical support aid is later used. The users are able to gain a hands-on experience that is crucial since SARs are tangible objects that clearly have a social influence. In comparison, other methods such as video demonstration cannot transport the feeling of social presence to the user and hence cannot consider the very relevant social acceptance factors.

The method also enables a holistic evaluation of performance, acceptance and impact factors as it can flexibly incorporate result-generating methods that provide qualitative and quantitative data to investigate a large variety of research questions centred on these domains. Combining and fusing results from different methods on the same research question can enhance the scientific power of insights gained. Additionally, where quantitative results provide insights that can be easily visualized and compared (e.g. between trial iterations), qualitative methods can cover the details and the complexity of social interaction and provide insights into other potential factors that influence in particular the acceptance of SARs.

The fact that several research questions can be evaluated at the same time during the holistic evaluation has a positive effect on the budget as several research goals can be covered within the same trial.

Because the environmental conditions at the test site as well as the SAR system can be controlled, this method is robust against any lack of technical performance from the prototypes, and even early and partly unstable robotic solutions can be tested which is crucial for SAR evaluations, as shown in chapter 2. As a method to further enhance the technical performance of a prototype and thereby enable the test with end-users, the Wizard of Oz [Kelley1984] method can be implemented. By means of this method, parts or even the whole technical system can be controlled not by algorithms acting autonomously, but by a human in the loop. Here the goal is to simulate not-yet-functional parts of the system in a way that simulates the real behaviour of the system in the future. It has to be taken care that the so-generated user experience is not superior to that achievable by the current technologic state-of-the-art to avoid biasing results on acceptance of the technology [Riek2012]. Further, for obvious reasons, the technical performance cannot be evaluated for simulated parts of technology.

3.10.2.2 Inherent challenges

Given that the SSUT method is not taking place in real life but in a setup that tries to mimic real life, the ecological validity is lower than in a true real-life setting. The users are invited to another location instead of the researchers going to their premises which makes it harder for the user to imagine how the system would work in his / her home environment. Descriptive stories can be used to set the participants into the context of use at home to enhance the ecological validity in this setup.

The need for a long-term evaluation to gather information about the impacts on such technology in real-life is well known and not well supported by the SSUT method as it seems unfeasible to invite users to stay within the controlled setup for periods of time longer than several hours or days. Within this dissertation, an approach was taken to invite users weekly over a duration of six weeks to simulate a long-term trial and gain information on potential impacts and the change in acceptance over time. This approach gave new insights but, due to the necessary iterations, is time consuming for the researchers and users and also regularly reminds the users of the test situation, enhancing the observer effect. Still, it proved feasible to conduct and might be the only solution to gain long-term acceptance results with prototypes not ripe enough to be used in a real-life context.

The evaluation of technical performance is limited as in this aspect the real-life context is diminished by the fact that the test environment is controlled. To ensure a sound demonstration of the prototype to the user, depending on the robustness of the provided solution, the environment has to be adapted to the system which reduces the realism and constrains the scientific value of the generated results, except those from functionalities that are not influenced by the controlled conditions. To minimize this effect, researchers during analysis can take into account which and how conditions were controlled to estimate the technical performance in a real-life setting.

3.10.3 Group discussions

Group discussions such as focus groups [Morgan1997] were conducted with secondary users to gain their impression of the prototypes. In a discussion, session experts from a specific field of therapy (e.g. physiotherapy), nursing staff or care management took part. The functionality of the prototype was demonstrated to the participants using a real-life demo of the robot to provide them with insights about the functionality and behaviour of the SAR. If a human user was needed during the demonstration, one of the

participants was asked to play that role while the others observed the scenario. A discussion was started after the demonstration of technology based on a pre-defined discussion plan including the specified research questions. The group discussions were audio recorded and transcribed verbatim and notes were taken.

3.11 Key methods used for data analysis

In the scope of the conducted evaluation phases, we gathered quantitative and qualitative data using subjective and objective methods. This section describes the main techniques used when analysing this data within this dissertation.

Quantitative data.

To gather quantitative subjective data, construct-based questionnaires were used. Most of the quantitative data was gathered using a five- or seven-point Likert scale. Despite the long on-going discussion as to whether Likert scale data should be analysed on an ordinal scale or an interval scale [Sullivan2013], it is very common and also recommended to analyse data using their means and standard deviation at the interval measurement scale [Bone2012], [Sullivan2013].

Within this dissertation to analyse quantitative data from the Likert-scale questionnaires, a descriptive approach was undertaken using mean, maximum, minimum and standard deviation. Results are mostly represented using either bar-charts showing mean values and standard deviation or box-whisker-plots showing mean, 9th percentile as minimum bar, 91st percentile as maximum bar, 25 and 75 percentile as well as outliers to provide the reader with a sense about typical values and their distribution along the group of test participants. The author considers this descriptive approach as particularly valuable considering the mostly small number of participants and hence small sample sizes that often do not allow the implementation of statistical testing.

In the cases of incorporated experiments, either two-sided t-tests in simple cases of comparing two conditions were implemented, or, in case of multiple dependent variables, a multivariate analysis of variance (MANOVA) was used to test for significant effects before investigating these effects further using multiple analyses of variances (ANOVA) on the individual constructs .

Qualitative data.

Qualitative data was obtained by recording and transcribing interviews, focus groups or workshops with secondary and tertiary users and user comments that were given during the implementation of the SSUT method (thinking aloud).

The qualitative data was structured using a thematic-analysis approach [Braun2006]. Thematic analysis is a widely used method in social sciences, psychology and HCI to analyse and interpret qualitative data by identifying and reporting common themes. A theme within the data is understood as any aspect that seems important to the researcher. It can emerge from any number of occurrences within the analysed data, the common ground being to capture something important from the data in relation to the research aim [Braun2006].

In contrast to grounded theory, discourse analysis or narrative analysis, thematic analysis is less defined and implemented using differing procedures across the literature [Braun2006]. As thematic analysis is an explorative method, it can be implemented without prior definition of concise research questions.

Within this dissertation, a combination of deductive analysis and inductive analysis was undertaken as the coding structure was known beforehand and based on the research questions and acceptance factors but open for additions that could be found during the analysis. If additional themes were found during the process, these were reported as they indicate potential new acceptance factors.

The following process of thematic analysis was derived from [Braun2006] and adapted to fit the needs of the research presented, in particular the use of evaluation factors as research questions, and therefore the implementation of the process in a less explorative and more descriptive manner.

1. Familiarize with the gathered data, transcription of verbal data, going through the material and gain initial ideas about possible patterns.
2. Generate initial codes, having the evaluation factors in mind and sort codes based on common contents.
3. Link found codes with existing themes based on evaluation factors or, in case they seem not to fit nicely with existing themes, develop new themes by grouping codes.
4. Reviewing themes and check for internal homogeneity and external heterogeneity.

5. Choosing names for new themes.
6. Reporting themes and selecting appropriate user quotes for representation.

Triangulation and validation of data.

The SSUT method in particular delivers qualitative and quantitative data using the same research questions, which hence can be fused to enhance the scientific validity of results.

For result fusion, the presented evaluation domains and factors were used as a grid into which structured qualitative data was sorted. Prior to this sorting, qualitative non-structured data was analysed (as explained in the section above). Structured qualitative data from open questions of questionnaires and structured interviews could directly be linked with the respective quantitative data concerning the same evaluation factors. Qualitative data that could not be introduced into the pre-set grid is described as separate results. The overall fusion takes place in two layers: The fusion layer, in which structured data is fused according to the evaluation factors, and the discussion layer, where all gathered results are discussed in relation to each other and where we try to conclude general knowledge out of the fusion of all results. (See also Figure 17 for a graphical representation that details how the different methods produced data of different qualities that were analysed and fused according to their qualities.) The term “structured” is used for data that was gathered having a specific evaluation factor in mind, whereas “unstructured data” was gathered in an explorative fashion and later introduced into the analysis grid if possible.

Additionally, discussion groups and semi-structured interviews were used to include secondary and tertiary user groups. Those results were fused at the discussion stage.

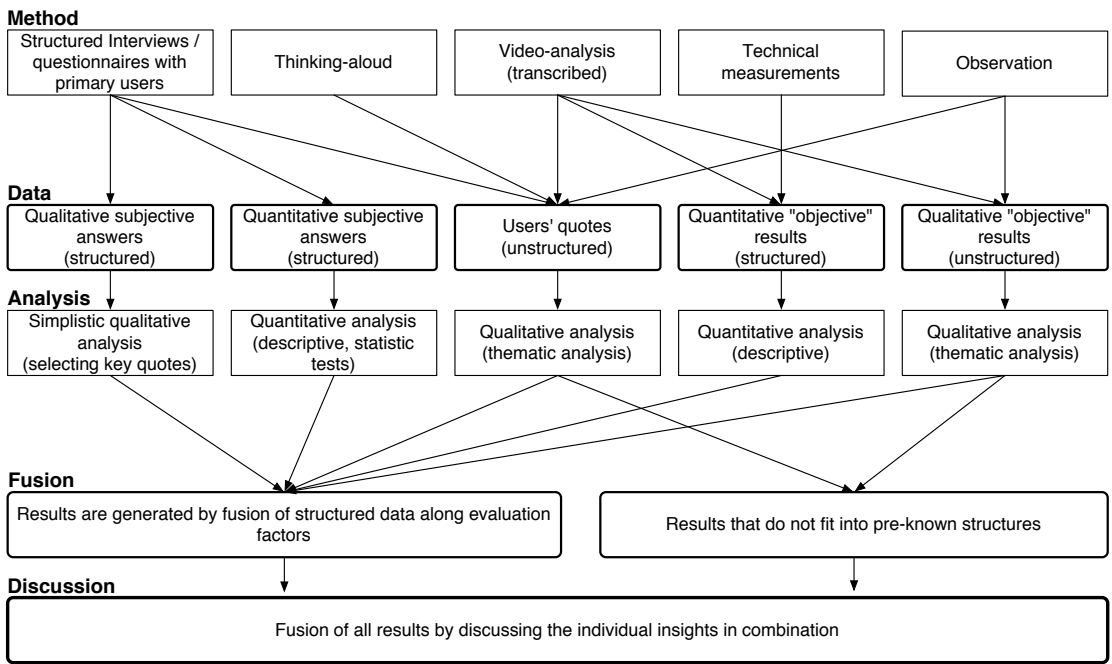


Figure 17: Overview on the used methods for data gathering, analysis and fusion.

this page intentionally left blank

4 An assistive robot to support vulnerable older users

This chapter describes the implementation of an evaluation methodology and the thereby generated results from the example of two short-term evaluation phases within a European research project.

4.1 Basic concept and idea

This chapter briefly introduces the research project and describes the idea of generating a mobile robotic assistant that supports older users in general, and users with COPD in particular, during their daily living at home. The intended functionalities of such a system were health and behaviour monitoring to support the self-care of users, ICT services to enable users to easily get in contact with relatives, friends and health providers, and smart home capabilities for additional comfort and safety.

As a unique aspect, one single and easy to use and understand SAR interface should be used to realize the intended functionalities, address the core user needs and thereby act both as an assistant and companion.

This basic idea was derived from the “KSERA” research project (funded within the 7th framework-programme).

Three central problems were addressed within this project:

1. How to realize the robotic mobility inside a constrained home environment with narrow pathways and space availability
2. How to facilitate ubiquitous monitoring of physiological and behavioural data by means of sensors inside a home
3. How to implement an acceptable HRI with older users and patients

The focus was later put on the third point, the issues around HRI, because the used anthropomorphic platform provided many interesting HRI research topics and strongly related expertise was present within the project team.

4.1.1 User needs, use cases and scenarios

Following a UCD approach, user needs and requirements were gathered by three project partners using different qualitative research methods such as focus groups, interviews and expert discussions with the user groups of older people, COPD patients, medical doctors (physicians) and care experts (care management and carers). A set of user

needs was compiled, potential use cases were derived and scenarios were developed that combine the use cases and set them into a coherent story for demonstration. The projects consortium chose a set of six scenarios based on the most pressing user needs and the technical feasibility as means to evaluate both the feasibility of the concept and the acceptance of the users.

1. Environmental warning. In the morning after the user wakes up, or in case of critical environmental conditions, the robot approaches the user and informs them about the current weather and indoor air conditions. This aspect was considered particularly relevant as COPD patients are sensitive to worsening air-quality conditions.
2. Entertainment. Upon user request, the SAR system is able to play music via the on-board loudspeakers
3. Task-oriented training. In configurable intervals, the system approaches the user and motivates them to perform physical exercises.
4. Medical measurement. In configurable intervals, the system approaches the user to motivate them to perform a medical measurement using a pulse-oximeter (a device that measures the O₂ saturation of the blood and heart rate). In case the readings are out of the norm, the system recommends calling a doctor and upon request, initiates a call.
5. Video telephony. Upon request, the SAR system triggers a voice-over-IP call to a pre-configured number. By means of a mini-beamer, the SAR is able to project the image of the call's opponent on a free spot on a wall and relays the voice via its onboard loudspeakers and microphones.
6. Environmental control. Upon request, the SAR system is able to use machine-to-machine communication to control appliances at the user's home. In particular, it is able to open/close doors and trigger the lights.

4.1.2 The prototype system

The developed system consists of an AAL smart-home environment with typical functionalities to enhance the safety, comfort and autonomy of the user, and the SAR platform "Nao" from *Aldebaran robotics* as a main interface.³⁹ Additionally, a touch-based user interface was integrated to augment the interaction capabilities of the robot and as a means for comparison between HCI and HRI concepts.

³⁹ www.aldebaran.com

The employed SAR system is capable of performing complex movements such as dancing, walking, standing up after a fall and gestures in a human-like way. Further, the platform includes a text-to-speech engine for audio output and is capable of simple mimics and emotional expressions by using coloured LEDs located on the head (mainly eyes and ears). The robot was augmented with speech recognition and the option to display images or videos via a mini-beamer mounted to the robot's back to enhance the possible interaction channels.

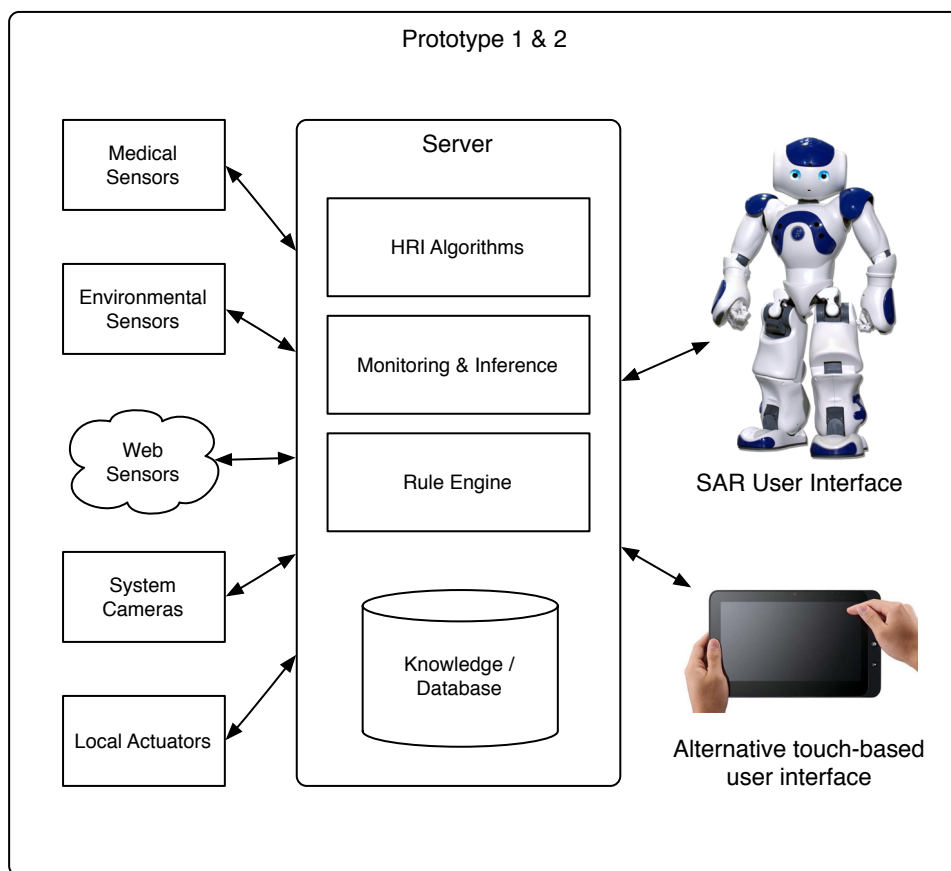


Figure 18: Overview of the first (E1) and second (E2) prototype, adapted from [Werner2013].

Figure 18 provides a schematic overview of the system's main components. As inputs, several sensors were integrated that provide information on the user's health status, the environmental conditions in- and outdoors and the localization of the robot within the test environment. The server uses the inputs to control the interaction with the SAR, the alternative touch-screen and the local actuators which were used to control the lights and entrance door in the test-environment. Only the SAR, the touch-based user interface and the system cameras are visible to the user, other components are integrated and hidden in the test environment.

4.2 Implementation of the evaluation framework for the first prototype (E1)

This part of the chapter describes how the initial evaluation framework, as described in chapter 3, was detailed and instantiated to generate an evaluation model for the evaluation of the first prototype.

4.2.1 Evaluation goals

The main aims of the evaluation of the first prototype were derived out of the aforementioned phase of user requirements gathering.

- a) Gain information about the applicability of the approach to use a SAR within a real-life setting, its potentials and limitations.
- b) Gain information about the applicability and validity of chosen research methods.
- c) Gain information on the acceptance of primary users of the approach, the perception of social impact and potential influences on the QoL.

4.2.2 Evaluation model

The evaluation framework was used as presented in chapter 3 and modified to focus on an early evaluation with a first technical integration and technical performance measurement in an LL setting. We focussed on the SSUT method to estimate the applicability for a robotic prototype and the special primary user group, while secondary and tertiary users were only invited to take part in an interview. Additionally, no long-term aspects were studied as this was considered secondary to the establishment of acceptance factors and early first insights on performance and acceptance. Mainly specifically developed acceptance factors were used to gain experience as to which factors are relevant for the assessment of acceptance and how they contribute to a greater picture of SAR acceptance.

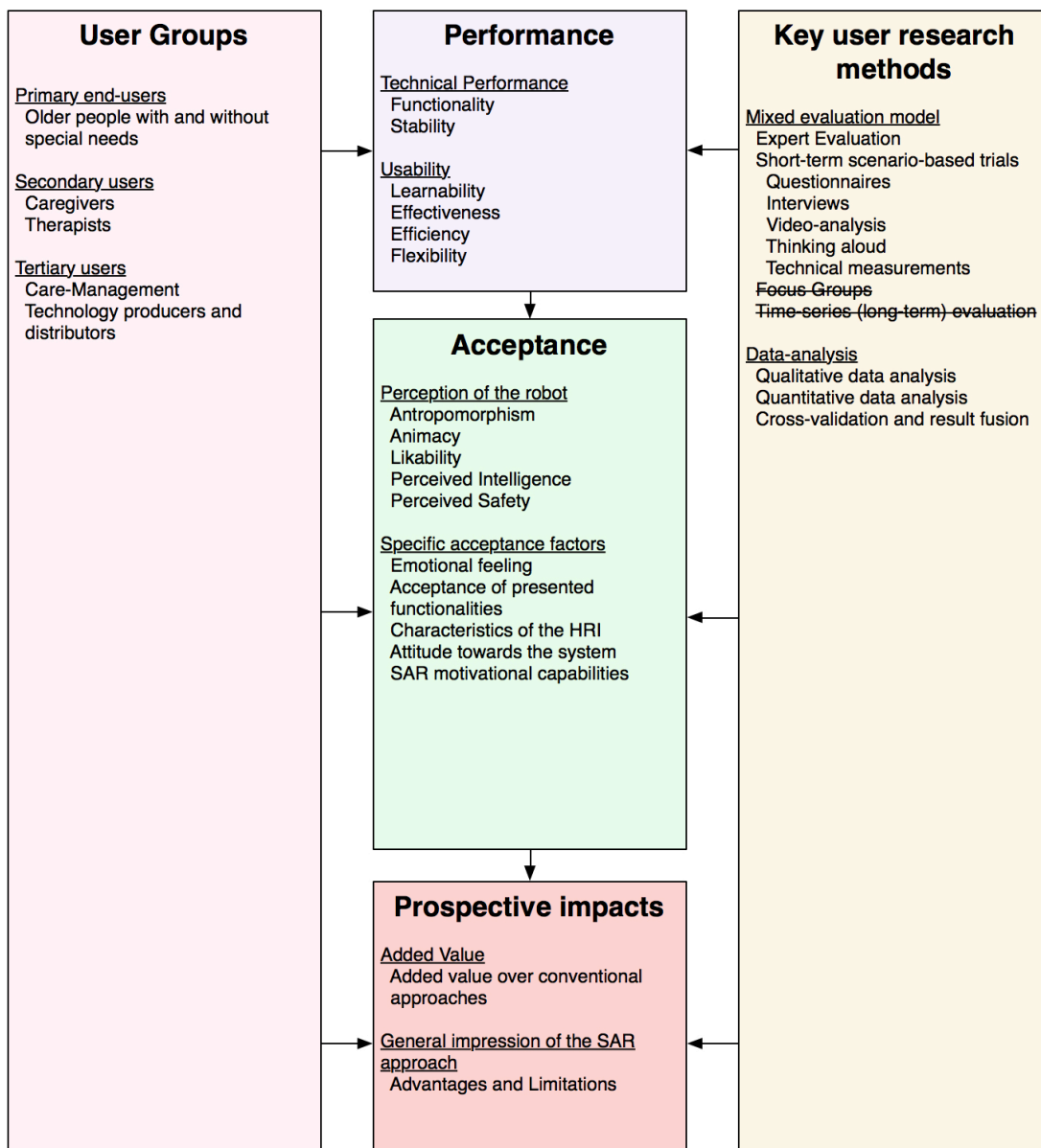


Figure 19: Evaluation domains and methods used for the evaluation of the first prototype.

Figure 19 gives an overview of the targeted user group, evaluation domains and factors and key-user research methods used during the evaluation of the first prototype. Striked-through items indicate methods and user groups that are part of the presented initial model but were not included in the evaluation in order to focus on the most important aspects needed to develop a later stage of prototype.

4.2.3 Evaluation methodology

To enhance the readability and allow the reader to link between research questions and results within this and following chapters, labels for research questions were defined

starting with RQ (research question), the evaluation phase they were used in (E1, E2 or E3), a letter for the domain (P for performance, A for acceptance, I for impacts) and a running number.

Performance

The following research questions were identified and used to drive the analysis of the systems performance:

- RQ_E1_P1: Which general technical issues exist in the second prototype that could negatively influence a later adoption of the system?
- RQ_E1_P2: Which usability issues exist in the second prototype?

Technical performance

The technical performance was assessed during and after the conduction of SSUT trials by the experimenters on a functional level by using black-box testing. Score sheets were prepared in which the experimenters marked correctly and incorrectly conducted test cases live during the experiment. In case this could not be undertaken in time or the situation was unclear and thus hard to interpret, the audio/video log and logging data of the system was used retrospectively. An example of the used score sheet can be found in the annex section 3.

Usability

The usability evaluation was undertaken by fusing the results of the technical performance evaluation regarding the time taken by the prototype to reach a goal with comments of the users about the duration of tasks and respective usability questions that targeted the learnability, efficiency and perceived flexibility of the system.

Acceptance

Perception of the robot

The perception of the robot by the users was explored by using the following research question:

RQ_E1_A1: How is the robot perceived by the user?

The Godspeed questionnaire was used to assess the perception of the robot as described in section 3.6.2.

Specific factors

The emotional feeling towards the robot, the acceptance with particular regard to the individual scenarios and functionalities presented, characteristics of the presented HRI such as conversational and movement abilities, and the general attitude towards the robot were evaluated by using customized questionnaires.

- Emotional feeling

This factor assesses how the use of the system influences the emotional feeling of the test participant; the following research question was used to design the evaluation.

RQ_E1_A2: How do users feel right after the test?

The emotional response towards the system was evaluated by means of a post-test, predefined, specifically developed 7-point Likert-scale questionnaire, asking whether the system is fun, whether it makes the users feel happy and if it fulfilled the initial expectancy or whether the interaction was unpleasant, if the users disliked the system or found it to be boring.

- Acceptance of presented functionalities

Individual acceptance questionnaires were developed for the specifics of the particular functionalities for training support and physical parameter measurements which were shown within the demonstration scenarios with the goal of gathering the satisfaction level concerning these particular functionalities. The following research question was used to design the questionnaire:

RQ_E1_A3: To what extent are users satisfied with the implemented set of functionalities?

- Characteristics of the HRI

RQ_E1_A4: To what extent are users satisfied with the implemented HRI characteristics and the robot's attitude?

As characteristics of the HRI, the quality of the SAR's movements, speech output and recognition were measured.

- Perceived attitude towards the robot

This factor evaluates how participants perceive the robot and whether this perception fits the role of the robot and the goals of the system.

- SAR motivational capabilities

This factor estimates the user's extrinsic motivation after the robot has shown them motivating behaviour. The motivational capabilities of robots are one of the core aspects of SARs that differentiate them from alternative technical systems, which makes this factor important to consider when trying to find out about the potential benefits of a SAR system. The following research question was used:

RQ_E1_A5: To what extent do users feel motivated by the system?

To partly answer the question, we can measure objectively whether the users actually performed the tasks the robot asked them to do. This factor is important in health-care scenarios as robots might have the ability to recommend certain behaviour such as training exercises. The motivational capabilities of the SAR were evaluated regarding the task-oriented physical training.

Prospective impacts

As a secondary goal of this first evaluation phase, we wanted to gain information about which impacts could arise out of the use of SAR technology.

RQ_E1_I1: Which beneficial effects for the support of older people at home can be expected?

Added value

To gain information on the added value of the solution, a comparative experiment was undertaken. Within the experiment, a touch-screen pc was used to present the participants with a user-interface that was capable of guiding the user through the same scenarios as were previously shown by the robot, but without the intervention of the robot. The participants were then asked a questionnaire containing predefined specifically developed open and Likert-scale questions for comparison of the user experience with the two systems.

General impression of the SAR approach

A predefined and specifically developed questionnaire was used to evaluate the general impression of the SAR approach, its advantages and limitations. (See also annex section 3 for examples of questionnaires used within the first evaluation phase.)

4.2.4 User group

The inclusion criteria for taking part in the trials of the first prototype were to be cognitively healthy and physically able to conduct mild physical exercises in a sitting position and a minimum age of 70 years. Further, the users had to sign an informed consent document and agree to audio and video recordings that may be used for scientific purposes.

16 users were recruited (three male, 13 female, aged from 71 to 90, average age 80 years), half of them at a test site in Schwechat, Austria, the other half at a test site near Tel-Aviv, Israel. Austrian participants were recruited by the local trial site and selecting them seeking diversity of age and technology affinity. A custom made questionnaire was used to assess the technology affinity. Within this dissertation, only the eight Austrian participants are considered since only they were evaluated by the author.

4.2.5 Test setting

As a test setting, the LL room at the “Seniorenzentrum Schwechat” was used (as described in section 3.8) and an additional comparably equipped test room in a sheltered housing facility near Tel Aviv, Israel. The comparability of test sites was ensured by means of a detailed installation and set-up plan and an installation meeting in which the experimenters that conducted the evaluation in Austria helped to install the system in Israel.

4.2.6 Evaluation procedure and test flow

In order to enhance the comparability between test sites, it was important to define the test flow prior testing and describe it in detail to allow the researchers who conducted the trials to perform the test with comparable methods. The researchers at both sites conducted the tests by performing the following steps:

1. Prior to every trial, the researchers verified the technical prototype, prepared the test documents (informed consent, questionnaires, evaluation forms) and set the system to a defined starting state with a defined local position and posture of the robot. The environmental conditions (curtains closed, all furniture such as chairs in the same position) were controlled and the tester interface, including cameras and microphones, were arranged. A checklist was used to document these activities.
2. Upon the arrival of the participant, the researchers explained the project and its goals to the user again (participants were already informed during the

recruitment). The role of the participant was discussed, highlighting the fact that it was not the user who was being tested but the technical system. Finally, the participant gave the formal agreement by signing the informed consent document.

3. Three pre-test questionnaires were filled out together with the user to gain information about the user's current emotional state and general technology acceptance. The questionnaires WHOQOL-BREF [WHO1996], a predefined, specifically developed questionnaire on the participant's attitude towards technology (see annex section 1), and the PANAS standardized questionnaire [Watson1988] were used for that purpose.
4. Every visible part of the system was explained to the user, in particular the robot and the medical measurement device which had to be handled correctly by the user during one of the test scenarios.
5. Three test-scenarios were presented to the user.
 - a. The user performs a medical measurement instructed by the robot.
 - b. The user performs physical training together with the robot.
 - c. The system simulates a response to unhealthy environmental conditions.

Each test-scenario was introduced by a short user-story that explained a typical situation in daily life, when the scenario is triggered. Within each test scenario, the robot walked to the user from a fixed starting position, interacted with the user to accomplish a task, and walked back to the starting position. In the first iteration, each test-scenario was shown twice in order to let the user experience different interaction paths the scenario could take and to allow for a higher number of results for the later performance analysis. During half of the conducted tests, a PC-based system was used as an alternative to the robot for every test scenario. The PC system facilitated a touch screen and was capable of guiding the user through the same scenarios, showing comparable content to the robot by providing visual outputs.

The user was observed during the test by an experimenter who was outside the room and watching the scene through the open door.

6. Directly after the test run, a series of questionnaires was filled out together with the user.
7. Finally, the participant was discharged and informed about the next steps of user involvement and the research work.

4.2.7 Analysis of the gathered data

The analysis of the gathered data was undertaken as described in the initial evaluation framework within chapter 3.11.

4.3 Evaluation results of the first prototype (E1)

The evaluation of the first prototype had two main goals:

1. Identify the technical performance of the integrated technical prototype within a close to real-life setting
2. Gain early information on the usability and acceptance of the prototype within user trial sessions.

These two goals conflict with each other as the first goal would require testing of the prototype under environmental conditions which are as realistic as possible, whereas the second goal requires controlling the environmental conditions in a way that allows the prototype to perform as if it were the final solution. This trial tried to balance the two main goals and to keep the environment as realistic as possible while ensuring that the system performs stably enough to allow users to gain a first experience in using the system.

4.3.1 Summary of performance results

To ensure the functional performance of the system during the user trials, the system was tested in the laboratory as well as at the trial sites during and after the installation. Further, the trial location was partly adapted to circumvent potentially critical situations to allow the users to perceive the prototype in the intended way and imagine in the closest way possible its use in their own home. Allowing the user to experience the prototype was treated as of higher relevance than testing the system under real conditions.

4.3.1.1 Technical performance

The technical performance of the system during the conduction of trials was measured in order to measure the dependability of the system within a close to real-life setting and to gain information about technical issues that might influence the user's perception of the system and thereby possibly bias the analysis of acceptance and impacts.

RQ_E1_P1: Which general technical issues exist in the second prototype that could negatively influence a later adoption of the system?

As part of the trials, performance tests for the following technical domains were undertaken.

Functionalities for localization and navigation of the robotic system

We assessed the performance of localization and navigation algorithms by analysing the system's output that indicated the deviation between the real location and the automatically calculated location of robot and user. In case the offset was greater than 0.5 metres, the localization was considered to be incorrect.

We found that the technical system was capable of correctly localizing the user and the robot within the test environment in 48 of 56 cases (86%). Consequently, the navigation from the pre-set starting point to the user performed correctly in only 31 of 38 cases (82%) as it is dependent on the localization. It has to be noted that navigation did not take place in all cases where the user or robot was localized, since localization was also undertaken in other scenarios such as when interacting with the user. The reasons for errors were the performance rate of the particle filter algorithms responsible for detecting user and robot by analysing the camera feed of a ceiling-mounted camera. As the algorithms used movement within the camera's view to detect people and the robot, the location of either of the two was sometimes lost in cases when the user or robot did not move much.

This issue is partly due to the trial setup as the technical system was developed for the real-life case where it would be able to locate the user continuously. Since the system was started and active only for the time of the test, it had to be initialized after each start by a movement from the user. To avoid this issue, the users were asked within the trial procedure to walk into the trial room to allow the system to detect them.

As a common error, the automatic location detection misinterpreted the direction in which user or robot were facing. This is an issue as the navigation towards the user takes the user's field of view into account and tries to approach the user always from the front, as this was found to be socially acceptable behaviour.

Functionalities for interaction and communication (HRI)

The system used voice-based interaction including speech recognition as the main communicational means. Although the command-set for speech recognition was kept

minimal and limited to a few phrases, it was found that the actual recognition worked correctly in only 38% of cases (10 of 28 cases). Because of this major limitation, the speech recognition was simulated using a Wizard of Oz technique in all trials.

The speech output of the system worked well, however since a text-to-speech engine was used, the rather mechanical melody and intonation of speech led to a poor voice understandability; in 13% of cases (28 of 216), the users could either not understand the robot's requests correctly (based on analysis of users' comments and behaviour) or complained directly about poor understandability.

To support voice interaction, the robot was programmed to move the head in a position that is typically interpreted as looking towards the user during interaction phases. The technological functionality behind this behaviour was realized by means of a face detection algorithm. This algorithm worked according to specifications in 79% of cases (18 of 24). The algorithm could not detect the face in cases when the user wore glasses, which accounted for the majority of errors. In two of 24 cases, the algorithm mistakenly interpreted objects in the room as faces.

In a further attempt to enhance the human-robot speech dialog, the gaze direction of the user was analysed. This way the robot could react to the user's behaviour and raise their attention in cases when the user was not looking at it during the conversation. In 52% of cases (12 of 23), this algorithm miscalculated the user's head orientation, which led to unwanted gesturing by the robot (waving the arms) attempting to garner attention which already be regained.

Summary of the functional performance of the system

The system used for the trials can be considered an early and complex prototype. As such, the performance within a close to real-life situation of several sub-systems was still far from optimal and also inflicted an unknown dimension with the measurement of acceptance or even impact factors. Nevertheless, the results on the functional performance pose important findings for the redevelopment of the prototype for later user trials focusing on acceptance and potential impacts.

4.3.1.2 Usability

Additional to issues of functional performance, the usability factors of learnability, efficiency and ease of use were evaluated based on the following research question:

RQ E1 P2: Which usability issues exist in the second prototype?

Efficiency

During the assessment of functional performance, the time needed by the robot for navigation towards and from the user was measured as an efficiency factor of usability. It was found that the navigation times varied strongly between iterations depending on the performance of the dynamic localization routines. On average, it took the robot 35 seconds to move from its starting point to the user sitting 2.5m away which can be considered as problematic, in particular when thinking about the likability of larger distances in daily practices, e.g. in larger homes.

The face-detection algorithm used a search method that involved the robot turning its head to search for the user's face at the most probable locations. This face-search routine proved to be time intensive, with a median time taken of 39 seconds (best – three seconds, worst – 86 seconds) until the user's face was found. Since the conversation with the user was only started after the face was detected, this was influencing the user acceptance, as shown by users' comments who found the timing to be too slow.

Learnability and ease of use

By using a post-test Likert-scale questionnaire, the general impression of older users regarding the system was evaluated. The 16 users asked found the system easy to use and also easily learnable. The voice understandability received mixed ratings, for reasons already covered in the assessment of technical performance.

The voice interaction with the robot was particularly appreciated for quick feedback with low complexity; however, users found that the presentation of more complex data such as the weather report or long-term data on their physical performance would be better comprehensible when displayed on a screen, e.g. as a chart, because the robot's utterance were sometimes hard to follow and remember for a longer time. This finding is also backed up by the rate of communication errors, which were highest during the report of environmental conditions because of the lengthy, data-heavy spoken output.

4.3.2 Summary of acceptance results

Despite the bias caused by technical issues that were perceivable by the test participants, acceptance factors were measured and analysed, also in order to gain information and training on the process of measurement and analysis for later evaluation phases. The main aim was to gain information on the relevance of the

developed acceptance factors, build hypotheses on how they relate to each other, try-out how qualitative and quantitative results can be best fused, and find potential methodological issues. As a secondary aim, these first results should also give early insights into the acceptance of the evaluated SAR system and thereby drive the next stages of design and development.

Perception of the robot

RQ E1 A1: How is the robot perceived by the user?

Regarding aspects of HRI, the perception of the SAR was evaluated by implementing the Godspeed questionnaire.

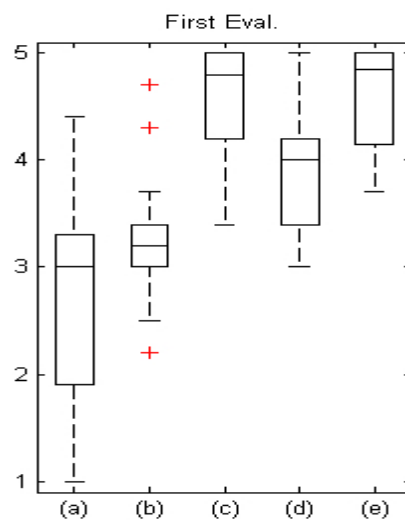


Figure 20: Acceptance factors regarding the perception of the SAR, adapted from [Werner2012], (a) anthropomorphism, (b) animacy, (c) likeability, (d) perceived intelligence, (e) perceived safety.

As is shown in the boxplot diagram of Figure 20, the users were at variance when asked about the anthropomorphism of the presented prototype. On the one hand, users noted that the robot looks like a small child but on the other hand, the behaviour of the robot did not match the appearance very well. Additionally, some users principally found that a robot cannot be conscious, which was one of the items within this construct. Others were more open and found that the system showed certain human-like aspects, which explains the high variance shown in the “anthropomorphism” factor (a). Issues in the interactivity of the SAR such as the named issues in timing and voice interaction influenced the score for “animacy” (b) negatively. Additionally, the system was perceived as rather static and robot-like. The SAR was perceived as being very likeable and safe, as is shown by the corresponding high scores of “likeability” (c) and “safety”

(e) because of the nice appearance and small size which did not evoke anxious reactions. The system was perceived to be principally “intelligent” (d); however, certain malfunctions that were perceived by the users – such as navigational issues in which the robot turned in circles or stood in front of the user without being able to find the face – likely had a negative impact, which also explains the shown variance.

Specific acceptance factors

RQ_E1_A2: How do users feel right after the test?

As a result, it was found that users experience the system to be mostly fun to use (median 5 of 7) and as having a positive influence on their emotions (median 5 of 7). Within the test duration, the users did not perceive the system to be boring, however this was not to be expected due to the novel nature of the system and the short time-frame of the trials of only one to two hours per participant.

As a related result, the users found the system to be highly entertaining (mean 4.5 out of 5) which is likely to have a significant influence on the overall perception of the system and also opens future application areas, as it opens the possibility to use the system within serious gaming applications.

The system only partly lived up to the initial expectations of the users (median 3 of 5). This is partly due to the humanoid appearance of the robot, which is known to cause high functional expectations as users expect it to behave similarly to a human, including being able to show similar interaction and multi-purpose functional skills that cannot be met by current technology. Many participants expressed their disappointment regarding the behaviour of the system, which they found should be more human-like. Therefore, it can be expected that a more human-like HRI that matches the design of the system would enhance the overall perception and acceptance.

RQ_E1_A5: To what extent do users feel motivated by the system?

The motivational capabilities of the system were assessed both by questionnaire and by observing the participants’ behaviour during the trials. The behavioural analysis showed that users were very motivated to comply with tasks given by the robot (100% compliance when asked by the robot to perform a medical measurement); however, this is likely to be strongly influenced by the test situation itself in which users also were, to some extent, socially bound to comply with the test tasks.

Subjectively the participants found themselves “quite” (4 out of 5) motivated by the robot in general, whereas the support of physical training was particularly found to be highly motivating (mean 4.6 out of 5).

RQ E1 A3: To what extent are users satisfied with the implemented set of functionalities?

Regarding the physical training, almost all participants gave positive feedback about their motivation to conduct the training with the robot as trainer. Participants were very willing to mimic the robot’s movements and also found they could easily follow them despite the fact that the robot did not perform the training movements entirely correctly due to mechanical restrictions. Regarding the mechanical limitations, users found that it would be helpful to conduct regular training sessions with a human trainer in order to refresh the memory and correct training mistakes. Several participants particularly emphasized that the reminder to perform physical training measurements seemed to be very helpful for motivation as the robot could directly approach and remind them.

RQ E1 A4: To what extent are users satisfied with the implemented HRI characteristics and the robot’s attitude?

Most participants rated the speed of interaction and the walking speed as slow. Also, the quality of movements was given mediocre marks in terms of elegance. A faster and more human-like movement would be beneficial for further developments as it would fit to the humanoid appearance of the robot and hence better meet the expectations of the user group.

The participants showed a very positive attitude towards the system; the robot was perceived as a friendly, happy, mindful and intelligent companion (all ratings above 4 out of 5), which are all aspects of confidence and trust in the system that are considered by acceptance models (UTAUT, Almere model) to influence the intention to use in later real life. The ratings were most probably negatively influenced by perceived technical malfunctions during the trial, in particular the rating for perceived intelligence as users commented negatively when, e.g. the robot tried to find the user’s face for a longer time but could not find it in the end.

4.3.3 Applicability of the solution and acceptance from the view of care experts (secondary users)

Two care experts were interviewed regarding their general impression on the presented SAR solution after the finalization of trials with primary users. Expert A is the manager

of a senior-citizen centre; expert B is an experienced social worker from the same senior-citizen centre. During this interview, results from the already finished primary-user trials were also discussed. An interview guide was used and focused on the experts' reasoning around a potential introduction to the senior citizen centre and on their opinions on the acceptance by older users in general.

Acceptance of the solution

Both experts themselves obviously found the SAR solution very likable, laughed a lot during the demonstration, commented often on the nice and cute look of the robot and generally had a lot of fun exploring the solution. For them, the participation in the study seemed to be rather an event and a welcome diversion from their work. In particular, the robotic movements received exciting reviews as both participants were surprised by the capabilities. In this sense, their review is biased by the novelty effect of the solution.

Expert A found the solution to be very motivating, in particular during the physical exercises; expert B agreed but commented that this motivational effect would decline as soon as one finds out that the robot is not able to assess the user's movements. During the discussion, both experts found the motivational effect to be stronger as when compared with a video showing a human trainer. Expert B questioned the duration of the motivational effect and whether this could rectify the probable high price of the solution.

Expert A found it would be necessary that the robot understands who it is interacting with; in particular, a confusion with a visitor was discussed and seen as problematic.

Both experts found that the robot's behaviour does not fit its appearance; in particular, they agreed that the chosen voice does not fit the SAR's design as it sounded to them to be female, whereas they interpreted the blue colour of the robot as male.

Expert B compared the system with a toy that would be interesting in the beginning, but as soon as one has seen all its functionality, this novelty effect would wear off and then the system would be perceived as slow and inefficient to use.

Expert A found that talking to the robot feels inevitable, which would be a positive effect because it shows that the system builds up a level of trust instantly.

Expert A stated that the robot should be able to help in manifold tasks, showing a large set of different functionalities. Those should then be adapted to the individual needs of the current user.

Both came up with ideas for additional functionalities:

- A reminder for older adults with dementia to not leave the flat during the night-time in case the robot detected movement.
- Organizing a taxi upon request.
- Triggering appliances within the home, such as the doors or the lights.

Applicability within a care institution

When asked about the applicability of the solution within the senior-citizen centre, both agreed that the SAR prototype as presented would not fit into their institution because carers already do the job and hence a robot would not be needed. Expert B detailed that the system would be “too much” for an institution with carers and was concerned that the SAR could be seen as replacing human carers. Another concern raised by expert B was that the dementia patients at the institution could get confused and tell their relatives about “little men walking through the rooms”.

The applicability was further discussed with the institutions ward for older adults with dementia specifically in mind. Both experts found that dementia patients could be afraid or confused if the system approached them to try to interact. Expert A found that dementia patients could want to dress or feed the robot like a puppet, because it could evoke a maternal instinct. Expert B warned that the system should not be anthropomorphized too much and made clear that she thinks the SAR should only be used by older people who know that they interact with a robot and that helps them with a certain set of tasks. Otherwise, the robot would not be used appropriately and it would be a waste of money. Both experts agreed that the solution should not be used with dementia patients in general.

The system would be rather interesting for usage at home and for users who are alone most of the time, lacking social contacts.

4.3.4 Summary of results regarding prospective impacts

To gain early information on the potential impacts of the provided SAR system, we used the following research question:

RQ E1 I1: Which beneficial effects for the support of older people at home can be expected?

Some potential impacts became apparent during the analysis of acceptance results and the added values by comparison with a touch-screen device. Mainly the capability of the

system to motivate users to perform strenuous tasks was highlighted within qualitative comments by the users, as the system seemed to be “funnier than a PC”, “pleasant and cheering up” and also demonstrated physical exercises in a very understandable way. In general, the system was perceived as a nice, polite, little helper and the sympathy towards the system was higher than towards a regular touch-screen-based PC. Users saw the fact that the robot can move as a strong advantage as they gained the idea it might be able to find them in an emergency situation, despite this not being a scenario that was demonstrated.

The main impacts found are hence those of SARs in general, which are derived from the capability to provide entertainment, to motivate due to their social presence and HRI capabilities, and to provide additional safety. Regarding the impacts, we also built the hypothesis that a positive impact on the QoL of a user can be achieved by a system that fulfils the users’ needs and is accepted by the user. In that sense, impacts can be expected for all implemented scenarios that are accepted by the users.

4.3.5 Recommendations for further development

The following points were derived from the evaluation results and summarize the recommendations for the most important enhancements of the next prototype.

The performance of several technical functionalities needs to be improved. In particular, the following systematic errors should be reduced:

- The performance of the localization and navigation algorithms should allow the system to move reliably to the front of the user.
- The performance of speech recognition should at least allow “yes” and “no” answers to be reliably given over the communication distance of 1-2 metres with a rate of 95% correctly understood words.
- The performance of face-direction detection should either be greatly enhanced or the feature should be excluded since it hindered the natural interaction flow by introducing unnecessary delays.
- The face-detection functionality should not cause delays in interaction.

Enhancing technical stability is more important than enhancing functionality for the next prototype since within the complex robotic system, the instabilities of all modules add up, increasing the statistical chance of an error that prevents the demonstration to a user altogether.

The overall speed of interaction should be enhanced; the system should respond more quickly and delays during interaction should be avoided where possible.

The interaction with the user should be more natural and interaction design should take care of possible errors both from HRI (e.g. understanding issues) and the system (e.g. input not received in time, technical error) in order to cope with the high expectations of users.

- By using a more sophisticated interaction design, the robot should act in a more intelligent and human-like way.
- Mimics and gestures should be used in order to enhance the animation of the system and make the robot more vivid.
- The speech output should be less repetitive and for each meaning, three alternative outputs should be generated to make the system appear more intelligent.

The duration of HRI during the trials should be prolonged; ideally the user should be able to interact with the robot for the whole duration of a trial (approx. one hour) without the intervention of a researcher in order to allow the user to experience a coherent story of typical usage at home.

4.3.6 Contribution to HRI design based on evaluation results

Based on the identified need to enhance the SAR design, a method for the interaction design was developed and named interaction flows (IF). The goal of this method was to avoid the development of a straight-forward simple interaction path based on the view of developers which would not allow for a diverse interaction, and instead develop a HRI that fits the users' requests regarding the robot's humanoid, animated and intelligent behaviour.

Interaction flows were designed to be a natural interaction between a human and a robot considering all system-inherent interaction channels and thereby guide the technical development of prototypes based on actual user needs.

Interaction flows use a specially developed graphical representation similar to unified-mark-up-language (UML) interaction diagrams to define the most probable interaction paths. In contrast to existing use-case diagrams as can be found within the UML (e.g. compare [Collins-Cope1999] or [Longo2015]), interaction flows focus on multi-modal interaction as used in HRI.

Within this dissertation, IFs were used for the purpose of modelling the human-robot interaction facilitating the interaction capabilities of the used platform: “voice”, “gestures” and “mimics” in all further iterations. By implementing IFs, the behaviour and thereby also the represented character of the used robot could be modelled along with the text output. In order to design plausible and lively robot behaviour, it is important to model behaviour considering all input and output modalities the robotic platform provides. The strength of using this method lies in the simultaneous representation of all input and output modalities within a single and simple to draw block as shown in Figure 21.

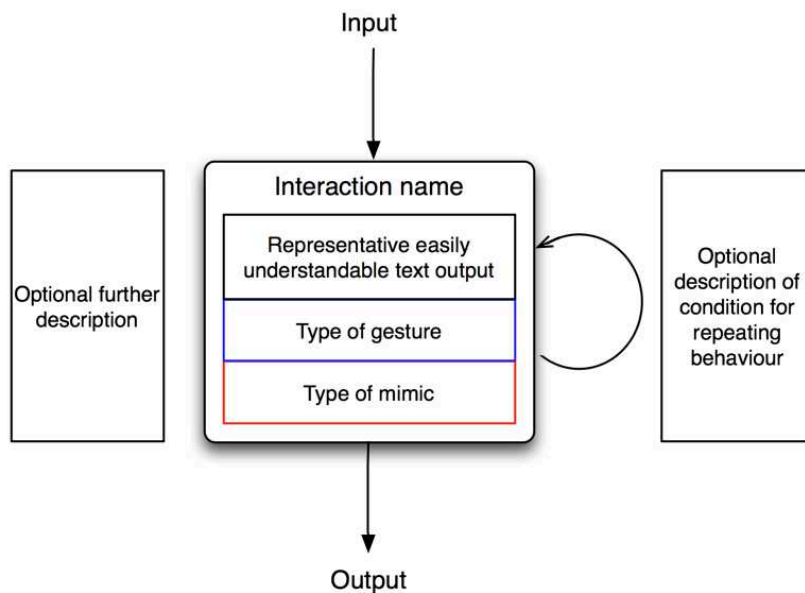


Figure 21: Example of an "interaction block". The coloured boxes are placeholders for interaction capabilities; in our case top-down: voice, gestures and mimics.

An interaction flow is composed of several interaction blocks; an example of which is shown in Figure 21. An interaction block itself is composed of a unique name, an input, at least one output that is based on a condition which is typically depending on a user’s input, an additional description and most importantly, the description of the modelled output channels. The output channels are generic and to be adapted depending on the specific robotic platform in question.

The IF design method is meant to be simple, quick to draft, easily reusable and quickly understandable by all partners within an R&D project to share a common understanding of the robot’s future behaviour during the design phase. For that reason, the name of blocks, the text for voice output, the description of gestures and mimics has to be

descriptive (e.g. descriptive names have to be used) and easy to understand. The exact voice output and the detailed description of the individual gestures and mimics are not part of the presented IF charts (to keep them simple to understand), but of a separate document that should be provided with the charts.

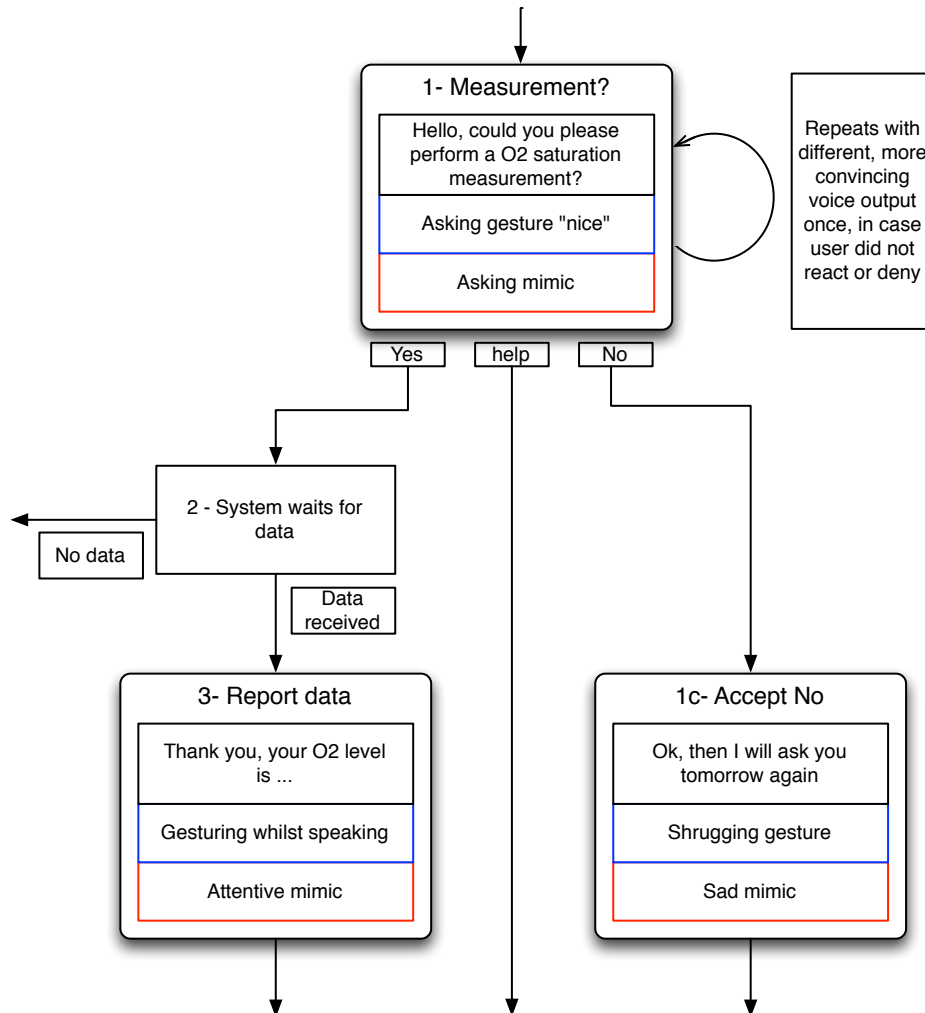


Figure 22: Example fragment of an interaction flow.

Because the IF method is flexible to use, complex interaction can be modelled and visually represented as is also demonstrated with a further example of an actually used interaction flow that can be found in the section 2 of the ANNEX.

4.3.7 Lessons learned regarding the methodological approach

This sub-chapter summarizes lessons learned during the evaluation procedure and the analysis of data. Conclusions were drawn and used during the further development of methodologies for user evaluation of the following robotic prototypes.

Evaluation procedure

From our observations when conducting the trials we found that the prototype system is complex to install and use and that a learning phase for the technicians that control the system during trials seems to be needed. For future trials, it was planned to conduct a series of test sessions with colleagues or students to simulate the human participants within the exact same location and the final technical setup in order to let the technical support team learn the peculiarities of the system and how to handle them during the trial phase. To allow comparative studies between trial sites, this lessons-learned phase should be integrated into the test plan and distributed between the trial sites. As a final step, prior to the main trial phase, a pre-test with an actual older user should be conducted to verify the system's capability for the trial with older and vulnerable users (patients).

Regarding the pre- and post-test questionnaires, we found that some questions are not well suited as they were hard for older participants to understand (we had to ask the questions several times) and to some extent repetitive, since similar validated questionnaires were combined but could not be fused without compromising the validity of each separate questionnaire that needed to be analysed on its own. In one case, an ethical issue surfaced when conducting the WHOQOL-BREF questionnaire [WHO1996] which also asks the older participants about their satisfaction with their sex- lives, which was found to result in the users refusing to answer the question and caused an awkward situation for the testers.

It was also found by the experimenters that the post-test questionnaire was time-consuming because, in most cases, the questionnaire could not be given to the participants but had to be read out-loud and partly explained to the participants in order to make sure they fully understood the questions. For that reason, it was advised to refrain from extending the questionnaire for the consecutive evaluation phases of later prototypes.

During the evaluation of the first prototype, it was found that the test with participants took more time than budgeted because of technical issues present in the prototype. As the technical setup needed to be rebooted and functionally tested for every participant, only a maximum of two trials per day could be realized. Further, the impact on financial resources was also higher than expected as three researchers were needed during each trial (two technicians to control the SAR system, the audio/video recordings and

evaluate the functionality of the prototype, and one researcher that introduced, interviewed and interacted with the trial participants in cases of system errors).

Data analysis

It was partly difficult to interpret the meaning of quantitative results as baselines were not yet published for the used questionnaires, in particular the Almere model. This is due to a lack of experience with the used questionnaires, also in particular regarding our self-developed questionnaires. As hardly any comparable evaluation results from other systems were available, the scope and value of our interpretations were limited. Very helpful in this aspect was the fusion with qualitative data, which often provided insights into the desires of the users. One example for this would be the Godspeed questionnaire and the results on the anthropomorphism of the robot. It is unclear whether a high or low anthropomorphism would be beneficial, also considering the (debatable) uncanny valley theory [Mori1970]. Here, it was considered that neither a very high nor very low value would be beneficial but a value that fits what users expect given the appearance of the system, which in this case meant that the target value should be increased for the next evaluation as users perceived that the system should act more like a human.

As the acceptance results were derived from subjective data only, it was decided to also try to integrate more objective measurements in the next evaluation phase to gain a higher scientific significance by means of result fusion.

The acceptance factors were initially derived from the UTAUT model and augmented with additional factors based on a literature analysis. In the meantime, Heerink et al. [Heerink2010] proposed a model for HRI evaluation that we decided should be adopted in our future evaluation phases to allow for an eventual future comparison of results, and to base the evaluation on a validated metric.

General

The evaluation procedure was pre-defined in a way that participants were presented with the developed demonstration scenarios in a given order. Additionally, the users received information on how to behave (e.g. when to enter the room, where to sit and how to use the measurement equipment). This was needed to compensate for technical issues such as that the system only being able to detect a moving user. To avoid overseeing issues that could arise in a future real-life situation, it was decided that the next evaluation phase should also include “free interaction sessions” in which users should be able to interact with the robot as they like. Therefore, the users should only be

given a set of keywords needed to interact with the robot, the flow of events should then be handed over to the user. To guarantee the technical performance, experimenters should support the systems by correcting system decisions live during the interaction by means of a Wizard of Oz technique.

As an effect of the lessons learned within the evaluation of the first prototype, it was decided that the next evaluation phase should not be undertaken within a true real-life setting at users' homes as initially planned, but again within an LL setting as it will likely not be feasible within the course of the project to realize a prototype that fulfils the high safety and performance requirements of a longer-term trial in users' homes.

Given that robots are new to the public and that the presented humanoid approach in particular likely introduces a strong novelty effect, it would be important to perform trials over a duration longer than two hours to gain information on the long-term acceptance of the system. However, as this was not feasible at the time we conducted the study, another approach such as repeating invitations to the same users over the course of three weeks (the duration is based on research suggesting the novelty effect wears off after days rather than weeks, compare also Koch et al. [Koch2018]) should be aimed at, to at least gain some information on how acceptance rates vary over time.

4.4 Implementation of the evaluation framework for the second prototype

4.4.1 Evaluation goals

This evaluation phase was the final evaluation within the "KSERA" project. In comparison to the above reported results of E1, this evaluation phase was considered to be rather summative than formative. We therefore put the focus of the evaluation phase onto the assessment of acceptance factors and tried to gather first insights and give prognoses on future impacts of the technology.

The main aims of the second evaluation phase with the second prototype were:

- a) Assessing the systems performance under LL conditions and how they compare to the results of E1 given the recommendations after the first evaluation phase.
- b) Gaining information on the acceptance and added values of the technology on end-users and secondary users and how results compare to those of earlier evaluation phases given the recommendations on the design of the robot
- c) Gaining experience with the developed revised evaluation model

4.4.2 Evaluation model

The initial evaluation framework as presented in chapter 3, together with its modifications from the first trials in (E1), was used as a base to define the evaluation model for the second evaluation phase (E2). Secondary and tertiary user groups were now included into the evaluation to gain a wider understanding on the evaluated domains by means of interviews and focus groups. The technical performance and usability measurements were kept unchanged within the model, as these constitute moderating factors to the acceptance and impact results. Further, it was intended to compare the results between the trial phases. Most acceptance factors were left unchanged for the same reasoning but the acceptance domain was augmented with constructs from the Almere Model to gain a deeper understanding of acceptance. To gain an understanding on the development of user acceptance over time, the users were asked three times within three weeks to interact with the system, which made it possible to analyse results within a time series.

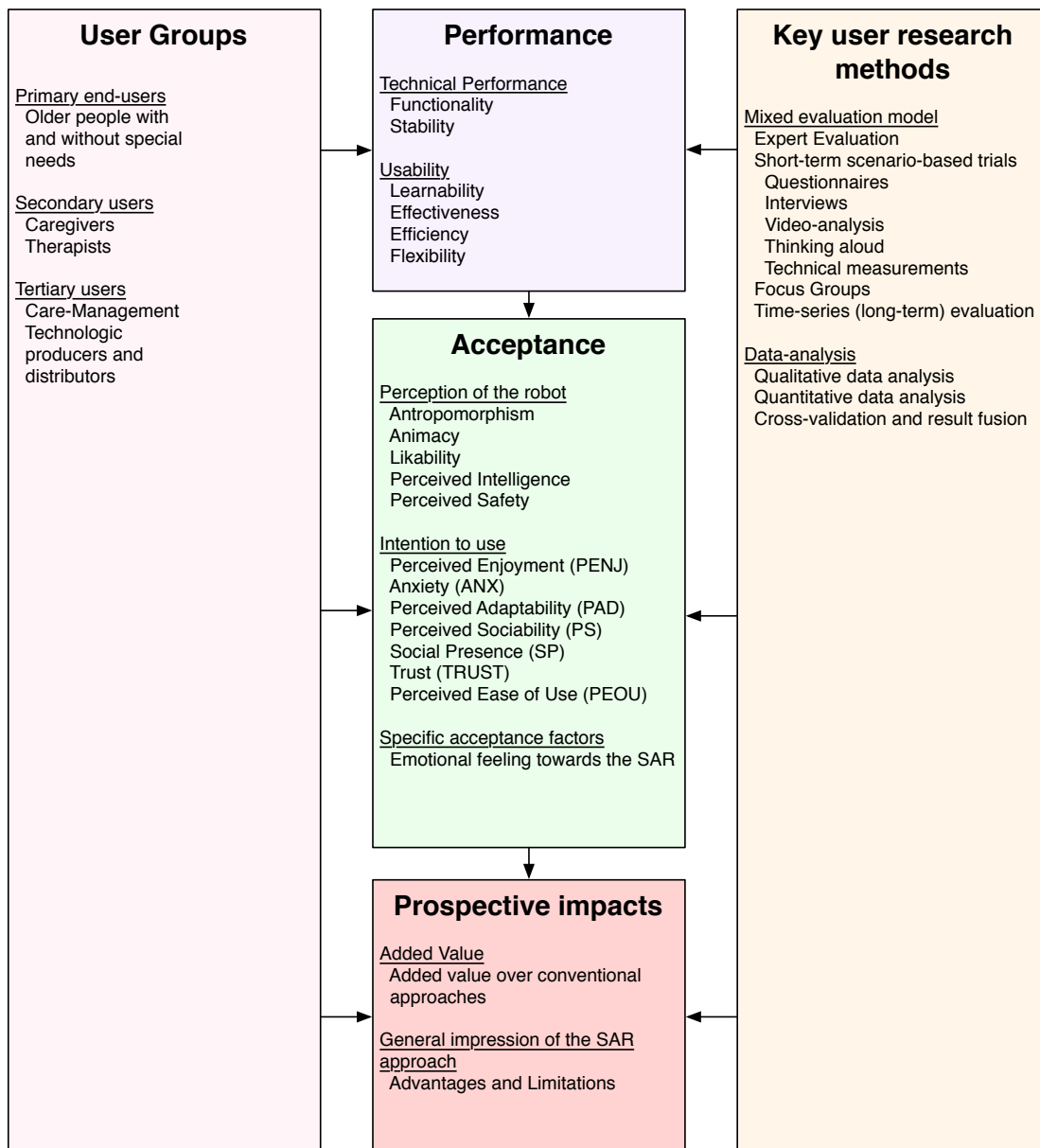


Figure 23: Evaluation domains and methods used for the evaluation of the second prototype.

4.4.3 Evaluation methodology

Performance assessment

The performance assessment was undertaken using the same methodology as during E1. The score sheets for the experimenters were adapted to the new test cases. Because the methodology behind this assessment was black-box testing, which is unaffected by internal system changes, the score sheets were only updated in parts regarding the new functionalities of the robot, leaving the other parts aside and available for comparison between the prototypes.

The usability evaluation was undertaken using pre-defined customized questionnaires and fused with qualitative comments of the thinking-aloud process, with results of the observation as well as technical-performance data regarding efficiency.

The following research questions were used to guide the analysis of performance results:

- RQ_E2_P1: Which general technical issues exist in the second prototype that could negatively influence a later adoption of the system?
- RQ_E2_P2: To what extent do technical issues influence acceptance factors?
- RQ_E2_P3: How does the systems technical performance compare to the first prototype?
- RQ_E2_P4: Which usability issues exist in the second prototype?

Acceptance evaluation

The following research questions were used to guide the analysis:

- RQ_E2_A1: How do the identified acceptance factors score when using the SAR approach with all implemented functionalities?
- RQ_E2_A2: How do acceptance factors compare between the first two prototypes?
- RQ_E2_A3: Which general limitations exist hindering a later usage at home?
- RQ_E2_A4: Which limitations exist concerning the different implemented user-need domains?

The acceptance of the primary user group was evaluated using the in chapter 3.10.2 described SSUT method. Within this method, questionnaires were implemented which were in most cases descriptive but also contained open questions to obtain qualitative data. Furthermore, the thinking-aloud method was used during the demonstration of scenarios to gain qualitative data, and video analysis was performed to verify the data gathered and to estimate the emotional feeling of the participants.

To assess the perception of the robot, the Godspeed questionnaire [Bartneck2008] was implemented as described within E1 in section 4.2.3.

The intention to use rates whether test participants would want to use the system without considering a future price. This can be considered the strongest acceptance factor as the current acceptance models try to include all other relevant factors. However, when directly asking the users, one needs to be aware of certain biases such

as interviewer bias, social desirability and the fact that participants of studies value technical systems higher as “they have to be important” if researchers ask about them.

To assess the intention to use, selected constructs of Heerink’s Almere model were used. As we already experienced during the conduction of E1 (see also chapter 4.3.7) that the older users had difficulties understanding some of the questionnaire items, and due to the fact that we observed the tests to be straining for older people, we decided to only implement parts of the Almere model and augment it with a custom-developed questionnaire that covered specific acceptance questions related to the specific functionalities and appearance of the used prototype.

The Almere model uses statements that participants can agree upon using a Likert scale from 1-5 (totally disagree, disagree, don’t know, agree, totally agree). The statements were presented to the participating users in a random order.

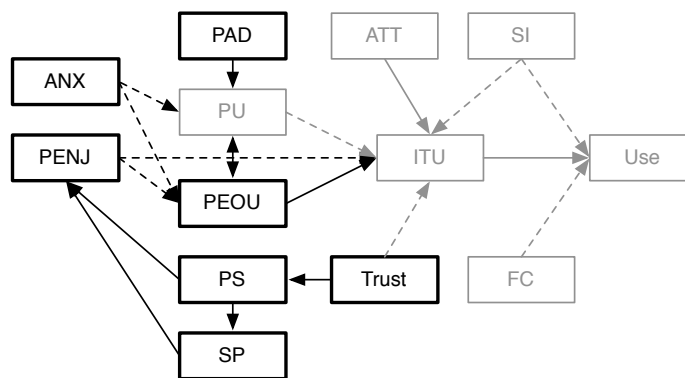


Figure 24: Relations of constructs within the Almere Model [Heerink2010]. Greyed-out constructs were not used during the trials. Figure originally published by the same author in [Torta2014].

The following constructs were used during this evaluation phase

- Perceived enjoyment (PENJ)

The factor evaluates feelings of joy and pleasure when using the system. According to Heerink et al., the perceived enjoyment (PENJ) when using a technical system positively contributes to the willingness to use the system in the future [Heerink2010]. In addition; Sun et al. [Sun2006] found that this factor also positively influences the ease of use and thereby the intention to use the system. This means that functionalities that support entertainment; or more generally are fun to use; positively influence the acceptance of the system. This becomes particularly important for SAR systems which may use their multi-modal capabilities to facilitate entertainment functionalities, even during otherwise tedious tasks.

Questions within this construct were derived from earlier questionnaires of E1 and are therefore comparable with earlier results:

- The system was amusing and I enjoyed interacting with it.
- Using the system was boring and did not interest me.
 - Trust in the system (TRUST).

This construct models the belief that the system performs with personal integrity and reliability. This is important since the users need to trust the robot in order to follow the advice given, such as the corrective statements made during execution of exercises. Trust is claimed to have a direct influence on the intention to use the system and on perceived sociability.

Questions within this construct:

- I would trust the robot if it gave me advice.
- I would follow the advice the robot gives me.
 - Perceived ease of use (PEOU)

This construct represents the degree to which a user believes that using the system is free of effort. This score is closely related to the usability of the system and is directly related to the intended use of the system and is hence a very important acceptance factor.

Questions within this construct:

- I think I will know quickly how to use the robot.
- I find the robot easy to use.
- I think I can use the robot without any help.
- I think I can use the robot when there is someone around to help me.
- I think I can use the robot when I have a good manual.
 - Perceived Adaptability (PAD)

This construct rates the perceived ability of the system to adapt to the needs of the user. Since user abilities such as motor and sensory abilities constantly change, assistive devices, particularly if designed for older users, need to constantly adapt towards the current users' needs. If users perceive a system to be adaptive to their needs, they will find it more useful which has a positive influence on the acceptance of the system.

Questions within this construct:

- I think the robot can be adaptive to what I need.
- I think the robot will only do what I need at that particular moment.
- I think the robot will help me when I consider it to be necessary.
 - Anxiety towards the system

This construct measures whether or not the system evokes anxious or negative emotional reactions during usage and influences PU and PEOU (see also Figure 11).

Questions within this construct:

- If I should use the robot, I would be afraid to make mistakes with it.
- If I should use the robot, I would be afraid to break something.
- I find the robot scary.
- I find the robot intimidating.
 - Social presence

This construct evaluates whether or not users sense a social entity when interacting with the system. This was shown for SARs [Fasola2012] and has an influence on the acceptance of the system, and the motivation to use the system and follow the system's advice.

Questions within this construct:

- When interacting with the robot, I felt like I'm talking to a real person.
- It sometimes felt as if the robot was really looking at me.
- I can imagine the robot to be a living creature.
- I often think the robot is not a real person.
- Sometimes the robot seems to have real feelings.
 - Perceived sociability

This construct is especially important for a social assistive robot since it measures the perceived ability of the system to perform sociable behaviour, which is an important factor when using the human-human model for HRI. It is not uncommon for people to interpret technological behaviour as social behaviour and build up a social bond with technical items such as smart phones, cameras or personal computers (Shibata 2003).

Questions within this construct:

- I consider the robot a pleasant conversation partner.
- I find the robot pleasant to interact with.
- I feel the robot understands me.
- I think the robot is nice.

Specific acceptance factors

As a specific factor, the user's emotional feeling during the robot's performance of scenarios was evaluated retrospectively by means of video analysis. We introduced a grid that allowed us to analyse emotions based on participant behaviour and speech. Several typical emotions such as feeling uncomfortable, amused or interested were described in the grid. Local researchers, who had the same cultural background and therefore could also understand the body language, undertook the video analysis.

Evaluation of prospective impacts

General impression of the SAR approach and its added value

We evaluated the general impression and its added value of the SAR approach by using customized pre-designed questionnaires that included open and Likert-scale questions and linked the results with qualitative results from the thinking-aloud process during the SSUT. Examples of the used questionnaires can be found in ANNEX section 3.

Focus groups with secondary users

Three focus groups with secondary users were conducted during the evaluation of the second prototype (E2). The groups were homogeneous in the sense of a focus group and conducted using pre-defined guides and questions.

- Group 1 consisted of 12 health professionals
- Group 2 consisted of five formal and informal carers
- Group 3 consisted of three therapists and trainers

The evaluation methods for secondary users are not described in detail but referenced as the respective trials were not mainly conducted by the author and are already described in [KSERA2012a].

4.4.4 User group

4.4.4.1 Primary users

The local trial site (the Seniorenzentrum Schwechat) recruited the primary users based on pre-defined exclusion and inclusion criteria. Users needed to be cognitively healthy and physically able to perform the given training exercises in a sitting position. Users with untypically high technological experience, which was assessed using a pre-defined custom-developed questionnaire, were excluded from the study to avoid the well-known bias of convenience sampling, since such “expert users” are more likely to volunteer to take part in technology-acceptance studies and high technologic experience influences the acceptance positively (compare also [Broadbent2009]). Additionally, users had to agree to taking part in the study and signed an informed consent document.

Eight primary users were recruited in Schwechat and took part in the evaluation, with an average age of 77 years (70-95). Eight participants took part in the first trial iteration, six took part in the second iteration (two participants quit due to the time needed to participate and lost interest) and two participants took part in four additional iterations, which were conducted to gain insights into long-term acceptance.

4.4.4.2 Secondary users

Secondary users were recruited by a project partner and consisted of 12 health professionals, five representatives from formal and informal care who were affiliated with the primary users, and three representatives from therapists and trainers.

4.4.5 Test setting

The test setting corresponds with the test setting used for the first evaluation as described in section 4.2.5 and section 3.8 in general.

4.4.6 Evaluation procedure and test flow

The evaluation was conducted in a similar way as during the first evaluation phase to allow results to be compared between evaluation phases. This section therefore mainly describes the differences.

The test-setting was prepared and the functionality of the prototype was assessed based on a protocol. After the users arrived, they again gave their formal agreement and signed an informed consent document. The same questionnaires as during the first evaluation

phase were used for the assessment of the users' current emotional state (PANAS) [Terraciano2003] and QoL [WHO1996].

The main difference between the trials was the selection of scenarios to be presented. Only three scenarios were presented during the first evaluation phase. Within the second phase (E2) the following five scenarios were presented:

- Scenario 1: Environmental information. Within this scenario, the system informs the user about the environmental parameters.
- Scenario 2: Entertainment. During this scenario, the robot entertains the user by playing a piece of music.
- Scenario 3: Medical measurement and scenario 4: Physical training. Within these scenarios, the SAR asks the user to perform a physical measurement (measuring O₂ saturation and pulse) and recommends a physical training based on the received medical data.
- Scenario 5: Video telephony. A video call to a friend is executed (simulated by an experimenter) facilitating a LED beamer mounted on the robot.

As within the first evaluation phase, each scenario was conducted twice to give the user the chance to choose different interaction paths and explore the functionalities of the system. In contrast to the first evaluation phase, the second iteration of scenarios was not undertaken directly after the first but within a block (also referred to as a “free interaction session”). We developed this methodology as a direct outcome of the results of the first evaluation session in order to allow the user to experience the system as it would behave in a real-life setting, including the necessity for voice commands to trigger a scenario, and without the interruption of experimenters between demonstration runs to explain the user story. Within this free interaction session, the participants received a piece of paper with the voice commands the system could understand and were otherwise alone with the SAR in the room. This free interaction session took about 15-20 minutes. Within this session, the navigation of the robot and the speech input were both simulated by a Wizard of Oz technique to provide the flexibility needed. The technical system was in control of all other actions including the HRI (gestures, mimics, voice output, facial detection), the flow of events and the technical communication (video-telephony).

After the “free interaction session”, one experimenter entered the room again and went through the acceptance questionnaires together with the user (which were partly redeveloped after the first evaluation phase).

We selected two users randomly from the trial participants to take part in four additional sessions (summing up to six sessions in total per person) to investigate the influence of experience with the system on the acceptance factors. For the third to sixth iterations, only scenario 3 – medical measurement and physical training – were shown. For these additional iterations the pre-/post-questionnaires regarding the QoL, attitude towards technology and current emotional state were also not conducted to reduce the impact on the participant’s time schedule.

4.5 Summary of evaluation results of the second prototype (E2)

In total, 22 iterations of the SSUT method (8 participants x 1 trial + 6 participants x 1 trial + 2 participants x 4 trials) with eight different older users leading to over 10 hours of HRI recorded material were analysed to generate the following results.

The following sub-chapters 4.5.1 and 4.5.2 provide a summary of the author’s published results from the findings of evaluation phase 2 (E2). Only a small fraction of the study results are presented within this section to allow the reader to follow the lines of methodological development, derive design conclusions and how the individual studies played together and contribute to the summary of findings presented in chapter 8. The results can be found within the publications [Werner2013], [Torta2014] and [KSERA2012a].

4.5.1 Performance results

4.5.1.1 Technical performance

The main aim of the technical performance analysis was to gain information about the feasibility of the technical approach, the feasibility of integration of the prototype into a real environment and an estimation of technical risks for the conduction of future trials at users’ homes. A secondary aim was to gain a detailed understanding about which trial iterations could be conducted without technical issues and which have to be omitted from acceptance evaluation due to possible influences of perceived technical malfunctions by trial participants.

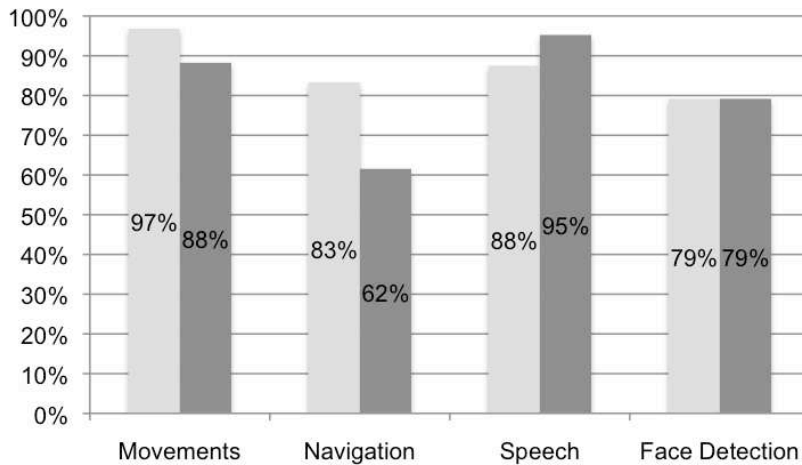


Figure 25: Comparison of performance results between E1 (light grey) and E2 (dark grey), adapted from [Werner2013], the chart presents exact measurements, hence no error bars are given.

Figure 25 shows the percentage of specific functional clusters over all 22 Austrian trial runs (including trial runs that showed malfunctions). Each functional cluster is composed of several related assessed functionalities; in total, 529 functional results from the first prototype and 306 results from the second prototype were analysed and grouped together for visualization.

Within the movements cluster, all of the robot's movements were summarized, including movements for gestures, walking, face recognition (head turning), standing up and sitting down. Generally, these movements worked fine during the evaluation as they were pre-recorded and their conduction only depended upon the main program's correct execution timing. It has to be noted that the score decreased in E2 because of the significantly lower navigational performance. The reason lies behind the higher rate of corner cases due to navigational problems, e.g. the robot not being able to turn the head correctly because it was already misplaced and not directly in the front of the user.

The navigation cluster shows a significantly lower score after the first evaluation because the navigational algorithms were re-developed after E1 to reach a higher performance, which worked fine in the laboratory but showed a lower robustness towards the environmental conditions in the LL. In particular, routines for localizing the user and the Nao often showed a deviation of over 0.5m from the real location, which was treated as a "malfunction".

The speech cluster is composed of word output and speech understandability and shows a slight improvement after E1 which is likely due to a slightly enhanced phrasing of

words (words that were found to be difficult to understand were replaced after E1 and some words were written differently so that the text-to-speech engine correctly pronounced them) and because the volume was turned up.

The face detection cluster performed equally well in both evaluation phases. The performance of the implemented face detection varied, mainly due to different faces and had a particular issues with users that wore glasses.

Influence on the acceptance rates

By comparing the acceptance results of users that perceived technical malfunctions with those that did not, we can give an estimation on the impact of technical problems on acceptance rates. We could not compensate for inter-individual differences and given the small number of participants and the fact that negative experiences in a trial run will likely influence the acceptance of later trials, we can only give very rough estimations. When comparing the acceptance data regarding the intended use of the robot (see acceptance results) between all trials and trials without perceived technical issues, it shows that most constructs perform around 10% lower in the condition including malfunctions. In particular, the construct “anxiety” towards the solution performed 27% lower in comparison, showing that users were worried that the system could harm them because of technical issues. To compensate for this effect, trial runs in which users perceived technical issues were omitted from the data prior to analysis of acceptance factors, if not stated otherwise.

Summary of the functional performance of the system

Despite setting the main goal after E1 to enhance the functional performance of the system within a LL condition, the measured performance in some functional clusters actually decreased. One of the main reasons behind this was the fact that within the development team, researchers whose main interests were to develop new publishable algorithms undertook the development of a prototype that needed to be tested with real users in a close to real-life environment. This resulted in a conflict of interest between scientific excellence and real-life applicability.

Several general technological issues could be found that cannot be solved within a research study but which require research on very specific topics.

Autonomous navigation in real-life environments is difficult to solve, in particular when using a biped robot, and has not been solved sufficiently so far in general. As part of the

problem, the ability of robots to perceive the environment is limited. The SAR would have to recognize the users and differentiate them from static objects and other moving users or animals. Within the environment, objects exist that are hard to recognize by the robot's sensors (as they may be sonar or laser based) such as glass surfaces or very small surfaces like chair legs. Current navigational challenges include moving over steps (such as doorsteps), stairs and carpets, and navigational aspects such as tracking users and navigating through cluttered environments, as was also found in other projects [Payr2015].

The implemented speech recognition was not sufficient and hence replaced by a Wizard of Oz method. In separate experiments, it was found that other commercial recognition engines were also not able to allow a dialog over a typical distance of two meters as the sound quality strongly decreases over distance.

It could be shown that the perception of technical issues by the users had a negative influence on the users' acceptance.

4.5.1.2 Usability

The participants found the system in general easy to learn and easy to use, as is shown by the quantitative results of the specifically developed questionnaire in Figure 26. One user found that the system would not at all be easy to use without help, one was unsure, the rest found it would be very easy to use. All users (n = 8) found it would be very easy to use if provided with a manual.

"If I have it for a longer time, I think I could use it."

The qualitative analysis of users' comments and behaviour during the demonstration showed the following main usability relevant themes:

Understandability. The speech output was optimized after E1 regarding volume and pronunciation, but some users still had problems in understanding the robot.

"Some words are hard to understand."

"He is hard to understand when noise is around, such as a phone ringing."

"The words are loud and clear."

Functional performance. Users that perceived technical issues partly commented on them, which shows that there is an influence on acceptance.

"The responding to commands worked better last time."

“The functionality is currently rather limited.”

“Today it takes him a long time to receive the values.”

Speed. The slow walking speed and the speed it took the robot to find and recognize the user’s face were still annoying to most users. Participants had to wait for 20-30 seconds for the robot to approach their position; many experienced smaller timing delays of 5-10 seconds during the interaction, in particular when the SAR tried to find their face or waited to receive data. Finally, participants had to wait until the robot walked back to its starting position, which again took 20-30 seconds depending on the exact path chosen.

“You have to grow used to how everything is slower.”

“Nao walks first left, then right, then he has to find my face...”

Interaction. The SAR system was only able to react on user commands at specific times, not during the whole interaction. If participants asked the robot during task execution, the system ignored them. This led also to situations in which the SAR interrupted the user by voice output since it could not notice that the user just spoke.

Unclear current state of the system. The SAR did not always inform the user about the current and next planned steps; the information given was also partly vague and followed by a delay of 5-10 seconds, which made the participant unsure as to whether the system was actually still active and what it was going to do next.

“And what do you do now?”

Recall of commands. Around one third of the users spoke the voice commands wrongly, despite reading them from a sheet of paper, and only seven commands were used in total. Most commonly, the name of the robot “Nao” was misspelled (“Noa”, “Nano”). This did not lead to an issue in interaction as the speech recognition was simulated by the experimenters but tells us that the recognition system has to be very tolerant for a real-life use.

“Nano, can you play music?”

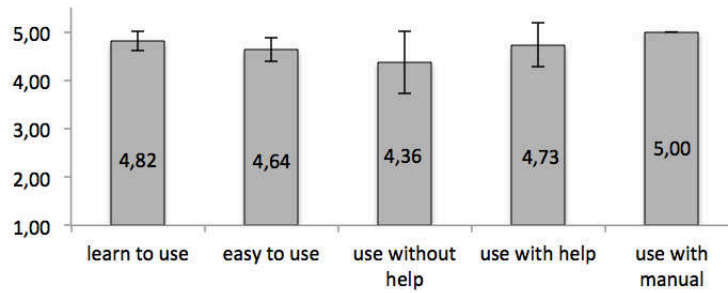


Figure 26: Quantitative results to the custom-developed questionnaire regarding the perceived ease of use. Error bars indicate the standard deviation, data from all iterations of E2 excluding participants that perceived technical malfunctions (score from 1 to 5, 1 = "not at all, 5 = "totally").

Identified limitations of the SAR approach

This subsection discusses the limitations of the approach in general, rather than the limitations of the second prototype.

Robustness. Given the limited robustness, trial participants stated that the system would only be usable if the robustness could be increased to a point where the system becomes completely reliable. Users found this particularly important because the robot provided safety features (medical warning, reminding, environmental warning) on which users would become dependent.

Limited size. The limited size was perceived as problematic for future scenarios, because users wanted to look into the SARs face and therefore had to bend over to communicate with it. One suggestion was to put the robot on a cupboard or next to the bed, although this would impede mobility.

Limited speed. The limited walking speed not only reduces the efficiency of the system, but also limits its general applicability. In an emergency situation, the system would have to react as quickly as possible. It could take several minutes or worse to evaluate the situation by means of the robotic platform's sensors, in particular if the robot is in another room or even on another floor.

Autonomy. The autonomy of the robot was found to be limited because of the limited operational time of around 45 minutes per battery charge. This also impacts the range of operation given the relatively slow walking speed.

Single-user approach. As the prototype system was limited to one user within the room, users noted that this would not be sufficient in a real-life situation where guests could visit, animals could be present or the user's partner may be around.

Limited in- and output channels. We initially defined that speech recognition should achieve a recognition rate of at least 95% to not conflict with usability; this was not even closely achieved during the technical pre-trials. Also, the face recognition was limited to a scenario where the user was sitting on a chair. Users lying on a bed cannot be targeted by the solution as the face would be blocked by the bed.

4.5.2 Acceptance results

Perception of the robot

The perception of the robot by the user was evaluated using the five constructs from the Godspeed questionnaire – anthropomorphism (a), animacy (b), likeability (c), perceived intelligence (d) and perceived safety (e).

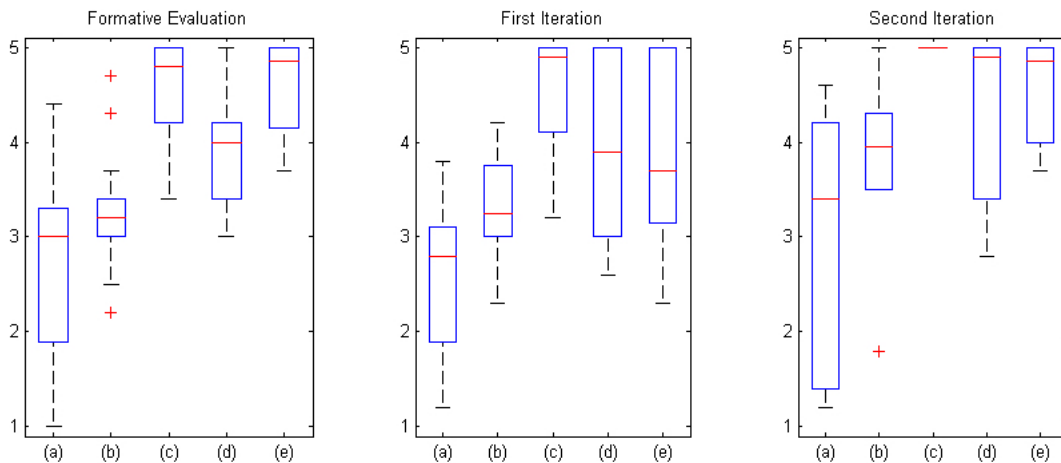


Figure 27: Perception of the robot for E1 (left) and E2 first iteration (middle) and E2 second iteration (Wizard of Oz, right), previously published in [KSERA2012a], all data, including technical issues.

Figure 27 gives a comparison of the used constructs between the first prototype (E1) and the two trial iterations of E2. The data from the two trial iterations were not summarized, to be able to analyse the influence of technical-performance issues on the results. Performance issues were rare in the second trial run, as here the navigation of the robot was not autonomous but controlled by the experimenters.

The overall perception of the robot seems to become better over the prototype iterations. A one-way MANOVA was calculated to determine the effects of the different development phases (in particular regarding the different HRI concepts) on the perception of the robot. Therefore; the constructs of the perception of the SAR were used as dependent variables and the iterations (first iteration of E1, second iteration of

E1, first iteration of E2) as independent variables. The second iteration of E2 was not included in the comparison to avoid a possible bias by measuring the effect of the Wizard's more human-like navigation. The MANOVA revealed no significant difference for any construct $F(10,34) = 0.77, p = 0.68$. Given the insignificant results, univariate ANOVAs were not calculated.

Regarding anthropomorphism (a) users had trouble ranking the robot between human-like and machine-like. Some were consequent in stating that a robot can principally not be human-like, whereas others claimed the system as quite close to a human because of its realistic movements and resemblance to a small child. This is also shown in the quantitative analysis by the wide deviation of results over all trial iterations.

"When he looks at me, I have the impression he understands me."

"For me, robots are a mechanical object."

"The voice is impersonal (mechanical)."

The animacy (b) of the robot was similarly perceived in all trial iterations. From qualitative results, we know that users expected a higher interactivity and in particular a more vivid human-like communication. Because the robot could not understand users' general phrases, it did not respond as would have been expected.

"It is somehow funny how he dangles around."

"It is very different than when talking to a human, I do not know whether he understands me."

"You have to be aware that he does not always answer. I am used to getting an answer when I talk to somebody."

The SAR was perceived to be very likable. Likability (c) was a clear strength of the system, as is shown by the high median of quantitative results in all iterations and by a large amount of related positive qualitative comments.

"I think he is funny."

"I just like him."

"I did not feel uncomfortable being alone with the robot; it was rather funny."

In the autonomous mode, the SAR was perceived to be rather intelligent (d). Within the second iteration, this construct was evaluated much more positively, which is most likely biased by the fact that the robot's navigation was remote controlled (Wizard of

Oz). From qualitative comments during E1 and the first iteration of E2, we know that users were rather surprised about the shown intelligence of the system, but on the other hand, often found the intelligence to not suffice for the system to help in daily situations in the long term.

“Considering he is a machine, he seems to be intelligent.”

“Before I follow his advice, I would think about it.”

“If I really need him, he has to react correctly, otherwise I get grumpy.”

Because of the small size and the weak and slow movements, the system was perceived to be safe (e), the users did not vocalize any concerns about safety and the related quantitative results regarding anxiety were also very positive (median above 4.8 out of 5 in all iterations). Interestingly, the here presented quantitative results of the Godspeed construct from the first iteration of E2 were less positive and show a high deviation. The reason was found to be a single problematic questionnaire item within the construct (quiescent – surprised), which the participants found hard to understand and in this case, was answered likely in error by two participants.

“As long as he is smaller than me, I am not scared.”

Intention to use

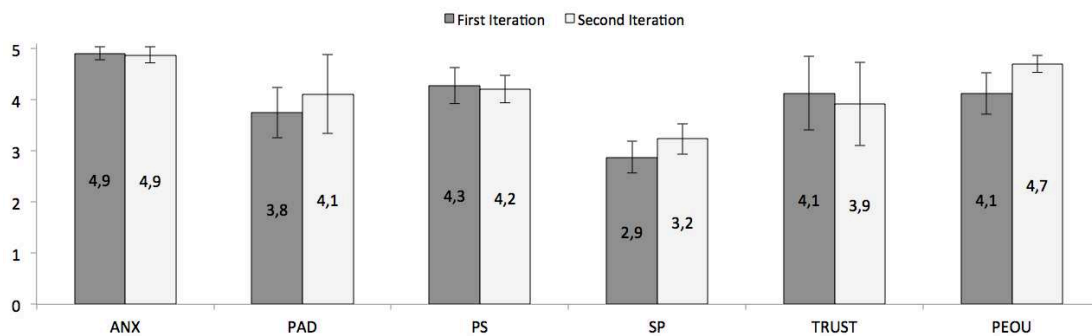


Figure 28: Mean values of the constructs within intention to use during the first and second iteration of E2. The chart was updated from a published version by the same author in [Torta2014].

The analysis of results is presented in detail in [Torta2014], the main insights from the analysis of these short-term trials were:

Participants showed no anxiety towards the solution, which is mostly due to the small size which seemingly made the robot incapable of harming them.

The adaptability was highly rated but also limits such as the maximum volume (which limits the adaptability to compensate hearing loss) were noticed. Given the multi-modal nature of the prototype, participants expected that such issues could be solved in the future.

The factors for perceived sociability were also rated highly, stating that the robot was nice to interact with. From qualitative comments, we know that users found the robot to be sociable and that it could act as a friend.

“For people who are very alone, I could imagine that he cheers them up, that they become friends.”

“I can image, if you are old [...] there are many who talk to puppets...”

Although the usage of a small social humanoid robot seems to be beneficial for acceptance as it lowers the anxiety towards the overall system, we found that the small size also might lower the social presence as users treated the robot more like a stuffed animal or pet than as a human.

The users mainly said they trusted the advice of the system. From comments, we know that this is not related to the appearance of the robot but rather to the participants' reasoning that they expected such a system would be well developed and therefore able to help them in the ways it was designed for.

The participants mostly found that the system to be easy to use, even without help, and easily learnable.

By performing a MANOVA, it was found that the differences between the trial iterations were not significant, despite the SAR being perceived within the second iteration to be slightly more adaptable and easier to use, which is most possibly due to the non-autonomous navigation during this iteration.

By comparing the results of the same constructs over the longer term (six iterations) and fusing them with the qualitative comments of the two participants, we found weak evidence indicating that the PEOU might increase over time when users have gotten used to the control of the robot. Also, a growing relation with the robot might enhance the perceived sociability of the robot. On the downside, participants expected that the PENJ of the system might decline over longer usage as the novelty of robots in general is what makes the solution interesting. The results are limited because they are based on two long-term participants only.

Specific acceptance factors

We measured the emotional feeling right after the test to gain information about the emotional influence of using the system. Figure 29 details the respective quantitative results.

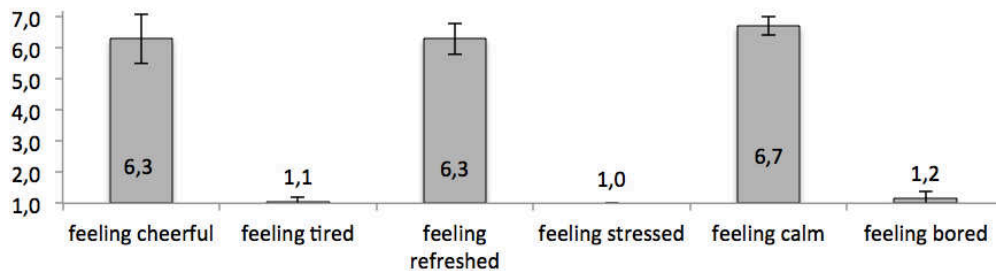


Figure 29: Emotional feeling right after the test on a scale from 1 = not at all to 7 = very much, error bars indicate the standard deviation.

After the test, all users stated they felt rather cheerful, refreshed and calm, nobody felt tired, stressed or bored.

These quantitative results can only partly be confirmed by qualitative data from observations. All observed users of the second iteration of E2 (n = 6) showed signs of boredom during the demonstration, in particular during phases in which the robot walked towards and from the starting position which took around 20-30 seconds each time without any interaction happening. One participant yawned twice when waiting for the robot to walk back after the task, one user took a look at their wristwatch, one participant was clearly frustrated at having to wait so long and also commented on it.

To the SAR: *“Shall we go on now?”*

“You are somewhat slow, right?”

“Have you found my face finally?”

Most participants also showed positive emotions during the demonstration. Two users hummed when music was playing, one bounced and sang a song; most clearly enjoyed the music. Two of the six participants of the second iteration of E2 were clearly positively surprised by the capabilities of the robot.

“Well, now look at this, is that real?” (enthusiastic)

“Wow, you did that really well!” (astounded)

4.5.3 Impacts and added value of the presented solution

We asked participants about their general impression of the system directly after the interaction. From 14 trial runs (eight participants in the first iteration, six participants within the second iteration, long-term participants were not included in the analysis in order not to give those participants additional weight) we received in total 14 comments, of which four can be seen as critical, two were classified as neutral comments and eight comments were classified as positive towards the solution.

One user who commented negatively was concerned that the overall approach is technically not feasible and the maintenance cannot be guaranteed in the long run. Another user thought that the interaction, in particular the speech dialogs, should be improved and one stated it would be a hybrid between a living thing and a machine.

“There will always be errors and it will not work. It will need technical support all the time and can never replace a human.”

“There is room for improvement, it should talk more in a more humanlike voice and less broken up.”

The first statement can be seen as negative expectable impact because the user expects that the complex technical system will have an influence on his time and budget or, in case the system malfunctions in a critical situation, even his health.

Users that gave neutral comments stated that the system would be a funny time-killer, is fun to use but likely will not be used by many people. One participant stated that it might be good if developed well.

“It’s better than having nobody else but you have to be mentally fit to use it, I find it funny but others might ask what they should do with it.”

“It’s good if developed well; it gives the feeling that you can chat with it, the small size is good, otherwise you could be afraid of it..”

The second statement indicates an expected impact on the QoL as it assumes an entertainment value from the SAR itself.

Several users commented briefly and positively about the system. Examples are:

“I see it positively.”

“...a funny happy one...”

“I like the idea very much.”

“Great, it’s exciting.”

Again, the excitement is an indicator of an expected impact. Users find it entertaining and exciting to use the robot which impacts positively on the QoL.

Further analysis on the general impression of the SAR is reported in [KSERA2012a]. In particular, an analysis of added value in comparison with related approaches, the limitations regarding the particular chosen use cases and general limitations of a future at home usage are reported in chapter 3.5.1. As a summary, a condensed result table including the most relevant pros and cons found is provided here.

Table 9: Condensed results regarding the used scenarios from workshops with primary and secondary users.

Scenario	Pro	Contra
1. Environmental information	The presentation of abstracted information was seen very positively.	The abstraction needs to be adaptive to the user’s needs. The analysis whether or not an emergency situation occurred needs to be user and context specific.
3. Medical measurement	Primary and secondary users appreciated the functionality, the general idea and the implementation of it.	The system should be personalized to the user to allow an autonomous interpretation of the values instead of direct reporting. This feature was not implemented on purpose to circumvent ethical issues regarding liability.
4. Physical training	The approach was seen as beneficial by primary users and health and care experts as the demonstrated training movements can easily be understood and followed, also when compared to traditional systems.	The system should be able to analyse the performed training of the user and give feedback and corrections.
5. Video telephony	The idea of having a mobile interface that approaches the user instead of the user having to come to the system was seen positively, as well as the overall idea of including video communication to support the connection with relatives, friends and doctors.	The solution seems not to be preferable over the state-of-the-art (tablet with Skype). Participants stated there would hardly be any free areas in their homes for the mobile beamer to project to. Privacy concerns were stated as the SAR facilitated an on-board camera.

4.6 Summary and discussion of evaluation results

4.6.1 Performance, acceptance and added values

This chapter summarizes all results directly generated from user involvement within evaluation phase 2 (E2).

Performance

The following research questions were targeted:

- RQ_E2_P1: Which general technical issues exist in the second prototype that could negatively influence a later adoption of the system?

It was found that in particular, the slow speed, the limited performance of input channels, and the limited autonomy to 45 minutes of operation have to be tackled in future research to allow later adoption of the system.

- RQ_E2_P2: How do technical issues influence acceptance factors?

It could be stated that technical issues persist and negatively influence the acceptance results. To compensate for that effect, trial runs in which technical malfunctions were perceived were excluded from further analysis of acceptance.

- RQ_E2_P3: How does the system's technical performance compare to the first prototype?

It was found that the technical performance of the second prototype was generally comparable to the first prototype. The technical performance was partly lower or equal, but the functionality was higher given that five instead of three scenarios could be shown to the users and the complexity of scenarios was higher.

- RQ_2_P4: Which usability issues exist in the second prototype?

A number of usability issues were found and reported. The most important issues were related to limitations of the interaction, robot size, walking speed, autonomy and robustness of the system. The general design as a single-user system was also criticized as unrealistic.

Acceptance

- RQ_E2_A1: How do the identified acceptance factors score when using the SAR approach with all implemented functionalities?

Perception of the robot. The user group's opinion was divided when trying to estimate whether the system is machine- or human-like. The robot was perceived as animated with the potential for improvement regarding the interaction capabilities where users found that important functionality is missing. The SAR was perceived to be very likable, funny, safe and rather intelligent.

Intention to use. Participants showed no anxiety towards the solution because of the small size which, on the other side, likely decreases the perceived social presence. Heerink's Almere model suggests that this is a negative effect towards acceptance; however we think it might also have a positive side as a solution with lower social presence might have a lower impact on ethical concerns. Participants found the system could act as a friend and trusted the system. Generally users agreed that the system would be easy to use and might become easier after they had gotten to know it better. On the other side, they expected the PENJ to decline over time because of novelty effects.

General impression

Most trial participants found the presented SAR to be beneficial as an assistive solution. Critical comments were received regarding the high technical complexity which might lead to short maintenance intervals, the interaction capabilities of the robot and that the system seems to be entertaining but not so useful as an assistive device.

- RQ_E2_A2: How do acceptance factors compare between the first two prototypes?

We found that there seems to be slight improvements in particular with the factors of PENJ, PEOU and PS, but all changes are statistically insignificant, meaning that the acceptance and thereby – following the Almere model – the intention to use the system in the future is similar between the two PT1 and PT2 prototypes.

- RQ_E2_A3: Which general limitations exist which could hinder a later usage at home?

We found a number of general limitations of the approach. The robustness of the system is low and one participant was not convinced that the system would be robust enough in practice. The size and walking speed was seen as critical in terms of usability and functional capabilities. The current autonomy of the robot of 45 minutes' operation on one battery charge was seen as low, also because no charging station was included. The

single-user approach was seen as unrealistic and the interaction capabilities were seen as limited in terms of functionality and performance.

- RQ_E2_A4: Which limitations exist concerning the different implemented user-need domains?

We found that the environmental information functionality should be customizable to the user and react specifically to context. Similarly, the system should adapt to the user regarding the medical warning and therefore should incorporate user detection. The physical-training scenario was found to be limited because the system was not able to observe and correct the user in case he/she performed the exercises wrongly, which in a worst-case scenario might lead to injuries. The video-telephony scenario seemed not to be preferable to current alternative technological solutions such as a tablet.

Added values

Direct impacts could not be measured due to the early stage of the prototype and the corresponding methodologies involved. Based on the users' qualitative statements, it can be summarized that users do not expect the system in its shown version to become a product due to technical issues and the presumed technical support needed. If technical issues can be fixed, users understood the system rather as an entertainment device than a real help for daily life and care, and hence estimated the impact on their daily life and work to be low.

4.6.2 Lessons learned regarding the methodological approach

Questionnaire changes. Parts of the used acceptance questionnaire were developed for E1 and have now been reused in order to be able to compare results between the prototypes. This also includes factors that are similar to those included in Heerink's Almere model, such as the factor of PENJ. In an effort to reduce the time needed by the participant to answer questionnaires after the test, we did not ask both our questionnaire and the full Almere model to be completed. Given that the Almere model is used by several other research groups and the experiences with the factors from the Almere model are therefore constantly increasing, we decided to drop our own efforts in further developing the respective parts of our custom questionnaires and incorporate the full Almere model in future evaluations.

Observation bias. Participants were informed that they were being watched and recorded within the informed-consent procedure. This seemed to have an effect during

the trials to which some users made reference (“*Would it work if there wasn’t three people outside controlling the system?*”), because they seemed to behave in an unrealistically patient and tolerant way to technical issues. In this sense, the invited users seemed to play a role as test participants in which they tried to actively provide us with input such as comments and feedback with the goal of helping to develop and enhance the system. For these reasons, we expect that the acceptance results are positively biased in comparison to a hypothetical real-life usage.

Test flow. After the demonstration, users were guided through the prepared questionnaires which also included open questions. In fact, much information could be gathered because users often provided us with their reasoning behind their chosen answers and by analysing the open questions. After the demonstrations, users were very willing to talk about their experience with the robot, partly because for them, taking part in the studies must have been an exciting event. It could be wise to not directly start with an open question followed by closed questionnaire items after the demonstration, but rather foresee around 15 minutes of a semi-structured interview to gather more qualitative data regarding the user experience.

Mixture of methods. The mixture of methods and gathered data quality (qualitative and quantitative) at this stage of research worked well and was in our view crucial to gain valid results for the following reasons:

- The qualitative data were used to make the quantitative statistics interpretable. Without the qualitative data, it would have been impossible to gather the meaning behind some questionnaire results. As a particular example, most literature suggests that a high level of social skills by the robot is preferable for acceptance. We found, on the contrary, that the social skills should not be as high as possible but rather match the functionality and appearance of the robot. Hence the quantitative scale alone does not give any meaning, as it is unclear what the target value could be.
- The quantitative data provide an overview on the measured factors and their distribution within the group of participants, whilst qualitative data provide deeper insights into specific aspects. Because of this quantitative overview, we were able to cognitively step back during the analysis of qualitative data and keep the overall picture in mind. This seems to be an important factor as we think we likely would have put too much meaning into single user statements.

- As intended, we could validate the results by comparing the outcomes of different methods. We also received differing results from different methods showing that this form of validation is crucial for the quality of insights gathered.

Lessons learned from implementing the “free-interaction session”

Although the participants were asked to and able to interact freely with the robot, all of them simply used the set of commands to trigger each functionality one after the other. None experimented freely, e.g. by moving around in the room, interrupting the robot, ignoring the robot or touching the robot. Most participants called the commands even in the order they were presented on the sheet. Consequently, the main difference to a regular SSUT evaluation was that no experimenter entered the room between demonstrations to introduce the upcoming scenario and so the user was not provided with the intended context of use.

For technical reasons, the robot had to start a new scenario from a particular starting point. This led to the impression that the user is not interacting with the system in one stretch but within several smaller phases that were interrupted by the robot moving back and forth to the starting position. This presumably gave the user the feeling to try out different functionalities that otherwise would have been distributed around the day. We assume this makes a difference regarding the ecological validity of the test because users had to imagine how the system would work in daily practice instead of immediately experiencing it. The aim behind the free-interaction session was to increase the ecological validity of results.

For this reasoning and because the users could not be introduced to the context of use, we do not see the ecological validity of this form of trials to be generally preferable to regular SSUT trials.

4.7 Heuristics for further design and development (design principles)

Based on the presented results, experiences and lessons learned, we can build a set of heuristics to guide future developments of SARs. These heuristics can be understood as a SAR-specific add-on to the well-established design heuristics presented by Jacob Nielson, which describe similar aims but were not developed with a SAR in mind [Nielson1994].

Role and personality of the robot

For E1 and E2, a friendly, helping, caring robot was designed. The acceptance results tell us that users liked the attitude of the robot which is in line with earlier results [Heerink2008b], but some participants also expressed concerns that over a long-time experience, users might get attached to the system and treat it as friend, leading to ethical issues if users confuse the system with a real person and expect similar emotional and social capabilities that the system cannot provide. Also, we know from earlier studies [Goetz2002] that despite users preferring a friendly and extrovert robot, when it comes to compliance, users rather follow the advice of a strict robot. Additionally, we know from presented qualitative results that most users saw the system rather like a tool that should competently support them in case of need or emergency, rather than as a friend who they would expect to chat with. Here SARs are different in regard to their application area to purely social robots, such as those used in therapy of older adults with dementia such as PARO [Wada2007] and should not be confused. Social robots are used particularly because of effects of emotional attachment and should not replace communication with human therapists but facilitate it. Hence, they should (and, as far as we know, also are) only be used in therapy sessions together with therapists, not like SARs at home.

As a conclusion from the reasoning explained above, we can create a heuristic for future designs of personalities:

H1: The robot's personality should be designed in a strict, precise and polite, positively motivational and functionally oriented manner that complies with the role of the SAR and resembles its function as a tool, not as a companion, servant or master, in order to generate an acceptable solution and minimize the chance of emotional attachment.

Form follows feasibility

We found in both evaluation phases that the interaction capabilities of the used SAR solution do not fully meet the users' expectations. The expectations are exceptionally high because a humanoid platform was chosen, suggesting human-like interaction capabilities. In this respect, the appearance of the robot leads to particular expected behavioural and functional capabilities. Naturally we should aim to develop systems that do not trigger unrealistic expectations and must already consider this during the design of the robot or when choosing the robotic platform respectively. Our recommendation for future designs of SARs regarding the appearance therefore is:

H2: The appearance of the robot should match the currently limited technical capabilities in terms of functionality and interaction design (form follows feasibility).

HRI follows form

In the current case, an anthropomorphic platform was chosen, particularly in order to become able to conduct HRI research, assuming that technology will one day be able to realize HRI in a suitable way by resembling human-human communication. For this case, it is necessary to target a sophisticated HRI experience that comes at least close to what users expect in terms of interaction to achieve user satisfaction.

H3: The behaviour, personality and interaction capability of the robot should match the appearance of the robot.

In cases of an anthropomorphic platform as used within this dissertation, users expect a fluid multi-modal interaction that follows the human-human model and includes well-timed solutions for making and breaking eye contact, understanding and simulation of non-verbal gestures, mimics and turn taking.

Performance over functionality

To become technically able to realize such an advanced HRI, it seems advisable to reduce the functionality to a minimum and focus on a single use case. Within such a single use case, complex technology that is currently above the state-of-the-art, such as safe autonomous navigation, might not be needed to realize a scenario that (from the users' view) makes sense and is assistive. Such a use case should provide in particular a minimal reliance on input channels, as technical solutions here are mostly not reliable enough for a real-world usage in safety-critical applications.

H4: Focussing on a single use case will lead to a more robust system, implying higher acceptance and lead in the long run to a working set of use cases that could be combined to achieve a multi-purpose SAR system.

Adaptive to changing user needs

Users expressed concerns that the PENJ of using the system might decline over long-term use due to the repetitive nature of the implemented interaction. Also, it is well known that users' needs change over time as they recover from a sickness or become more dependent on personal help due to worsening age-related conditions.

H5: A system that is targeted for long-term support at users' homes has to be adaptive to changing user needs and requirements.

Target individual user needs

Within the conducted evaluations, a single-user setup was assumed, partly in an attempt to enhance the technical performance. This was noted by participants and seen as unrealistic as many users do not live alone and, if they do, they still get visits from friends, relatives and neighbours.

H6: A SAR system should be able to recognize the user and adapt its functionality with respect to the currently active user.

Express what will happen

During the evaluation, situations often arose in which the user was not aware of why the robot behaved the way it did. This is partly due to the fact that it takes time for robots to move and during this time, the user might wonder what the robot is up to. Users expect to be informed about the current state and what will happen next. This information can be given using either of the output channels; hence in our case, by using gestures, mimics, body posture, movements, sounds or vocally.

H7: The SAR system should take care that the user is informed about its current state and planned next steps.

Efficient, speedy and well-timed interaction

Because robots move and movement takes time, the speed of interaction can easily suffer by implementing position changes and gestures. In all evaluation phases, we could see that the slow timing annoyed participants as it made task completion inefficient.

H8: The interaction with the robot should be efficient and free from delays.

5 Concept of a SAR for physiotherapy and refinement of the methodological framework

5.1 Basic concept and idea

Physical training is a commonly prescribed therapy, both for rehabilitation and the prevention of physical deficits. It is known to enhance mobility and increase the independence of patients. The success of this therapy depends largely on the motivation and training competence of the patients, which in daily practice varies strongly between individuals along with their training schedules .

As a basic idea, humanoid SARs could enhance both the training quality and quantity by means of their unique robotic abilities. As the base for a successful training is a regular, efficient and independently conducted training schedule, strategies to enhance motivation lead to an increase in both the quality and quantity of the training, which implies an increase in training efficiency. Given that SARs have the potential to motivate users by means of their HRI abilities, they could be used for motivational support and thereby enhance the training efficiency.

Furthermore, the humanoid robot Nao in particular is capable of performing human-like movements that also facilitate the demonstration of training-exercises, similar to how this is done by training instructors. If such a system is able to perform the training exercises in a way that allows the users to conduct efficient training by mirroring the behaviour of the robotic-training instructor, the training competence can be enhanced, leading to an increase in training quality.

As an additional feature, such a system could also measure the user's performance of the conducted training, give feedback and thereby motivate and allow the user to correct the training movements which also can positively influence the training quality.

The goal of this study was to evaluate the applicability of SARs for use within ICT-based physical training of older people at home. A prototype was created by using state-of-the-art HRI techniques and optimizing them based on the experience from earlier studies with the same robot and for the use case of physical training at home.

Older women and men with an age of at least 65 years who were still able to live independently at home were included in the study. Since reduced functional, physical and cognitive abilities are common within this age group, seniors with limited mobility,

balance and cognition were also included in the target group if they were still able to perform simple physical exercises on their own.

5.1.1 Background regarding physical therapy

Ageing comes along with a structural decline of the neuromuscular system, including the postural stability, mobility and strength. The decline already begins from the age of 25 and becomes typically apparent at the age of around 40, as for a long time it can be well compensated by experience and physical reserves. In particular, a decline of balance control, general and specific mobility as well as strength of the lower extremities is typical during ageing. These limitations over time lead to a higher risk of bone fractures, falls, dependencies during conduction of activities of daily living (ADLs) and more frequent hospitalization; compare also [Guralnik1994], [Guralnik1995], [Myers1996], [Tinetti2003].

The ability to conduct basic ADLs such as washing, eating, dressing and walking is essential to live independently at home. Physiotherapeutic training can be used as a means to prolong the ability to conduct these basic ADLs and hence works towards a more independent living at home, leading to a higher QoL among the older population and reduced care costs for society [Harada95].

Significance of physiotherapeutic training at home for the wellbeing of older adults

In a study with 50-65 year old persons, King et al. [King1991] found that training at home is as effective as group training with a trainer. Over a longer duration of 12 months, the at-home group even profited slightly more than the guided training group. In a similar study, Helbostad et al. found that daily autonomous training at home has a positive effect on the functional abilities of the participants, whereas additional group training showed no added effects. Further, the same authors found that the effects of three month of training wore off within the following six months without training [Helbostad2004].

Another study [Schwenk2008] compared the effects of physiotherapeutic training at home in a between-groups design, split between older users with and without dementia, and found that both groups profited from the training, though the positive effects in the dementia group only lasted over the longer term if the group members received regular and repeated guidance in conducting the exercises.

Courtney et al. investigated the effects of an intense post-rehabilitation support at home [Courtney2009] and could show that an individualized physical training in combination with intense home care for 24 weeks after hospital leads to a significant reduction in hospital readmissions.

A meta-study [Sherrington2008] investigated the effects of physical training on the risk of falls of older people and found that physical training can lead to a reduction of falls. The study recommends training balance and a general-intensity physical-training program. However, Littbrand et al. found that the intensity of the training program has little effect on the outcome regarding the enhancement of independent conduction of ADLs [Littbrand2009].

Requirements and issues of undertaking physiotherapy

In order to maintain the physical and mental health status and prevent a decline due to ageing, physiotherapy has to be conducted regularly. Because physical training is exhausting and takes time, people typically do not pursue the training goals over the longer term. Hence, motivation is an essential part to enhance the compliance and thereby success of physiotherapy.

Motivation is composed of extrinsic and intrinsic factors. Extrinsic motivation can be generated, for example, by praise, aggression, rewards or music. Intrinsic motivation is defined as resulting from one's own feelings and thoughts and can be generated by individuals themselves, e.g. if an activity is perceived as fun, meaningful or challenging.

Regarding ICT-based training the following motivational factors are relevant:

- Expectation of success: If an individual has a high expectation on the benefits of a training program, he or she will be motivated to conduct the training.
- Rewards for the conduction of physical training: The positive effects of physiotherapy typically take time to be recognizable, hence it can help to offer other rewards such as virtual points or new training exercises.
- Feedback on the training progress: Correct and plausible feedback on the training progress can have a positive effect on the motivation in the same manner as a reward. Feedback on the success of training (e.g. by visual screens showing reached points/levels) makes the increase in skills easily graspable by the user and so enhances the user's motivation.

5.1.2 Design principles

During the design phase of the project, an expert workshop was conducted with a care staff member of a senior-citizen centre, a physiotherapist and a motivational trainer. Within the workshop, the design of robot-supported physiotherapeutic training was discussed regarding general aspects and the content of the training, including training exercises and HRI aspects.

Training procedure: The experts found the robotic-training aid could best help with therapeutic work if the robot supports the training at home between regular training sessions with the therapist. This gives the possibility to potentially correct training exercises which are undertaken wrongly and to give feedback by a therapist in addition to the feedback of the robotic system.

During the home-training session, the experts found it could be helpful and motivating if the robot initiates the contact with the user and prompts him or her to perform the training. This could already start in the morning by coming directly to the bed and asking the user to perform easy wake-up movements within a morning routine.

Training exercises: The group found there are a number of around 10 simple training exercises that could easily be a starting set for a robotic support system as they can be undertaken by a large number of users without high risks. Since the exercises are well known by most users, they can also be used as a form of familiar introduction for the users which will then lead them towards more specialized and possibly harder ones at later stages. One expert found that “sitting gymnastics” could be part of the training as they work very well with the target group in her experience, since many mobility limitations are circumvented. In general, it was strongly recommended to use music when possible for motivation and to make the training more fun.

5.1.3 The prototype system

The prototype for E3 was developed within the nationally funded project “PhysicAAL” and comprised of the humanoid robot “Nao” from *Aldebaran* (now *Softbank*), a product-grade *Microsoft* Kinect optical sensor,⁴⁰ and a server system that includes the software for steering the robotic prototype and analysing the movements of the human user.

⁴⁰ <https://developer.microsoft.com/de-de/windows/kinect>

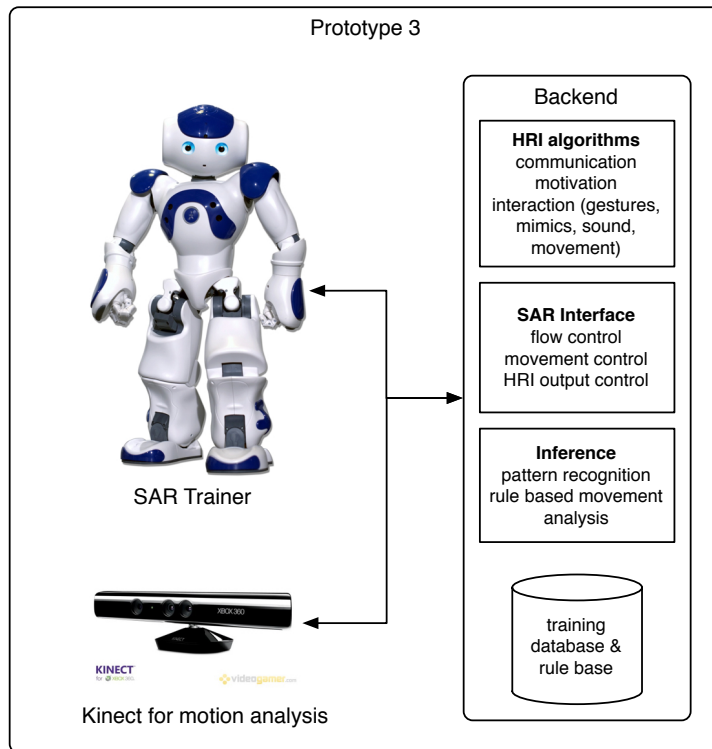


Figure 30: Overview of the 3rd prototype, adapted from [Werner2013b].

5.2 Evaluation model

The evaluation model was taken from E2 and adapted based on the experiences with E1 and E2 as well as the technical requirements of the new prototype (PT3). Given that the main functionality of the system is to support physical training, the effectivity from a therapeutic view was included as an evaluation factor. Based on the lessons learned in E2, we implemented the whole set of acceptance factors from the Almere model, the corresponding individually developed evaluation factors were discarded accordingly. Due to time constraints within the project, a long-term evaluation could not be implemented.

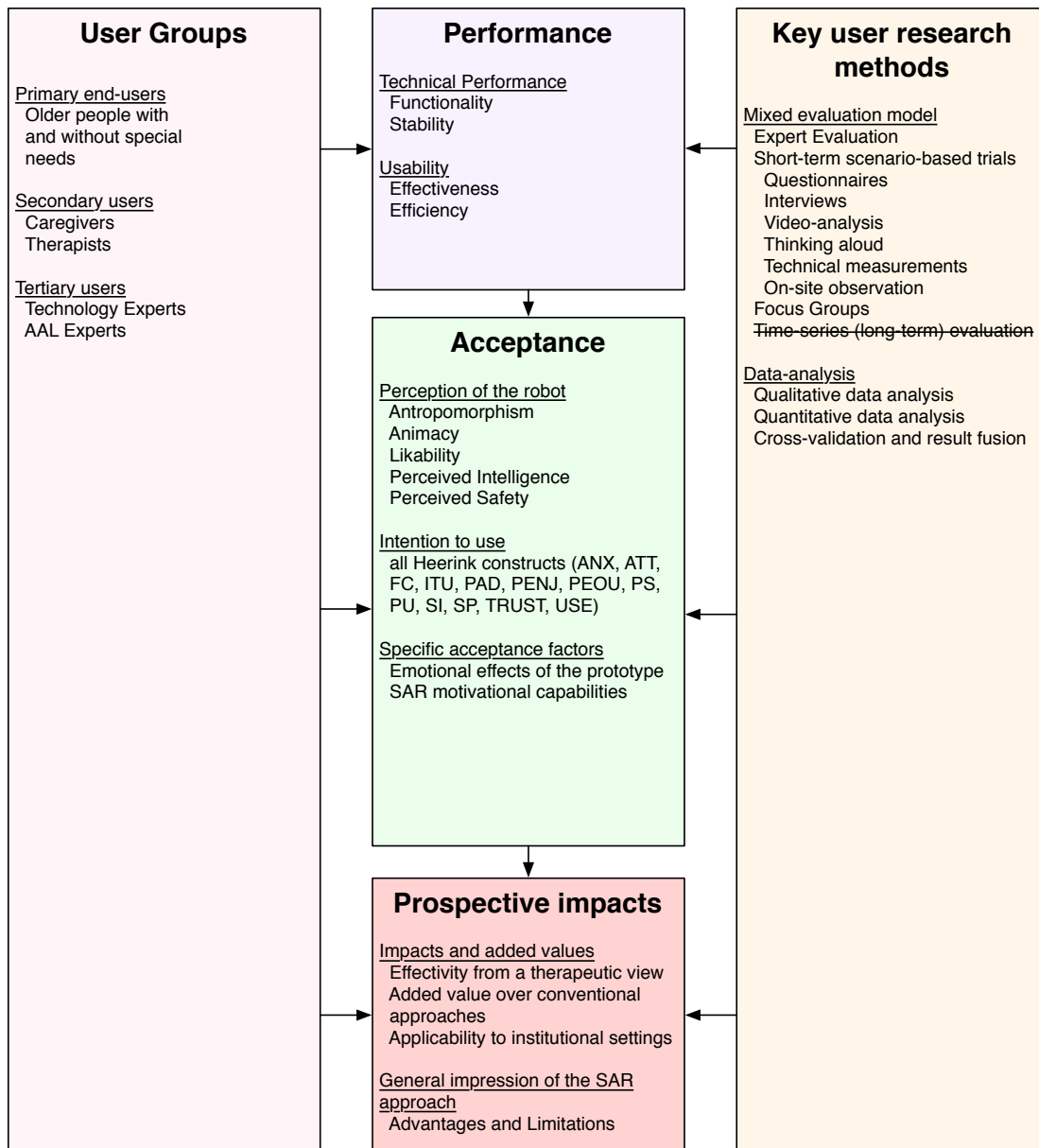


Figure 31: Evaluation model for the evaluation of a SAR system for physiotherapy support within E3.

The following Figure 32 details the methods used within the third evaluation phase. It lists the evaluation domains and their subdomains (as also depicted in Figure 31) and gives the evaluation factors per sub-domain. The used methods are linked with evaluation factors in order to display how different evaluation methodologies were fused and triangulated to gain insights into the individual factors. Data qualities are displayed and were used to check whether different data qualities can be obtained for factors to enhance the quality of results presented.

Evaluation (sub-)domains	Evaluation factors	Methods							Data quality			
		Questionnaires	Interviews	Video-analysis	Thinking aloud	Technical meas.	Observation	Group Discussion	Qualitative	Quantitative	Objective	Subjective
Performance												
<i>Technical performance</i>	Functionality	x		x		x	x			x	x	x
	Stability	x		x		x	x			x	x	x
<i>Usability</i>	Effectiveness	x	x		x					x	x	x
	Efficiency	x	x		x					x	x	x
Acceptance												
<i>Perception of the robot</i>	Antropomorphism	x	x		x					x	x	x
	Animacy	x	x		x					x	x	x
	Likability	x	x	x	x					x	x	x
	Perceived Intelligence	x	x		x					x	x	x
	Perceived Safety	x	x		x					x	x	x
<i>Intention to use</i>	Intention to use (13 constructs)	x	x		x					(x)	x	x
	<i>Specific acceptance factors</i>	Emotional effects of the prototype	x	x	x	x					x	x
Motivational capabilities		x	x	x	x					x	x	x
Prospective Impacts												
<i>Impacts and added values</i>	Therapeutic effectivity	x	x	x						x	x	x
	Comparison with non-robotic aids	x	x		x					x	x	x
	Applicability within an inst. setting			x						x		x
<i>General impression of the SAR</i>	Advantages and Limitations								x	x		x

Figure 32: Evaluation matrix detailing the used evaluation methods and how they interlink with evaluation domains.

5.3 Evaluation factors and research questions

The following section describes the evaluation factors and research questions used within the main trials in detail. Pre-trials were conducted to evaluate the technical robustness using a subset of research questions, as reported in section 5.4.2.

5.3.1 Performance

The following research questions were used to assess performance and usability:

- RQ_E3_P1: To what extent is the system in the current state usable; what usability issues exist?
- RQ_E3_P2: Is the system able to perform correctly from a technical viewpoint under real-life conditions?
- RQ_E3_P3: What are the technical limitations of the approach?

The performance assessment was undertaken as described in section 3.6.1 and 5.3.1; the score sheets for the experimenters were adapted to the new test cases.

5.3.2 Acceptance

The following research questions were used to assess acceptance factors:

- RQ_E3_A1: Is a robotic training assistant accepted for regular training at home by older users?
- RQ_E3_A2: How do users feel right after training with a socially assistive humanoid robot? What emotional influences can be expected?
- RQ_E3_A3: To what extent does the system motivate users to perform the training?

Perception of the robot

To assess the perception of the robot, the Godspeed questionnaire [Bartneck2008] was implemented, as described within E1 in section 4.2.3.

Intention to use

The comprehensive full Almere model was implemented, as discussed within the evaluation results and implications for further methodological development in E2. The model comprises elements of usability, social and environmental factors and hence can be seen as the core of acceptance evaluation within E3.

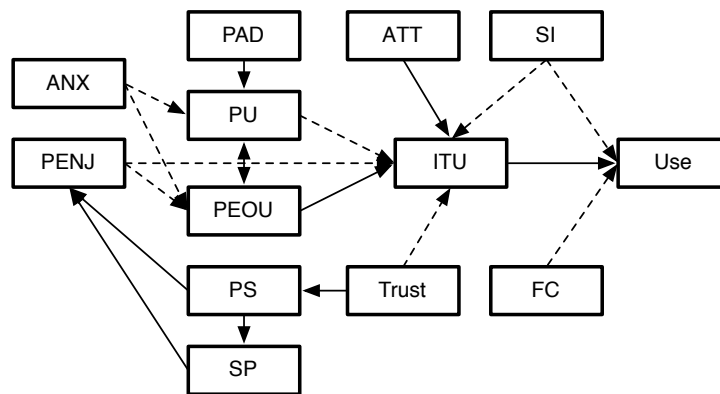


Figure 33: Full Almere model with its constructs and their links to predict the use of a system.

The following table gives an overview of the Almere model constructs used for the subjective evaluation of acceptance of the SAR system.

Table 10: Overview of evaluation factors used during the evaluation of the intention to use the system, based on [Heerink2010].

ANX	Anxiety	Evoking anxious or emotional reactions when it comes to using the system
ATT	Attitude	Positive or negative feelings about the appliance of the technology

FC	Facilitating conditions	Factors in the environment that facilitate use of the system
ITU	Intention to use	The intention to use the system over a longer period of time
PAD	Perceived adaptability	The perceived ability of the system to adapt to the needs of the user
PENJ	Perceived enjoyment	Feelings of joy/pleasure associated with the use of the system
PEOU	Perceived ease of use	The degree to which one believes that using the system would be free of effort
PS	Perceived sociability	The perceived ability of the system to perform sociable behaviour
PU	Perceived usefulness	The degree to which a person believes that the system would be assistive
SI	Social influence	The person's perception that people who are important to him/her think he/she should or should not use the system
SP	Social presence	The experience of sensing a social entity when interacting with the system
TRUST	Trust	The belief that the system performs with personal integrity and reliability
USE	Use/Usage	The planned use of the system over a longer period of time

General impression

We evaluated the general impression of the SAR approach by using customized pre-designed questionnaires that included Likert-scale questions during the SSUT. The questionnaire items are presented when reporting the results. The general impression was evaluated with the primary target group during the pre-trials and with secondary users during a workshop concluding the main trials.

Specific acceptance factors

As specific acceptance factors, the emotional feeling towards the SAR solution and the motivational capabilities of the system were assessed. The motivational capabilities are especially important considering how the system aims to motivate users to perform training exercises and hence strongly influence the potential impacts of the system.

Specific acceptance factors were evaluated by using customized pre-designed questionnaires that included open and Likert-scale questions and linked the results with qualitative results from the thinking-aloud process during the SSUT. Examples of the used questionnaires can be found in ANNEX section 4.

5.3.3 Impacts and added values

The following research questions were used to assess impacts and added values:

RQ_E3_I1: What added value does the solution provide over currently used, similar, non-robotic training aids?

RQ_E3_I2: Is the SAR system effective from a therapeutic perspective?

RQ_E3_I3: Could the SAR solution be integrated into current institutional care?

Regarding RQ_E3_I1, we incorporated a comparative experiment using four different technical training aids as an independent variable and the acceptance factors related to these systems as a dependent variable. A custom questionnaire was developed for comparison of the systems including Likert-scale and open questionnaire items. The questionnaire can be found in ANNEX section 5.

Regarding RQ_E3_I2, we asked two physiotherapists as part of the secondary user group to analyse the video-recordings of the conducted 12 SSUT sessions.

The physiotherapists evaluated the following factors quantitatively using pre-defined custom-created questionnaires and by commenting on the same questions qualitatively.

- Understandability and correctness of the system’s demonstration of exercises.
- Soundness of the corrective training feedback that was issued by the system during and after the training.
- Performance of the user during the training.

Regarding RQ_I3, an interview with two secondary users, both institutional carers with experience in the physical training of older users, was conducted with the aim of gathering their views on the usefulness of the developed PT3 solution for the support of physical training in an institutional setting.

The prototype was explained to the participants in the same way as to the members of the primary user group. In contrast, only one participant was able to experience the physical-training scenario by simulating the user, the other one observed the demonstration. A structured interview including open questions was conducted after the training.

5.4 Evaluation methodology

The evaluation can be split into three parts: the laboratory evaluation, a pre-trial with a group of users and the main evaluation phase which included SSUT sessions with all recruited primary users (n = 12).

Table 11 gives an overview of the methods used and dimensions studied within the three phases.

Table 11: Methods used and evaluation domains studied within the three phases of E3.

Research phase	Methods and techniques	Studied dimensions
Laboratory validation	– Lab tests	– Technical performance
	– Usability checklist	– System safety
Pre-trial	– On-site observation	– Usability
	– Direct observation, questionnaires, on-site interviews	– Selected acceptance factors
		– Technical performance
Main evaluation phase		– Usability
	– Short-term scenario-based user trials (SSUT)	– Perception of the robot
		– Intention to use
		– General impression of the SAR approach
		– Specific acceptance factors
		– Impacts and added values

The individual phases are described in the following sub-sections.

5.4.1 Methodology of the laboratory validation

To validate the SAR system and ensure the technical performance during the main evaluation phase, a laboratory validation was conducted. The focus of the laboratory evaluation phase was put on usability issues, system safety and technical robustness. The laboratory evaluation was conducted without external participants. Mainly black-box testing of the system’s individual functionalities was undertaken. Additionally, experts within the project team performed a heuristic evaluation. As the intention was mainly to verify the system’s performance for later user trials, results are not provided in this dissertation but the inclusion of such laboratory pre-trials is essential from a methodological point of view to ensure the system’s robustness during user interaction.

5.4.2 Methodology of pre-trials

Within a pre-pilot study, a group of 14 (n = 14) seniors from a visiting senior citizen club was invited to a gymnasium in the Schwechat senior-citizen centre for a 15 minute demonstration of the prototype robotic system. The group had an average age of 69 years and consisted of seven women and seven men. After an explanation of the system, the robotic trainer demonstrated physical exercises and the older users mimicked the movements. A specifically tailored questionnaire was asked directly after the demonstration and a focus-group session was conducted to generate qualitative results.

The questionnaire, composed to get a comprehensive overview of the seniors' impression during the pre-pilot study, was comprised of 25 questions in five sections within the categories:

- Motivation and training support
- Usefulness
- General impression
- Characteristics of the robot trainer
- Training preferences

Test setup

The pre-trials were conducted in a gymnasium within the “Living Lab Schwechat”. The gymnasium was not specifically prepared to fit the needs of the technical system; only the SAR system's components were integrated into the gymnasium and tested beforehand.

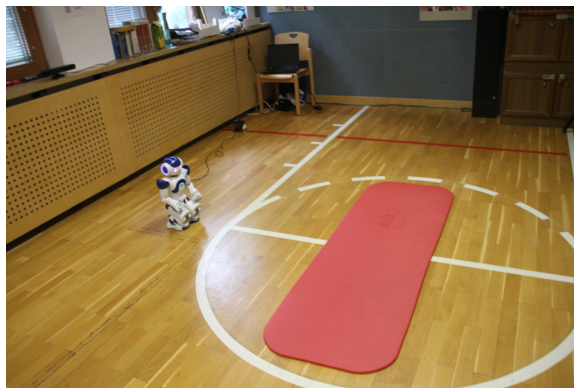


Figure 34: Impression of the gymnasium used for testing.

The gymnasium, which is depicted in Figure 34, was mostly empty when used. Chairs were placed within one third of the large room to create a demonstration arena similar to a conference setup. The technical equipment could not be hidden as within the LL room but was barely recognizable due to the small size and the location within a corner of the large room (in the image top centre). The environmental conditions could be controlled in the same manner as described above. The third version of the system used here was already capable of performing autonomously after a single press of a button, hence no technical experimenters were needed to control parts of the system and neither camera equipment for recording the trials nor an external setup for experimenters was necessary.

5.4.3 Methodology of the main user trials

5.4.3.1 User group

The criteria to take part in the user trials have been being at least 65 years old, cognitively healthy and physically able to conduct at least parts of the training exercises. Users were recruited by calling through a list of seniors living in Schwechat and seeking for diversity in age and technology affinity. The users had to agree to the audio and video recording of the trials and signed an informed-consent document (see also annex section 7). 12 persons (n = 12) with an average age of 74 years (66-92) took part in the trials; this included three men and nine women.

5.4.3.2 Trial setup

The trials took place in one room of the Schwechat senior-citizen centre in Austria. The room was equipped with furniture typical to that in a local living room to simulate a real user's apartment and with the SAR prototype system to become the test environment. Figure 35 gives a topographic view on the (since E2 adapted) test environment that now hosts a workout area to conduct the physical training.

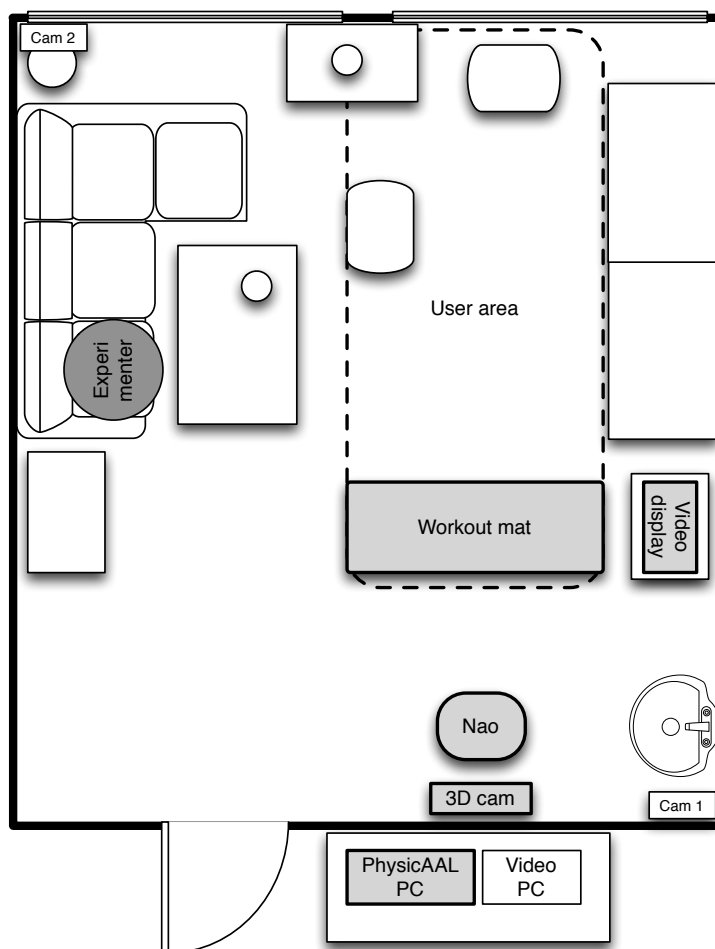


Figure 35: Top view of the test environment at the Schwechat senior-citizen centre. SAR system components and test components are shown.

The test environment was equipped with the following components:

- 3D camera for motion analysis and exercise evaluation
- Nao robot for exercise demonstration, feedback and motivation
- Workout map for exercises in a lying position
- Video display to show alternative ways of training support such as video-based training support and Wii-based training support.

The marked area shows the position of the user during the demonstration of exercises. The possible position is restricted due to the visual coverage of the 3D camera used for motion analysis.

The robot remained static regarding its position during the demonstrated scenario.

Three cameras and two microphones were used to record the trials. The camera placement is shown in the map as indicated with the labels *cam 1* and *cam 2*. The third

camera source was the *Microsoft Kinect*. Figure 36 shows a screenshot of the experimenter interface including the view from three perspectives and the trial site in the “Living Lab Schwechat” in the Schwechat senior-citizen centre. The video stream from the Kinect camera also contains an overlay of the recognized user position as technical feedback for the experimenters. The picture also gives an impression of the real-life condition (concerning furniture, arrangement, light and sound conditions).

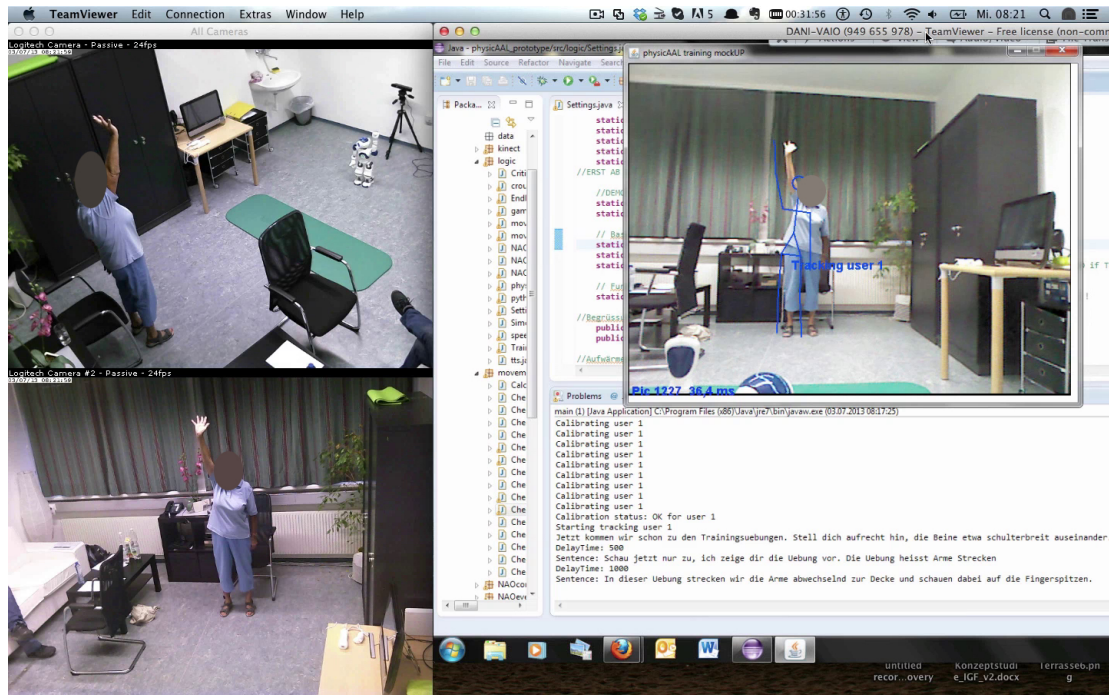


Figure 36: User interface for the experimenters as used during E3.

5.4.3.3 Test procedure and test flow

Preparations (no user present)

The system was reset to the same initial state prior to every test session. The observer prepared the documents needed to conduct the session (informed consent, questionnaires, interview script). The whole setting was predisposed to welcome the participant.

Trial-participant reception

In this phase, the project and the goal of the end evaluation were explained to the participants, their role during the tests was discussed; in particular, the briefing included instructions to interact with the system (e.g. input modalities, expected output) and instructions to allow researchers to collect data (e.g. thinking aloud, observer’s

role). Finally, the formal agreement of the participant was given within the informed-consent document which was composed of a general explanation of the project that included the idea behind the project, the goals of the trials, what is going to be tested, information regarding the following test procedure, an explanation of the informed-consent document and the signing procedure of the informed-consent document.

Explanation of the SAR system

After signing the informed-consent document, the experimenters explained every visible part of the system to the user. They explained the Nao robotic system, the implemented use cases, how to interact with the robot, and the test setup including the cameras, microphones and the place of the experimenter outside the test environment.

Test phase

Two members of the research team were needed at this point, one technical experimenter and one usability experimenter.

The technical experimenter controlled the system, started and stopped audio/video recordings and took notes in case technical malfunctions occurred.

The usability experimenter observed the scene directly (e.g. through an open door) but avoided distracting the system in any way. The usability experimenter noted any general comments of the test user from the thinking-aloud process, helped in case the system malfunctioned, and helped the test subject in case of questions during the test.

User assessment – physical state of the end users

To ensure the safety of the participants during the exercises, an occupational therapist from within the experimenter team assessed the physical capabilities of the participant prior to the training. The training schedule was adapted to the individual needs of the participants. In case the participant was not able to perform certain exercises (e.g. not able to lie down and get up again, not able to stretch arms to the ceiling) those exercises were omitted from the test setup. In case more than one exercise could not be performed by the user, the experimenters chose an alternative exercise from a base set that fitted the users' needs based on a predefined exchange list, in order to keep the user experience as similar as possible between the users.

Demonstration of SAR training

A full demonstration of SAR training based on the chosen trainings set, which took about 15-20 minutes per participant, was conducted.

Acceptance questionnaires

Directly after the training, the gathering of results was started by an initial open question about the user's opinion on the training. Afterwards, a questionnaire including open qualitative and quantitative questions was either handed over to the participant or read out-loud for the participant in case the person had trouble filling out the questionnaires alone.

Explanation of the paper version of training exercises

An alternative paper version of physiotherapeutic training (including text and images) was shown and explained to the participant and a predefined specifically developed questionnaire regarding acceptance of the paper-supported training was filled out by the participant (annex section 5).

Demonstration of video training

As an example of video-supported training, a selected scene from an online web-based training video was shown to the participants and a predefined specifically developed questionnaire on the acceptance of such video-supported training was filled out afterwards (annex section 5). Figure 37 shows a screenshot of the used training video.



Figure 37: Screenshot of training video used for comparison (Source: YouTube.com).

Demonstration of Wii Exercise

A commercial Wii-based training game was shown to the users. The users took part in one selected exercise by using the Wii mote as the input device and a TV screen as the display. Afterwards, a predefined specifically developed questionnaire regarding acceptance of the Wii training was filled out by the participant. (annex section 5)



Figure 38: Screenshot of the used training programme (EA Sports Active on the Nintendo Wii),

Source: <http://www.technologytell.com>.

Comparative questionnaire

A questionnaire for comparative analysis of the four different options (paper based vs. video training vs. Wii training vs. SAR training) was filled out by the participants. Qualitative comments given during the training and afterwards were noted.

Discharge and planning of future tests

Finally the test session was concluded by thanking the user for participation and the user was invited to take part in a focus-group session to discuss the results of data analysis.

5.4.3.4 Data analysis

The analysis of results gathered from questionnaires and the thinking-aloud process was conducted, as described in chapter 3.11.

Additionally, retrospective video analysis was used in order to gather quantitative objective data, such as the number of laughs during the demonstration of the robot, the percentage of exercises followed, the number of sentences talked to the robot and the total execution time. In this case the respective factors were gathered by reviewing the

video recordings of the demonstration of video training and transcribing the specific events.

5.4.3.5 Data manipulation

The following data were removed from the data set before analysis because of obvious mismatches between the user's interpretation of the question and their intentional meaning:

Within the Godspeed questionnaire, the users found one particular question hard to interpret. The question in the construct "perceived safety": "quiescent – surprised" was very often misinterpreted. "Surprised" can be either positively surprised by the high functionality of the system or negatively surprised in the sense of "shocked" and hence seems not to be a good marker for "perceived safety". Similarly, "quiescent" (German translation – "still") is not a negative attribute as suggested in the questionnaire but could also be positive in the meaning of "calm" as the robotic system was not perceived as agitating. This question was omitted from the construct.

this page intentionally left blank

6 Results of the evaluation of an assistive robot to support the physical therapy of older users (E3)

This chapter describes the evaluation results of the SAR concept and prototype described in chapter 5 using the methodology as described in chapter 3 and section 5.4.

6.1 Results from pre-trials

Technical performance analysis

The system has been integrated into a gymnasium within a setup time of two hours. The technical performance was validated by running the demonstration scenario three times. No technical errors could be found; hence the system was considered safe and functionally sound to be used with a group of test participants.

Motivation and training-support functions

Figure 39 shows the results of the “motivation” and “training support” categories on a 5-point Likert scale, illustrated in a boxplot. All 14 seniors state that Nao “almost” or “quite” motivates them to do exercises (Q1), they are motivated “almost” or “quite” more than when compared to having a personal trainer (Q2). Regarding the training support, all seniors state that the exercises are shown in a “quite” or “very” understandable way (Q3) and that mimicking the exercises helps them “quite” or “very much” to perform them in a better way than when only having a basic description (Q4).

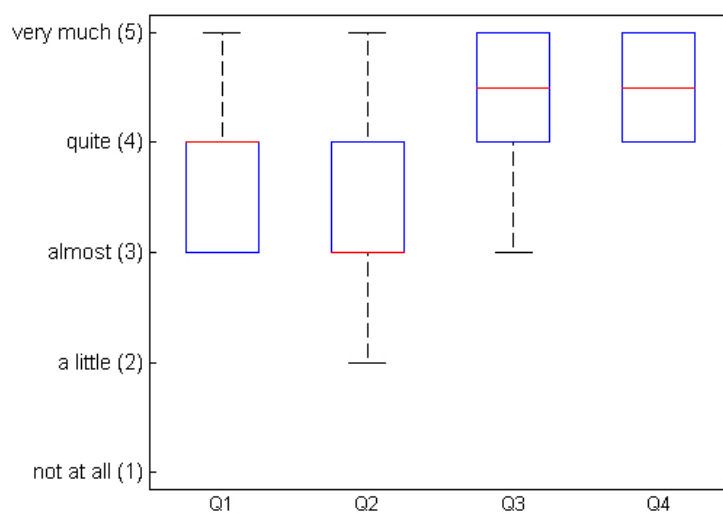


Figure 39: Results of motivation and training support (n = 14).

- Q1** Nao is motivating me to do the exercises.
- Q2** I am more motivated compared to when only having a personal trainer.
- Q3** The exercises were described and shown in an understandable way.
- Q4** Mimicking the exercises makes me perform them in a better way than when only having the basic description.

Usefulness and intended use of the prototype

Figure 40 shows the results of the PU of the training system on a 5-point Likert scale, illustrated in a boxplot. All participants, except two, rate the SAR system as beneficial for themselves (Q5). The PU results for people of all ages shows an increase of benefits with an increasing age of the potential target audience (Q6-Q9). Being asked how beneficial the participants imagine the solution could be for certain age groups, the results for the target group of 10-30 year olds (Q6) spread from “not very beneficial” to “beneficial” with its median at “beneficial”. The result for 31-50 year olds (Q7) have their median at “beneficial” with two outliers and the median for the 51-70 year olds (Q8) and the over 70 year olds (Q9) is “very beneficial”.

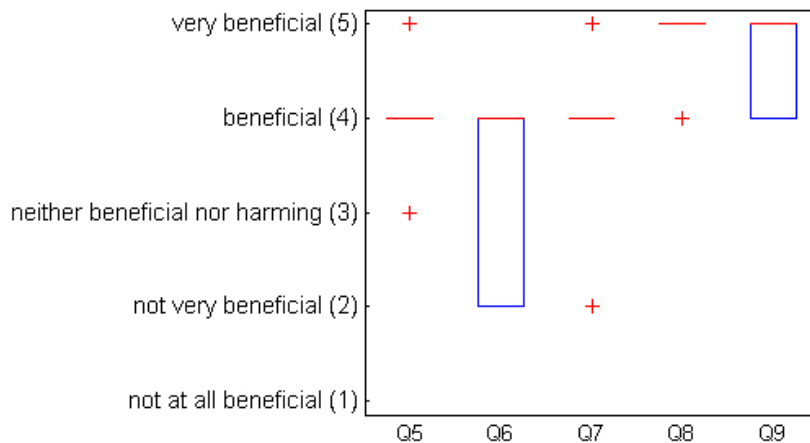


Figure 40: Results of usefulness (Q5-Q9) of the training system (n = 14).

- Q5** How beneficial do you believe the SAR system is for you?
- Q6** How beneficial do you believe the SAR system is for a 10–30 year old?
- Q7** How beneficial do you believe the SAR system is for a 31–50 year old?
- Q8** How beneficial do you believe the SAR system is for a 51–70 year old?

Q9 How beneficial do you believe the SAR system is for an over 70 year old?

When asking users about their intention to use the SAR system, all of them answered they would like to use the system at least once a week. Nine out of 14 interviewees stated that they would use the SAR system even three times a week, four persons would use it once a week, and one person every day.

Q10 How often would you use the SAR system if you had it at home?

Figure 41 shows the results of the questions regarding the general impression of the SAR on a 5-point Likert scale, illustrated in a boxplot. All participants “agree” or “strongly agree” that the SAR system met their expectations (Q13), that it amused them and they enjoyed interacting with it (Q11). In addition, all except one “disagreed” or “strongly disagreed” that the SAR system was boring and did not interest them (Q12).

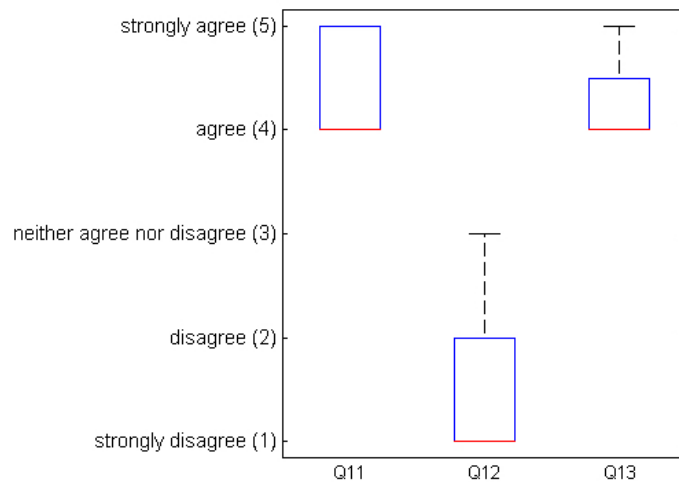


Figure 41: Results of general impression (Q11-Q13) of the training system (n = 14).

Q11 The SAR system was amusing and I enjoyed interacting with it.

Q12 Using the SAR system was boring and did not interest me.

Q13 The SAR system has met my expectations.

General characteristics of the robotic trainer

The characteristics of Nao concerning the movements, the exercise description and the training feedback were surveyed by rating the following properties from (1) to (5):

Nao's movements

Q14 fast (1) – slow (5)

Q15 graceful (1) – clumsy (5)

Q16 human-like (1) – machine-like (5)

Nao's exercise description

Q17 too detailed (1) – insufficient (5)

Q18 easy to understand (1) – hard to understand (5)

Nao's feedback

Q19 proper (1) – improper (5)

Q20 too detailed (1) – insufficient (5)

Q21 easy to understand (1) – hard to understand (5)

In Figure 42, the results of Nao's characteristics on a 5-point semantic differential scale, answered by a group of 11 participants ($n = 11$), are illustrated in a boxplot. Nao's movements are rated with a score of 2-3 for its speed (Q14) as well as for the characteristics "human-like – machine-like" (Q16). The dexterity (Q15) is rated with a score of 3 – neither graceful nor clumsy.

The level of detail of the exercise description (Q17) is rated with a median of 2 and an extreme at 4. The articulation (Q18) of Nao's exercise description is rated with a median of 2 with an extreme and outliers at 1 and 3.

The assessment of Nao's training feedback resulted in a score of 2-3 for the level of detail (Q20) as well as for its articulation (Q21) and the properness (Q19), with extrema rated with 1 at question Q19 and Q21.

By consideration of the positive and negative qualities of all of Nao's surveyed characteristics, the majority of votes lies at 2 or 3, which points to positive or neutral qualities.

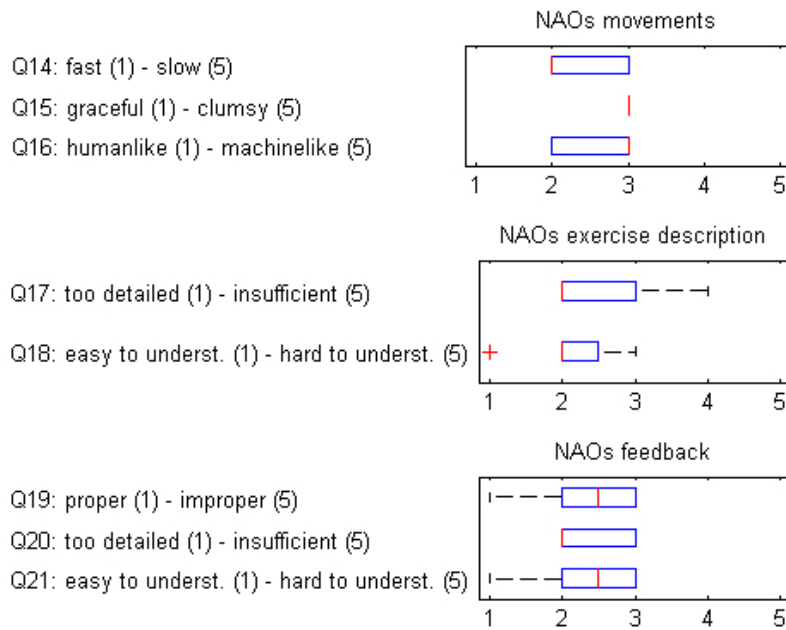


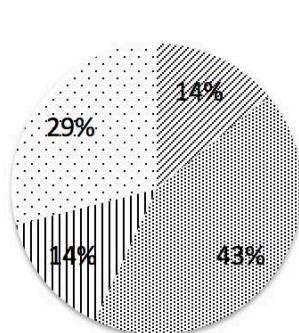
Figure 42: Results of the characteristics of the robot trainer (Q14-Q21), (n = 11).

Training preferences

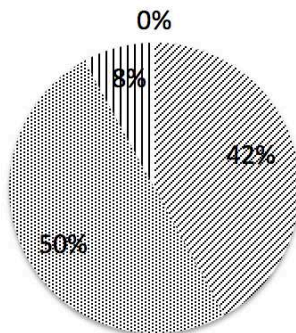
The last part of the questionnaire is about the participants' system preferences for a daily physical training routine at home. The options of the training systems are "SAR", "training video with instruction", "paper-based instruction" and "nothing".

The following figure shows the results of comparative questions regarding the training preferences, answered by a group of 12 senior citizens (n = 12). 86% of the participants prefer conventional systems (video, paper, nothing) for their regular training at home (Q22). As reasons, they stated that it seems more realistic to them and with video or paper instruction, one can better benefit by choosing from a variety of exercises (Q23). Concerning the motivational skills of the different training systems, the video and the robot solution are nearly equally motivating with a small preference for the video version (42% vs. 50%) (Q24). In contrast, the majority of the users (77%) would prefer the robot if they could choose one system to take home for a month (Q25).

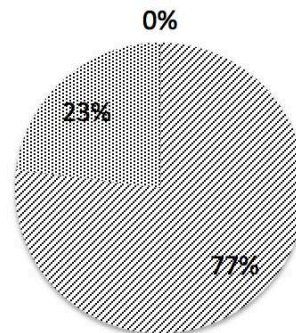
Q22: I prefer the following system for my daily training at home.



Q24: What system seemed to motivate you the most?



Q25: If you had the choice of taking one system to your home for one month – which one would you choose?



 robot
  video
  paper
  nothing

Figure 43: Comparative analysis of different ways of training support (Q22, Q24, Q25), (n = 12).

Q22 I prefer the following system for my daily training at home.

Q23 Why would you prefer this system?

Q24 What system seemed to motivate you the most?

Q25 If you had the choice of taking one system to your home for one month – which one would you choose?

Summary and conclusions of pre-trials

The ability to conduct the pre-pilot within a real-life setting of an unmodified gymnasium presents a major step towards the real-life applicability of the system and serves as proof of concept of the idea, the implementation and the feasibility of integration into new settings.

Acceptance results from the pre-pilot show that users found the SAR system highly motivating and supporting (mean score 4 on a 5-point Likert scale) and (very) beneficial (mean score 4.2 on a 5-point Likert scale) for themselves and people of all ages. Ten out of the 14 participants stated that they would like to use the training system at least

three times a week. The SAR system was amusing and met the expectations of all participants. They did like to interact with the system and all participants stated it was neither boring nor uninteresting to them.

Nao's characteristics concerning its movement, the training description and giving feedback was experienced as neutral or positive (mean score 2.5 on a 5 point scale) When compared to similar training support systems such as video- and paper-based support, users preferred the video course over the robot trainer (6 video, 2 robot, 2 paper, 4 no training support aid) for regular training at home, but 10 out of 12 would prefer to take the robotic system if offered the chance to test one of the systems for the duration of one month at home.

The first results of the pre-pilot indicate that, after a 15 minute demonstration of the system, potential end-users accept the idea of physical training with a socially assistive robotic trainer and find the robotic prototype entertaining and motivating. The fact that users preferred video-supported training over the robotic solution but would opt for the robotic solution for testing suggests that the initial excitement of the new and innovative solution was high but the long-term expectations are not equivalent to the video-supported training.

6.2 Evaluation results of the main evaluation phase

6.2.1 Performance results, technical performance and usability

The following research questions were used to drive the evaluation:

RQ_E3_P1: To what extent is the system in the current state usable; what usability issues exist?

RQ_E3_P2: Is the system able to perform correctly from a technical viewpoint under real-life conditions?

RQ_E3_P3: What are the technical limitations of the approach?

Usability results regarding the training explanation and feedback

In addition to the Almere model constructs of "PEOU", "PU", and "PAD" (see section 6.2.2.2) a usability questionnaire was designed that better covers the factors of effectiveness and efficiency of the training demonstration and training feedback. Further qualitative analysis of user comments was undertaken.

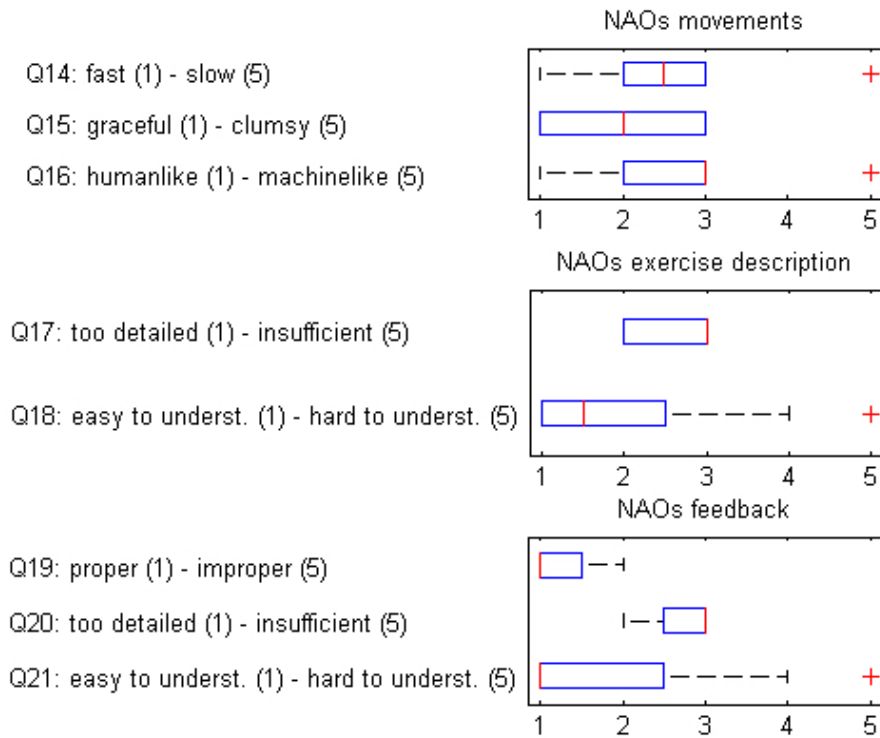


Figure 44: Usability results regarding training explanation and feedback.

As shown in Figure 44, the trial participants stated that the SAR system seems to be rather fast, graceful and humanlike (Q14-Q16), which corresponds well with previous results from the Almere model and the Godspeed questionnaire.

The exercise description was given orally by the robot via the text-to-speech functionality. According to the trial participants, the amount of text spoken by the robot was rated well (Q17). The quality of the speech output (Q18) was rated with a median of 1.5 meaning that the majority of users found what the robot was saying easy to understand. As shown by the high standard deviation, some users, in particular users with hearing aids, found the system very hard to understand, which corresponds also with the objective analysis of the video material and qualitative user comments.

Trial participants rated Nao's feedback in general as "proper" (Q19), similar to the exercise description, the amount of feedback given was rated to be slightly higher than necessary (Q20). Since the feedback was to most extent given orally, the feedback understandability suffered the same problems as the understandability of the training instructions and resulted in a nearly identical score.

Technical results gathered during the trial execution

The stability of the system is an important usability factor. The stability was tested in laboratory trials prior to the user trials and found to be sufficient to conduct trials with users.

Since the SAR system is in a prototype state, some technical problems still occurred during the user trials.

In one case, the system was unable to localize the user – which is necessary in order to analyse the user’s training movements – because the user was dressed in colours very similar to the wall and furniture behind her, making it difficult for the pattern recognition to differentiate between the person and the background. The problem could be solved during the trial by asking the user to put on some different outerwear (a light jacket). The respective user gave a qualitative statement for this issue:

“It is a problem if the system malfunctions, how should I get help if no technician is around?” – (“Das ist aber dumm wenn das System nicht korrekt funktioniert, wie soll man sich denn in solch’ einem Fall helfen, wenn kein Techniker da ist?”)

In one case, the trial needed to be postponed after the user arrived because of a technical malfunction of the interface between the notebook used and the Kinect sensor. Another notebook was used for the subsequent trials that did not show this instability.

In one case, the trial needed to be postponed before the user arrived because the used Nao robot showed a physical malfunction; a gear wheel was broken and replaced within a week’s time.

In one case, a minor deviation of the feedback text was spoken because of a software bug, which was corrected immediately after the trial. The user did not recognize the malfunction as such as the text still made sense.

In one case, the applause sound after the training was not produced by the robot system for unknown reasons. The user could not recognize the malfunction as such as she/he did not know what was to be expected.

In one case, the trial was delayed by 5-10 minutes because of initial technical problems. The user did not recognize the malfunction as such as the experimenter was able to cover the time.

General technical issues of robotic prototypes in real life - e.g. within smart homes

Navigation. This very complex AI issue was avoided by keeping the robot in one place. Navigating a biped robot inside an unmodified living environment is a current research challenge that in general is not sufficiently solved to date.

Speech recognition. Current best-performing speech-recognition engines, such as *Apple SIRI*,⁴¹ *Google Voice*,⁴² or *Dragon speech*,⁴³ were not able to sufficiently allow a dialog over a distance of two meters with the Nao robot. Since this is a research topic that needs resources far out of scope of this dissertation, the speech recognition was not used for HRI.

Pattern-recognition algorithms. Several visual methods are used to detect the user's face during face tracking and to closely follow the user's movements. These pattern-recognition algorithms facilitate cameras as environmental sensors and are prone to errors due to changing lighting conditions. These conditions were kept static during the trials in order to maintain comparable trial results, which might pose a usability issue in real life.

Autonomy of the robot. The used robot, Nao, has several issues that negatively influence the autonomy. Firstly, the maximum time of battery operation is only 45 minutes. This was sufficient to conduct the trials but would require timely recharging of the robot in a real-life situation.

Secondly, the operation system of the robot is provided by *Aldebaran* and still under development, hence also prone to errors. Technical supervision is constantly necessary to operate the robot because of possible malfunctions of the operation system which might also lead to a crash of the robot.

In addition, the used motors and heat-dissipation system do not allow the robot to be operated over long time periods since the robot can easily overheat.

Space needed. When asked the facilitating-conditions questionnaire within the Almere model, users sometimes commented that they would not have the space necessary to store the robot.

⁴¹ <http://www.apple.com/de/ios/siri/>

⁴² <http://www.google.com/googlevoice/about.html>

⁴³ www.nuance.com

6.2.2 Acceptance results

To what extent older users accept a humanoid robotic trainer for regular physical training at home was one of the main research questions of the third evaluation phase (E3). In order to evaluate this question, quantitative questionnaires were analysed and augmented with qualitative statements received during the training and the interviews from the trial participants. Within our first research question we ask:

RQ E3 A1: Is a robotic training assistant accepted for regular training at home by older users?

Three quantitative questionnaires were used to assess this question; the Godspeed questionnaire primarily evaluates HRI and the user's feelings towards the robot. The Almere model covers a wider spectrum of acceptance factors also including social influences, usability and facilitating conditions and is partly overlapping with the Godspeed model which gives the possibility for cross testing. In addition to these two publicly available questionnaires, which were also used for reference with similar studies, an ad-hoc questionnaire was designed that specifically evaluates the intention to use the system in the future.

6.2.2.1 Perception of the SAR

This section reports the acceptance factors related to the perception of the SAR, which are based on the Godspeed questionnaire.

Table 12: Descriptive statistics of Godspeed constructs.

Descriptive statistics Godspeed						
	min	max	mean	std	cron. α	
Anthropomorphism	1.80	4.20	3.15	1.00	0.89	
Animacy	2.33	5.00	3.67	0.99	0.93	
Likability	3.60	5.00	4.60	0.65	0.89	
Perceived Intelligence	1.80	5.00	4.08	1.24	0.97	
Perceived Safety	3.50	5.00	4.83	0.46	0.81	

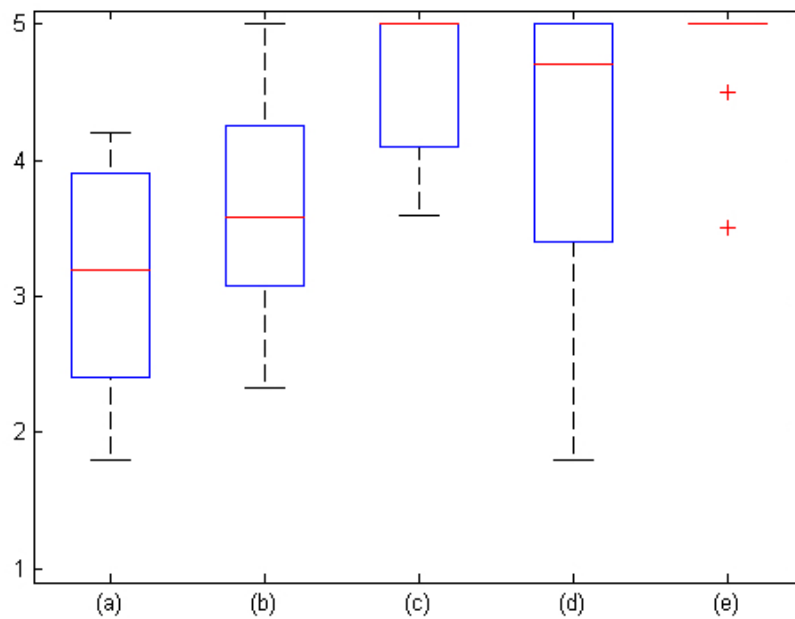


Figure 45: Godspeed constructs for E3: (a) anthropomorphism, (b) animacy, (c) likability, (d) perceived intelligence, (e) perceived safety.

The values of Cronbach’s alpha were calculated and found to be clearly above 0.7 for all constructs, so the internal consistency reliability is fine for a summarized analysis as follows.

Anthropomorphism (a)

This construct received a mediocre score meaning that the system showed partly human-like behaviour. The variance among the 12 trial participants is high because some users perceived the system to be much more human-like than others. When analysing the questions that form the construct, it shows that nearly all users rated the system exactly between the attributes “real” and “unreal”. The differentiation between “conscious/unconscious” received the most mixed reviews, resulting in the highest variance. Three users said that the system clearly is unconscious (rating 1), which was the expected answer; on the contrary, three users stated that the system is nearly conscious (rating 4). The best rating was achieved by the pair “moving rigidly/moving elegantly” since most people got the impression that the robots movements were rather fast, smooth and elegant.

Qualitative comments of the users related to this construct were:

“It is a crossover between a machine and a human being.” – (*“Er ist ein Mittelding zwischen Mensch und Maschine.”*)

"He is a machine but if I have nobody else ... it is a diversion for me." – (“*Er ist eine Maschine aber wenn ich niemanden hab ... ist es für mich doch eine Abwechslung.*”)

"I find the robot fascinating because he is so human-like." – (“*Finde den Roboter faszinierend, weil er so menschlich ist.*”)

After activating the eye LEDs: *"Now he is awake!"* – (“*Er ist wach!*”)

"It is hard to tell whether he is human- or machine-like since he is a machine but acts like a human." – (“*Menschlich oder Maschinenähnlich ist schwierig zu sagen weil er ja eine Maschine ist, aber er macht's menschlich.*”)

Animacy (b) rates to what extent the robot appears to be a living, organic and interactive individual rather than an inanimate machine.

This score was rated between animate and inanimate by most users. Some gave more positive marks; one trial participant even rated all related questions with the top mark. This means that the trial participants rated the SAR system to be more animated than inanimate.

The pair “dead/alive” was rated most positively within the construct by the users (avg. = 4.17, std. = 0.83). Interestingly all trial participants rated the pair “dead/alive” higher than 3, meaning that all favoured the attribute “alive” for a clearly non-living object.

The user responses were divided regarding the pairs “mechanical/organic” and “artificial/life-like”. Both attribute pairs received the best and worst results (avg. = 2.67, std. = 1.30 and avg. = 3.17, std. = 1.19 respectively).

Although no interactive speech dialog was implemented, the attribute pair “inert/interactive” scored with high values (mean = 3.83, std. = 0.94). To further raise this score, a more interactive communication also involving speech recognition and non-functional elements such as glimpsing with eyes would likely be beneficial.

Qualitative comments of the users related to this construct were:

Regarding the attribute pair “dead/alive”: *"Dead is rubbish."* – (“*Tot ist ein Blödsinn.*”)

Confident: *"No, he is alive."* – (“*Na, der is lebendig.*”)

Likeability (c) shows to what extent the robot appears to be likable, kind, pleasant and nice.

This score achieved very high ratings from all users (overall avg. = 4.6, std. = 0.65) The score is comprised of five attribute pairs of which all received ratings higher than 4.3 and none reached a standard deviation higher than 0.89

Many qualitative comments were noted regarding this construct:

"He is nice." – ("Er is lieb.")

"Well, what is nice, for a machine?" (laughs) – ("Naja was ist nett, an einer Maschine?" (lacht))

"I somehow like him." – ("Irgendwie gefällt mir der.")

"He is quite delightful." – ("Er ist eh entzückend.")

The likability of the system is a clear strength that also supports the motivational abilities. The users found the nice appearance of the robot that resembles a small child to be very appealing. The participants also commented positively on the small size with regards to the friendly, kind and harmless look that the small size supports.

Perceived Intelligence (d) rates the competence, intelligence, responsibility and sensibility of the robot's appearance and actions.

This construct scored very highly with a mean of 4.08 but showed also the highest standard deviation with 1.24. When analysing the construct in detail, it shows that users were divided on all questions within the construct, not agreeing if the system is truly intelligent or not. There seems to be an inverse correlation between the technical background of the users and the perceived intelligence in a way that technically experienced users, mostly men, experienced the system as a programmed computer which cannot be intelligent, whereas other participants overrated the system and stated that the robot is intelligent. This theory is also supported by [Siino2005] who found that engineers and male administrators tended to view an autonomous hospital robot as a controllable machine whereas female administrators and low-status staff viewed the robot as a human male.

Qualitative comments of the users related to this construct were:

"He is intelligent – it seems so when he is looking at me." – ("Ist schon intelligent, wenn er einen so beobachtet.")

“He is knowledgeable.” – (“Wissend ist er auch.”)

“He is responsible since he tells you what to do.” – (“Verantwortlich ist er auch, er sagt dir nämlich was tun sollst.”)

Perceived Safety (e) rates the three attribute duplets “anxious – relaxed”, “calm – agitated” and “quiescent – surprised”.

The trial participants clearly perceived the system as very safe which is shown by a high mean of 4.83 (std. = 0.63).

Hardly any comments were received regarding the safety of the system. The robotic prototype did not move or walk from its initial position which likely had a positive effect on this constructs results since moving systems are more likely to pose a threat by increasing the chance to stumble upon, damage the interior or pump into.

The following related user comment was received:

“What if the system malfunctions – how can I help myself if there is no technician around?”
– (“Das ist aber dumm, wenn das System nicht korrekt funktioniert, wie soll man sich denn in solch einem Fall helfen, wenn kein Techniker da ist?”)

Comparison with similar systems

The most comparable system is the prototype of the predecessor system, as tested in E2. The E2 prototype focused on a similar user group and facilitated the same robot. The study methodology and the metrics used were partly the same since the third SAR system built upon the knowledge gained within E2.

One goal of PT3 was to take one of the use cases identified and realized within E2, optimize it and analyse its potential in detail. Hence, lessons learnt from user trials within E2 were implemented into PT3 such as:

- Faster interaction flow to avoid boredom
- Smoother physical movements of the robot to increase animacy and anthropomorphism scores
- Usage of “mimics” and gestures to enhance speech dialogs

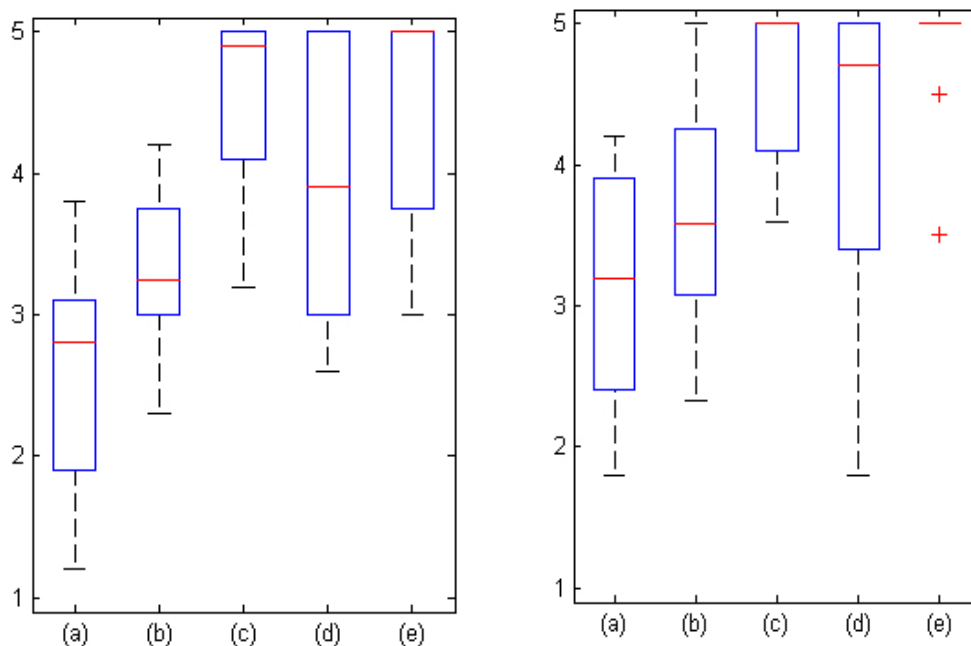


Figure 46: Godspeed comparison with the first iteration of E2 (left) and E3 (right).
(a) anthropomorphism, (b) animacy, (c) likability, (d) perceived intelligence, (e) perceived safety.

When comparing the results of the Godspeed questionnaire over the two evaluation phases, it shows that slightly higher ratings were achieved for all constructs with the PT3 system. A MANOVA was calculated using the evaluation phase as an independent variable and the acceptance factors as a dependent variable which revealed no significant difference for any construct $F(5,14) = 1.984, p = 0.14$. Given the insignificant results, individual ANOVAs were not calculated to avoid a probable type 1 error.

As effort was particularly undertaken to make the system's behaviour more vivid by reducing delays during interaction, we expected to achieve a significant difference in the questionnaire items that measure the vividness of the robot's performance. We therefore additionally conducted two two-sided t-tests testing for the H_0 Hypothesis = "The PT3 system is perceived as acting more vividly than the PT2 system". We tested the two questionnaire items "rigidly – elegantly" and "stagnant – lively" and found that the PT3 system was perceived to act more elegantly than PT2 with a highly significant difference ($p=0.006$) but there was no significant difference for the rating of the "stagnant-lively" pair ($p=0.06$).

As the user group between E1 and E2 (eight users) and E3 (12 users) changed, the measured quantitative results could be biased due to the users' inter-individual

differences. To compensate this bias, we took a closer look at the qualitative results from those five users who took part in E1/E2 and E3 respectively.

During the test, three of those five users commented on their own initiative on the robot's performance which they considered to be clearly superior to the previously shown prototype.

To the robot: *"Well, you have learnt a lot!"* (since last time) – (Zum SAR: *"Na, du hast ja einiges gelernt."*)

"I am surprised what he can do already because I've known him for some time." – (*"Ich bin überrascht, dass er schon so viel kann, denn ich kenne ihn ja schon länger"*)

"Well, he has made some progress." – (*"Na der hat ja Fortschritte gemacht."*)

"I think you've developed this pretty well." – (*"Ich glaub ihr habt's es eh schon ganz gut entwickelt, muss man schon sagen."*)

Based on qualitative results, we can see that the design philosophy and action taken to improve the system with the knowledge gained from the E2 user trials were successful. Based on the quantitative and qualitative detailed analysis, the approach to more closely follow the human-human model by also including gestures and "mimics" via LED patterns seems to have led to an improvement of the respective scores for animacy and anthropomorphism.

Comparison with other assistive robots.

Similar studies that facilitated the Godspeed test and provide the data to the public are scarce, but two robotic projects could be found for comparison in order to get a slight impression about typical Godspeed results for assistive robots.

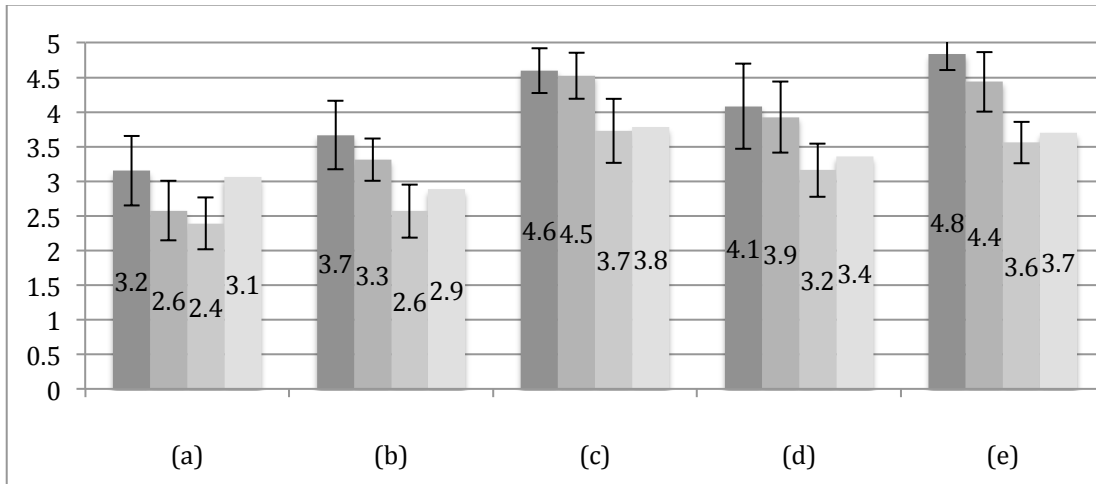


Figure 47: Godspeed construct means for comparison with other robotic systems (from left to right: E3, E2, James Robot, Magabot)- (a) anthropomorphism, (b) animacy, (c) likability, (d) perceived intelligence, (e) perceived safety – error bars indicate one standard deviation.



Figure 48: Nao robot used for E1 to E3 (left), James Robot (middle) [Foster2012] and Magabot (right) for comparison.

As shown in Figure 47, the mean scores of the E3 Godspeed constructs are slightly higher than those of all other comparable systems. In particular, the constructs of likability and safety are strengths of the PT2 and PT3 SAR systems and are mostly due to the nice-looking and small child-size appearance of the robotic prototype. Interestingly the wheel-based robot “Magabot” – with a screen as the main user interface – achieves similar results in anthropomorphism to an anthropomorphic humanoid biped robot. This seems unlikely at first; when analysing the data further, it shows that the high result is influenced by very high scores in behavioural parameters such as the attributes “unconscious / conscious”, so the robot might have achieved these results by showing intelligent behaviour. Additionally, this comparison is limited as the user groups and

specific test settings differ between the projects and we cannot estimate this influence on the results.

6.2.2.2 *Intention to use*

Here we present the results concerning the acceptance and intended use. We give an overview of the quantitative results of the Almere model constructs and of corresponding qualitative measurements. The quantitative results are presented in descriptive statistics and have to be interpreted whilst accounting for related variables.

Table 13: Descriptive statistics of the Almere Model.

Descriptive statistics Almere Model					
	min	max	mean	std	cron. α
ANX	4.5	5.0	4.9	0.27	0.10
ATT	3.0	5.0	4.5	0.72	0.79
FC	3.0	5.0	4.3	0.93	0.56
ITU	4.0	5.0	4.9	0.29	-
PAD	3.3	5.0	4.6	0.59	0.76
PENJ	3.6	5.0	4.7	0.51	0.79
PEOU	3.4	5.0	4.5	0.72	0.54
PS	1.5	5.0	4.1	1.18	0.78
PU	3.3	5.0	4.4	0.69	0.85
SI	2.5	5.0	4.4	0.78	0.78
SP	1.8	4.4	3.0	1.21	0.36
TRUST	2.0	5.0	4.3	0.92	0.94

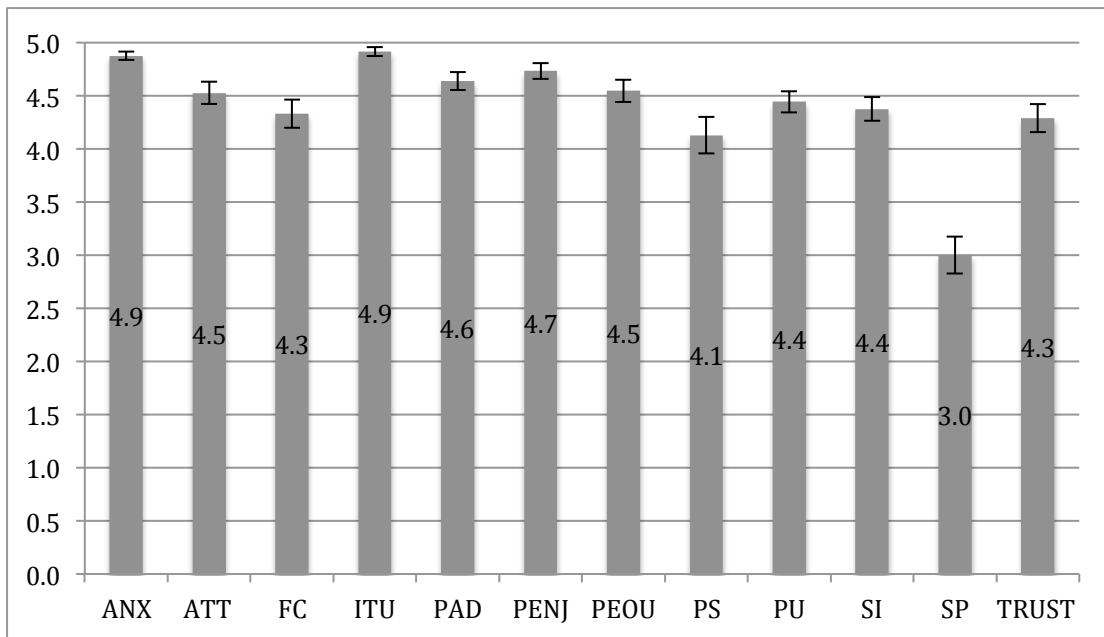


Figure 49: Mean scores of the Almere model constructs, error indicators show the standard error of the mean.

Anxiety (ANX). This construct measures whether or not the system evokes anxious or negative emotional reactions during usage and influences PU and PEOU (see also Figure 11). The statements within this construct are:

ANX1: If I should use the robot, I would be afraid to make mistakes with it.

ANX2: If I should use the robot, I would be afraid to break something.

ANX3: I find the robot scary.

ANX4: I find the robot intimidating.

The constructs statements' results show a very low consensus among all trial participants and hence suggest that the SAR system does not evoke anxious reactions. None of the test participants found the robot scary or intimidating; three users were a little concerned about making mistakes or damaging the robot and gave a score of ≥ 3 . The Cronbach alpha score of 0.1 tells that the consistency of statements within the construct is unacceptable. A factor analysis shows that ANX1 and ANX2 are unrelated to ANX3 and ANX4 which suggests that the anxiety construct within the Almere questionnaire should be further validated.

Attitude towards technology (ATT). This construct gives information about the user's feelings regarding the appliance of the robot. The attitude towards the robot has a direct influence on the intention to use the system. The statements within this construct are:

ATT1: I think it is a good idea to use the robot.

ATT2: I think the robot would make life more interesting.

ATT3: It is good to use the robot.

The results of this construct show high mean values (mean = 4.53, std. = 0.72). When analysing the single questions, it shows that users agree very much that it seems to be a good idea to use the robot (mean = 4.75) and also that the robot would make life more interesting (mean = 4.25).

Facilitating conditions (FC). This construct evaluates whether the environmental conditions allow the usage of the robot.

FC1: I have everything I need to use the robot.

FC2: I know enough about the robot to make good use of it.

During the trials it was shown that many users had a problem understanding these two statements correctly. For one side, it was not clear what users would need to use the robot when applied to their daily routine. The experimenters explained this with examples such as: having enough space to place the robot; providing a power outlet close to the robot; having free space to place a PC or laptop.

The second statement FC2 was hard to answer for the trial participants because they only gathered experience in how to use the robot but not in how to maintain it, such as replacing batteries, setting it up, powering on and off, etc.

For these reasons, the construct received slightly mixed reviews depending on the understanding of the statements by the users, resulting in a higher variance of FC2 in particular (mean. = 4.0, std. = 1.21).

Intention to use (ITU). This construct represents the intention to use the system without considering a potential price. This construct has a strong direct influence on the future use of the system and hence is a very important acceptance marker [Heerink2010].

ITU1: If I had the robot at home, I would use it within the next few days.

The statement within this construct had to be altered from the original proposal by Heerink because the test did not take place at the users' homes but within a laboratory mimicking a user's home. We assess the prospective use whereas Heerink suggested to measure the real use at home.

This construct shows a very high score of mean = 4.92, std. = 0.29. Only one out of twelve users did not completely agree with this statement and rated it with the score 4. This is a very significant acceptance result of this study since it shows that the robot was not only accepted very well by the whole trial user group but also that they were motivated to perform physical training at home with the system. This result is most likely influenced by the novelty effect as users might have wanted to try out this new technology as soon as possible.

Perceived adaptability (PAD). This construct rates the perceived ability of the system to adapt to the needs of the user. Since user abilities such as the motor and sensory abilities constantly change, assistive devices – particularly if designed for older users – need to constantly adapt towards the current users' needs. If users perceive a system to be adaptive to their needs, they will find it more useful which has a positive influence on the acceptance of the system. The construct includes the following statements:

PAD1: I think the robot can be adaptive to what I need.

PAD2: I think the robot will only do what I need at that particular moment.

PAD3: I think the robot will help me when I consider it to be necessary.

During the trials, the user experienced how the robot adapts to their behaviour since the robot constantly adapted in relation to the movements made by the user. The robot commented on the quality of the execution of the physical exercises and changed the interaction flow based on user behaviour.

Most users thought that the robot will adapt to their needs (mean = 4.67, std. = 0.49). The answers to the statement PAD3 were more diverse since PT3 is focused on only one application, whereas the statement suggests that the robot is used in a multifunctional manner and could help the user with several needs (mean = 4.42, std. = 0.9).

Perceived enjoyment (PENJ). The construct evaluates feelings of joy and pleasure when using the system.

PENJ1: I enjoy the robot talking to me.

PENJ2: I enjoy doing things with the robot.

PENJ3: I find the robot enjoyable.

PENJ4: I find the robot fascinating.

PENJ5: I find the robot boring.

Most users enjoyed talking and working out together with the robot, with a score of at least 3 (PENJ1 mean = 4.58 PENJ2 mean = 4.45). No user found the system boring (mean = 5). All users found the robot fascinating, with a score of at least 3 out of 5.

Perceived ease of use (PEOU). This construct represents the degree to which a user believes that using the system is free of effort. This score is closely related to the usability of the system and is directly related to the intended use of the system and is hence a very important acceptance factor. The statements within this construct are:

PEOU1: I think I will quickly know how to use the robot.

PEOU2: I find the robot easy to use.

PEOU3: I think I can use the robot without any help.

PEOU4: I think I can use the robot when there is someone around to help me.

PEOU5: I think I can use the robot when I have a good manual.

All users agree that the robot is easy to use with a score of at least 3. The possibility to ask somebody for help (PEOU4 mean = 4.82) only slightly improves the score in comparison with using the robot without help (PEOU3 mean = 4.5).

Perceived Sociability (PS). This construct is especially important for a social assistive robot since it measures the perceived ability of the system to perform sociable behaviour which is an important factor when using the human-human model for HRI. It is not uncommon for people to interpret behaviour by technology as social behaviour and to build up a social bond with technical items such as smart phones, cameras or personal computers [Shibata2003]. The construct includes the following statements:

PS1: I consider the robot a pleasant conversational partner.

PS2: I find the robot pleasant to interact with.

PS3: I feel the robot understands me.

PS4: I think the robot is nice.

Participating users agreed that the robot is nice (PS4 mean = 4.42). The user group was divided regarding the statements PS1 to PS3. PS1 in particular received mixed ratings (mean = 3.74, std. = 1.14) and users commented that it seems not to be a true conversational partner.

This has to do with the design of the robotic behaviour. The system was not designed to be entertaining through conversation but to be a physiotherapeutic trainer. Further, no speech recognition was implemented; hence the robot could not understand the users'

words and react upon them, which led to a lower score and high variance for the statement PS3 (mean = 4.18, std. = 1.4).

This construct was altered slightly since it contains general statements that not only measure sociability but general usability.

Perceived Usefulness (PU). This construct measures the degree to which the user thinks that the system would assist and be useful for personal use. The construct comprises the following statements:

PU1: I think the robot is useful to me.

PU2: It would be convenient for me to have the robot.

PU3: I think the robot can help me with many things.

All participants stated that the robot would be useful for them personally at a level of at least 4 of 5. The statement PU3 received slightly mixed ratings because the system was developed for only one use case and hence can only help with physical training, not “many things”. Despite this fact, many participants imagined what the robot could possibly do for them in the future and answered the statement very positively (mean = 4,17)

Social influence (SI). This construct measures the users’ perception that people who are important to them would like them to use the system. This is an important factor for extrinsic motivation since a high social influence can result in high personal rewards from friends and relatives. The following statements are parts of this construct:

SI1: I think people around me would like me to use the robot.

SI2: I think I would make a good impression if I should use the robot.

Both statements received very high ratings (SI1 mean = 4.87, SI2 mean = 4.08) from the users for two reasons:

1. Several participants stated that they would like to show the robot to friends and relatives in order to show them the newest technical possibilities. In this respect, the robotic system has a similar influence as any high-tech gadget.
2. Many potential users have a bad conscience because they know they should perform regular physical activity and physical training (which often was also prescribed to them by doctors) but do not find the time and motivation to undertake the effort. Hence, users think they could become a role model for others if they manage to overcome their weaker selves.

Related qualitative user comments:

"I would use him because I could show him to my children." – ("Tät ihn schon verwenden, weil da würd ich ja meine Kinder einladen dazu.")

Social presence (SP). This construct evaluates whether or not users sense a social entity when interacting with the system. This was shown for SARs and has an influence on the acceptance of the system, the motivation to use the system and follow the system's advice. The construct includes the following statements:

SP1: When interacting with the robot, I feel like I'm talking to a real person.

SP2: It sometimes felt as if the robot was really looking at me.

SP3: I can imagine the robot to be a living creature.

SP4: I often think the robot is not a real person.

SP5: Sometimes the robot seems to have real feelings.

This construct scored with the lowest score of all measurements (mean = 3.0, std. = 1.21). The trial participants commonly only agreed to SP2 and often referred to the fact that the robot uses face tracking during the training. SP1 received mixed ratings, ranging from 1 to 5 with a mean of 3.18, which corresponds to the results of the Godspeed construct of anthropomorphism. To the users, interacting with the robot was enjoyable and entertaining but only partly feel like talking to a real human. It seems plausible that the size, shape and material of the robot has a strong influence on this construct.

Interestingly the statement SP3 received very low ratings, between 1 and 4 resulting in a mean of 2.25 and a standard deviation of 1.22 – although a very similar attribute pair within the Godspeed questionnaire (dead / living) received high scores. When analysing qualitative user comments in detail, there were two reasons for this difference:

1. The two statements have a subtle difference in meaning, in that the meaning of "living creature" (in German "*lebendiges Wesen*") implies an animate being which possibly also has a soul. Whereas living could also be interpreted as lively (German "*lebendig*") which can also be an attribute for a technological object that acts in a fast and agile manner.
2. The attribute pair "dead / living" often led the users to the deduction that the robot is too animate and lively to be dead – so it has to be living. In this manner, the word "dead" is strong enough to shift the user opinion towards the opposite, which was living.

Trust (Trust). This construct models the belief that the system performs with personal integrity and reliability. This construct is important for PT3 since the users need to trust the robot in order to follow the advice given, such as the corrective statements during the exercise execution. Trust is claimed to have a direct influence on the intention to use the system and to perceived sociability. The statements to measure this construct are:

TRUST1: I would trust the robot if it gave me advice.

TRUST2: I would follow the advice the robot gives me.

Most of the trial participants trusted the robotic system very much, resulting in a mean of 4.29 (std. = 0.92). Some users would also follow the advice of the robot (TRUST2 mean = 4.42) although they would not fully trust the robot (TRUST1 mean = 4.17).

Comparison with similar assistive robots.

The Almere model was developed in 2010 by Heerink [Heerink2010] and since then, only a few studies have used the model and, aside from “KSERA”, only one comparable study could be found that provided result data [Xu2012]. Xu et al. used an assistive 1.5m tall humanoid robot called “Mika” to study the effect of scenario media (see Figure 51). A group of 15 older users was shown Wizard of Oz demonstrations of scenarios that described how a robot could help an older adult at home to handle the chores of daily life and engage in social life. Structured interviews and the Almere questionnaire were used to evaluate the users’ perceptions. The size and age of the user group and the methodology used is very similar to E3. The comparison between the scores of the E3 and E2 and the scores of the Mika trials are reported in Figure 50.

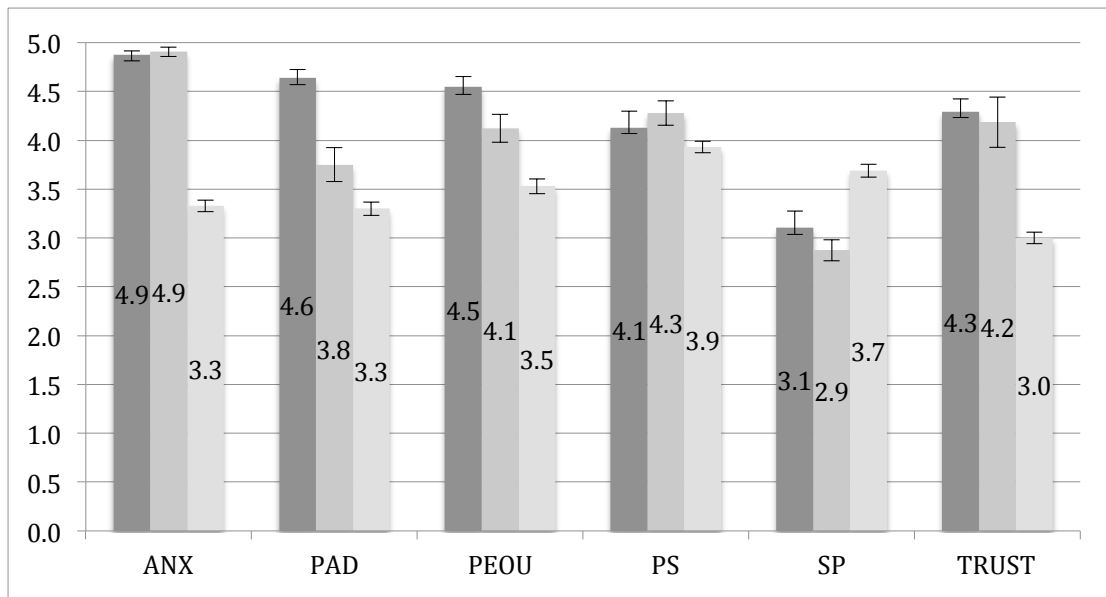


Figure 50: Comparison of mean scores of selected constructs with similar assistive robotic projects, error bars indicate the standard error of the mean. E3: dark grey, E2: grey, Mika trials: light grey.

The mean Anxiety (ANX) score for E3 (4.88) and E2 (4.89) were both higher ($p < .05$) than the one for Mika (3.33), suggesting that users were less anxious of the PT3 and PT2 systems than the Mika robot. This tends to reinforce the conclusion that the robot's size (1.5m vs. 60cm) has an influence on the level of anxiety.

The mean adaptiveness (PAD) score of the PT3 system was rated significantly higher ($p < .05$) than that for the Mika system.

The mean PEOU score was rated significantly higher ($p < .05$) for the PT3 system in comparison with the Mika system. This is an important finding since, despite the fact that the Mika evaluation used a Wizard of Oz technique, the users found the PT3 system, which acts fully autonomously, easier to use.

The SP mean score is not significantly different for the PT3 or PT2 system than for the Mika system; hence no conclusions can be drawn regarding SP.

The mean Trust score for the PT3 system (4.29) and the PT2 system (4.04) was significantly higher ($p < .05$) than the one for Mika (3.0), suggesting that users trusted the PT2 / PT3 more than the Mika robot. We believe this might be due to the smaller size and more human-like appearance of the Nao robot.

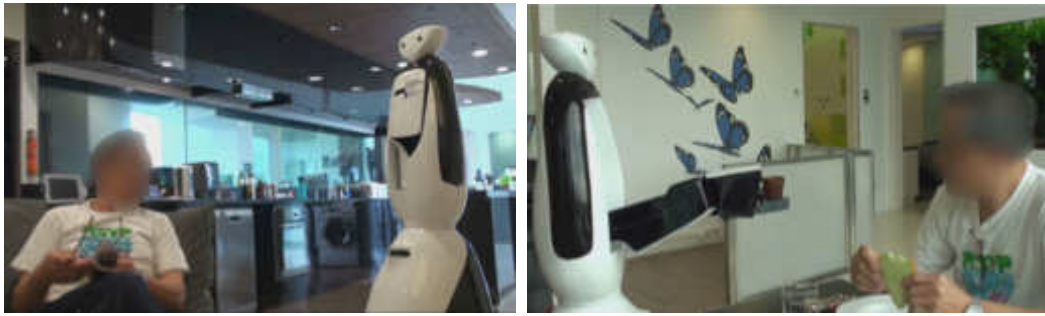


Figure 51: Mika robot, image taken from [Xu2012].

Additional analysis of intended future use and perceived usefulness

As described in the methodology section, aside from the external metrics also used for reference with other projects, questionnaires were used in addition and to complement results gathered with the Almere model and Godspeed questionnaire. Q5 to Q9 are measures for the PU.

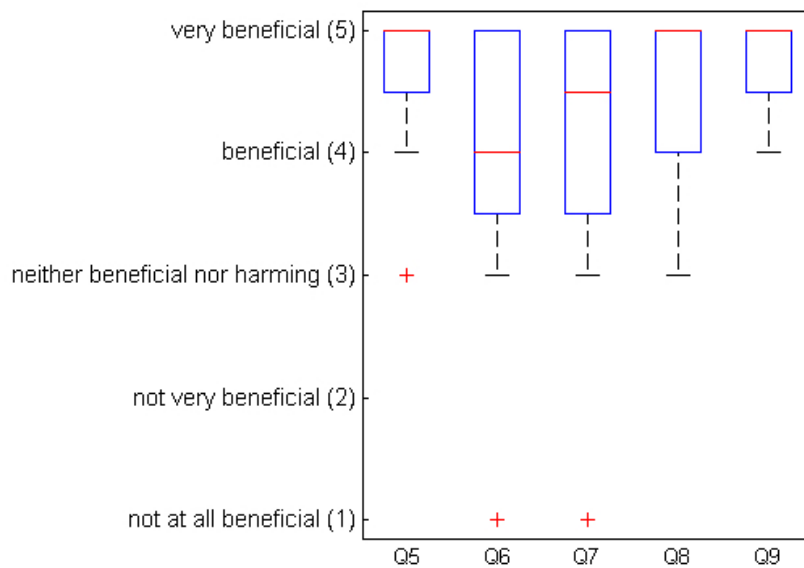


Figure 52: Results of usefulness (Q5-Q9) of the training system (n = 12).

- Q5** How beneficial do you believe is the SAR system for you?
- Q6** How beneficial do you believe is the SAR system for a 10–30 year old?
- Q7** How beneficial do you believe is the SAR system for a 31–50 year old?
- Q8** How beneficial do you believe is the SAR system for a 51–70 year old?

Q9 How beneficial do you believe is the SAR system for an over 70 year old?

Figure 52 shows the results of the PU of the training system on a 5-point Likert scale, illustrated in a boxplot. All trial participants, except one, rate the SAR system as beneficial for themselves (Q5). The system is perceived to be targeted for old people, as shown by the scores of Q6-Q9 which tend to be slightly higher for older ages. Users expect the system to be also beneficial for young users. Trial participants also commonly referred to younger people that like to play video-based training games and supposed that they would also like the SAR system.

Figure 53 answers the question whether or not the system fulfilled the expectations of the users.

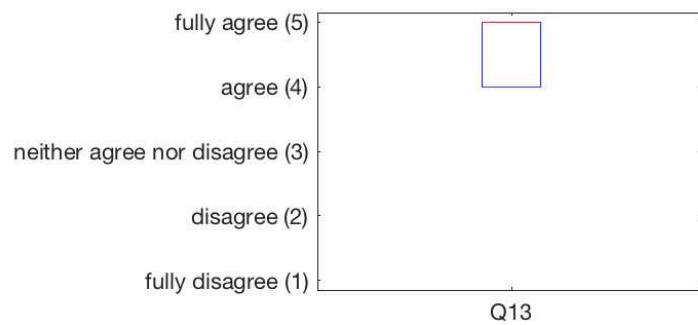


Figure 53: Score for fulfilment of user expectations given in absolute numbers.

Q13 The SAR system has met my expectations.

All users either agree (n = 5 of 12) or even fully agree (n = 7 of 11) that the system met their expectations.

6.2.2.3 Emotional effects of the prototype system

The post-test feeling is used to evaluate the acceptance of the system and is considered an important marker since it is completely uninfluenced by the experimenters. The post-test feeling was evaluated by transcribing the first qualitative subjective comments of the trial participant right after the training and by conducting a very short quantitative questionnaire immediately after the training.

The following research question was used as a base for this evaluation.

RQ A2: How do users feel right after training with a socially assistive humanoid robot?

What emotional influences can be expected?

Trial participants gave the following comments right after the training without any intervention from the experimenters:

- *"I think this is great!"* – (*"Also ich finde das super!"*)
- *"It is good that the exercise performance is assessed, this way you know if you do something wrong."* – (*"Es ist gut, dass die Übungen erkannt werden, so dass man weiß ob man es richtig oder falsch macht."*)
- *"The exercises are not bad. The first time, if you do not know the exercise and if you do not train regularly, it is exhausting because everything is new and you have to take care of many things."* – (*"Übungen sind nicht schlecht. Beim 1. Mal wenn man die Übungen nicht kennt, und wenn man auch nie Übungen macht, ist es anstrengend weil alles neu ist und man auf vieles aufpassen muss."*)
- Talking to Nao: *"Well, you've learnt a lot."* – (Zu Nao: *"Na Du hast ja einiges gelernt."*)
- *"It is good, but I think you will not find a lot of people that are able to do the training."* *"Sometimes I could not understand Nao, the speech is not very clear. If you are a little bit nervous, it is possible that you misunderstand something."* – (*"Zu „Das ist gut, aber ich glaube Sie werden nicht viele finden die das wirklich machen können.“ „Manchmal habe ich Nao nicht gut verstanden, sehr deutlich ist es nicht. Wenn man ein bisschen nervös ist, kanns schon sein, dass man etwas missversteht."*)
- *"Yes, this is funny."* – (*"Ja das ist lustig."*)
- *"It was great, it sure is different when he motivates you."* – (*"Es war super, Ist ganz was anderes wenn er motiviert."*)
- *"It is amazing what he can already do."* – (*"Na toll was der schon kann."*)
- *"That's it? Great!"* – (*"Das wars? Na super!"*)
- *"Thank you, I forgot one exercise ..."* – (*"Danke, eine Übung hab'ich mir leider nicht gemerkt ..."*)
- *"If I had him, I would surely train more often."* *"The other time, when I had this display, I also trained every day."* (Annotation: User participated in an earlier study for physical training without a robot) – (*"Wenn ich den hätte, würde ich sicher mehr Training machen." "Damals mit dem Schirm hab ich auch alle Tage trainiert."*)

Figure 54 gives the boxplot of the results of a 7-point Likert scale questionnaire regarding selected emotional statements, which was answered by the trial participants right after the training.

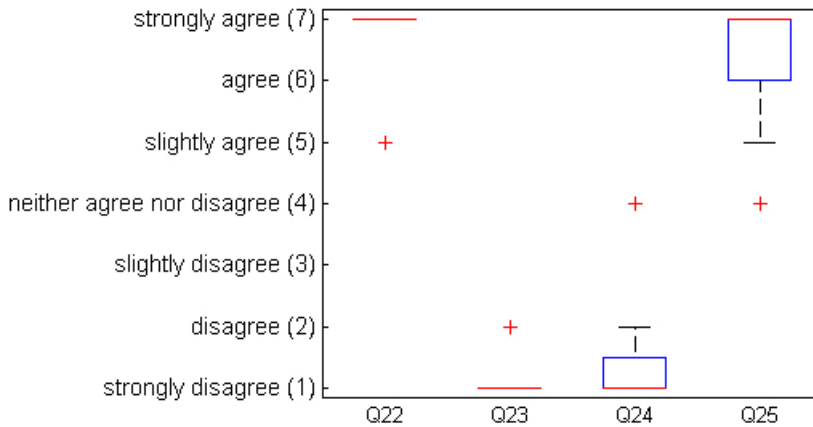


Figure 54: Results of the post-test feeling questionnaire (n = 12).

Q22 It is fun.

Q23 It is unpleasant.

Q24 I do not like it.

Q25 It makes me happy.

All but one trial participant stated that they “strongly agree” that the system is fun. One participant only slightly agreed to this statement.

Ten out of twelve trial participants “strongly disagree” that the system is unpleasant, two participants only “disagree”.

Ten out of twelve users strongly disagree that they “do not like” the system and the robot; the remaining two “disagree” and “neither agree nor disagree”.

Most trial participants strongly agree that the system makes them happy (7 of 12). Three participants “agree”, one “slightly agrees”, one “neither agrees nor disagrees” with this statement.

6.2.2.4 Motivational abilities of the prototype system

The main goal of the SAR system is to motivate users to perform physical training at home regularly, efficiently and independently in order to enhance therapy success.

Whether or not the system is able to motivate the user to perform the training regularly was asked using qualitative interviews and quantitative questionnaires.

The following research questions were derived as a base for the evaluation:

RQ_E3_A3: To what extent is the system motivating users to perform the training?

Figure 55 shows the results of the categories motivation on a 5-point Likert scale, illustrated in a boxplot.

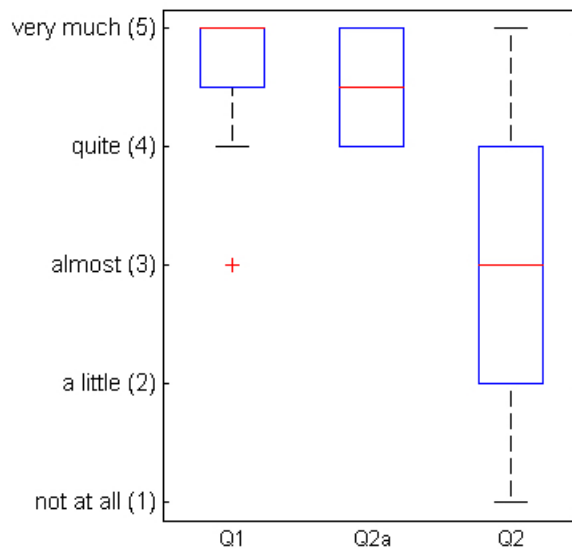


Figure 55: Results of motivation and training support (n = 12).

Q1 Nao is motivating me to do the exercises.

Q2a I am more motivated than when training with a standard training plan.

Q2 I am more motivated compared to only having a personal trainer.

As shown in Figure 55, most (nine of 12) users stated that they felt very much motivated by the robot to perform the exercises (Q1). Two users stated they were “quite” motivated; one user stated he/she felt “almost” motivated.

When asked to rate the level of motivation in comparison with a standard training plan (on paper with description and images), all users stated they would be quite or very much more motivated to train with the SAR system (Q2a).

When compared with a human physical trainer or therapist, user ratings were more mixed, some rated the therapist higher, others the SAR system, as shown by the high variance of Q2.

Related qualitative user comments to Q2 were:

"The robot has the advantage that no human observes you whilst you are training." – ("Der Roboter ist besser wenn man nicht beobachtet werden will.")

"Well, the small guy obviously can't show the exercises perfectly, it is not a human but only a robot" – ("Naja der Kasperl kann natürlich die Übungen nicht ganz richtig vorzeigen, es ist ja kein Mensch sondern nur ein Roboter")

"A human is of course something different to a robot – this is just in case of an emergency" – ("Natürlich ist ein Mensch was anderes als ein Roboter, das ist nur im Notfall")

The intended future use (Q10) is a very commonly used acceptance factor that also gives information about the current motivation to use the system in the future.

Q10 How often would you use the SAR system if you had it at home?

Test participants were asked about the intended frequency of use. Seven users would like to use the system three times a week, five users even every day. Hence, interestingly all users would like to use it at home at least three times a week (Q10) which implies a very high motivation to perform physical training with the system.

6.2.3 Impacts and added values

6.2.3.1 Comparison with non-robotic training aids

Any robotic solution has to compete against current methods in use and on the market. Given that robotic technology is most likely more expensive than current systems, there should be other gains that make users want such technology. In this context, a comparative study was undertaken to set the robotic system into the context of existing training support aids.

As a main research question behind this study we used:

RQ_E3_I1: What added value does the solution provide over currently used, similar, non-robotic training aids?

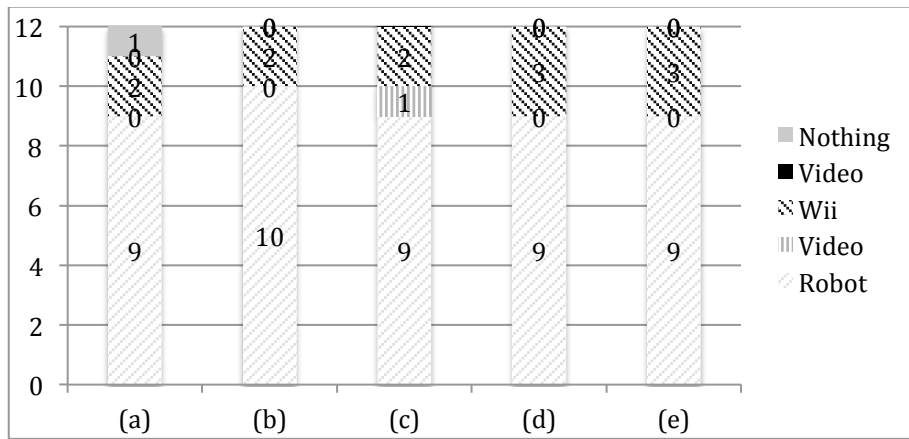


Figure 56: Comparative analysis of quantitative results of the different training support options.

- (a)** I prefer the following support for my daily training at home.
- (b)** Which system could motivate you best?
- (c)** Which system best displays the training exercises?
- (d)** Which system gave the best feedback?
- (e)** Which system would you choose if you could take one of them home for a month?

After a try-out session with all options, most of the participants preferred the robotic system across all questions. Nine out of twelve prefer the robotic system for their training, 10 out of 12 find the robotic system to be most motivational, nine out of 12 found when the robot displays the training exercises best and gave the best feedback. Also, nine out of twelve would choose the robot if they could take one of the systems home.

To answer why most people prefer the robotic system, a set of open questions were asked.

- (f)** What do you think are the weakest points regarding this training aid?
- (g)** What do you think are the most important positive points of this training aid?
- (h)** Why do you prefer this training aid?

The qualitative analysis gives the following main points to answer the reasoning behind the participants' preference.

1. The video-based and paper-based training do not provide any form of feedback. The users found the feedback to be very important in order not to perform the training in a potentially unsafe way, which is why most participants did not consider these training-support systems any further and concentrated on the comparison between the robotic system and the Wii console.
2. The robot-based training appeared to be more “natural” than the interaction with the Wii console. One user argued *“I find it unnatural to look at the screen as compared to the robot”* (*“Finde das unnatürlicher wenn ich do eineschau in die Flimmerkiste, als wenn ich ma den Roboter anschaue”*). Another user simply stated that she does not like computers as a reason, implying that the robot would in her eyes not qualify as a computer because the interaction is different than a typical HCI (*“Ich mag keine Computer”*).
3. Although a PC-based system could also initiate contact with the user and motivate them to perform the training, this functionality was imagined to work well with the robot. Users found the robot could come to them or initiate interaction when they should be motivated to perform the training (although this was not part of the test system). One user said: *“The robot could motivate me whereas I would have to turn on the Wii console first”* (*“Der Roboter könnte mich motivieren, die Wii muss ich erst wieder einschalten”*).
4. The display capabilities of the Wii console were the reason why three of 12 users found the Wii gave the best feedback. The Wii system is able to present graphs of the users’ performance whereas the robot can only give this information acoustically. One user found: *“You can see the enhancement over time for a better overview.”* (*“Man sieht die Steigerung über die Zeit für mehr Überblick”*) Another one: *“You can see how many calories have been burned”* (*“Man sieht den Kalorienverbrauch”*).

As a follow-up question to measure their motivation in using the various systems, the participants were asked how often they would use the different training aids in their daily routine.

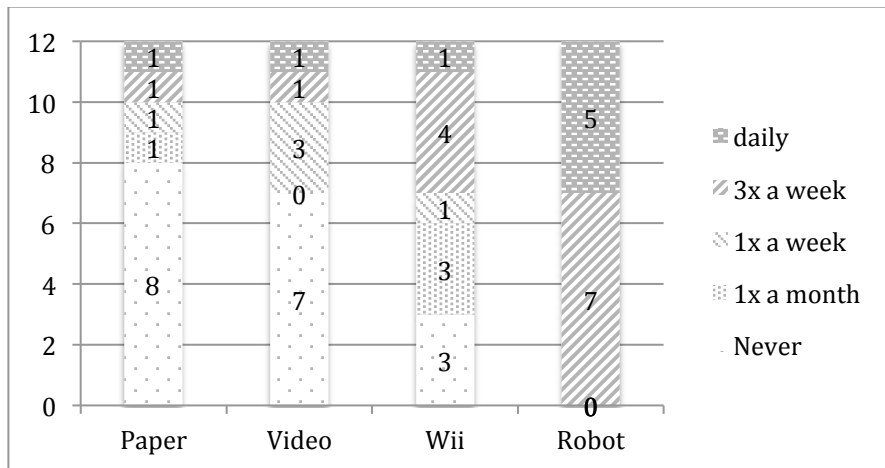


Figure 57: Planned usage of the comparison systems.

As Figure 57 shows, most users would not use the paper-based training instructions and over half of the participants would not use the video-based training support. Some of the users would use the Wii console-based training as suggested, three times a week. The motivation to use the robot-based training was highest; seven out of 12 would want to use it three times a week and five more would even plan to use it every day.

6.2.3.2 Effectivity of the robotic training system from a therapeutic viewpoint

The following research question was used to obtain the results:

RQ_I2: Is the SAR system effective from a therapeutic perspective?

The analysis and results regarding the effectivity from a therapeutic viewpoint were already published in [Krainer2014]. The results are presented as a summary here for the sake of completeness and to provide the reader with the big picture.

Physiotherapists partly found the quality and timing of the robot’s verbal outputs not precise enough and confusing. Verbal corrections were partly not recognized by the users. It seemed to the physiotherapists to be important to vary the spoken corrective phrases to provide users with additional help in case they could not follow the advice of the robot in the first place.

It was hard for users to watch the robot during the execution of exercises in a lying position. The spoken instructions during these exercises alone were not sufficient to let users follow the training schedule.

From physiotherapists’ qualitative comments, we can deduce that users followed the behaviour of the robot rather than the spoken instructions. Hence, users performed all

movements by mimicking the robot's movement, even if the robot said otherwise. This became obvious in an exercise where the robot instructed that the exercises should be performed as fast as possible (*"Do the exercise as fast as possible, faster than me."*) Only half of the users followed this request, the other half performed the exercises at the same speed as the robot.

The autonomous evaluation of the trainees' performances during the training only partly corresponds with the performance estimations given by physiotherapists. Physiotherapists were on average stricter in their assessment. Interestingly, the quantitative results of the two therapists who were asked also disagree regarding this aspect.

6.2.3.3 Usefulness within an institutional setting from the viewpoint of secondary users (carers)

The following research question was used to investigate the usefulness within an institutional setting:

RQ 13: Could the SAR solution be integrated into current institutional care?

Both carers found the system to be motivating and found it likely that older users would like to train with it regularly. *"The motivation would last three to four weeks, as compared to three to four days without any training system."* Both carers found the system would be suitable for group training as an add-on to a therapist, to simply show the exercises without correcting individual users, which then could become the therapist's role.

When asked about the usefulness of the shown SAR system, one carer found it would only be useful for patients in therapy rather than healthy seniors as those could still do other sport activities. The other carer found it is useful for seniors who are not afraid of technical systems in general and provided that the users are competently introduced to the capabilities of the system.

Both carers offered to include the system in their respective group training sessions for a test session.

6.2.4 Discussion of results with secondary users

After the conduction of E3, the results were discussed with a round of experts from care, therapy and technology (AAL) within an interdisciplinary workshop.

The goal of the workshop was to disseminate the study results and to discuss the insights with experts and potential secondary users.

6.2.4.1 Methodology of the final workshop

The workshop took three hours and was split into three parts:

1. Introduction and demonstration of the SAR prototype within the SSUT setting

After a short welcome, we gave a general introduction to the topic of assistive robotics for older adults. A live demonstration (the same as that used during the SSUT trials) in which a member of the workshop participants played the role of a user, was used as the starting point for the group discussions.

2. Presentation of the gathered results and insights

The gathered results of the conducted studies within E3 were presented to the participants.

3. Group discussion

As a first step, the participants were split into groups (group 1: "technology" and group 2: "therapy and care"). The following questions were used to trigger the discussion on the positive and negative aspects around the presented SAR prototype:

Q1: How could the SAR system be improved?

Q2: What existing issues are there which need to be solved for the system to become a product?

As a second step, a participant of the group presented the results to the overall audience and all results were again discussed in a second round with all participants.

A description of the workshop and agenda can be found in the ANNEX section 6

6.2.4.2 Participants of the workshop

The workshop was promoted by emailing national contacts and using the mailing list of the national association for active and assisted living (AAL Austria).⁴⁴

Seven participants were recruited. Among the participants were two experts from the care of older users, one physiotherapist and four persons with technological background in eHealth and AAL.

⁴⁴ www.aal.at

6.2.4.3 Results of the discussion

Q1: How could the SAR system be improved?

Enhance the audio output for users with hearing aids. As was also discussed within the results' analysis of the presented studies, the quality of the synthesized speech is non-optimal. Participants suggested improving this issue using e.g. pre-recorded real speech.

Reduction of the complexity of interaction. As can also be seen from the results of E3, some users had troubles following the multi-modal interaction due to their inherent complexity. Therefore, participants recommended evaluating if the consecutive speech and movement of the robot can be inhibited and instead a sequential flow of events can be implemented.

Regarding the feedback given during the training, it was thought that the SAR system should not report all found training mistakes but just the most important ones until the user corrected the mistake or at least tried to do so, thus reducing the number of utterances by the robot and thereby the complexity of output.

To enhance the demonstration of exercises and to make it easier for users to mimic the robot, very simple exercises should be chosen. As an example, one exercise should only target one joint.

Participants noted that the complexity of interaction might be well suited after some time of usage as the ease of use likely increases over time and users learn how to interpret the SAR's behaviour and interact with it.

Q2: What existing issues are there which need to be solved for the system to become a product?

Ethical issues. As the system advises the user autonomously, it could harm the user by giving bad advice. The solution found within the group was that the user still has to accept responsibility regarding the correct performance of exercises and usage of the system, in a similar way with the case of a car-route guidance system which also so advises its users.

Financial costs. A fictive price of €10,000 was given to the participants (current cost of robot is around €5,000). This was naturally considered to be too high for end-users. Still, the price was considered not to be off-putting, considering in particular institutional customers such as care and therapy centres.

System design as simple and portable solution. The complete system currently consists of several components (PC, Kinect, SAR). For productization, the technical components of the solution should be fused to the SAR only to become mobile and easily integratable into real environments.

Adaptability of the SAR system to changing user needs. The adaptability of the shown SAR prototype is limited to the selection of exercises by the physiotherapist. The system should allow the exercises to be customized to the patient by means of a simple-to-use graphical user interface that allows the selection of specific exercises for specific functional disabilities and age groups from a large set of exercises. A product has also to be capable of adjusting the individual exercises based on the functional limitations and capabilities of the patient. The exercises and training feedback should be adaptable to different levels of difficulty.

The adaptability of exercises should be limited as the relevance and clinical evidence of exercises that were customized by the individual physiotherapist might be questionable.

All parameters such as the SAR's output channels, the number of corrections, complexity of movements and motivational behaviour of the SAR should be customizable within a setup menu.

Training of secondary users regarding the usage of the SAR system.

The occupational qualification of physiotherapists should integrate training on such systems, in order to let physiotherapists understand the potentials and limitations of SAR systems and allow them to use such systems correctly.

6.3 Summary and discussion of evaluation results

To summarize the results we can relate to and answer the initially defined research questions in the following sub-chapters.

6.3.1 Summary of usability results

RQ_E3_P1: To what extent is the system in the current state usable; what usability issues exist?

The usability of the system was rated highly, as also shown by the scores of the Almere model: PU (PU mean = 4.4), PEOU (PEOU mean = 4.5) and perceived adaptability (PAD mean = 4.6).

Some users complained about the speed of execution of the exercises, which to them seemed to be rather fast with a slightly too detailed voice output from the robot; others about the audio quality of the voice output. Usability problems discovered include:

- Trial execution speed slightly too high.
- Dialog of robot speech contains slightly too much text.
- The placement of the robot was suboptimal for exercises in a lying position as the users needed to turn the head to watch the robot which conflicted with the exercise movements and provided only a poor view of the robot.
- The system needs to be highly adaptable to the users' physical abilities. In the presented version, the adaptation to the users' needs was undertaken by the researchers who set up the training schedule.
- For exercises that were not already previously known by the users, the users found it rather difficult to correctly perform the exercises purely from the robot's text and movements which resulted in partly ineffective training. (See also section 6.2.3.2 – Effectivity of the robotic training system from a therapeutic viewpoint)
- The multi-modal HRI seems to overstrain the users' attention. It seems likely that this effect will degrade over time after the users get to know the SAR system.

Given the uncovered usability issues, it seems clear that a final product should be customizable in terms of exercise presentation speed and exercise complexity. It is also clear that the training itself would have to be adaptable to the individual users' needs, in an ideal case by the users' physician or physical therapist.

Regarding the placement of the robot, it can be said that this is an issue that cannot be easily solved as the optimal placement would have to be different depending on the exercise. This restriction however is inherent to any type of training where the user mimics a (real, virtual or robotic) trainer.

RQ E3 P2: Is the system able to perform correctly from a technical viewpoint under real-life conditions?

Some technical issues occurred during the trials. In summary, out of 12 user trials, three needed to be postponed for technical reasons. Only one user recognized technical issues during the trials (localization problem), all others experienced a working system. The questionnaire results given by the user who experienced technical problems were

analysed separately regarding typical influences of perceived technical malfunctions such as influences on perceived usability and acceptance but were found to be within the range of other participants' answers.

RQ_E3_P3: What are the technical limitations of the SAR approach?

Many general technical issues of mobile robotics in smart homes could be circumvented by keeping the robot in a static position and by avoiding a truly interactive speech dialog with the user. Instead, interactivity was achieved by measuring body language which could be shown to be a robust technique within this setting; also because the used "Kinect" motion sensors are product-grade hardware.

Still, users commented on the space needed by the robot and it seems clear that the current autonomy of the robot in terms of battery life might become an issue over longer-term usage. The most challenging issues remain the limited stability of the robotic platform and its overall frailty, which makes hardware issues to be expected over longer usage and currently hinders the execution of long-term trials in real-users' homes.

6.3.2 Summary of acceptance results

All used metrics such as the Godspeed questionnaire, the Almere questionnaire and the specifically developed acceptance questionnaire show high levels of acceptance from the trial participants towards the SAR system.

When analysing the Godspeed results, it becomes clear that the trial participants did not experience the SAR system to be what it is – a clearly inanimate computer – but because of the shape, look and behaviour of the robot, assume at least a mediocre level of human-like animacy and anthropomorphism which puts it "*somewhere between a dead object and a human*" [trial participant] in these aspects. By comparison with an earlier prototype ("KSERA"), it could be shown that the design concept of implementing a fluid interaction without significant delays and further interaction cues such as gestures and mimics contributed positively, in particular to the anthropomorphism and animacy scores and hence to the acceptance of the system. The design concept of implementing a nice and kind but still strict attitude for the robot together with the child-sized friendly appearance of the used robotic platform also led to very high likability measures in comparison with earlier and similar studies and prototypes. The question as to whether or not the system is intelligent polarized the users but still resulted in higher scores than those received for similar robotic prototypes. Safety issues were not a concern of the

user group, the system received high safety ratings which we believe is a benefit of a small and static robot over a human-sized moving robot.

The Almere model augments the results of the Godspeed questionnaire by adding further acceptance measures such as the perceived sociability, social presence and social influence. All but one construct of the model resulted in very high scores. The high scores for “anxiety”, “trust” and “perceived safety” support the corresponding result (perceived safety) of the Godspeed questionnaire and show a strength of the SAR system. This is a particularly important finding as older people are known to be less comfortable with new technology in general and robots in particular [Arras2005]. Our research hence supports the findings of [Libin2004] that older people prefer unthreatening robots with a female voice, small size, and which are slow-moving and less autonomous.

The PENJ score corresponds well with the Godspeed score for likability. The lowest score was received for the construct “social presence” which is a weak point of the SAR system, as also shown in the comparative analysis with similar projects. We believe that this weakness is due to the robotic sound of the speech output, the small size of the robot and the synthetic material of the robot’s housing. In particular, the robotic speech output was sometimes named as displeasing by trial participants. It has to be said that the questionnaire of the Almere model partly suffered from low comprehensibility, as stated by the trial participants, resulting in a higher variance of answers given. Some of the constructs also suffer a low or even unacceptable Cronbach alpha score which states that the internal-consistency reliability among a group of items is low, which might be related to the problem described above of questions that were hard to correctly interpret by the users.

The results of the questionnaire used to assess the intended future use and the PU of the system acknowledge the high results of the Almere model constructs “perceived usefulness” and “intention to use”. The system was perceived as being intended mostly for older users and as “very beneficial” for the trial participants. This result has to be interpreted knowing that the price for purchasing the system was not considered as a factor for the acceptance evaluation since the system is in the state of a very early prototype and the study focused on the evaluation of the future potential of SARs for physical training.

All users agree or even strongly agree that the training with the SAR system met their expectations, which is a very promising and interesting result given the earlier results of E1 and E2 and the general fact that a humanoid robot was used which is known to raise expectations which are too high to fulfil (as the humanoid shape leads users to expect to be able to communicate with the system in a similar way as with a human), see also [Broadbent2009]. This is one of the reasons current research projects often prefer non-humanoid robots.

Based on the gathered results, we conclude RQ_A1 in that a large audience among the user group of older people would accept to use the SAR system if it was provided for free.

Emotional effects of the solution

Right after the training, most trial participants were very excited about the new experience of training with a robot and commonly commented that they had found the training to be fun, which answers RQ_A2. Users were mostly amazed about the behaviour and functionality of the robot, in a positive sense. Some participants directly started to give feedback about details that could be optimized, such as the quality of the speech output and the training intensity. Two participants stated after the training that they found the system motivating. Two other participants stated that they found the training to be too ambitious.

The quantitative post-test results show that nearly all users found the system to be fun and even strongly agree that using the system has a positive influence on their mood. In addition, users to most extent disagree that they do not like the system or find the system unpleasant. These scores fit very well with the acceptance scores of the Godspeed construct “likability” and the Almere construct “PENJ”, as shown in the analysis of the perception of the robot and its intended use.

Motivational abilities of the prototype

As an answer to RQ_A3, trial participants stated that they felt very much motivated by the system to conduct the training and that the motivation is one of the key benefits of the system since the robot could come and ask you to perform the training daily. Compared to a standard training plan, users are confident that training with the SAR system is more motivating. When compared with a human trainer, users gave mixed reviews. Some users stated that a machine could never replace a human and a human

trainer could give better/more detailed feedback. Others favoured the robotic trainer because they found it would be more patient and they would feel less observed and under pressure to perform, compared to when with a human trainer.

All users would like to use the training system at least three times a week, which is much higher than their current average frequency of training, which further shows a high motivation to use the system.

The results of the quantitative and qualitative evaluations strongly support the result of the Almere model construct (“intention to use”), which received a score of 4.9 out of 5 on a 5-point Likert scale.

The results of the SAR user-motivation evaluation clearly suggest that the tested system has a strength in motivating users to perform otherwise unenjoyable behaviour.

6.3.3 Summary of impacts and added values

Discussion of results regarding the comparison of training aids

The comparison of training aids showed that most participants preferred the robotic system with regard to aspects of acceptance, motivation and intention to use at home. As was expected, most participants would not use paper-based training aids which currently are standard in physiotherapy; also, video-based aids were mostly declined due to the lack of interactivity and feedback. Instead, participants focused on the comparison of the Wii console-based training game and the presented robotic training aid. The natural interaction and the non-technical design as well as the PU of the robotic system were considered strong positive points for the acceptance of the SAR solution. The possibility to approach the user and initiate interaction to motivate the undertaking of the exercises was considered a valuable differentiation point to other solutions. On the other hand, the robotic system was considered to need more physical space than might be available at home and the lack of a visual-screen interface for display of key training results was criticized. Overall, most participants preferred the robot for their training at home and intended to use it significantly more often than the other presented solutions. The provided high frequencies of use (most participants wanted to use the system at least three times a week) have to be understood considering the gap between intention and actual behaviour. Sheeran found in a review of six prospective studies of cognitive predictors of health behaviours that 47% of intenders did not act on their intention [Sheeran2002].

Discussion of results regarding the therapeutic effectiveness

The autonomous performance assessment by the SAR system showed results that were similar to but only partly corresponded with the opinion of physiotherapists. Here additional input from many physiotherapists would be needed to train the system to provide better feedback.

As could be shown, participants followed the movements of the robot as well as they could, partly even ignoring the robot's verbal instructions. We can interpret this effect as a clear indication that the demonstration of exercises by a SAR has an added value over verbal or written instructions.

According to physiotherapists, the usage of a SAR as a trainer for physical exercises for older people is possible and works well [Krainer2014]. A user interface for physiotherapists would be beneficial to easily adapt the training exercises to their patients' individual needs.

6.3.4 Summary of results from discussion with secondary users and AAL experts

The group described the system as an innovative solution that shows the right direction for the development of SARs and could eventually enter the market but which still needs a thorough rework to achieve the necessary functionality and robustness.

The participants mostly provided concise ideas for the improvement of further prototypes and ideas for additional use cases. Additionally, ethical issues were discussed and the necessity of a high rate of adaptivity towards the individual users' needs was highlighted.

Surrounding conditions regarding the application of SARs such as training courses for therapists as well as the economic potential of the solution, the target price and how it could be lowered, was discussed.

7 Summary and discussion

The aim of this chapter is to summarize the insights gained within this dissertation, set them in relation to what is already known and discuss them critically by putting them within the context of other research to verify or falsify the knowledge gained by other views.

7.1 Summary of evaluation results

This section summarizes and discusses the results of the conducted three main evaluation phases, which were presented in chapters 4 and 6. The results per evaluation phase can be found in the individual summary of the results sections of these chapters. Here we take a top-down view on all presented studies and try to gain a better understanding on the generalizability of results over the course of the studies.

7.1.1 Summary of performance results

The assessment of technical performance was undertaken in all evaluation phases (E1-E3) to gain insights on the applicability of the solution from a technical viewpoint and to understand the influences of technical issues on usability, acceptance and impact factors. We want to present the summary along the initially defined research questions.

RQ1: To what extent are current SAR systems applicable under real-life conditions from a technological perspective?

Over the course of this dissertation, three SAR prototypes were evaluated. By conducting sets of user trials, we could show that SAR systems are applicable to realistic settings if certain technological restrictions are considered and compensated for.

While we had to rely on a Wizard of Oz methodology to cope with the specific technological functionalities of navigation, localization and speech recognition issues during our studies in E1 and E2, the last prototype showed that when considering the current technical limitations and restricting the SAR functionality to technically feasible use cases, **it is already possible to realize SAR solutions that are helpful to users**, are accepted by them to a great extent and which can be integrated into realistic settings. If compared with other studies (e.g. [Pripfl2016], [Pigini2013] and [Schröter2014]), E3 seems to be one of the few success stories in which technical issues were not major limitations of the quality of presented studies. Papadopoulos et al.

reported that out of 12 reviewed studies eight reported of technical issues influencing user studies, back this point up [Papadopoulos2019].

Although initially planned, we were not able to perform real-life trials in users' homes in any of the evaluation phases due to considerations of safety, which were also related to the limited robustness of the prototypes. This is in line with most literature (compare also the state of the art section 2.2.2). However, we were able to demonstrate the feasibility of integration and the applicability of PT3 on the real setting of a care institution's gymnasium. Similar integrations to remote care facilities were rare at the time of our study, but are at the time of writing becoming more numerous as shown by more recent studies that were conducted in different forms of residential care facilities [Loi2018], [Khosla2017], [Hebesberger2017].

Our proposed strategy of enhancing the robustness of solutions by reducing solutions to a limited set of technically already robust functionalities, is backed up by the fact that the company *QBMT (Zora Robotics)* was already able to productize a solution based on the same robotic platform as used in E1-E3 by implementing a very similar concept as shown in PT3. The company had already stated in mid-2015 that they integrated 88 pieces of this solution into care institutions in France, Belgium and Netherlands and also received positive feedback from the secondary-user group of caregivers [Deblieck2015], [Payr2015].

RQ1a: To what extent is current SAR technology able to satisfy relevant user needs?

Within this dissertation, we took the approach to realize a set of use cases to assess the potentials of a SAR solution to target specific needs relevant for the target group. The user needs addressed were also chosen based on the feasibility of the technical implementation. Therefore, the developed SAR technology cannot satisfy all relevant user needs, but we found use cases, such as the demonstration of physical training, that were accepted very well by the target group and posed a clear advantage over alternative solutions. This result is in line with similar studies of Juan Fasola and Maja Mataric [Fasola2011] who found that the particular use case of coaching physical training is better accepted when presented by a real humanoid SAR than by its virtual representative and Mann et al. who found that people are more likely to participate into the training and follow directions provided by technology when approached by a robot in comparison to a computer tablet [Mann2015].

We also could show that one hurdle was that some of our demonstrated use cases did not pose a clear added value over the current state of technology; in particular when comparing the solutions with a similar implementation on portable touch-based devices. As SARs do not provide physical assistance, several of the tested use cases such as video telephony and reminders could alternatively be realized by touch-based portable UIs such as tablets and smart-phones which are cheaper, easily available and already well accepted. To compensate the presumably higher price, we think that the added value of a SAR solution must be clear to the potential customer. Here one clear added value could be shown for the use case of guiding physical training as the physically present SAR solution was preferred over any of the compared technologies. In addition, we need to consider that the versatility to offer a large set of potential use cases is a strength of the SAR, as well as how it presents them by using an easy to understand and socially meaningful communication. Hence, use cases that could easily be realized with other, more ubiquitous technologies can make sense if packaged together with functionalities that make use of the specific capabilities of SARs.

RQ1b: Which flaws and challenges need to be solved on a technological base in order to allow an acceptable HRI?

- Reliability of technical components

All evaluated SAR systems were prototypes which were developed to test the feasibility of the chosen approaches and therefore the focus was not laid on product-grade reliability but on functionality. Reliability issues were common, in particular during E1 and E2, but also the third prototype used in E3 showed reliability issues that led to the postponement of three trials out of 11. In all trials, a team of technical experimenters was needed to guarantee the robot's performance by validating the functionality just before the trial, and within E1 and E2 also by supplementing unreliable functionality through Wizard of OZ techniques.

We know from other projects and products that very simple approaches (as taken in "SERA" [Heylen2012]) and approaches in which the SAR acts semi-autonomously on the commands of e.g. a caregiver [Deblieck2015] have reached a level of technical robustness that allows long-term operation in real environments.

For more complex solutions, the achievable reliability depends on the technical state-of-the-art of the particular needed functionalities. Functionalities such as navigation in cluttered environments of users' homes, perception abilities such as speech, face and

gesture recognition, as well as autonomous decision making on uncertain data would be needed to realize the initial idea of a mobile multi-purpose SAR that shows human-like interaction capabilities in users' homes. These are still within a state of research and hence currently not providing the robustness needed to realize an assistive solution that vulnerable users can rely on. Papadopoulos et al. found in their review the same key technologies currently missing [Papadopoulos2019]. At the time of writing in 2020, these limitations are still hindering research, but given recent developments, it seems breakthroughs in autonomous navigation will soon allow the development of assistive robotic systems that are indeed capable of including interactive services that rely on autonomous navigation in cluttered environments. One example, which is currently under development but enhances the state-of-the-art is the "temi" robot.⁴⁵

- HRI challenges

All evaluated prototypes were based on the 57cm tall, biped, humanoid robotic platform "Nao". This design choice was taken considering the initial idea of a multi-purpose assistive companion robot, which also fulfils the desire of creating a human counterpart, as already described in section 1.1.2. However, this choice comes at a price as users expect the system to behave according to its appearance and hence only accept such a solution well if it is capable of interacting with them in a way fitting the design, which according to our results has to be similar to the interaction between real humans for a humanoid robot. Whether or not the interaction between SARs and humans should be similar to the interaction between humans can be controversially discussed for ethical reasons (see also section 7.4b), but it seems clear that human like communication enhances the user-acceptance as this is also reported in studies such as in [Bedaf2017], [Bickmore2005] regarding virtual agents, or [Breazeal2002] who even suggests that robot movements should be programmed by directly imitating humans.

Participants suggested that the PENJ when using the system would likely decline over time due to the repetitive nature of the interaction. This indeed could already be shown for the few undertaken long-term studies such as [Fernaesus2010] and [DeGraaf2016]. In order to cope with this demand, we suggest that a larger set of behaviours and voice-interaction flows is needed that allows the system to interact more diversely and naturally by wisely choosing the right interactive behaviour. In addition, memory functions that let the SAR choose answers based on earlier interactions with the same

⁴⁵ <https://www.robotemi.com>

user need to be implemented. To achieve this, the SAR would need to be able to detect and recognize the user, allowing also adaptation to the specific user's preferences. Additionally, the SAR would need cognitive functionalities that facilitate decision-making based on the context of use and the actual intentions of users in order to choose answers wisely.

The HRI of the presented prototype was also limited regarding aspects of turn taking. The facilitated turn-taking concept built upon listening to the users' commands at specific times, recognizing the speech of the user and acting according to it. On one hand, this functionality did not perform well within the evaluation due to speech recognition issues, but even more importantly, it was conceptually limited to interpreting speech only, leaving other interaction cues such as eye contact, gestures, mimics or body posture aside. As an example, during the trials a participant tried to send the SAR back to the starting position by gesturing. In this respect, the perception abilities of SARs have to be increased to allow reliable detection and interpretation of additional interaction cues to supplement speech-based interaction. This point is backed up by Al-Shamayleh's review of vision based gesture recognition techniques, which makes clear that systems are currently still in the state of research and not available as product grade systems that would be needed for real-world use cases [AlShamayleh2018].

Several further drawbacks could be uncovered during the evaluation phases regarding speech-based interaction. In all phases from E1 to E3, some users found the robot's voice hard to understand, a problem also found in several other studies [Papadopoulos2019]. During E1 and E2, participants verbalized this phenomenon; in E3 it became obvious, as users did not always follow what the robot said but instead mimicked the robots' movements, which seemed to be easier for them. Here we see that the additional provision of a second output modality led to the inclusion of parts of the user group that was formally excluded because of age-related hearing issues. This presents another strong point for the multi-modal capabilities of humanoid SARs, and we suggest making use of them.

The voice output of the SAR was also hard to remember for some participants. In E1 and E2, we received the feedback that longer texts in particular, such as given when reporting on weather, were not only hard to understand but also hard to remember due to their uni-modality. The invited primary users also partly had difficulties in remembering the voice commands needed to trigger the interaction with the SAR in E1 and E2, despite having them written on a piece of paper in their hands. In E3, the

interaction design was much simpler as the robot could be started by a press of a button and after that, no commands had to be remembered, as users only had to follow the SAR's instructions immediately. Here we can learn that when dealing with older users, it seems important that the interaction design does not rely on the users remembering a set of commands. Instead a user interface that also allows the user to control the robot by at least another reliable input channel such as buttons seems advisable.

In the phases E1 and E2, the interaction with the robot was considered to be too slow and technical delays hindered a smooth and efficient interaction. Therefore, PT3 was developed in such a way as to take care of avoiding delays, enabling quick interaction with users and using all provided output channels (voice, sounds, gestures, movements and simple mimics). The resulting interaction received better acceptance and was even considered by some participants from both primary and secondary user groups as to be slightly overwhelming for first-time users.

- Further challenges

Primary and secondary E3 participants argued that the space needed (approx. 2 x 3 metres) for the presented solution would be a problematic point considering typical users' homes and their limited space availability. In addition, primary users stated that they would like to be able to stow the robot away in periods when it was not needed, which is generally possible as it is a mobile solution but conflicts with the system's ability to engage the user pro-actively.

Within this section we could show that a number of technical issues are still present, which are hindering the further development of SAR based use-cases. This is and was already done by other researchers such as Papadopoulos et al. [Papadopoulos2019]. We discussed them here in order to highlight issues and thereby draw the attention of developers towards them. To reach our goal of assessing the potentials of SARs it seems important that technical limitations can be alleviated. Otherwise we are limited to measure the potentials of a small sub-set of possible functionalities as could be shown within this dissertation by the use case of demonstrations for physical training.

7.1.2 Summary of acceptance results

RQ2: To what extent do both older users and their carers accept socially assistive robotic solutions for the support of older users at home?

Acceptance models consider a number of factors related to the PEOU, PU and social interaction as relevant for the acceptance, which is typically defined as the user's intention to use the system in the future [Davis1989], [Venkatesh2000], [Heerink2010].

Within this dissertation, it is argued to augment the traditional acceptance model by further including acceptance factors related to the main advantages of SARs over other (e.g. touchscreen-based) technologies. In particular, the social perception of the robot, the motivational capabilities and emotional influences of the SAR, and further specific functionality-dependent factors as detailed in the following paragraphs.

- Intention to use (ITU)

Heerink et al. found that the ITU of an assistive robot indicates a later use of the product, as participants who showed a higher ITU also used their prototype for a longer duration in a real-life trial [Heerink2008a]. But obviously there is a gap between the ITU and the actual use, which was already investigated by Bogozzi et al.. "Clearly one's intention to use the computer, which leads people to acquire the systems in the first place, does not assure that sustained usage will occur" [Bagozzi1992]. So although we cannot exactly predict the usage of a SAR by evaluating acceptance factors, we can provide estimations that also can be used in a formative way to guide the design of SAR solutions.

Heerink's Almere model was used to estimate the ITU and at the same time gain information on a well-elaborated and wide set of SAR-relevant acceptance factors.

We found an increasing acceptance leading to a higher ITU over the course of the three developed prototypes, showing that the overall approach of conducting formative evaluations was successful. From the beginning (E1), we have seen that participants showed a positive attitude towards the robot despite partially severe technical issues and doubts regarding the applicability in real life. The robot was perceived as friendly, happy and mindful, and hence did not invoke anxious reactions but was trusted by the participants. If users showed any anxious reactions, then only in relation to the fear they could break the SAR and might then become dependent on technical support.

In all evaluation phases, participants found the systems to be enjoyable and easy to use. Users expected that over time, the enjoyment in using the system could decline as it is

partly based on the novelty of the solution. The ease of use was found to likely increase over time when users came to learn how to handle the system.

Participants stated that the system shows sociable behaviour and also expected that users might develop a kind of friendship with the device over time. However, several participants were concerned that this should not be a goal for ethical reasons as the robot is not a social being and hence can only simulate friendship. This ethical dilemma is discussed further in section 7.4b.

Interestingly participants in E3 did not perceive the SAR as a stigmatizing technology but rather as a high-tech gadget that they would like to show to their family and friends. We consider this to be one of the most important features of SARs as they are perceived as novel and trendsetting whereas many current assistive technologies (e.g. compare fall-detection wristbands) are not accepted as they are perceived as stigmatizing.

In E2 and E3, we found that the social presence of the presented SAR solution is also rather low compared to other SAR robots. We expect this to be related to the small size, the shape and material (e.g. not fur, but plastic), which, apart from the general shape, does not resemble a human but rather an artificial creature. Users described the system to be lively but not a living creature. Despite acceptance research suggesting that social presence is a positive acceptance factor [Heerink2010a], we argue that a low social presence might be beneficial for the global acceptance of SAR solutions on the long run as social presence enhances the emotional attachment to the device, which is seen critically by several participants for ethical reasons.

PT3 in particular was seen as beneficial to use and the ITU was rated significantly higher than for comparative non-robotic solutions. Regarding PT1 and PT2, users commented that they find the robot beneficial to use but are also partly unclear about the added value in comparison with current technologies such as touch-screen tablets. PT3 was the only solution that met the high expectations of primary users towards an anthropomorphic robot. We believe this is in large parts due to us focussing on the development of a single use cases until it reaches a high technology readiness and the fact that the used system is by design well suited for the particular use case of demonstrating physical training as also discussed in section 7.1.1.

- Perception of the robot

The anthropomorphism of the SARs could be increased during the development of this dissertation. When comparing PT2 and PT3, we found clear qualitative preferences for

the newer prototype and highly significant statistical evidence regarding the display of movements, which were described as being more elegant in PT3. This tells us that the approach to designing a more vivid interaction and following the developed SAR design heuristics were helpful in this aspect and led to a more acceptable solution.

In all evaluation phases (E1-E3), participants had trouble rating the human-likeness of the SAR system. Some clarified this trouble by stating that what they see is clearly a machine that acts like a human, hence they cannot decide between these two representations. Several users found that a SAR cannot be largely described as conscious or human-like; that it would be a mixture between a human and a machine. Others approached the system more openly and saw certain human-like aspects and thought that the system was close to being conscious. Parts of this user group were even fascinated by the human-likeness. We therefore argue that anthropomorphic SARs are polarizing and might only be well accepted by user groups with a certain level of technology or, more specifically, robot acceptance. This result is in line with the reasoning of Goudey & Bonnin who found that an anthropomorphic appearance improves acceptance by people with practical experience of similar technologies (smartphones), but it reduces acceptance of other people [Goudey2016].

In E1 and E2, users found the animacy of the prototype to be rather static and “robot-like”, arguing that they expected a higher interactivity and more vivid human-like behaviour. After a re-design in E3 in which these comments were targeted, users had difficulties in rating the robot’s attributes between “organic” or “lifelike” as those attributes tend to polarize (as discussed above), but the majority clearly favoured the term “alive” over the term “dead” for the robot.

In all evaluation phases (E1-E3), participants concurrently found the system to be very likable or even lovely, making the likability a clear strength of the solution. In particular, the nice appearance of the “Nao” robotic platform influenced this result positively as users commented on how it was especially likeable and resembles a small child. We reason that the appearance of the robot is a strong acceptance factor and in this case was well chosen.

As an unexpected result, the user group was divided regarding the perceived intelligence of the SAR solution in all evaluation phases. In particular, participants with low technical experience overrated the intelligence of the robot and even sometimes found it to be truly intelligent, whereas more experienced users were often positively

surprised by the shown intelligence but partly doubted that the system possesses the necessary intelligence to truly help in day-to-day scenarios over the long term. Hence, we learned that more experienced users rated the intelligence lower which is in line with the earlier results of Siino et al. [Siino2005]. Because the target group of older adults becomes technically more experienced over time, we could argue that this acceptance factor will decrease in the future. However, we assume that the current rapid advancements in AI will cover this effect up, leading to a higher perceived intelligence and acceptance of SARs in the future.

In all evaluation phases (E1-E3), users found the system to be safe as they could not imagine that it would physically hurt them because of its small size, limited strength and likable appearance. Given our results, we consider the perceived safety to be a strength of the solution in comparison with other taller SAR systems. This finding is in line with earlier results of Broadbent et al. who found that older people prefer robots to be “unthreatening, with a female voice, small size, slow-moving and less autonomous, with a serious aspect and single colour” [Broadbent2009, p322].

- Motivational capabilities

The motivational capabilities of the SAR were assessed in E1 and E3 focussing on the motivation to perform physical exercises. Whereas participants in E1 found the SAR to be generally only quite motivational, they found themselves very motivated during the conduction of the physical training scenario. This scenario was later extended in E3 and studied in detail, where we found that not only 10 of 12 users found the PT3 solution to be motivating them “very much” during the training, but that nearly half of the user group (5/12) felt quite or even “very much” more motivated than with a human physical trainer. This surprising finding is also consistent with earlier results from E3 pre-trials and qualitative comments. Although users commented that the SAR could never replace a human, which is in line with other studies such as Sparrow & Sparrow [Sparrow2006] and Decker [Decker2008], some stated particular advantages, including that it would give “*better, more detailed feedback*” and that they do not feel as observed as with a real trainer, giving them more personal freedom to train when and how they want. In that sense, the idea of designing the SAR as a tool rather than a companion was well picked up by the participants.

Secondary users acknowledged the high motivational capabilities and found that the system could motivate users to train for “*three weeks instead of three days*” when

compared with the current state of the art where users receive either no support or a paper-based training guide for home training. This finding is interesting in three dimensions. Firstly, it fortifies the encouraging results on motivational capabilities; secondly, it indicates that the system also has a therapeutic effect in the eyes of secondary users, and thirdly, it tells us that the motivational effect is expected to wear off after some time, which is consistent with our findings on PENJ and the novelty effect over the long term.

- Emotional influences

In all evaluation phases, the vast majority of participants subjectively found that using the SAR system has a positive influence on their emotional state as it is fun to use and not boring. By triangulating data we found that the participants were too positive in their subjective estimation, as all participants in E2 showed clear signs of boredom (yawning, looking at a watch, commenting on the slow behaviour) during the demonstration of the scenarios. In E3, seven out of 12 users even strongly agreed that training with the SAR is fun and made them happy; most of the rest agreed to varying degrees. We can summarize that using the system for about one hour had a positive effect on the emotional state of most participants. This is in line with other authors who report interacting with a SAR to be an enjoyable experience and hence having a positive emotional effect, compare [Bedaf2017], [Beuscher2017] or [Loi2018].

- Further acceptance factors

Ethical issues can influence acceptance and are discussed in detail in section 7.4

The price, distribution model, marketing strategies or payment options were not investigated in detail, although the overall financial impact on the customer has an obvious influence on the intended use of the solution. We considered it too early to give price estimations for the developed prototype systems. In a discussion round with experts from technology, care and therapy during E3, we discussed a high potential price of €10,000 for the solution, which the experts found to be too high for individual consumers but not off-putting when a care institution is the target group in mind. This finding is backed up by the fact that the company *Zora Robotics* was already able to sell or rent robots with similar use-cases and prices to care institutions [Debliek2015].

RQ2a: How do acceptance rates compare between robotic solutions and technological but non-robotic solutions?

SAR solutions do not only have to be assistive and accepted, they also have to compete against other solutions that tackle the same problems on the market. We compared the presented SAR solutions against a touch-screen tablet in E1 and against a set of solutions targeting the same goal of supporting physical training in E3. Throughout the evaluation phases, we gathered some sceptic comments regarding the added value, in particular compared to technologies like a smartphone or tablet as they could implement a similar functionality and satisfy the same user needs as SARs, in particular regarding the functionalities of reminding, motivating, entertaining, communicating and serving as an information centre (compare also 2.1.1 – Typical functionalities of SARs).

In E1, SARs were preferred over touch-screen tablets as users found the SAR to be more motivating and to be “*funnier than a PC*” because using it seemed “*pleasant and cheering*”. Further, users commented that it was easier for them to exercise by simply following the robot’s movements. In E2, we found that in particular the scenario for video telephony suffered limitations in the SAR version because the robot had to find and move towards a free spot on the wall to project the video stream, which made a participant think a tablet would be simpler and more effective to use. In E3, a clear majority of users preferred the SAR training system to all comparative systems, including a commercial training system from *Nintendo*, as they found the robot better at motivating, at giving feedback on the quality of training and at displaying the exercises. Furthermore, the participants found it more natural to interact with the SAR in a human-like way as compared to the interaction with the *Nintendo* Wii console.

We also found that a user with strong scepticism against PCs profited from the SAR system, which was not perceived as a PC, which adds to the point that SARs would be beneficial in particular for the target group of users with low affinity to PCs. Our findings in this matter are backed up by other researchers who found that the physical presence of robots can enhance the acceptance compared with PCs [Bedaf2017], [Breazeal2002].

In particular, smartphones or small tablets question the SAR’s movement abilities, as these devices can always stay with the user and thereby reduce the necessity of the device to move itself in general. Nevertheless, the ability to move towards the user was considered a main advantage of the SAR, mainly because users considered it could be able to find them and trigger an alarm in an emergency situation. Additionally,

participants in E1 and E3 stated that the SAR system could motivate them better when walking towards them and initiating a conversation as compared to a PC or tablet-based system which they would have to turn-on first.

The main drawbacks of the robotic solution when compared to computer tablets turned out to be the high-assumed price, the technical complexity and associated maintenance efforts, the space needed by the solution and the lack of a visual display to present high-density information such as a summary of training results or the weather report.

In E3, we found the ITU and hence the acceptance of the presented SAR solution to be superior to all compared solutions. Most users would not use the paper- or video-based training aids and only five of 12 participants would use the *Nintendo* Wii-based training, whereas all 12 participants would want to use the SAR system in the recommended training intensity of three training sessions a week.

To sum up we can conclude that SAR systems can compete with regular technologies regarding specific use cases depending on their price, technical robustness, and practical matters (e.g. space needed).

RQ2b: What behaviour of SARs is socially accepted?

During E1 and E2, many participants expressed their disappointment with the behaviour of the presented SAR prototype because they expected it to be capable of acting in a more human-like way. We figured this to be due to the anthropomorphic and even child-like shape of the robot. This confirms previous results of Tapus et al. [Tapus2008] and we hence recommend in section 4.7 – “Heuristics for further design and development (design principles)” that the behaviour and personality of a SAR has to be designed to match the appearance of the robot.

Additionally, literature suggests that a socially expressive behaviour is beneficial for acceptance [Heerink2008b], [Broadbent2009]; however our results contradict these findings as we found that not only did a more strict functionality-oriented personality result in similar acceptance rates (comparison of E2 and E3) but we argue that it is even more acceptable from an ethical point of view since users might not get as easily emotionally attached to the SAR, which was a major concern of participants. We hence recommend (see also more detailed in section 4.7 – “Heuristics for further design and development (design principles)”) that the robot’s personality should be designed in a functionally-oriented manner that mainly resembles the SAR’s function as a tool rather than a companion. This reasoning is backed up by a group of researchers at the UK

Engineering and Physical Sciences Research Council (EPSRC) who drafted rules for robotics including rule four that says: “Robots are manufactured artefacts. They should not be designed in a deceptive way to exploit vulnerable users; instead their machine nature should be transparent” [EPSRC2010]. Compare also section 7.4 reg. ethical issues.

RQ2c: How can solutions be integrated into the daily life of users and daily work of carers?

Contrary to our initial plans, the prototypes for E1 and E2 could not be integrated into the homes of older users due to technical restrictions regarding their safety and robustness under real-world conditions.

Regarding the use within an institutional care setting, care experts during E1 were sceptic about the usefulness of PT1 as they found the solution would better fit the needs of users with a lack of social contacts at home. Within the institutions, human carers are present that can cope with the same tasks and the robot could also be seen as replacing them.

The integration into the homes of older users additionally faced conceptual issues. The SAR system needs space to be stored and charged, as well as space within the dwelling itself to move. Some participants stated they would not be able to provide these prerequisites. Whereas this could be solvable by adapting the users’ homes, we also found stronger limitations to the tested solution regarding obstacle-avoidance capabilities. Within the conducted trials, we had to significantly alter the environmental conditions (sound and light settings as well as furniture and placement thereof) within the test environment to become able to reliably perform the trials with the autonomously moving robot. Hence we learned that the technologic state-of-the-art is insufficient to integrate an autonomous, biped, anthropomorphic robot into the daily life and homes of primary users. Asking two care experts in E1 about the applicability within an institutional care scenario revealed similar results but for non-technical reasons. We therefore considered stripping the SAR solution of technically non-robust functionalities and facilitated the development of a solution that stays in place, which led to the development of PT3.

PT3 could be easily integrated into a therapeutic setting within the gymnasium of a senior-citizen centre. Also, two therapists were willing to integrate the system into a test session during their regular training. Such a test session was not conducted but we

know from the “Zora” project [Debliek2015] that the integration of a similar solution was successful by letting the robot act, as suggested also by our secondary users – as an assistant to the therapists that merely displays the exercises to be conducted within a group training session whereas the therapist is then able to concentrate on correcting the execution of exercises to enhance the quality of training.

7.1.3 Summary of prospective impacts

RQ3: Do SAR robots have beneficial effects for the support of older people at home and if yes, which?

Although no long-term real-life study could be undertaken and hence impacts could not be measured, during analysis we found several effects of using the technology, providing us with the possibility to derive prospective impacts. The following list of found effects serves as an answer to RQ3.

Effects on the emotions of users

As already described in section 7.1.2, the usage of the SAR solution had a predominantly positive effect on the emotions of the test participants in all evaluation phases. Using the system for around one hour was described as being fun and made users feel happy, in particular during E3; hence the system can have an entertaining value for the group of older users. Even during E1, this positive effect on the users’ emotions was explicit, although the demonstrated system showed severe technical issues. The effect did occur in all user groups.

Novelty effect

Some participants were positively excited after using the system because they found a SAR to be innovative and were astounded by its capabilities. Against our expectations, over the course of six iterations a decrement of this effect could not be measured; however we know from qualitative statements that primary and secondary users expected the PENJ to wear-off after some time, so maybe the methodology of repetitive interaction over the course of six weeks was not enough exposure to measure the downturn.

In contrast to literature, which describes the novelty effect mainly as a bias and threat to the external validity of evaluation results (see e.g. [Onwuegbuzie2007]), we consider this effect to be one of the most important impacts of applying SAR solutions and suggest considering its positive side. Depending on the variability of the SAR’s

implemented functionality, the novelty effect might wear-off after longer durations of usage; however until this time, this effect alone could motivate the target groups to use the system and let them profit from other impacts derived from its assistive functionality. When designing SAR solutions, we should therefore consider targeting functionalities that are already effective over the short term. The support of physical exercises (as shown in PT3, for example) might already generate a positive impact on the patient's health within the duration of the novelty effect. Obviously scenarios for long-term care are therefore less suitable for SAR solutions. One viable business model could be to lend SAR solutions for a limited duration to institutions providing therapy and care and hand the system over to a different institution as soon as the novelty effect wears off.

Initial social effects of SARs

Derived from the novelty effect, we found indications that SAR systems lead to an increase in the social participation of users as it gave our participants a novel topic to talk about with their relatives and friends. Because SAR systems are seen as novel and relevant in the future, several participants expected to gain attention from members of their social environment when using such a system and becoming able to demonstrate it possibly at their own homes. This idea is also backed up by [Deblieck2015] who speaks of rising social contacts and visitors after introducing their SAR solution to institutional care centres. We expect this to be an initial effect of the introduction of the system to the users and to wear off within weeks as soon as the novelty effect decreases.

Effects from assistive functionalities

A wide array of effects on the QoL can be logically argued by considering the assistive functionalities of the SAR. Here we assume that a SAR solution, which is accepted and used by stakeholders, generates the impacts that are known to arise from the respective assistive functionality. During our research activities, we could show that the system is able to motivate the conduction of physical exercises, to support health self-care, to provide additional safety and to entertain.

We know that exercising enhances mobility and the ability to autonomously conduct ADL [Harada95] and therefore increase the independence of patients leading to a higher QoL and a more active life. SARs can contribute to the therapy form of physical training as they are better able to motivate users to conduct the training than any other of the compared technical solutions and are able to demonstrate the exercises in a very

understandable way. Additionally, the tested PT3 solution was found to be feasible and working well despite usability issues that seemed to be solvable within further development. Considering our results, we can expect that the implementation of a SAR solution as a tool to support physical exercises has a positive impact on the quality and quantity of home-based training. Given the novelty effect and its impact on motivation to conduct the training, we expect that the positive effect of current solutions to be limited in time, to several weeks only.

Additionally, we expect an impact regarding the efficiency of care provided if the PT3 solution can be integrated into the daily work of institutional carers. An approach could be shown by successfully integrating the system into a gymnasium of a care institution. Furthermore, two carers were willing to integrate the system into their physical therapy with older users. Given these promising first steps, we expect that a temporal integration into institutional physiotherapeutic group training is possible, leading to a greater quality and efficiency of the care process as the tasks can be shared between the therapist and the SAR in such a way that the SAR could demonstrate the exercises and the therapist would then be free to move between participants to correct individual posture and exercise performance.

Assuming that the technical issue of mobility hindering the integration of a SAR into a user's home can be overcome in the future, we can also expect impacts from the other presented functionalities. By means of medical reminders and health measurements, the quality of self-care and awareness of one's own health could be improved.

The PT1 and PT2 systems provided safety to users by reminding and warning about critical conditions. In particular, COPD diseases are known to progressively worsen during exposure to low air quality. By implementing a respective warning, the worsening of this chronic disease and the corresponding limitations in personal autonomy can be slowed down, directly increasing the QoL.

Participants additionally expected an impact on their safety from using the SAR due to the mobility of the system as they found the robot could be able to search and find them after an injury such as a fall. However, this scenario was not implemented and hence remains a potential impact needing further investigation.

Potential adverse effects

Within all evaluation phases (E1-E3), we gathered user comments regarding the dependency of users upon a technical system and, as a side effect, upon technicians. If

users become dependent on the functionality of the SAR system, a malfunction of the system could have a negative impact on their health and thereby their QoL. See also section 7.4 regarding this ethical dilemma.

We can expect that users develop an emotional attachment to the robot – this was raised by participants within all evaluation phases and known from the literature. Within our studies, we could already find mild indications for emotional attachment from one particular user who often commented on the nice appearance of the SAR. Hence we expect that the emotional attachment to SAR systems, even if a rather strict personality was kept during the design-criteria development, could occur more often and to a higher intensity as compared with traditional technologies such as smartphones. As stated, this leads to ethical issues as SARs are not living creatures and not able to return these emotions.

During our studies of E1 and E2, we found indications that the system – despite its novelty and aforementioned related positive effects – could have a stigmatizing effect on users. One participant commented that others might think she would be insane to talk to a robot. Additionally, the assistive functionality of the SAR can be perceived as stigmatizing if the system primarily offers this functionality to older users or if the functionality helps with tabooed diseases such as, for example, the reminder functionality to compensate memory disorders. Whether or not an assistive solution is perceived as stigmatizing depends also on how it is marketed. If the SAR solution includes functionalities for entertainment and additional comfort, which are used primarily for marketing, we expect that the stigmatizing effect would be lower as when marketed as a device that increases safety and reduces memory disorders.

Some participants were concerned the SAR could lead to social isolation as carers could reduce the number of visits knowing the older user already is being taken care of by a robot. Our results show that this would be possible; in particular, PT3 could lead to less visits of the human physiotherapist if the user and/or therapist are convinced that a further interaction is not necessary because the robot is overtaking the task well enough. This would not be in the intention of the developed technology and we also consider this as misuse. In case such tendencies become apparent, one solution could be considering the development of solutions only as semi-autonomous devices that support the carer's or therapist's work and can be applied only in combination with the secondary user.

7.2 Summary of evaluation methods

This section is aligned with the initially proposed research questions, which were:

RQ_M1: Which current research methods can be used to assess a SAR?

Section 2.2 “User-centred evaluation methods” details the current state-of-the-art and hence answers this research question. As a summary, it can be said that the research methods differ based on the specific technological readiness of the prototypes, the test setting and the specific research aim. The methods used in current research have in common that they take a user-centred approach and facilitate the hands-on experience of users with technical prototypes to let them experience the interaction with a SAR.

As the research undertaken within this dissertation spans several years, the state-of-the-art of research methods regarding SAR technologies improved significantly over time. This section gives a brief lookback at the state of the methods available at the beginning of the dissertation and describes those that were used mainly during the course of research.

At the beginning of the research, it was unclear to what extent a new technology like a SAR could be integrated into the daily life of users, whether it would pose a safety risk and how it would perform from a technological viewpoint within realistic settings. Hence it was clear that a part of the evaluation had to target technical questions on whether the technology developed could serve as a proof of concept, whether it would be able to integrate such technology into real-user environments, what technical performance could be expected, and which technical issues exist for future research. The methods for technical performance analysis were derived from these research questions and consisted mainly of available methods for black-box testing of the developed functionalities. The usability of a SAR, which uses human-like interaction to support therapy and care tasks, was largely unknown. Here we have mainly adopted methods from traditional usability evaluations such as interviews, questionnaires and the thinking-aloud method, implemented them in an LL context and evaluated typical usability criteria such as learnability and effectiveness as proposed by Jacob Nielsen [Nielsen1994].

Due to many related research questions, we considered the acceptance of the solution to be the second major research domain. To what extent such solutions would be accepted and used by the specific target group of older users and their carers – despite a general

lack of technology acceptance and known scepticism towards robotics among the target groups – was our main interest. Literature such as [Bickmore2005] or [Breazeal2002] suggested that SARs could have beneficial effects on acceptance such as the ability to interact in a multi-modal manner and the ability to motivate due to the social presence, but it remained unclear to what extent these effects would be relevant. In order to investigate the identified research questions, in the beginning a large set of relevant factors was collected based on literature review and existing acceptance models such as the UTAUT and TAM models. In the following evaluation phases of E2 and E3, our acceptance model could be incorporated to a great extent into the Almere model and the Goodspeed questionnaire series, which had been well adopted by the research community at this time. For our evaluation framework, additional acceptance factors were incorporated in all evaluation phases regarding study-specific research questions such as the question on the motivational capabilities of the prototypes and emotional effects on users. Hence we could show that the methodology that we developed and used is flexible to the aims of the specific study and allows researchers to augment specific research questions.

The third main research domain identified was related to the impacts of the implementation of the technology. Questions related to impacts are obviously relevant because stakeholders need to have evidence of the solution's effectiveness in order to invest in the development of products or be willing to use them. However, measuring the impacts of SARs is specifically challenging because evidence of therapeutic effects typically requires the long-term application of the intervention due to human variability and the time it takes to generate measurable effects. Extensive evaluations such as randomized controlled trials are still only feasible regarding very specific technically limited settings and research questions due to technical constraints. Many authors take up this approach and try to investigate specific details, e.g. of factors influencing the interaction (see also section 2.1.1 for examples), but thereby are not able to provide a holistic top-down view, which was the aim of this dissertation. For this reason, instead of directly assessing the impacts, we tried to gather prospective impacts by mostly using qualitative methods such as interviews to gain an understanding of impacts that the user groups expect of the presented systems in the future. We are aware that this is a limitation of our methodology, as impacts that users think a technology might have in the future might greatly differ from impacts the technology really evokes in the field. Still, given the current state of technology, we think this is a helpful strategy to gather

information and support decisions on whether or not a solution should be developed further to a point where its technology readiness allows to conduct real-life long-term field-trials, which, as we and others [Bajones2019], [Broekens2009] stress, are necessary to obtain robust results on impacts.

RQ_M2: Which methods can be used and how can they be adapted to allow an evaluation of a SAR in settings as close to real life as possible?

We built our research on existing methods that were derived from usability engineering. These methods (questionnaires, interviews, video-analysis, thinking aloud, technical measurements, observations and group discussions) were originally developed for usability analysis in a laboratory or real-life context. Hence, they could already be used in an LL context but needed to be adapted to reflect the SAR-specific research questions and evaluation factors.

Within the undertaken research, we evaluated the possibility of assessing the SAR solutions within real-life contexts and found a way to do so by implementing E3 within a senior-citizen centre as a proof of concept. Due to the technical constraints of early prototypes and complex software components, we found that a LL setting was ideally suited to test all prototypes within E1 to E3. The LL, in our case a room at a senior-citizen centre that mimicked a living room, is a stable and secure place to conduct trials that allows the environmental conditions to be controlled and hence parts of the context of use. It thereby also gives the option of altering environmental conditions for testing purposes and allows for the testing of the feasibility of integration of the prototype in a realistic setting. The results on the integration into real-life settings are valuable as they can show open issues, but they do not serve as validation of the concept because additional real-world challenges are to be expected in the field (compare also [Bajones2019]).

The LL setting proved to be well suited to provide a consistent user experience and to test the functional performance of the system; however the results generated have to be viewed under the condition that the environmental context was controlled and should hence not be confused with the ecological validity of real-life trials.

The following methods were selected to fit the test setting of the LL. The methods are well known from usability research and applicable in either a laboratory setting or in the field. To use them within an LL, we found the following specifics to be relevant:

- > Questionnaires and semi-structured interviews

Semi-structured interviews and questionnaires, including open-ended questions, were conducted within the LL environment and directly after the demonstration of test scenarios. As these methods are not location specific, no adaptations had to be undertaken but the execution profited from the LL context due to the installed audio-recording equipment and because users could easily relate to objects in the room, such as the still present SAR and e.g. the locations it operated before the interview.

- > Video-analysis and observation

To allow video analysis of the SAR and the whole scene (including the user) to be used, we needed to develop a methodology that records not only the user and robot during their interaction but also the technical data gathered by the system and the output generated from the system. In particular, the reasoning of the localization and navigational algorithms of the system could not be provided in plain data but needed visual inspection on an augmented-reality display which was recorded together with the view of two opposing cameras and two microphones within one file for audio/video logging. We presented our solutions in chapter 3.8 (see Figure 14) and chapter 5.4 (see Figure 36). Because these solutions capture all possible data (audio, video from different perspectives and technical outputs of the system) within one video stream, we think that these were particular helpful approaches. The only drawback that we encountered during analysis was the limited resolution of the embedded camera-streams, which did not allow analysing the face of users, which could have been helpful to gain information on their emotional state. Another beneficial option could be to use semi-transparent mirrors to directly observe the scene. Our solution is limited in this aspect, but can be used in situations where semi-transparent mirrors cannot be implemented due to constraints of the room or building used. Other researchers often used a camera that was mounted in the robots head itself, creating a mobile solution without static cameras [Mucchiani2017], [Shiarlis2015]. In contrast, the presented solution can be used in static settings only, but provides stable images of high quality from different perspectives.

- > Thinking aloud

The thinking-aloud process [Kuusela2000] was well facilitated by the recording of audio and video data within the LL, as it proved to be difficult for the researchers to note the users' quotes alongside all the many other necessary tasks. This process is state of the

art, often applied in user studies [Weiss2009], including studies with SARs [Ramachandran2018] and has a sound theoretical basis [Charters2003]. As a critique, this method increases the cognitive load of participants and might become difficult if the task for the users is demanding [Jääskeläinen2010], [Charters2003] and provides only a simplified portrayal as that only reveals what becomes conscious to users [KaiYang2015]. Still we found this to be a helpful addition to our set of methods and valuable data source for triangulation.

> Technical measurements

Measurements for technical performance assessment were undertaken prior to the implementation of the SSUT method and during the SSUT in case of user-specific functionalities. For measurements that were conducted during the execution of user trials, we were restricted to measurements undertaken by the system itself or measurements that we could undertake based on the video recordings, such as time measurements. In particular, assessing the navigational performance during the trials was difficult because measurements could not be done physically in the room but only using the video recording of the augmented-reality display that showed the robot's location within the room.

> Observations

In addition to video recording and analysis, observations were undertaken by a researcher in most trials. Within the LL context, it was neither possible nor desired to observe users without their explicit knowledge because users were explicitly invited to the test and into the test site, which made it clear that an observation would take place. Additionally, research ethics prohibited the recording of participants without their knowledge. Consequently, users were fully aware of their role as test participants and an observational bias became likely during the trials of E2 as users were very tolerant of technical malfunctions because they understood their roles to be supporting the development of a system rather than evaluating it. Observations are state of the art within evaluations of SAR systems, compare also [Mucchiani2007], [Vroon2015], [Rherl2012], [Schröter2014], [Kosman2013], [Pérez2014].

> Group discussions

Discussions with groups of experts or primary users were conducted within the LL setting. This was helpful as the prototype could be shown and explored based on the same scenarios, as shown within the single-user trials, to give the participants a good

impression of the robot and its behaviour. Group discussions, or more specific focus groups [Morgan1997] are state of the art within SAR research, compare also [Cesta2012a], [UWE2013] or [Pigini2012].

RQ_M3: Which methods can be used and how can they be used to safely involve vulnerable older users (patients) and let them experience the interaction with a SAR?

Inviting vulnerable users into a remote test setting can be difficult as it requires the users' ability to travel to the test site as part of the inclusion criteria. Although we tested mostly with healthy older adults, we still had a (small) number of drop-outs due to the time and effort needed to take part in the trials. For these and ethical reasons, we did not involve COPD patients within the "Living Lab in Schwechat". Evaluations that stand aside of this dissertation but used the same methodology and involved COPD patients were conducted in Tel Aviv, Israel, as a proof of the applicability of the developed methodological framework with COPD patients. These evaluations were conducted at a care facility that also housed COPD patients and was hence able to care for them during user involvement.

Some questions initially posed within the questionnaires in E1 were not well-suited for the target group because they were found to be hard to understand. Additionally, the large number of questions resulted in a lengthy and tiring process for the participants. Part of the issue was that the questionnaires had to be read out-loud for some of the participants because of problems with eyesight. As we could narrow our focus from E1 to E3, the number of questions could be reduced significantly. Due to the specific needs of the target group, a researcher was present and able to help the users at any time during the trials.

The test-setting of the LL inside a senior-citizen centre was chosen in order to increase the ecological validity by giving the users the feeling of entering a typical assisted living setting that was intended for this technology.

The methodologies used throughout the course of this dissertation were used considering the involvement of vulnerable user groups and the final set, as used and described in E3 (see chapter 5.4 "Evaluation methodology"), was found to be well chosen.

RQ_M4: How can existing methods be synthesized together to form a reusable evaluation framework that facilitates a holistic evaluation?

A holistic evaluation approach was accomplished by combining qualitative and quantitative evaluation methods for the three research domains of performance, acceptance and impacts. Our evaluation framework specifies the evaluation domains, which are generic and from our point of view reusable within the field of AAL, evaluation factors, user groups and key user research methods, which are specific to the underlying aim and technology of the intended study. The evaluation framework is presented in its initial version in section 3.5 (see in particular Figure 8 for an overview), later revised in section 4.4 for E2 and revised again for E3 in section 5.2.

The presented evaluation framework was specifically developed for applications within the field of AAL, the specific robotic type of SARs, the target group of older people, their special needs and requirements and the impacts on care and care systems relevant to this group. Further we intend it to be used within real-life-like settings, but not purely laboratory trials or real-life field trials. It differentiates by the aforementioned specifics from the well-known USUS evaluation framework for HRI [Weiss2009], which provides a valuable guideline for HRI evaluations. In contrast to the presented framework, the USUS evaluation framework has a focus on HRI evaluation, addresses all target groups, considers laboratory and field trials and different kinds of robotics. Additionally the present framework includes a domain for the evaluation of performance, as this has proved to be the limiting factor for SAR development. It therefore has a stronger focus on aspects of technical performance, which are within the USUS framework not made explicit but included within the usability domain. At the same time both models share several characteristics. Both work with similar concepts of evaluation domains and include factors for usability, acceptance, user experience and impacts. Both recommend specific mostly qualitative key research methods including questionnaires, focus groups and interviews. These similarities back up our approach, which can be seen as an incremental step towards the evaluation of SARs in living lab settings for the purpose of assisting older user groups.

To supplement the evaluation framework, a method was developed and called “short-term scenario-based user trials” (SSUT). The SSUT is the key user research method of our evaluation framework and is described in detail in section 3.10.2. It provides a structured and detailed workflow for researchers to follow and gather qualitative and quantitative data using subjective and objective methods.

We found over the course of three evaluation phases that the SSUT method is flexible with the asked research questions, saves time and budget as methods can be cleverly combined and triangulated instead of simply conducted sequentially (e.g. questions in questionnaires or interviews can be asked only once but later be analysed from different viewpoints), and provides data that can be used from different perspectives. As an example, video analysis was used to support many research questions and analysed results were fused with results from thinking aloud, technical assessments and acceptance analysis. Hence by implementing the SSUT, several research aims can be covered within one user trial. It provides as much ecological validity as possible given the limitations of the applicability of early prototypes in real settings.

Short-term scenario-based user trials under controlled conditions are to our research not novel but the currently dominating methodology for evaluations in simulated real-life environments as many authors rely on it (e.g. [Kosman2013], [Lucia2013], [Ihsen2013], [Fischinger2014]), compare also the state of the art section 2.2.2.2. We see our contribution in the detailed description of its implementation and combination of research methods (see section 3.10.2) with the aim to lead towards a further standardization of approaches and to provide a reference for use within similar settings.

7.3 Limitations of the presented research

No undertaken research can be unbiased; within this dissertation, a number of biases could be identified and also partly compensated for.

User-selection bias. The targeted user groups were pre-defined within the respective undertaken studies. Gathering users from the defined user group randomly is prone to the problem that usually only users with a certain level of technological affinity agree to participate in a technology validation study. We hence tried to actively select critical users by tapping the knowledge of our LL partner (Seniorenzentrum Schwechat) who selected users based on our inclusion and exclusion criteria and the requirement to select a heterogeneous sample that also included people sceptic of new technologies. Additionally, we assessed the technology usage and acceptance of our test participants to gain an understanding of whether the individual results could be biased by the particular user's affinity to technologies. Nevertheless, we expect that this bias cannot be fully compensated for, as users also gained experience over the course of their participation and very critical users are nearly impossible to recruit. Here, we again compensated for the problem by triangulating the data gathered from users with those

gathered from formal carers who, due to their regular contact with users, have a better overview on the overall population and could hence state if they found that particular implemented functionalities or behaviours would suffer low acceptance from parts of the user group.

Due to the time and effort needed by users to take part within research trials, it is more likely for younger, fitter users to volunteer to participate. We again tried to compensate for this bias by going to the users in a care facility and actively selecting older participants and participants with disabilities to enhance the heterogeneity of our test groups; we think we achieved a good result considering the wide age span (typically between 65 and over 90 years of age, see also the individual descriptions of the test groups) and the large number of different age-dependent deficiencies among our test groups.

A researcher bias occurs when the researcher undertaking the study has “personal a priori assumptions” [Onweuegbuzie2003] which might be consciously or subconsciously transferred to the participants during the conduction of trials or applied during the analysis of results. This kind of bias is commonly present in qualitative research as here, typically, the researcher personally collects and interprets the data. Within qualitative research, this bias is considered as unavoidable [Strauss2014].

The evaluator (author) himself has a culturally rooted personal bias towards robots, the test participants and test scenarios. Test scenarios were co-developed by the author and given the technical background, a high affinity to technological solutions is present, making a bias towards a better acceptance and performance of SAR solutions plausible.

To diminish this bias, the following measures were undertaken:

1. Trials were not conducted by the author alone but always within a team of researchers to control for subconscious influence of the participants. In fact, during the trials researchers reminded each other on how to enhance the communication with the users to avoid influencing them.
2. The analysis of data was either undertaken by more than one researcher or the results were discussed with colleagues who were present during the data-gathering process.
3. Results were published and discussed with other researchers (see list of publications).

4. The author's personal bias is detailed to the reader in order to let her or him interpret the results' validity considering the author's personality.

The well-known observer effect might have influenced the presented results. According to Young et al., users might treat the robot in a more socially appropriate way when knowing they are being watched by the experimenters [Young2010]. We cannot estimate the strength of the influence of the observer effect in our results but tried to compensate for it by being aware of the effect and having it in mind, also during the interpretation of results. In addition, we tried to include an analysis of the users' behaviour and shown emotions during the trials in E2 that could also provide clues of the users' real opinion and acceptance of the system. However, this analysis proved to be limited due to the difficulty of correctly interpreting a user's emotions by observing a video.

The developed methodology is limited regarding the avoidance of the novelty effect, given that users can be in contact with the robot over repeated iterations but not feasibly over longer durations. Current research suggests that the usage and social acceptance of robots change over time within a timeframe of two or three months. This influence might have positive and negative factors, i.e. enhancing the acceptance and impacts because users have become accustomed or even attached to the new system and have learned how to use it, or users have neglected the new technology over a period of time [Young2010]. In that respect, the presented methodology presents a snapshot in time on the users' acceptance and potential impacts rather than a prediction of future long-term use.

As we implemented the user-centred design process (see chapter 3.1), we faced the same challenges as already presented in the state-of-the-art section regarding the time and efforts needed to conduct one full cycle of design, development and evaluation. Over the course of this dissertation and within six years of research, we only managed to develop three prototypes and implement the research methodology in three different evaluation phases. The main reason behind this is the funding scheme which depends on successfully winning research grants, but it also shows that the methodology of user trials itself is time, and thereby cost, intensive which limits its applicability.

For this reason, the developed holistic evaluation methodology needs further testing and its development is expected to carry on in future research. Here it is accepted that method development usually carries on over several years as it typically takes longer

than a PhD dissertation [Cairns2011]. Aside of this dissertation, parts of the developed evaluation methodology have already been successfully implemented into other research projects, most prominently into the European project “ReMIND”, where at the time of writing, it is being used and further developed.

We know from research that there is a significant gap between what users want and accept and what they are willing to pay for [WPU2013]. Given that we were testing early prototypes, we could not give a price for the solutions and hence, we could not evaluate the future use of the system which is a common flaw of used acceptance models that typically measure “intention” and not “behaviour” [Bagozzi2007].

Within all evaluation phases, the same robotic platform (“Nao”) was used. The chosen platform is the most commonly used robot in current research in Europe; however the results and methodology has to be validated in future research with other robotic prototypes to gain information on the transferability of results to other platforms. A first step towards this validation is currently being taken by the same author outside of the scope of this dissertation within the “ReMIND” European research project which facilitates a SAR on a wheeled platform.

It is hard to compare the results with others because different studies use different methodologies and differing robotic solutions, hence for comparisons, we could not define the source of the differences in the results. For that reason, we tried to build our evaluation model on accepted and well-used questionnaires such as the Almere model and Heerink’s Godspeed questionnaire, so that one day we have a large database of results that will allow meta-analysis. First attempts to do so have already been undertaken by Astrid Weiss and Christoph Bartneck who published a meta-analysis of results gathered by using the Godspeed Questionnaire [Weiss2015].

The sample sizes were initially planned to be 16 users in E1, 16 users in E2 and 12 in E3, but ultimate results only had eight users in E1, eight users in E2 and 11 users in E3. This seems to lead to an issue as eight participants are typically considered too few when taking into account the quantitative methods included in the mixed evaluation model. However, many results could be cross-validated between methods, across data qualities and also between the different evaluation phases, which often found similar results. We therefore think the gathered information is very valuable despite the low number of users.

The methodology presented builds on LL trials instead of real-life trials to compensate for absent technical functionalities and robustness. Since we could not compare the results with results from real-life trials, the influence of the restricted environment within the LL on the test results must remain unclear. However, from logical reasoning we know (compare also chapter 3.4) that the LL setup has a positive influence on acceptance measurements because the environmental conditions can be controlled, leading to a higher performance of the prototype; the setting is devoid of any distractions such as noise, pets or other people; and the setting is known outstandingly well to the developers beforehand which enables them to build customized solutions that would not be possible in real-users' homes.

Hence the intention of the LL approach was to create the illusion of a future system to gain information on what extent this future would be preferable to the user groups. This information is definitively needed as current research justifies high research costs by the potential advantages of future SAR technologies. We tried to be as realistic as possible when creating this illusion by estimating the future developmental progress. When simulating the speech recognition by a Wizard of Oz technique, we assumed that one day in the (near) future this technology would be capable enough to robustly recognize short sentences even over the distance of 2-3 meters (at the time of writing, latest advancements in this particular field have reached that point already). When we controlled the light settings in the LL, we assumed that either the technology for localization, navigation and user recognition becomes so robust that such control would not be necessary anymore, or that the lights in users' homes can be controlled in a similar way. We further assumed that the floor plans and furniture can be controlled in a way that lets the robot safely navigate through the premises, which excludes the presence of thick carpets and objects out of range of the used SAR's sensors, or that a future robot will not show the same limitations. Last but not least, we assumed that in future scenarios, either no other people or pets are around or that we will be able to develop functionalities for SARs that enable multi-user support.

By making these assumptions, we limited the ecological validity of the results and acknowledged that we cannot give evidence on the applicability of the shown SAR technology for the integration into real-life settings, but could give many valuable insights into what we could expect in case science and technology solves the mentioned restrictions and that we provide many detailed points that should be considered in future developments.

7.4 Summary of ethical, social and legal aspects of assistive robotics

The study of ethical and legal aspects was not an explicit aim of this dissertation, but these aspects were covered implicitly as they influence the acceptance of solutions and were hence often mentioned in user comments during the trials.

Regarding ethical aspects, we have to differentiate between:

a) Research ethics

Research ethics standards, such as the implementation of an “informed consent” procedure, defining exit strategies and insurance of personal safety of users, are available and well distributed at the Technische Universität Wien where this dissertation was written. However, during the time of our research, the institution did not host an ethical committee nor did the city of Vienna’s ethical commission consider itself competent enough to decide on user studies with assistive technologies. Typically, in larger studies of assistive robotics, an ethical advisor or even an advisory board takes part that can guide the ethically sound research design and react in case they experience or suspect ethical violations. An ethical advisory board was established within the “KSERA” project and gave valuable recommendations on the user involvement within the trials undertaken in E1 and E2. Because the research design and user-involvement procedures used in E3 were very similar to the ones used in the earlier user interventions, no additional ethical consultation was undertaken. In all the studies mentioned here, a number of precautions were undertaken to adhere to high ethical standards, as described in the section “Principles of SAR evaluation within a LL environment”.

b) Ethical implications of the usage of SAR robots in future care scenarios

Animacy, anthropomorphism and social attachment

The participants in the conducted studies could well discriminate between a human being and the used robot and were well aware that the presented system is a machine. Nonetheless, many tended to treat the SAR system like a living creature, a pet or even a human. An ethically relevant situation would arise in cases when the user is no longer able to discriminate well between the robot and a living creature, possibly considering the SAR a true friend or companion [Coeckelbergh2012]. This might be the case in particular for users with dementia. Within the studies presented here, participants were concerned that users might perceive the robot as a friend and conversational partner

and also expressed their concern because they found the idea of a human having a social conversation just for the sake of this communication scary. We think the reason for this concern lies in the common understanding of honesty and propose that within our society, a system that shows emotions despite obviously being not able to feel them is seen as dishonest, thereby negatively influencing acceptance.

This finding indicates that one of the initial goals – to build an artificial companion for older users – was clearly declined by test participants. Future developments of SAR solutions should hence be very careful regarding the usage of HRI to stimulate the feeling of companionship.

Because the social influence is one of the key capabilities of SAR robots, this restriction limits the use of SARs in general. If a SAR should not use its social presence and multi-modal HRI channels to convince a user to perform otherwise unwanted tasks, then SARs would lose one of their main unique selling points.

A solution to this ethical dilemma could be to design the SAR's look and behaviour to resemble not a friendly companion but a rather strict tool. However, there seems to be a trade-off between the acceptance of the solution (here studies suggest implementing a friendly extrovert solution) and the ethical implications of interacting with such a companion and possibly confusing it with a real person or even creating a stronger bond with it.

As we saw this dilemma already during evaluation of the second prototype (E2), we sought a solution by designing the third prototype to act in an animated and vivid way but to show a strict and functionality-oriented personality. The evaluation results suggest high acceptance ratings, clearly higher than previous solutions on most tested factors. Although there likely are several factors that caused the increase in acceptance, there were no indications that the acceptance suffered because of this design choice. Therefore, we suggest future developments to aim for robot personalities that are rather strict and functionality oriented instead of a friendly, extrovert, lovely one, to minimize the aforementioned ethical issues. This finding is also backed up by [Riek2014] who recommends carefully considering the human tendency to attach to robotic systems even during the design. Oliver Bendel even goes a step further and recommends informing the user about the fact that she/he is talking to a machine and not a person in a repeating manner [Bendel2016].

In general, this suggests that the ethical values and ideals of engineers are relevant as they have to take care that non-ethical systems are not developed. Therefore, ethical guidance for the development of socially assistive robotics seems to be needed. The UK EPSRC presented a first draft of such a guidance document in the form of “five principles of robotics” [EPSRC2010]. Additionally, the European Commission published ethics guidelines for a trustworthy AI which are also applicable for SAR systems and will support future developments [EC2019]. Within this dissertation, we provided additional heuristics within the described design principles that also include ethical points within their reasoning.

Dependence upon the solution

Some participants asked what they could do if, in a situation in which they had already become dependent on the system, the robot suddenly malfunctioned and could therefore no longer support them. A system or device that is able to assist users during their daily living also creates a dependency as soon as the user has integrated the device into her or his daily routine because it requires a lot of effort to change the routine back to the original one, if that is even possible, due to a deterioration of the user’s condition. A current well-known example are sat-navs that guide their users very well on the road but on the other hand, reduce the user’s ability to navigate on their own as the ability to remember routes can decline when not trained regularly. For that practical reason, deployed systems need to be stable and reliable and their permanent functionality needs to be guaranteed and supported in the same way as with other systems we depend upon, e.g. with fast breakdown services such as for cars on motorways, heating and cooling systems at home, toilet repairs, etc.

Physical harm from the solution

Mobile autonomous robots such as the used SAR systems introduce a number of physical dangers into the homes of users. These issues have to be discussed during the development of prototypes and eliminated as far as possible by the design of the solutions.

Users could trip over the (moving) robot and fall or hurt themselves. This is especially relevant considering the target group might have sensory limitations such as reduced eyesight, restricted mobility and cognitive capabilities.

Users could get hurt by the robot’s movements, despite the fact that the used robot is very small, because moving joints could pinch them.

The robot could incidentally create a fire and smoke danger if moving towards a hot surface such as a heating radiator, fireplace or stove. This issue is not hypothetical after a media report about a robotic vacuum cleaner that drove over a kitchen hotplate and set on fire [Mirror2013].

Privacy issues

As the main ability of SARs is to gather data and infer on it, SAR systems do typically gather large amounts of different data that are not only personal but can also be health related. To aggravate the issues, the systems are able to follow the user and it is not obvious to the users when exactly the system records data. This can lead to unintentional or intentional misuse of data. In one reported case, information was given by a participant to the robot, obviously not knowing that care staff was taking part at the conversation using the robot's microphones [Payr2015]. Experimenters in E1 reported another related issue where they found that there was a risk that the small robot could film under the skirt of a sitting woman.

Replacing caregivers

One idea behind the investigated solutions was to support the secondary users and enhance their QoL. Care providers in particular showed a fear that a robot could one day replace them in their job. Given the currently limited technical possibilities to support the care of users and the fact that there is a constant demand for carers [Fuchs2013], this fear seems to be rather unjustified at the time of writing. On the opposite side, because of time constraints, personal care time and therefore the possibilities of social support from human caregivers are scarce [Wohlmannstetter2016]. If a SAR robot could take over the time-consuming tasks of carers, there could be a chance to enhance the quality of personal care if carers get the possibility to spend more time with their patients. Nevertheless, this ethical topic needs to be kept in mind when developing future solutions that might, due to advances in AI and robotics, be actually capable of replacing the number of care tasks.

Cultural differences in behaviour

Because human-human interaction differs depending on culture, HRI also has to be aware of different cultures and how to interact appropriately accordingly. An autonomous detection of the culture of a human currently seems unfeasible, but future systems could either ask or be otherwise presented with the knowledge about the cultural background of the human counterpart.

c) Autonomous ethical decision making

If a SAR has the functionality to protect the user in case of emergencies, such as prototype 1 and 2, they need an ethical decision-making system to decide on how to react in an ethically correct fashion. Should the robot convince the user to take medication / perform exercises / perform physical measurements? What should the robot do in case the user refuses to comply? Allowing the user to skip a medication or measurement could cause harm. But insisting to comply would impinge the user's autonomy. Should it try to convince the user by threatening to call a doctor or a responsible relative? Should it actually call a doctor or relative? Or should it postpone the action for a later second reminder? For such ethically relevant questions, a system needs to be implemented that behaves in the best interest of the user. One way to gain information about the best interest of the user would be to involve stakeholders and decide based on the majority of views. Similarly, Anderson and Anderson describe the principles of an ethical decision-making system for the particular scenario of an AAL robot [Anderson2015]. At the time of writing, the Massachusetts Institute of Technology (MIT) uses the same technique to feed a system for autonomous ethical decision making that in the future should support autonomous cars when deciding on optimal driving pathways [Awad2018].

d) Positive influences of SARs regarding ethical challenges

On the positive side, within the presented studies it was also noted that robotic systems might be able to alleviate existing ethical problems.

Some users prefer a robot for toileting support over a human because they said needing assistance for that task made them feel humiliated . Also, Oliver Bendel highlighted the fact that a robotic solution decreases the dependency on human caregivers which enhances the QoL in cases where the user would not want to ask a caregiver e.g. because he or she needs support very often or the task doesn't seem important enough to ask a human [Bandel2016].

The current situation of residential care in Austria is full of ethical issues that could be partly alleviated by implementing SAR solutions. Most of today's care is undertaken by informal carers, who are mostly old themselves and suffer from the impact on their time and independence. Caring for a close friend or relative can create strong psychological and emotional strains and may lead to social isolation. Caring in general is often physically challenging, costly and stressful [Payr2015].

Neither is the current situation in formal home care less ethically challenging. Due to the high need for care personnel, in 2006 the Austrian Government passed a bill to legalize the employment of foreign care workers. Before that time, they opted to illegally hire care workers for 24-hour care to a point where around 50% of total care workers (almost all of whom were 24-hour carers) were illegally employed as legal formal care personnel is unaffordable for many Austrian families [DA2007]. These workers typically commute from eastern countries such as Hungary, Slovakia and Romania. Ethical issues arise out of the facts that the workers have to leave their own families for parts of the month to commute to Austria and are dependent upon their family hosts during their time of stay as they often barely understand the language, social and legal system. Their qualifications to conduct care are diverse and difficult to compare to Austrian standards and hence the quality of care provided is mostly unclear.

SARs could alleviate the ethical issues of formal and informal carers by facilitating parts of their work. Ideally tasks that are tedious, time-consuming or physically challenging should be taken over to enhance the work situation of carers.

As a conclusion, SARs introduce new ethical issues that we need to be aware of and try to tackle, and they also could ease existing issues. Although we cannot generally compare the ethical pros and contras because they differ in nature and their strengths depend on the users, their life situation and context of use, it seems plausible – also from an ethical viewpoint – that it could be beneficial in particular cases to introduce a SAR system.

8 Conclusions and proposed future steps

This dissertation aimed at investigating the potentials of socially assistive robotics for older users. A set of evaluation methodologies was developed and composed into a holistic evaluation framework. This framework was then used on the developed prototypes to gain the results.

Two central research aims were used to drive the research:

- a) As a main aim, we wanted to gain insights on the extent to which SAR solutions are a valid approach to support the care of older users, whether they gain acceptance among user groups and if they have beneficial effects.
- b) As a secondary aim, we wanted to understand which research methods can be used to evaluate SAR systems and how can they be implemented considering the specificities of the used systems and the targeted user groups.

Several findings could be achieved that contribute to fulfilling the research goals. The detailed contributions were presented in chapter 1.4 and comprise: i) insights on the performance, acceptance and prospective impacts of using SARs to support care, ii) findings on how to enhance the methodology to evaluate SAR systems, iii) a set of design principles to support the design of SAR systems as well as a method to support the design of HRI flows and iv) insights on ethical, social and legal aspects of the implementation of SARs.

Additionally, within this dissertation, we could give a proof of concept of SARs in real-life settings and show that SAR prototypes for older people are applicable in a constrained LL setting and also within a specific institutional real-life setting. Despite initial plans, we were unable to implement any of the prototypes in a residential setting for reasons of safety and performance, but we consider this a possibility for future R&D if the application scenario is chosen wisely and avoids technically fragile functionalities.

Ad i) insights on the performance, acceptance and prospective impacts of using SARs to support care

Regarding the performance of tested systems, a number of technological flaws and challenges could be identified, such as the need for robust navigation and localization techniques and better perception abilities and algorithms for autonomous decision-making. For future research, we can recommend excluding technically challenging

functionality from the design of prototypes in order to achieve a solution that can safely be tested with user groups.

By listening to our user groups, we also learned about many challenges regarding the interactive functionalities of SARs, such as the limited capability of users to remember key phrases to trigger commands and the necessity of the usage of different in- and output modalities to include users with varying types and degrees of disabilities. Further work is needed in particular on the perceptual abilities of SARs to enhance input channels such as the recognition of gestures, speech and emotions in order to allow for an enjoyable and accepted HRI.

We found as positive acceptance factors that the tested systems were perceived as friendly, happy, enjoyable and easy to use. The systems were also perceived as sociable, which in turn evoked the fear to develop a kind of friendship that was seen as ethically challenging. In particular, showing emotions and suggesting feelings by the robot was seen critically both by users and experts for ethical reasons.

Because the acceptance and perception of the robot varied between individual users, we expect that technology affinity is a strong acceptance factor that currently limits the target group to those with a certain level of technology affinity and experience. We can recommend that future researchers define in particular the target group of older people in greater detail, considering also the level of technical affinity.

Within the undertaken research, we could confirm the proposed motivational influence and entertaining effect of the robot, in particular for the physical training scenario in which parts of the user groups even rated the robot as being more motivational than a human trainer. It can be argued that this motivational effect can be used to generate other positive effects such as an increase of physical fitness if users are motivated to train regularly.

On the counter side, we found that users expect this motivational effect to wear-off during long-term usage. This so called novelty effect can be seen as an issue as positive effects are not pertaining over time but it can also be dealt with, given that products could be designed to be lent for periods of time.

Within this dissertation, only specific user needs were targetable by technology, but in at least one scenario it could be shown that SARs can provide support in a unique way that was found by users to be superior to comparable solutions. We can recommend that future developments focus on a small set of well-defined user needs and realize a single

functionality rather than building a system capable of supporting a large set of needs and concentrate the resources on functionalities that are unique to SARs.

Within this dissertation, prospective impacts of the proposed solutions could be gathered. To measure impacts of the technology test designs such as clinical trials and randomized controlled trials would be an ambitious goal for future research. To realize this next step, the author is currently working on the European research project “ReMIND” (2018-2021), which targets a randomized controlled multi-centre clinical trial with a robotic solution.

Ad ii) findings on how to enhance the methodology to evaluate SAR systems

Research methods to evaluate SAR systems were reviewed and selected within this dissertation and experiences gained from their implementation were used to develop and iteratively enhance an evaluation framework, which was presented. We could show how traditional methods from usability engineering and qualitative and quantitative research can be adapted to become valuable tools for evaluation of SARs with vulnerable user groups in real-life or close to real-life settings. The limitations of this approach were discussed and have also to be considered in future research. This framework can be used by future researchers as was already done by the authors own follow-up research done alongside this dissertation.

Ad iii) a set of design principles to support the design of SAR systems as well as a method to support the design of HRI flows

Out of the results of the undertaken studies, general lessons learned could be drawn and be used to form a set of eight design principles. These principles can be understood as SAR-specific heuristics and are provided to future researchers and developers as a starting point when generating new designs or when redesigning SAR solutions.

As an additional tool, a graphical design method called interaction flows is provided as a flexible and quick to implement tool to support the development and description of use cases and scenarios that involve multi-modal interaction capabilities of SAR systems.

Ad iv) insights on ethical, social and legal aspects of the implementation of SARs

We can confirm there are effects from the ethical implications of SARs that developers of SAR systems should be aware of and which should be treated carefully. Users might obtain a social attachment to SARs, which, according to our reasoning, developers should try to avoid. We should not try to enhance acceptance by including emotions and

the demonstration of feelings. During future research, we should remember where the whole idea started from: to build a tool that serves us well, not a friend.

Future developments should also consider that users can become dependent upon the technological solution leading to the necessity of quick support in case of technical malfunctions. Further, solutions can be physically harming, even if they seem to be small and harmless. Privacy issues need to be considered as well as differing user needs due to cultural differences. Systems for autonomous ethical decision-making will be needed for future solutions. Guidelines for ethically sound functionalities of SARs are currently being developed and should be considered at early stages of R&D.

We also found positive effects of the implementation of robotics, as the current situation without robotic solutions faces ethical challenges that could be alleviated, e.g. by facilitating the work of care persons. Of course, ethical pros and cons cannot outweigh each other but have to be considered individually.

References

- [AALAZ] Austria, A. (2013). Zweck und Ziele von AAL Austria. Retrieved November 11, 2019, from <http://www.aal.at/ueber-aal/>
- [AlShamayleh2018] Al-Shamayleh, A. S., Ahmad, R., Abushariah, M. A., Alam, K. A., & Jomhari, N. (2018). A systematic literature review on vision based gesture recognition techniques. *Multimedia Tools and Applications*, 77(21), 28121-28184. doi: 10.1007/s11042-018-5971-z
- [Aldebaran2014] Aldebaran Robotics. (2014). Unveiling of NAO Evolution: A Stronger Robot and a More Comprehensive Operating System. Retrieved February 9, 2020, from <https://www.prnewswire.com/news-releases/unveiling-of-nao-evolution-a-stronger-robot-and-a-more-comprehensive-operating-system-263934131.html>
- [Amirabdollahian2013] Amirabdollahian, F., Akker, R. O. D., Bedaf, S., Bormann, R., Draper, H., Evers, V., ... Dautenhahn, K. (2013). Accompany: Acceptable robotiCs COMPanions for AgeiNG Years — Multidimensional aspects of human-system interactions. *6th International Conference on Human System Interactions (HSI)*. doi: 10.1109/hsi.2013.6577882
- [Arras2005] Arras, K. O., & Cerqui, D. (2005). Do we want to share our lives and bodies with robots? A 2000-people survey. *Autonomous Systems Lab (ASL), Swiss Federal Institute of Technology Lausanne (EPFL), Tech. Rep*, 1–605.
- [Awad2018] Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., ... & Rahwan, I. (2018). The moral machine experiment. *Nature*, 563(7729), 59-64.
- [Bagozzi1992] Bagozzi, R. P., Davis, F. D., & Warshaw, P. R. (1992). Development and test of a theory of technological learning and usage. *Human relations*, 45(7), 659-686.
- [Bajones2019] Bajones, M., Fischinger, D., Weiss, A., Puente, P. D. L., Wolf, D., Vincze, M., ... & Qammaz, A. (2019). Results of Field Trials with a Mobile Service Robot for Older Adults in 16 Private Households. *ACM Transactions on Human-Robot Interaction (THRI)*, 9(2), 1-27.
- [Bartneck2008] Bartneck, C., Kulić, D., Croft, E., & Zoghbi, S. (2008). Measurement Instruments for the Anthropomorphism, Animacy, Likeability, Perceived Intelligence, and Perceived Safety of Robots. *International Journal of Social Robotics*, 1(1), 71–81. doi: 10.1007/s12369-008-0001-3
- [Baxter2015] Baxter, K., Courage, C., & Caine, K. (2015). *Understanding your users: a practical guide to user research methods*. Amsterdam: Morgan Kaufmann.
- [Bedaf2017] Bedaf, S., Marti, P., Amirabdollahian, F., & de Witte, L. (2018). A multi-perspective evaluation of a service robot for seniors: the voice of different stakeholders. *Disability and Rehabilitation: Assistive Technology*, 13(6), 592-599.

- [Bemelmans2012] Bemelmans, R., Gelderblom, G. J., Jonker, P., & Witte, L. D. (2012). Socially Assistive Robots in Elderly Care: A Systematic Review into Effects and Effectiveness. *Journal of the American Medical Directors Association*, 13(2). doi: 10.1016/j.jamda.2010.10.002
- [Bendel2016] Bendel, O. (2016, May). "Die Maschine in der Moral" Presentation at the *Open session of the Austrian bioethics commission*. Vienna.
- [Beuscher2017] Beuscher, L. M., Fan, J., Sarkar, N., Dietrich, M. S., Newhouse, P. A., Miller, K. F., & Mion, L. C. (2017). Socially assistive robots: measuring older adults' perceptions. *Journal of gerontological nursing*, 43(12), 35-43.
- [Bickmore2005] Bickmore, T. W., & Picard, R. W. (2005). Establishing and maintaining long-term human-computer relationships. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 12(2), 293-327.
- [BMASGK2018] Bundesministerium für Arbeit, Soziales, Gesundheit und Konsumentenschutz. (2018). *Österreichischer Pflegevorsorgebericht*, Austria.
- [Bødker2015] Bødker, S. (2015). Third-wave HCI, 10 years later – participation and sharing. *Interactions*, 22(5), 24–31. doi: 10.1145/2804405
- [Braun2006] Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2), 77–101. doi: 10.1191/1478088706qp063oa
- [Breazeal2002] Breazeal, C., & Scassellati, B. (2002). Robots that imitate humans. *Trends in Cognitive Sciences*, 6(11), 481–487. doi: 10.1016/s1364-6613(02)02016-8
- [Broadbent2009] Broadbent, E., Stafford, R., & Macdonald, B. (2009). Acceptance of Healthcare Robots for the Older Population: Review and Future Directions. *International Journal of Social Robotics*, 1(4), 319–330. doi: 10.1007/s12369-009-0030-6
- [Broekens2009] Broekens, J., Heerink, M., & Rosendal, H. (2009). Assistive social robots in elderly care: a review. *Gerontechnology*, 8(2). doi: 10.4017/gt.2009.08.02.002.00
- [Brooks2006] Brooks, A. G., & Breazeal, C. (2006, March). Working with robots and objects: Revisiting deictic reference for achieving spatial common ground. In *Proceedings of the 1st ACM SIGCHI/SIGART conference on Human-robot interaction* (pp. 297-304).
- [Cesta2012] Cesta, A., Cortellessa, G., Orlandini, A., & Tiberio, L. (2012). Into the wild: Pushing a telepresence robot outside the lab. In *Proceedings of SRT 2012—Workshop on Social Robotic Telepresence* (pp. 7-14).
- [Cesta2012a] Cesta, A., Cortellessa, G., Orlandini, A., & Tiberio, L. (2012, February). Evaluating telepresence robots in the field. In *International Conference on Agents and Artificial Intelligence* (pp. 433-448). Springer, Berlin, Heidelberg.

- [Charters2003] Charters, E. (2003). The use of think-aloud methods in qualitative research an introduction to think-aloud methods. *Brock Education: A Journal of Educational Research and Practice*, 12(2).
- [Chidambaram2012] Chidambaram, V., Chiang, Y. H., & Mutlu, B. (2012, March). Designing persuasive robots: how robots might persuade people using vocal and nonverbal cues. In *Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction* (pp. 293-300).
- [Collins-Cope1999] Collins-Cope, M. (1999). RSI-A Structured Approach to Use Cases and HCI Design. *Personal Communication, Ratio Group Ltd.*
- [Colombo2000] Colombo, G., Joerg, M., Schreier, R., & Dietz, V. (2000). Treadmill training of paraplegic patients using a robotic orthosis. *Journal of rehabilitation research and development*, 37(6), 693-700.
- [Coradeschi2013] Coradeschi, S., Cesta, A., Cortellessa, G., Coraci, L., Gonzalez, J., Karlsson, L., ... & Pecora, F. (2013, June). Giraffplus: Combining social interaction and long term monitoring for promoting independent living. In *2013 6th international conference on Human System Interactions (HSI)* (pp. 578-585). IEEE.
- [Courtney2009] Courtney, M., Edwards, H., Chang, A., Parker, A., Finlayson, K., & Hamilton, K. (2009). Fewer emergency readmissions and better quality of life for older adults at risk of hospital readmission: A randomized controlled trial to determine the effectiveness of a 24-week exercise and telephone follow-up program. *Journal of the American Geriatrics Society*, 57(3), 395-402.
- [Cuijpers2015] Cuijpers, R. H., & Knops, M. A. (2015, October). Motions of robots matter! the social effects of idle and meaningful motions. In *International Conference on Social Robotics* (pp. 174-183). Springer, Cham.
- [Davis1989] Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS quarterly*, 319-340.
- [Dautenhahn2007a] Dautenhahn, K. (2007). Socially intelligent robots: dimensions of human-robot interaction. *Philosophical transactions of the royal society B: Biological sciences*, 362(1480), 679-704.
- [DeGraf2016] de Graaf, M. M., Allouch, S. B., & van Dijk, J. A. (2016, March). Long-term acceptance of social robots in domestic environments: insights from a user's perspective. In *2016 AAAI Spring Symposium Series*.
- [Deblieck2015] Deblieck, T., CEO of QBMT (developer and distributor of ZORA), personal interview on 09.06.2015.

- [Decker2008] Decker, M. (2008). Caregiving robots and ethical reflection: the perspective of interdisciplinary technology assessment. *AI & society*, 22(3), 315-330.
- [Duffy2003] Duffy, B. R. (2003). Anthropomorphism and the social robot. *Robotics and Autonomous Systems*, 42(3-4), 177-190. [https://doi.org/10.1016/S0921-8890\(02\)00374-3](https://doi.org/10.1016/S0921-8890(02)00374-3)
- [EC2019] High-level expert group on artificial intelligence. (2019). *Comments on the "Ethics Guidelines for Trustworthy AI"*. Brussels.
- [ENOLL2016] ENOLL. (2016). www.openlivinglabs.eu. Retrieved April 25, 2020 from <http://www.openlivinglabs.eu/aboutus>
- [Eriksson2005] Eriksson, M., Niitamo, V. P., & Kulkki, S. (2005). State-of-the-art in utilizing Living Labs approach to user-centric ICT innovation-a European approach. *Lulea: Center for Distance-spanning Technology. Lulea University of Technology Sweden: Lulea*. Retrieved from: http://www.vinnova.se/upload/dokument/Verksamhet/TITA/Stateofheart_LivingLabs_Eriksson2005.pdf
- [euRobotics2016] euRobotics. (2016). Robotics 2020 Multi-Annual Roadmap. February 9,2020 from https://www.eu-robotics.net/cms/upload/topic_groups/H2020_Robotics_Multi-Annual_Roadmap_ICT-2017B.pdf
- [Eurostat2017] Eurostat. (2017). Statistiken zu Eheschließungen und Scheidungen. Retrieved February 9, 2020 from https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Archive:Marriage_and_divorce_statistics/de
- [Eurostat2019] Eurostat. (2019). Pressemitteilung: Über 5 Millionen Geburten im Jahr 2017. Retrieved February 9, 2020 from <https://www.presseportal.de/pm/121298/4215685>
- [EvAALuation2017] J. Himmelsbach, et al. (2017). Indikatorenhandbuch für die Messung von Wirkungen und Effizienzsteigerungen (draft). Retrieved February 9, 2020 from https://www.ffg.at/sites/default/files/allgemeine_downloads/thematische%20programme/Energie/EvAALuation2_D4.2_Handbuch_final.pdf
- [Fasola2011] Fasola, J., & Mataric, M. (2011). Comparing physical and virtual embodiment in a socially assistive robot exercise coach for the elderly. *Center for Robotics and Embedded Systems, Los Angeles, CA*.
- [Fasola2012] Fasola, J., & Mataric, M. J. (2012). Using socially assistive human-robot interaction to motivate physical exercise for older adults. *Proceedings of the IEEE*, 100(8), 2512-2526.
- [Feil-Seifer2005] Feil-Seifer, D., & Mataric, M. J. (2005, June). Defining socially assistive robotics. In *9th International Conference on Rehabilitation Robotics, 2005. ICORR 2005*. (pp. 465-468). IEEE.

- [Feil-Seifer2007] Feil-Seifer, D., Skinner, K., & Matarić, M. J. (2007). Benchmarks for evaluating socially assistive robotics. *Interaction Studies*, 8(3), 423-439.
- [FER2014] 5elementsrobotics. (2014). Too many things to carry? Retrieved April 25, 2020, from <http://5elementsrobotics.com/>
- [Fernaesus2010] Fernaeus, Y., Håkansson, M., Jacobsson, M., & Ljungblad, S. (2010, June). How do you play with a robotic toy animal? A long-term study of Pleo. In *Proceedings of the 9th international Conference on interaction Design and Children* (pp. 39-48).
- [Ferri2011] Ferri, G., Manzi, A., Salvini, P., Mazzolai, B., Laschi, C., & Dario, P. (2011, May). DustCart, an autonomous robot for door-to-door garbage collection: From DustBot project to the experimentation in the small town of Peccioli. In *2011 IEEE International Conference on Robotics and Automation* (pp. 655-660). IEEE.
- [Fischinger2014] Fischinger, D., Einramhof, P., Papoutsakis, K., Wohlkinger, W., Mayer, P., Panek, P., ... & Vincze, M. (2016). Hobbit, a care robot supporting independent living at home: First prototype and lessons learned. *Robotics and Autonomous Systems*, 75, 60-78.
- [Forlizzi2007] Forlizzi, J. (2007, March). How robotic products become social products: an ethnographic study of cleaning in the home. In *2007 2nd ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (pp. 129-136). IEEE.
- [Foster2012] Foster, M. E., Gaschler, A., Giuliani, M., Isard, A., Pateraki, M., & Petrick, R. P. A. (2012). Two People Walk Into a Bar : Dynamic Multi-Party Social Interaction with a Robot Agent. *Icmi, (Icmi)*, 14–21. <https://doi.org/10.1145/2388676.2388680>
- [Frennert2013] Frennert, S., Efring, H., & Östlund, B. (2013, October). What older people expect of robots: A mixed methods approach. In *International conference on social robotics* (pp. 19-29). Springer, Cham.
- [Frokjar2000] Frøkjær, E., Hertzum, M., & Hornbæk, K. (2000, April). Measuring usability: are effectiveness, efficiency, and satisfaction really correlated?. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems* (pp. 345-352).
- [Fuchs2013] Fuchs, M., & Weyh, A. (2013). Impact of demographic change on the employment for carers in central Germany. An analysis of Saxony, Saxony-Anhalt, and Thuringia. *Bundesgesundheitsblatt, Gesundheitsforschung, Gesundheitsschutz*, 56(8), 1048-1055.
- [Ganster2010] Ganster, T., Eimler, S. C., Von Der Pütten, A. M., Hoffmann, L., Krämer, N. C., & von der Pütten, A. (2010). *Methodological considerations for long-term experience with robots and agents*.

- [Goetz2002] Goetz, J., & Kiesler, S. (2002). Cooperation with a robotic assistant. *CHI '02 Extended Abstracts on Human Factors in Computer Systems - CHI '02*, (November), 578. <https://doi.org/10.1145/506486.506492>
- [Goudey2016] Goudey, A., & Bonnin, G. (2016). Must smart objects look human? Study of the impact of anthropomorphism on the acceptance of companion robots. *Recherche et Applications en Marketing (English Edition)*, 31(2), 2-20.
- [Green2000] Green, A., Hüttenrauch, H., Norman, M., Oestreicher, L., & Eklundh, K. S. (2000). User centered design for intelligent service robots. *Robot and Human Communication - Proceedings of the IEEE International Workshop*.
- [Green2004] Green, A., Huttenrauch, H., & Eklundh, K. S. (2004, September). Applying the Wizard-of-Oz framework to cooperative service discovery and configuration. In *RO-MAN 2004. 13th IEEE International Workshop on Robot and Human Interactive Communication (IEEE Catalog No. 04TH8759)* (pp. 575-580). IEEE.
- [Green2006] Green, B. N., Johnson, C. D., & Adams, A. (2006). Writing narrative literature reviews for peer-reviewed journals: secrets of the trade. *Journal of chiropractic medicine*, 5(3), 101-117.
- [Gunawardena1997] Gunawardena, C. N., & Zittle, F. J. (1997). Social presence as a predictor of satisfaction within a computer-mediated conferencing environment. *American journal of distance education*, 11(3), 8-26.
- [Guralnik1994] Guralnik, J. M., Simonsick, E. M., Ferrucci, L., Glynn, R. J., Berkman, L. F., Blazer, D. G., ... & Wallace, R. B. (1994). A short physical performance battery assessing lower extremity function: association with self-reported disability and prediction of mortality and nursing home admission. *Journal of gerontology*, 49(2), M85-M94.
- [Guralnik1995] Guralnik, J. M., Ferrucci, L., Simonsick, E. M., Salive, M. E., & Wallace, R. B. (1995). Lower-extremity function in persons over the age of 70 years as a predictor of subsequent disability. *New England Journal of Medicine*, 332(9), 556-562.
- [Gustafsson2015] Gustafsson, C., Svanberg, C., & Müllersdorf, M. (2015). Using a robotic cat in dementia care: a pilot study. *Journal of gerontological nursing*, 41(10), 46-56.
- [Harada1995] Harada, N., Chiu, V., Fowler, E., Lee, M., & Reuben, D. B. (1995). Physical therapy to improve functioning of older people in residential care facilities. *Physical Therapy*, 75(9), 830-838.
- [Hassenzahl2003] Hassenzahl, M., Burmester, M., & Koller, F. (2003). AttrakDiff: Ein Fragebogen zur Messung wahrgenommener hedonischer und pragmatischer Qualität. In *Mensch & computer 2003* (pp. 187-196). Vieweg+ Teubner Verlag.

- [HCIInternational2011] HCI International (2011). HCI International 2011. Retrieved April 25, 2020, from <http://2011.hci.international/index.php?module=webpage&id=35>
- [Hebesberger2017] Hebesberger, D., Koertner, T., Gisinger, C., & Pripfl, J. (2017). A Long-Term Autonomous Robot at a Care Hospital: A Mixed Methods Study on Social Acceptance and Experiences of Staff and Older Adults. *International Journal of Social Robotics*, 9(3), 417–429. <https://doi.org/10.1007/s12369-016-0391-6>
- [Heerink2008] Heerink, M., Kröse, B., Evers, V., & Wielinga, B. (2008). The influence of social presence on acceptance of a companion robot by older people. *J. Phys. Agents*, vol. 2, no. 2, pp. 33–40
- [Heerink2008b] Heerink, M., Kröse, B., Wielinga, B., & Evers, V. (2008). The influence of perceived adaptiveness of a social agent on acceptance by elderly users. *Gerontechnology*, 7(2), 521–526. <https://doi.org/10.4017/gt.2008.07.02.057.00>
- [Heerink2009] Heerink, M., Krose, B., Evers, V., & Wielinga, B. (2009, September). Measuring acceptance of an assistive social robot: a suggested toolkit. In *RO-MAN 2009-The 18th IEEE International Symposium on Robot and Human Interactive Communication* (pp. 528-533). IEEE.
- [Heerink2010] Heerink, M., Kröse, B., Evers, V., & Wielinga, B. (2010). Assessing acceptance of assistive social agent technology by older adults: the almere model. *International journal of social robotics*, 2(4), 361-375.
- [Heerink2010a] Heerink, M., Kröse, B., Evers, V., & Wielinga, B. (2010). Relating conversational expressiveness to social presence and acceptance of an assistive social robot. *Virtual reality*, 14(1), 77-84.
- [Helbostad2004] Helbostad, J. L., Sletvold, O., & Moe-Nilssen, R. (2004). Effects of home exercises and group training on functional abilities in home-dwelling older persons with mobility and balance problems. A randomized study. *Aging clinical and experimental research*, 16(2), 113-121.
- [Heylen2012] Heylen, D., van Dijk, B., & Nijholt, A. (2012, January). Robotic Rabbit Companions: amusing or a nuisance?. *Journal on multimodal user interfaces*, 5(1-2), 53-59.
- [Hoffmann1819] E. T. A. Hoffmann. (1819). Die Serapionsbrüder. Gesammelte Erzählungen und Märchen. Second issue. Berlin 1819. Issued by G. Reimer. 614 S
- [Holzinger2005] Holzinger, A. (2005). Usability engineering methods for software developers. *Communications of the ACM*, 48(1), 71-74.
- [HWWI2015] HWWI. (2015). Niedrigste Geburtenrate weltweit. Retrieved February 9, 2020 from http://www.hwwi.org/fileadmin/hwwi/Mediencenter/Pressemitteilungen/2015_Pressemitteilungen/2015-05-29-BDO/20150529_PM_IBC_Geburtenrate_HWWI.pdf.

- [IFR] International Federation of Robotics, (n.d.). Service Robots. Retrieved February 9, 2020 from <http://www.ifr.org/service-robots/>
- [ISO9241_2019] ISO. (2019). ISO 9241-210:2019. Retrieved April 9, 2020, from <https://www.iso.org/standard/77520.html>
- [Ihsen2013] Ihsen, S., Scheibl, K., Schneider, W., Glende, S., & Kohl, F. (2013). *ALIAS D1.5, Analysis of pilot's second test-run with qualitative advices on how to improve specific functions / usability of the robot.*
- [Jääskeläinen2010] Jääskeläinen, R. (2010). Think-aloud protocol. *Handbook of translation studies, 1*, 371-374.
- [Jenkins2014] Jenkins, S., & Draper, H. (2014, October). Robots and the Division of Healthcare Responsibilities in the Homes of Older People. In *International Conference on Social Robotics* (pp. 176-185). Springer, Cham.
- [Jung2004] Jung, Y., & Lee, K. M. (2004). Effects of physical embodiment on social presence of social robots. *Proceedings of PRESENCE*, 80-87.
- [Kareborn2009] Bergvall-Kåreborn, B., & Ståhlbröst, A. (2009). Living Lab: an open and citizen-centric approach for innovation. *International Journal of Innovation and Regional Development, 1*(4), 356-370.
- [Kelley1984] Kelley, J. F. (1984). An iterative design methodology for user-friendly natural language office information applications. *ACM Transactions on Information Systems (TOIS), 2*(1), 26-41.
- [Khosla2017] Khosla, R., Nguyen, K., & Chu, M. T. (2017). Human robot engagement and acceptability in residential aged care. *International Journal of Human-Computer Interaction, 33*(6), 510-522.
- [Kidd2008] Kidd, C. D., & Breazeal, C. (2008, September). Robots at home: Understanding long-term human-robot interaction. In *2008 IEEE/RSJ International Conference on Intelligent Robots and Systems* (pp. 3230-3235). IEEE.
- [King1991] King, A. C., Haskell, W. L., Taylor, C. B., Kraemer, H. C., & DeBusk, R. F. (1991). Group- vs home-based exercise training in healthy older men and women: a community-based clinical trial. *Jama, 266*(11), 1535-1542.
- [Koch2018] Koch, M., von Luck, K., Schwarzer, J., & Draheim, S. (2018). The novelty effect in large display deployments—Experiences and lessons-learned for evaluating prototypes. In *Proceedings of 16th European Conference on Computer-Supported Cooperative Work- Exploratory Papers*. European Society for Socially Embedded Technologies (EUSSET).

- [Kosman2013] Kosman, R., Eertink, H., Van der Wal, C., Ebben, P., Reitsma, J., Quinones, P., & Isken, M. (2013). *Florence D6.6: Evaluation of the FLORENCE System*.
- [Krainer2014] Krainer, D., Werner, F. & Oberzaucher, J. (2014). Performance of a socially assistive robot as trainer for physical exercises for older people. In *Wohnen - Pflege - teilhabe - "Besser leben durch Technik" - 7. Deutscher AAL-Kongress*. Berlin, Germany.
- [KSERA2012] Oberzaucher, J., Werner, F., Lewy, H., Lemberger, J., & Werner, K. (2012). *Deliverable of the KSERA project, D5.3 Formative Evaluation*. Vienna, Austria.
- [KSERA2012a] Lemberger, J., Oberzaucher, J., Werner, F., Lewy, H., & Werner, K. (2012). *Deliverable D5.5 End Evaluation*. Vienna, Austria.
- [Kuusela2000] Kuusela, H., & Pallab, P. (2000). A comparison of concurrent and retrospective verbal protocol analysis. *The American journal of psychology*, 113(3), 387.
- [Lee2004] Lee, K. M. (2004). Presence, explicated. *Communication theory*, 14(1), 27-50.
- [Lee2006] Lee, K. M., Jung, Y., Kim, J., & Kim, S. R. (2006). Are physically embodied social agents better than disembodied social agents?: The effects of physical embodiment, tactile interaction, and people's loneliness in human-robot interaction. *International journal of human-computer studies*, 64(10), 962-973.
- [Leite2009] Leite, I., Martinho, C., Pereira, A., & Paiva, A. (2009, September). As time goes by: Long-term evaluation of social presence in robotic companions. In *RO-MAN 2009-The 18th IEEE International Symposium on Robot and Human Interactive Communication* (pp. 669-674). IEEE.
- [Libin2004] Libin, A. V., & Libin, E. V. (2004). Person-robot interactions from the robopsychologists' point of view: The robotic psychology and robototherapy approach. *Proceedings of the IEEE*, 92(11), 1789-1803.
- [Littbrand2009] Littbrand, H., Lundin-Olsson, L., Gustafson, Y., & Rosendahl, E. (2009). The effect of a high-intensity functional exercise program on activities of daily living: a randomized controlled trial in residential care facilities. *Journal of the American Geriatrics Society*, 57(10), 1741-1749.
- [Loi2018] Loi, S. M., Bennett, A., Pearce, M., Nguyen, K., Lautenschlager, N. T., Khosla, R., & Velakoulis, D. (2018). A pilot study exploring staff acceptability of a socially assistive robot in a residential care facility that accommodates people under 65 years old. *International psychogeriatrics*, 30(7), 1075-1080.
- [Lombard2009] Lombard, M., Ditton, T. B., & Weinstein, L. (2009, October). Measuring presence: the temple presence inventory. In *Proceedings of the 12th annual international workshop on presence* (pp. 1-15).

- [Longo2015] Longo, D. H., & Vilain, P. (2015). Creating User Scenarios through User Interaction Diagrams by Non-Technical Customers. In *SEKE* (pp. 330-335).
- [Lucia2013] Lucia, P., Marcus, M., David, F., Alexander, N., Renxi, Q., Claudia, S., ... Rafael, L. (2013). *SRS D6.2: User validation results*.
- [Mann2015] Mann, J. A., MacDonald, B. A., Kuo, I. H., Li, X., & Broadbent, E. (2015). People respond better to robots than computer tablets delivering healthcare instructions. *Computers in Human Behavior*, 43, 112-117.
- [Markoff2010] Markoff, J. (2010). Google cars drive themselves, in traffic. *The New York Times*. Retrieved February 9, 2020, from <https://www.nytimes.com/2010/10/10/science/10google.html>
- [Mucchiani2017] Mucchiani, C., Sharma, S., Johnson, M., Sefcik, J., Vivio, N., Huang, J., ... & Lau, T. (2017, September). Evaluating older adults' interaction with a mobile assistive robot. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (pp. 840-847). IEEE.
- [Myers1996] Myers, E. R., & Weiner, G. I. (1996). Keeping old bones whole. *Harvard Health Letter*, 21(11), 1-2.
- [McCraty1998] McCraty, R., Barrios-Choplin, B., Rozman, D., Atkinson, M., & Watkins, A. D. (1998). The impact of a new emotional self-management program on stress, emotions, heart rate variability, DHEA and cortisol. *Integrative Physiological and Behavioral Science*, 33(2), 151-170.
- [McCroskey1999] McCroskey, J. C., & Teven, J. J. (1999). Goodwill: A reexamination of the construct and its measurement. *Communications Monographs*, 66(1), 90-103.
- [Melenhorst2013] Melenhorst, M., Isken, M., Lowet, D., Van de Wal, C., & Eertink, H. (2013). *Florence D6.4: Report on the Testing and Evaluation Methodology for the Living Lab Testing*.
- [Merten2012] Merten, M., Bley, A., Schroeter, C., & Gross, H. M. (2012, May). A mobile robot platform for socially assistive home-care applications. In *ROBOTIK 2012; 7th German Conference on Robotics* (pp. 1-6). VDE.
- [Mirror2013] Mirror. (2013, November 13). World's first robot SUICIDE as family return to find cleaning gadget had turned to ash. Retrieved February 9, 2020, from <https://www.mirror.co.uk/news/weird-news/worlds-first-robot-suicide-family-2786901>
- [Morgan1997] Morgan, D. L. (1997). *The focus group guidebook* (Vol. 1). Sage publications.
- [Mori1970] Mori, M., MacDorman, K. F., & Kageki, N. (1970). The uncanny valley. *IEEE Robotics and Automation Magazine*, 19(2), 98-100. <https://doi.org/10.1109/MRA.2012.2192811>
- [NASA2015] NASA. (2015, May 6). Technology Readiness Level. Retrieved February 9, 2020, from <http://www.nasa.gov/content/technology-readiness-level/#.VOXJUlPTOXs>.

- [Nestle2014] Nestlé to use humanoid robot to sell Nescafé in Japan. (2014, October 29). Retrieved February 9, 2020, from <http://www.nestle.com/media/news/nestle-humanoid-robot-nescafe-japan>
- [Nielsen1990] Nielsen, J., & Molich, R. (1990, March). Heuristic evaluation of user interfaces. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (pp. 249-256).
- [Nielsen1994] Nielsen, J. (1994). *Usability engineering*. Morgan Kaufmann.
- [Onwuegbuzie2007] Onwuegbuzie, A. J., & Leech, N. L. (2007). Validity and qualitative research: An oxymoron?. *Quality & quantity*, 41(2), 233-249.
- [Operto2008] Operto, F. (2008). Florence D6.4: Report on the Testing and Evaluation Methodology for the Living Lab Testing. Springer Handbook of Robotics. <https://doi.org/10.1007/978-3-540-30301-5>
- [Östlund2015] Östlund, B., Olander, E., Jonsson, O., & Frennert, S. (2015). STS-inspired design to meet the challenges of modern aging. Welfare technology as a tool to promote user driven innovations or another way to keep older users hostage?. *Technological Forecasting and Social Change*, 93, 82-90.
- [Panek2015] P. Panek, P. Mayer. (2015, February 19). Interview with HRI experts at the Centre for Applied Assistive Technologies, Human Computer Interaction (HCI) Group of the Institute of Design & Assessment of Technology, TU Wien
- [Papadopoulos2019] Papadopoulos, R., Koulouglioti, C., Lazzarino, R., & Ali, S. (2019). A systematic review of enablers and barriers to the implementation of socially assistive humanoid robots in health and social care. <https://doi.org/10.1136/bmjopen-2019-033096>
- [Payr2013] Payr, S. (2013). Virtual butlers and real people: styles and practices in long-term use of a companion. In *Your Virtual Butler*(pp. 134-178). Springer, Berlin, Heidelberg.
- [Payr2015] Payr, S., Werner, F., & Werner, K. (2015). Potential of robotics for ambient assisted living. Report, published by the Austrian research promotion agency, Vienna.
- [Payr2015a] Payr, S., Werner, F., & Werner, K. (2015, June). AAL robotics: state of the field and challenges. In *eHealth* (pp. 117-124).
- [Pérez2014] Pérez, J. G., Lohse, M., & Evers, V. (2014). *Accompany D6.3, Acceptability of a home companion robot*. Public Report.
- [Pigini2012] Pigini, L., Facal, D., Mast, M., Blasi, L., López, R., & Arbeiter, G. (2012). *SRS Project D6.1: Testing site preparation and protocol development*. Public Report
- [Pigini2013] L. Pigini, M. Mast, D. Facal, A. Noyvirt, R. Qiu, S. Claudia, G. Alvaro, and L. Rafael. (2013). SRS Deliverable D6.2: User validation results. Public Report

- [Plasticpals2009] Plasticpals.com. (2009). Ifbot. Retrieved November 10, 2015, from <http://www.plasticpals.com/?p=1409>
- [Pollack2002] Pollack, M. E., Brown, L., Colbry, D., Orosz, C., Peintner, B., Ramakrishnan, S., ... & Thrun, S. (2002, August). Pearl: A mobile robotic assistant for the elderly. In *AAAI workshop on automation as eldercare* (Vol. 2002).
- [Potenziaal2015] Johanneum Research Forschungsgesellschaft. (2015). PotenziAAL-Pflege, Abschätzung des Marktpotenzials von Technologien aus dem Bereich 'Ambient Assisted Living' – Abschlussbericht. Vienna.
- [Powers2007] Powers, A., Kiesler, S., Fussell, S., & Torrey, C. (2007). Comparing a Computer Agent with a Humanoid Robot. *ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, (November 2015), 145–152. <https://doi.org/10.1145/1228716.1228736>
- [Pripfl2016] Pripfl, J., Kortner, T., Batko-Klein, D., Hebesberger, D., Weninger, M., Gisinger, C., ... Vincze, M. (2016). Results of a real world trial with a mobile social service robot for older adults. *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 497–498. <https://doi.org/10.1109/HRI.2016.7451824>
- [Purchase2012] Purchase, H. C. (2012). *Experimental human-computer interaction: a practical guide with visual examples*. Cambridge University Press.
- [Radio2017] Radio Consortium. (2017). Public Deliverable 6.11 User Evaluation Report of the Radio Project.
- [Radio2017a] Radio Consortium. (2017). Public Deliverable 6.4 Piloting Plan IV of the Radio Project. Retrieved February 9, 2020 from <http://radio-project.eu/downloads/documents/Radio-d6.04-PilotingPlan.pdf>
- [Ramachandran2018] Ramachandran, A., Huang, C. M., Gartland, E., & Scassellati, B. (2018, February). Thinking aloud with a tutoring robot to enhance learning. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction* (pp. 59-68).
- [Rehrl2012] Rehrl, T., Troncy, R., Bley, A., Ihsen, S., Scheibl, K., Schneider, W., ... & Wallhoff, F. (2012). The ambient adaptable living assistant is meeting its users. In *Proc. of AAL Forum 2012-Eindhoven*.
- [Rentschler2008] Rentschler, A. J., Simpson, R., Cooper, R. A., & Boninger, M. L. (2008). Clinical evaluation of Guido robotic walker. *Journal of Rehabilitation Research and Development*, 45(9), 1281–1293. <https://doi.org/10.1682/JRRD.2007.10.0160>
- [Riek2012] Riek, L. D. (2012). Wizard of oz studies in hri: a systematic review and new reporting guidelines. *Journal of Human-Robot Interaction*, 1(1), 119-136.

- [Riek2014] Riek, L., & Howard, D. (2014). A code of ethics for the human-robot interaction profession. *Proceedings of We Robot*.
- [Russel1980] Russell, D., Peplau, L. A., & Cutrona, C. E. (1980). The revised UCLA Loneliness Scale: concurrent and discriminant validity evidence. *Journal of personality and social psychology*, 39(3), 472.
- [Sabanovic2013] Sabanovic, S., Bennett, C. C., Chang, W., Huber, L., & Šabanović, S. (2013). PARO robot affects diverse interaction modalities in group sensory therapy for older adults with dementia. *IEEE International Conference on Rehabilitation Robotics: [Proceedings], 2013*, 6650427. <https://doi.org/10.1109/ICORR.2013.6650427>
- [Saerbeck2010] Saerbeck, M., Schut, T., Bartneck, C., & Janse, M. D. (2010, April). Expressive robots in education: varying the degree of social supportive behavior of a robotic tutor. In *Proceedings of the SIGCHI conference on human factors in computing systems*(pp. 1613-1622).
- [Savioke2015] Savioke - Meet Relay. (2015). Retrieved April 14, 2020, from <http://www.savioke.com/>
- [Schanze2010] Mascha Film. (2010). *Plug&Pray*. directed by Jens Schanze.
- [Schroter2014] Schröter, C., Müller, S., Volkhardt, M., Einhorn, E., Gross, H., Neuroinformatik, F., ... Bley, A. (2014). CompanionAble – ein robotischer Assistent und Begleiter für Menschen mit leichter kognitiver Beeinträchtigung, 216487(Aal).
- [Sheeran2002] Sheeran, P. (2002). Intention—behavior relations: a conceptual and empirical review. *European review of social psychology*, 12(1), 1-36.
- [Sherrington2008] Sherrington, C., Whitney, J. C., Lord, S. R., Herbert, R. D., Cumming, R. G., & Close, J. C. (2008). Effective exercise for the prevention of falls: a systematic review and meta-analysis. *Journal of the American Geriatrics Society*, 56(12), 2234-2243.
- [Schröter2013] Schroeter, C., Mueller, S., Volkhardt, M., Einhorn, E., Huijnen, C., van den Heuvel, H., ... & Gross, H. M. (2013, May). Realization and user evaluation of a companion robot for people with mild cognitive impairments. In *2013 IEEE International Conference on robotics and automation* (pp. 1153-1159). IEEE.
- [Schröter2014] Schröter, C., Müller, S., Volkhardt, M., Einhorn, E., Gross, H. M., Huijnen, C., ... & Bley, A. (2014). CompanionAble—ein robotischer Assistent und Begleiter für Menschen mit leichter kognitiver Beeinträchtigung. In *Proc. of 7th German AAL Conference (AAL 2014)*. VDE Verlag, 5 (Vol. 8).
- [Schwenk2008] Schwenk, M., Oster, P., & Hauer, K. (2008). Kraft-und Funktionstraining bei älteren Menschen mit dementieller Erkrankung. *Praxis Physiotherapie*, 2, 59-65.

- [Shibata2011] Shibata, T., & Wada, K. (2011). Robot therapy: a new approach for mental healthcare of the elderly – a mini-review. *Gerontology*, 57(4), 378-386.
- [Shiarlis2015] Shiarlis, K., Messias, J., van Someren, M., Whiteson, S., Kim, J., Vroon, J., ... & Pérez-Hurtado, I. (2015, May). Teresa: A socially intelligent semi-autonomous telepresence system. In *Workshop on machine learning for social robotics at ICRA-2015 in Seattle*.
- [Siino2005] Siino, R. M., & Hinds, P. J. (2005, April). Robots, Gender & Sensemaking: Sex Segregation's Impact On Workers Making Sense Of a Mobile Autonomous Robot. In *Proceedings of the 2005 IEEE international conference on robotics and automation*(pp. 2773-2778). IEEE.
- [Smith1995] Smith, J. A. (1995). Semi structured interviewing and qualitative analysis.
- [Sparrow2006] Sparrow, R., & Sparrow, L. (2006). In the hands of machines? The future of aged care. *Minds and Machines*, 16(2), 141-161.
- [Strauss2014] Corbin, J., & Strauss, A. (2014). *Basics of qualitative research: Techniques and procedures for developing grounded theory*. Fourth edition. Sage publications.
- [Sun2006] Sun, H., & Zhang, P. (2006). Causal relationships between perceived enjoyment and perceived ease of use: An alternative approach. *Journal of the Association for Information Systems*, 7(1), 24.
- [Takayama2009] Takayama, L., & Pantofaru, C. (2009). Influences on proxemic behaviors in human-robot interaction. *2009 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2009*, (2009), 5495–5502. <https://doi.org/10.1109/IROS.2009.5354145>
- [Tapus2008] Tapus, A., Țăpuș, C., & Matarić, M. J. (2008). User-robot personality matching and assistive robot behavior adaptation for post-stroke rehabilitation therapy. *Intelligent Service Robotics*, 1(2), 169.
- [Terraciano2003] Terraciano, A., McCrae, R. R., & Costa Jr, P. T. (2003). Factorial and construct validity of the Italian Positive and Negative Affect Schedule (PANAS). *European Journal of Psychological Assessment*, 19(2), 131.
- [Tamura2004] Tamura, T., Yonemitsu, S., Itoh, A., Oikawa, D., Kawakami, A., Higashi, Y., ... & Nakajima, K. (2004). Is an entertainment robot useful in the care of elderly people with severe dementia?. *The Journals of Gerontology Series A: Biological Sciences and Medical Sciences*, 59(1), M83-M85.
- [Tinetti2003] Tinetti, M. E. (2003). Preventing falls in elderly persons. *New England journal of medicine*, 348(1), 42-49.
- [Torta2012] Torta, E., van Heumen, J., Cuijpers, R. H., & Juola, J. F. (2012, October). How can a robot attract the attention of its human partner? a comparative study over different

modalities for attracting attention. In *International Conference on Social Robotics* (pp. 288-297). Springer, Berlin, Heidelberg.

[Torta2013] Torta, E., Oberzaucher, J., Werner, F., Cuijpers, R. H., & Juola, J. F. (2012). Attitudes towards socially assistive robots in intelligent homes: results from laboratory studies and field trials. *Journal of Human-Robot Interaction*, 1(2), 76-99.

[Torta2013b] Torta, E., Cuijpers, R. H., & Juola, J. F. (2013). Design of a parametric model of personal space for robotic social navigation. *International Journal of Social Robotics*, 5(3), 357-365.

[Torta2014] Torta, E., Werner, F., Johnson, D. O., Juola, J. F., Cuijpers, R. H., Bazzani, M., ... & Bregman, J. (2014). Evaluation of a small socially-assistive humanoid robot in intelligent homes for the care of the elderly. *Journal of Intelligent & Robotic Systems*, 76(1), 57-71.

[Coleman1993] Coleman, N. (1993). SUMI (Software Usability Measurement Inventory) as a knowledge elicitation tool for improving usability. *Ireland: Department of Applied Psychology, University College Cork*.

[UniB2005] Universität Bremen. (2005). Assistenzroboter Friend. Retrieved April 25, 2020, from <http://www.iat.uni-bremen.de/sixcms/detail.php?id=1090>

[UWE2010] University of the West of England. (2010). Mobiserv Project D7.3: Final system Prototype. Public Report.

[UWE2013] University of the West of England. (2013). Mobiserv Project D2.4: Evaluation Plan. Public Report.

[VanBreemen2005] van Breemen, A., Yan, X., & Meerbeek, B. (2005, July). iCat: an animated user-interface robot with personality. In *Proceedings of the fourth international joint conference on Autonomous agents and multiagent systems* (pp. 143-144).

[VandenBerg2004] Van den Berg, B., Brouwer, W. B., & Koopmanschap, M. A. (2004). Economic valuation of informal care. *The European Journal of Health Economics, formerly: HEPAC*, 5(1), 36-45.

[VanDijk2013] Van Dijk, E. T., Torta, E., & Cuijpers, R. H. (2013). Effects of eye contact and iconic gestures on message retention in human-robot interaction. *International Journal of Social Robotics*, 5(4), 491-501.

[Venkatesh2000] Venkatesh, V. (2000). Determinants of perceived ease of use: Integrating control, intrinsic motivation, and emotion into the technology acceptance model. *Information systems research*, 11(4), 342-365.

[Venkatesh2003] Venkatesh, V., Morris, M. G., Davis, G. B., & Davis, F. D. (2003). User acceptance of information technology: Toward a unified view. *MIS quarterly*, 425-478.

- [Venkatesh2008] Venkatesh, V., & Bala, H. (2008). Technology acceptance model 3 and a research agenda on interventions. *Decision sciences*, 39(2), 273-315.
- [Victores2014] Victores, J. G. (2014). An accessible interface for programming an assistive robot. *Journal of accessibility and design for all*, 4(3), 161-176.
- [Villarreal2011] Jiménez Villarreal, J., & Ljungblad, S. (2011, March). Experience centred design for a robotic eating aid. In *Proceedings of the 6th international conference on Human-robot interaction* (pp. 155-156). Lausanne, Switzerland.
- [Vroon2015] Jered Vroon, Gwenn Englebienne, V. E. (2015). Deliverable 3.2: Longitudinal Effects Report of the project "Teresa,".
- [Wada2007] Wada, K., & Shibata, T. (2007). Living with seal robots — its sociopsychological and physiological influences on the elderly at a care house. *IEEE transactions on robotics*, 23(5), 972-980.
- [Walters2009] Walters, M. L., Dautenhahn, K., Te Boekhorst, R., Koay, K. L., Syrdal, D. S., & Nehaniv, C. L. (2009). An empirical framework for human-robot proxemics. *Procs of new frontiers in human-robot interaction*.
- [Ware1996] Ware Jr, J. E., Kosinski, M., & Keller, S. D. (1996). A 12-Item Short-Form Health Survey: construction of scales and preliminary tests of reliability and validity. *Medical care*, 220-233.
- [Watson1988] Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: the PANAS scales. *Journal of personality and social psychology*, 54(6), 1063.
- [Weiss2009] Weiss, A., Bernhaupt, R., Lankes, M., & Tscheligi, M. (2009, April). The USUS evaluation framework for human-robot interaction. In *AISB2009: proceedings of the symposium on new frontiers in human-robot interaction* (Vol. 4, No. 1, pp. 11-26).
- [Weiss2015] Weiss, A., & Bartneck, C. (2015). Meta analysis of the usage of the Godspeed Questionnaire Series. *Proceedings - IEEE International Workshop on Robot and Human Interactive Communication, 2015-Novem*, 381-388.
<https://doi.org/10.1109/ROMAN.2015.7333568>
- [Werner2012] Werner, K., Oberzaucher, J., & Werner, F. (2012, July). Evaluation of human robot interaction factors of a socially assistive robot together with older people. In *2012 sixth International Conference on complex, intelligent, and software intensive systems* (pp. 455-460). IEEE.
- [Werner2012a] F. Werner, J. Oberzaucher, and K. Werner. (2012). Real-life Evaluation of a Socially Assistive Robot. *Gerontechnology*, vol. 11, no. 2, p. 382,

- [Werner2013] F. Werner, K. Werner, and J. Oberzaucher. (2013). Evaluation of the acceptance of a socially assistive robot by older users within the project KSERA. In *Proceedings: Lebensqualität im Wandel von Demografie und Technik-6. Deutscher AAL-Kongress*.
- [Werner2013a] F. Werner and D. Krainer. (2013). A Socially Assistive Robot to Support Physical Training of Older People – An End User Acceptance Study. In *Social Robotics, 5th international Conference, ICSR 2013, 2013*, pp. 562–563.
- [Werner2013b] Werner, F., Krainer, D., Oberzaucher, J., & Werner, K. (2013). Evaluation of the acceptance of a social assistive robot for physical training support together with older users and domain experts. *Assistive Technology: From Research to Practice*, 33, 137-142.
- [WHO1996] World Health Organization. (1996). *WHOQOL-BREF: introduction, administration, scoring and generic version of the assessment: field trial version, December 1996* (No. WHOQOL-BREF). Geneva: World Health Organization.
- [WHO2015a] 10 facts on healthy ageing in Europe. (2015). Retrieved April 14, 2020, from <http://www.euro.who.int/en/health-topics/Life-stages/healthy-ageing/data-and-statistics/10-facts-on-healthy-ageing-in-europe>
- [Wohlmannstetter2016] M. Wohlmannstetter. (2016). Einsatz von Robotern in der Pflege – aus der Sicht der Pflege. Presentation at the open session of the Austrian bioethics commission, Vienna, Austria.
- [WPU2013] WPU. (2013). Studie zur Geschäftsmodellentwicklung für den AAL-Markt unter Berücksichtigung der österreichischen Rahmenbedingungen. Vienna. Public Report.
- [WRSR2015] International Federation of Robotics. (2015, September). World Robotics – Service Robots 2015. Report published by the IFR. Frankfurt, Germany.
- [Xu2012] Xu, Q., Ng, J. S. L., Cheong, Y. L., Tan, O. Y., Wong, J. Bin, Tay, B. T. C., & Park, T. (2012). Effect of scenario media on human-robot interaction evaluation. In *Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction* (pp. 275–276).
- [Yesavage1982] Yesavage, J. A., Brink, T. L., Rose, T. L., Lum, O., Huang, V., Adey, M., & Leirer, V. O. (1982). Development and validation of a geriatric depression screening scale: a preliminary report. *Journal of psychiatric research*, 17(1), 37-49.
- [Zimet1988] Zimet, G. D., Dahlem, N. W., Zimet, S. G., & Farley, G. K. (1988). The multidimensional scale of perceived social support. *Journal of personality assessment*, 52(1), 30-41.

this page intentionally left blank

Annex

A1 Attitude towards technology questionnaire – usage of technical devices in daily life

1. Please indicate the technical devices you use during your daily living

- TV
- Mobile phone
- Computer
- Radio
- Medical Reminder
- Physical Measurement Tools

Others: _____

2. Do you have an internet connection in your home

- yes
- no

3. What is the TOTAL number of hours a week that you spend on a computer?

0	Less than 1	1-3	4-6	7-10	11-15	16-20	20+
---	-------------	-----	-----	------	-------	-------	-----

4. What is your most common activity when using a computer? (check ONLY ONE)

- Internet surfing
- Email
- Games
- Writing
- Getting information
- Social activities

Others? _____

5. Do you use the internet more or less often than your friends/relatives?

more less same as

6. How many hours a week do you spend doing the following activities:

Internet surfing / Searching for information on the internet

0	Less than 1	1-3	4-6	7-10	11-15	16-20	20+
---	-------------	-----	-----	------	-------	-------	-----

Using email

0	Less than 1	1-3	4-6	7-10	11-15	16-20	20+
---	-------------	-----	-----	------	-------	-------	-----

How many hours a week do you spend talking on the telephone?

0	Less than 1	1-3	4-6	7-10	11-15	16-20	20+
---	-------------	-----	-----	------	-------	-------	-----

How many hours a week do you spend watching television?

0	Less than 1	1-3	4-6	7-10	11-15	16-20	20+
---	-------------	-----	-----	------	-------	-------	-----

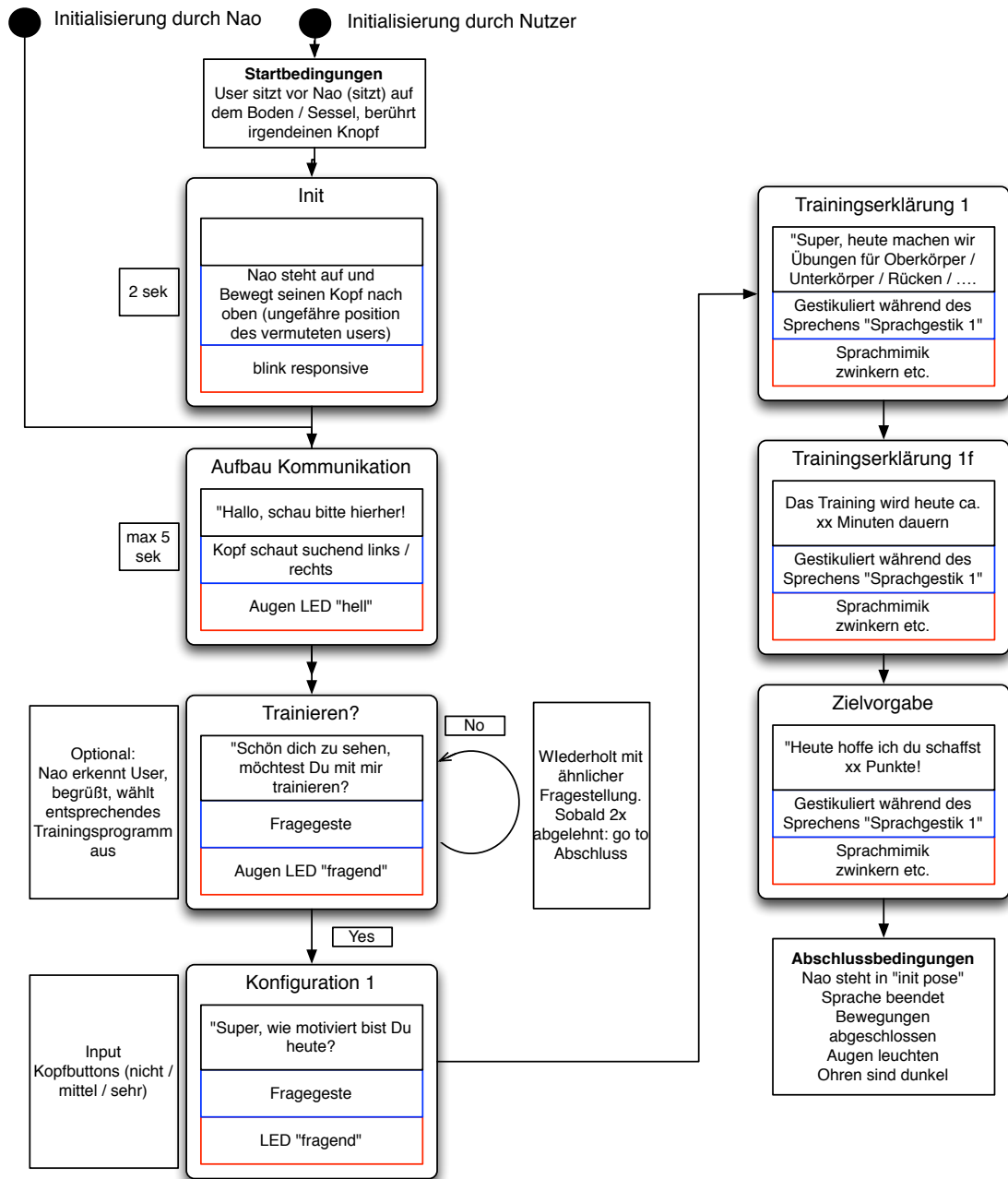
How many hours a week do you spend watching television with friends / family?

0	Less than 1	1-3	4-6	7-10	11-15	16-20	20+
---	-------------	-----	-----	------	-------	-------	-----

How many hours a week do you spend socializing with friends / family?

0	Less than 1	1-3	4-6	7-10	11-15	16-20	20+
---	-------------	-----	-----	------	-------	-------	-----

A2 Example of an interaction flow used within E3



A3 Examples of a custom developed questionnaire for satisfaction assessment within E1

Some statements are listed below. Please circle the number that fits your opinion best.

1. How beneficial do you believe is KSERA for you?

1	2	3	4	5
Not beneficial at all	Not very beneficial	Neither beneficial nor harming	Beneficial	Very beneficial

2. How often would you use KSERA if you had it at home?

1	2	3	4	5
Never	once a month	Once a week	3 times a week	Every day

3. Did KSERA make you feel more confident about using new technologies?

1	2	3	4	5
Not at all confident	A little confident	Neutral	Confident	Very confident

4. KSERA was amusing and I enjoyed interacting with it.

1	2	3	4	5
Strongly disagree	Disagree	Neither agree nor disagree	Agree	Strongly agree

5. I prefer spending my time by doing something else than interacting with KSERA

1	2	3	4	5
Strongly dissagree	Disagree	Neither agree nor disagree	Agree	Strongly agree

6. Using KSERA was boring and did not interest me.

1	2	3	4	5
Strongly dissagree	Disagree	Neither agree nor disagree	Agree	Strongly agree

7. KSERA has met my expectations.

1	2	3	4	5
Strongly dissagree	Disagree	Neither agree nor disagree	Agree	Strongly agree

The following score sheet represents an excerpt of an example as used for the first test case in the first iteration of the first evaluation of the first prototype.

Reliability Evaluation Form - Testcase 1		Testperson number (e.g., TP2):			
Testcase 1 Injection Point 0 start in tester interface					
STC 1	Nao navigates to TS / designated place	Method	First Testrun	Second Testrun	Third Testrun
A1	Is Nao performing movements in the correct way?	Visual	Y / N text description (in case of fa) from till duration	Y / N text description from till duration	Y / N text description from till duration
A2	How long does it take NAO to perform the intended movement? time taken in mm:ss		Y / N diff in m text	Y / N diff in m text	Y / N diff in m text
A3	Is the KSERA system calculating the right user position?	Visual by interpretation the KSERA UI	Y / N diff in m text	Y / N diff in m text	Y / N diff in m text
A4	Is the KSERA system calculating the right NAO position?	Visual by interpretation the KSERA UI	Y / N diff in m text	Y / N diff in m text	Y / N diff in m text
Break Criteria		Nao could not move (e.g., falls down, gets to hot, battery out, etc) Could not get to the user within 2 minutes	Other Problems	Other Problems	Other Problems
Testcase 1 Injection Point 1 start in tester interface in case STC1 failed					
STC 2	Nao detects focus of the TS	Method	First Testrun	Second Testrun	Third Testrun
A1	Is Nao performing movements in the correct way?	Visual	Y / N text description (in case of fa) from till duration	Y / N text description from till duration	Y / N text description from till duration
AX	Is KSERA able to detect the users face?	Visual (KSERA UI, NAOs camera)	Y / N text description (in case of fa) from till duration	Y / N text description (in case of fa) from till duration	Y / N text description (in case of fa) from till duration
XX	How long does it take the system to find the users face?	Observation until user image visible in t	Y / N diff in m text	Y / N diff in m text	Y / N diff in m text
Break Criteria		Nao does not detect the focus of the TP within 1 minute, although the testperson clearly responded	Other Problems	Other Problems	Other Problems
In case NAO did not raise the attention successfully (TP was briefed not to look at Nao)					
STC 3	Nao attracts attention	Method	First Testrun	Second Testrun	Third Testrun
A1	Is Nao performing movements in the correct way?	Visual	Y / N text description (in case of fa) from till duration	Y / N text description from till duration	Y / N text description from till duration

A4 Example of a customized questionnaire to assess motivational capabilities of the robotic system

1. NAO is motivating me to do the exercises.

1	2	3	4	5
Not at all	A little	Almost	Quite	Very much

2. I am more motivated, compared to only having a training plan

1	2	3	4	5
Not at all	A little	Almost	Quite	Very much

3. I am more motivated, compared to only having a personal trainer

1	2	3	4	5
Not at all	A little	Almost	Quite	Very much

4. The exercises were described and shown in an understandable way

1	2	3	4	5
Not at all	A little	Almost	Quite	Very much

5. Mimicking the exercises makes me performing them in a better way, than only having the basic description

1	2	3	4	5
Not at all	A little	Almost	Quite	Very much

6. The direct feedback after the exercise performance makes me feel good

1	2	3	4	5
Not at all	A little	Almost	Quite	Very much

A5 Examples of questionnaires used within evaluation phase 3

Emotionale Auswertung der persönlichen Gefühle kurz nach Testdurchführung

Anschließend sind einige Gefühle aufgelistet, die beschreiben sollen, wie Ihr Eindruck von den durchgeführten Testszenarien JETZT ist. Zu jedem Gefühl gibt es eine Skala, die angeben soll, wie sehr dieses Gefühl bei Ihnen zutrifft oder nicht. **Bitte kreisen Sie die jeweilige Nummer ein, dessen Antwort am ehesten beschreibt, was sie AKTUELL gegenüber dem Physicaal-System und den durchgeführten Szenarien empfinden!**

Denken Sie nicht lange über Ihre Antwort nach und behalten Sie im Hinterkopf: Es gibt keine richtigen oder falschen Antworten. Nur Ihre Meinung.

1. Es macht Spaß.

1	2	3	4	5	6	7
überhaupt nicht			Ein bisschen			sehr

2. Es ist unangenehm.

1	2	3	4	5	6	7
überhaupt nicht			Ein bisschen			sehr

3. Ich mag es nicht.

1	2	3	4	5	6	7
überhaupt nicht			Ein bisschen			sehr

4. Es macht mich gut gelaunt.

1	2	3	4	5	6	7
überhaupt nicht			Ein bisschen			sehr

5. Ich fühle mich stark.

1	2	3	4	5	6	7
überhaupt nicht			Ein bisschen			sehr

6. Ich bin müde.

1	2	3	4	5	6	7
überhaupt nicht			Ein bisschen			sehr

7. Ich fühle mich erfrischt.

1	2	3	4	5	6	7
überhaupt nicht			Ein bisschen			sehr

8. Ich fühle mich gestresst.

1	2	3	4	5	6	7
überhaupt nicht			Ein bisschen			sehr

9. Ich fühle mich ruhig.

1	2	3	4	5	6	7
überhaupt nicht			Ein bisschen			sehr

10. Ich fühle mich gelangweilt.

1	2	3	4	5	6	7
überhaupt nicht			Ein bisschen			sehr

Trainingsunterstützung

1. NAO motiviert mich, die Übungen durchzuführen.

1	2	3	4	5
Überhaupt nicht	kaum	Ein wenig	ziemlich	Sehr stark

2. Ich bin motivierter als wenn ich einem vorgegebenen Trainingsplan folgen müsste.

1	2	3	4	5
Überhaupt nicht	kaum	Ein wenig	Viel motivierter	Sehr viel motivierter

3. Ich bin motivierter als wenn ich mit einem persönlichen Trainer trainiere.

1	2	3	4	5
Überhaupt nicht	kaum	Ein wenig	Viel motivierter	Sehr viel motivierter

4. Die Übungen werden auf verständliche Weise präsentiert.

1	2	3	4	5
Überhaupt nicht	kaum	Ein wenig	ziemlich	Sehr stark

5. Die Übungen vorgezeigt zu bekommen führt dazu, dass ich die Übungen richtiger ausführe als wenn ich nur eine Beschreibung der Übungen hätte.

1	2	3	4	5
Überhaupt nicht	kaum	Ein wenig	ziemlich	Sehr stark

Nützlichkeit und geplante Nutzung des Systems

1. Wie nützlich denken Sie wäre PHYSICAAL aktuell für Sie persönlich?

1	2	3	4	5
Gar nicht nützlich	Nicht sehr nützlich	Weder noch	Nützlich	Sehr nützlich

1. Wie nützlich denken Sie wäre PHYSICAAL aktuell für 10-30 jährige?

1	2	3	4	5
Gar nicht nützlich	Nicht sehr nützlich	Weder noch	Nützlich	Sehr nützlich

1. Wie nützlich denken Sie wäre PHYSICAAL aktuell für 31-50 jährige?

1	2	3	4	5
Gar nicht nützlich	Nicht sehr nützlich	Weder noch	Nützlich	Sehr nützlich

1. Wie nützlich denken Sie wäre PHYSICAAL aktuell für 51-70 jährige?

1	2	3	4	5
Gar nicht nützlich	Nicht sehr nützlich	Weder noch	Nützlich	Sehr nützlich

1. Wie nützlich denken Sie wäre PHYSICAAL aktuell für über 70 jährige?

1	2	3	4	5
Gar nicht nützlich	Nicht sehr nützlich	Weder noch	Nützlich	Sehr nützlich

2. Wie oft würden Sie PHYSICAAL nützen, wenn Sie es zu Hause hätten?

1	2	3	4	5
Nie	Ein Mal im Monat	1x pro Woche	3x pro Woche	Jeden Tag

4. PHYSICAAL war unterhaltsam und die Interaktion mit dem System hat mir gefallen.

1	2	3	4	5
Stimme überhaupt nicht zu	Stimme nicht zu	Weder noch	Stimme zu	Stimme stark zu

5. PHYSICAAL zu benutzen war langweilig und hat mich nicht interessiert.

1	2	3	4	5
überhaupt nicht	Stimme nicht zu	Weder noch	Stimme zu	Stimme stark zu

6. PHYSICAAL hat meine Erwartungen erfüllt.

1	2	3	4	5
Stimme überhaupt nicht zu	Stimme nicht zu	Weder noch	Stimme zu	Stimme stark zu

7. Was sind aus ihrer Sicht die größten Kritikpunkte an dieser Trainingsform? (max. 3)

8. Was sind aus ihrer Sicht die größten Stärken dieser Trainingsform? (max. 3)

Eigenschaften des Roboters

1. "Wie würden Sie NAOs Bewegungen beschreiben?"

schnell 1 2 3 4 5 langsam

elegant 1 2 3 4 5 unbeholfen

menschlich 1 2 3 4 5 maschinenähnlich

2. "Wie würden Sie NAOs Trainingserklärungen beschreiben?"

zu ausführlich 1 2 3 4 5 zu wenig

gut verständlich 1 2 3 4 5 schlecht verständlich

3. "Wie würden Sie NAOs Feedback beschreiben?"

passend 1 2 3 4 5 unpassend

zu ausführlich 1 2 3 4 5 zu wenig

gut verständlich 1 2 3 4 5 schlecht verständlich

Akzeptanz des Systems – Heerink Fragebogen

Angstgefühl – (ANX)

Wenn ich den Roboter verwenden würde, hätte ich Angst, Fehler zu machen.

Überhaupt nicht 1 2 3 4 5 sehr

Wenn ich den Roboter verwenden würde, hätte ich Angst etwas kaputt zu machen.

Überhaupt nicht 1 2 3 4 5 sehr

Ich finde den Roboter erschreckend.

Überhaupt nicht 1 2 3 4 5 sehr

Ich finde den Roboter einschüchternd.

Überhaupt nicht 1 2 3 4 5 sehr

Technikakzeptanz – (ATT)

Ich denke es ist eine gute Idee den Roboter zu verwenden.

Überhaupt nicht 1 2 3 4 5 sehr

Der Roboter würde das Leben interessanter machen.

Überhaupt nicht 1 2 3 4 5 sehr

Es ist gut den Roboter zu verwenden.

Überhaupt nicht 1 2 3 4 5 sehr

Empfundene Nutzbarkeit – (FCC)

Ich habe alles was ich brauche um den Roboter zu nutzen.

Überhaupt nicht 1 2 3 4 5 sehr

Ich kenne mich mit dem Roboter gut genug aus um ihn zu verwenden.

Überhaupt nicht 1 2 3 4 5 sehr

Geplante Nutzung – (ITU)

Hätte ich den Roboter zu Hause würde ich ihn innerhalb der nächsten Tage verwenden.

Überhaupt nicht 1 2 3 4 5 sehr

Empfundene Anpassbarkeit – (PAD)

Ich glaube, der Roboter kann sich an meine Bedürfnisse anpassen.

Überhaupt nicht 1 2 3 4 5 sehr

Ich glaube, der Roboter wird nur tun was ich gerade brauche.

Überhaupt nicht 1 2 3 4 5 sehr

Ich glaube, der Roboter wird mich unterstützen, wenn ich es für notwendig halte.

Überhaupt nicht 1 2 3 4 5 sehr

Empfundenes Vergnügen – (PENJ)

Ich mag es wenn der Roboter zu mir spricht.

Überhaupt nicht 1 2 3 4 5 sehr

Ich genieße es mit dem Roboter Dinge zu tun.

Überhaupt nicht 1 2 3 4 5 sehr

In finde den Roboter unterhaltsam.

Überhaupt nicht 1 2 3 4 5 sehr

In finde den Roboter faszinierend.

Überhaupt nicht 1 2 3 4 5 sehr

In finde den Roboter langweilig.

Überhaupt nicht 1 2 3 4 5 sehr

Empfundener Bedienungskomfort – (PEOU)

Ich glaube, ich werde schnell wissen, wie der Roboter zu benutzen ist.

Überhaupt nicht 1 2 3 4 5 sehr

Ich finde den Roboter einfach zu verwenden.

Überhaupt nicht 1 2 3 4 5 sehr

Ich glaube, ich kann den Roboter benutzen, wenn jemand anderer da ist um mir zu helfen.

Überhaupt nicht 1 2 3 4 5 sehr

Ich glaube, ich kann den Roboter ohne Hilfe benutzen.

Überhaupt nicht 1 2 3 4 5 sehr

Ich glaube, ich kann den Roboter benutzen, wenn ich eine gute Anleitung habe.

Überhaupt nicht 1 2 3 4 5 sehr

Empfundene Freundschaftlichkeit – (PS)

Ich empfinde den Roboter als erfreulichen Gesprächspartner.

Überhaupt nicht 1 2 3 4 5 sehr

Ich finde es nett mit dem Roboter zu interagieren.

Überhaupt nicht 1 2 3 4 5 sehr

Mir kommt es vor als ob der Roboter mich versteht.

Überhaupt nicht 1 2 3 4 5 sehr

Ich denke der Roboter ist nett.

Überhaupt nicht 1 2 3 4 5 sehr

Empfundene Nutzbarkeit – (PU)

Ich denke der Roboter ist hilfreich für mich.

Überhaupt nicht 1 2 3 4 5 sehr

Es wäre praktisch für mich wenn ich den Roboter zu Hause hätte.

Überhaupt nicht 1 2 3 4 5 sehr

Der Roboter könnte mir mit vielen Dingen helfen.

Überhaupt nicht 1 2 3 4 5 sehr

Sozialer Einfluss – (SI)

Ich denke meine Umgebung möchte dass ich den Roboter verwende.

Überhaupt nicht 1 2 3 4 5 sehr

Ich denke ich würde eine guten Eindruck machen wenn ich den Roboter verwende.

Überhaupt nicht 1 2 3 4 5 sehr

Soziale Präsenz – (SP)

Wenn ich mit dem Roboter interagiere fühlt es sich an, wie wenn ich mit einer richtigen Person spreche.

Überhaupt nicht 1 2 3 4 5 sehr

Manchmal hat es sich so angefühlt, als ob der Roboter mich ansehen würde.

Überhaupt nicht 1 2 3 4 5 sehr

Ich kann mir vorstellen, dass der Roboter ein lebendiges Wesen ist.

Überhaupt nicht 1 2 3 4 5 sehr

Ich denke oft, dass der Roboter keine echte Person ist.

Überhaupt nicht 1 2 3 4 5 sehr

Manchmal scheint der Roboter Gefühle zu haben.

Überhaupt nicht 1 2 3 4 5 sehr

Vertrauen – (Trust)

Ich vertraue dem Rat des Roboters

Überhaupt nicht 1 2 3 4 5 sehr

Ich würde dem Rat des Roboters folgen

Überhaupt nicht 1 2 3 4 5 sehr

HRI – Godspeed Questionnaire

Godspeed I: Anthropomorphismus / Vermenschlichung

Bitte stufen Sie den Roboter nach folgender Skala ein:

Unecht	1	2	3	4	5	Natürlich
Wie eine Maschine	1	2	3	4	5	Wie ein Mensch
Hat kein Bewusstsein	1	2	3	4	5	Hat ein Bewusstsein
Künstlich	1	2	3	4	5	Lebensnahe
Bewegt sich steif	1	2	3	4	5	Bewegt sich flüssig

Godspeed II: Belebtheit

Tot	1	2	3	4	5	Lebendig
Unbewegt	1	2	3	4	5	Lebendig
Mechanisch	1	2	3	4	5	Organisch
Künstlich	1	2	3	4	5	Lebensnahe
Träge	1	2	3	4	5	Interaktiv
Apatisch	1	2	3	4	5	Reagierend

Godspeed III: Liebenswürdigkeit / Sympathie

Gefällt nicht	1	2	3	4	5	Gefällt
Unfreundlich	1	2	3	4	5	Freundlich
Unhöflich	1	2	3	4	5	Höflich
Unangenehm	1	2	3	4	5	Angenehm
Furchtbar	1	2	3	4	5	Nett

Godspeed IV: wahrgenommene Intelligenz

Inkompetent	1	2	3	4	5	Kompetent
Ungebildet	1	2	3	4	5	Wissend
Unverantwortlich	1	2	3	4	5	Verantwortlich
Unintelligent	1	2	3	4	5	Intelligent
Unvernünftig	1	2	3	4	5	Vernünftig

Godspeed V: individuelles Sicherheitsgefühl

Bitte stufen Sie Ihren persönlichen emotionalen Zustand ein:

Ängstlich	1	2	3	4	5	Entspannt
Unruhig	1	2	3	4	5	Ruhig
Überrascht	1	2	3	4	5	Still

Zu den einzelnen Trainingssystemen

Anleitung auf Papier

Wie oft würden Sie die Trainingsanleitung auf Papier nützen, wenn Sie sie zu Hause hätten?

1	2	3	4	5
Nie	Ein Mal im Monat	1x pro Woche	3x pro Woche	Jeden Tag

Was sind aus ihrer Sicht die größten Kritikpunkte an dieser Trainingsform? (max 3)

Was sind aus ihrer Sicht die größten Stärken dieser Trainingsform?

Videogeführtes Training

Wie oft würden Sie das Videogeführte Training nützen, wenn Sie es zu Hause hätten?

1	2	3	4	5
Nie	Ein Mal im Monat	1x pro Woche	3x pro Woche	Jeden Tag

Was sind aus ihrer Sicht die größten Kritikpunkte an dieser Trainingsform? (max 3)

Was sind aus ihrer Sicht die größten Stärken dieser Trainingsform? (max 3)

Nintendo Wii Training

Wie oft würden Sie das Wii Training nützen, wenn Sie es zu Hause hätten?

1	2	3	4	5
Nie	Ein Mal im Monat	1x pro Woche	3x pro Woche	Jeden Tag

Was sind aus ihrer Sicht die größten Kritikpunkte an dieser Trainingsform? (max 3)

Was sind aus ihrer Sicht die größten Stärken dieser Trainingsform? (max 3)

Präferenz verschiedener Trainingsmöglichkeiten

1. Ich bevorzuge folgende Hilfe für mein tägliches Training zu Hause

Physicaal Roboter	Trainingsvideo	Anleitung auf Papier	Nintendo Wii Training	Keines
-------------------	----------------	-------------------------	--------------------------	--------

2. "Warum bevorzugen Sie dieses System?"

3. "Welches System könnte Sie am stärksten motivieren?"

Physicaal Roboter	Trainingsvideo	Anleitung auf Papier	Nintendo Wii Training	Keines
-------------------	----------------	-------------------------	--------------------------	--------

4. "Welches System stellt die Übungen am besten dar?"

Physicaal Roboter	Trainingsvideo	Anleitung auf Papier	Nintendo Wii Training	Keines
-------------------	----------------	-------------------------	--------------------------	--------

5. "Welches System hat ihrer Meinung die beste Trainingsbeurteilung?"

Physicaal Roboter	Trainingsvideo	Anleitung auf Papier	Nintendo Wii Training	Keines
-------------------	----------------	-------------------------	--------------------------	--------

6. "Wenn Sie die Möglichkeit hätten, eines der Systeme für 1 Monat zu Hause zu verwenden, für welches würden Sie sich entscheiden?"

Physicaal Roboter	Trainingsvideo	Anleitung auf Papier	Nintendo Wii Training	Keines
-------------------	----------------	-------------------------	--------------------------	--------

A6 Invitation to the final workshop within E3

Workshop

Eignung von sozial assistiver Robotik in AAL und zur Unterstützung von physiotherapeutischem Training

Schwechat 12.9.2013, 9:00-12:00

Ort & Zeit

Seniorenzentrum Schwechat, Altkettenhoferstrasse 5, 2320 Schwechat, Österreich

Datum: Do 12.Sept 2013, 9:00-12:00

Ziele

Im Rahmen des Workshops werden Erkenntnisse aus den Endnutzerstudien des FP7 Projektes „KSERA“⁴⁶ und des nationalen Benefit Projektes „PhysicAAL“⁴⁷ weitergegeben welche den humanoiden, sozial assistiven Roboter „Nao“ von Aldebaran robotics einsetzen.

Ziel des Workshops ist es, einen Wissensaustausch zum multidisziplinären Thema der assistiven Robotik für AAL mit ExpertInnen aus den Bereichen HRI, AAL, Pflege, Robotik und Physiotherapie zu ermöglichen.

Mit Hilfe der Impulsvorträge und Gruppenarbeiten wird eine Diskussion über den Einsatz von Mensch-Roboter Interaktion in AAL im Allgemeinen sowie im Bereich des physiotherapeutischen Trainings für SeniorInnen im Speziellen angeregt.

⁴⁶ <http://ksera.ieis.tue.nl>

⁴⁷ <http://physicaal.raltec.at>

Kontakt

Der Workshop wird vom Forschungsinstitut für Rehabilitation und assistive Technologien "CEIT RALTEC" (www.ceit.at), sowie der Fachhochschule Kärnten (www.fh-kaernten.at) durchgeführt.

Kontaktperson: DI Mag. Franz-Lothar Werner – f.werner@ceit.at

Kosten

Der Workshop wird vom Bundesministerium für Transport, Innovation und Technologie im Rahmen des FFG-Programmes „Benefit“ gefördert; die Teilnahme ist kostenlos.

Registrierung unter:

<http://physicaal.raltec.at>



Agenda

Teil I: Einleitung & Demo

- Vorstellung der Projekte KSERA & PhysicAAL
- Vorstellung und Live Demo des PhysicAAL Prototypen

Teil II: Präsentation der Ergebnisse der Endnutzerstudien

- Präsentation der Kernaussagen des Projektes PhysicAAL und Ergebnisse der Endnutzertrials zu den Themenbereichen
 - Akzeptanz von sozial assistiver Robotik für SeniorInnen
 - Motivationsfähigkeit eines robotischen Trainers
 - Technische Realisierung und Tauglichkeit der erarbeiteten Lösung für einen Real-Einsatz zu Hause.
 - Interpretation der Ergebnisse aus physiotherapeutischer Sicht
- Präsentation von Erfahrungen bei der Durchführung von Endnutzertrials mit assistiver Robotik

Teil III: Geführte Gruppendiskussion

- Gruppenarbeiten zu Themengebieten von HRI in AAL
 - Aspekte von HRI für verletzbare Nutzergruppen
 - Potentiale assistiver Robotik in der Pflege
 - Physiotherapeutische Einsetzbarkeit von assistiver Robotik
- Zusammenfassung der Gruppendiskussionen & Präsentation der Ergebnisse

Hintergrundinformation PhysicAAL

Physisches Training zu Hause ist eine häufige Therapie zur Prävention und Rehabilitation von physischen Defiziten bei älteren Menschen. Der Erfolg des verordneten physischen Trainings ist jedoch maßgeblich abhängig von der Motivation und Trainingskompetenz der Patienten und damit einer hohen Varianz unterworfen.



In der Studie physicAAL wird die Eignung humanoider sozial assistiver Roboter (SAR) zum Zwecke eines IKT gestützten physischen Trainings für SeniorInnen im häuslichen Bereich erforscht. Es sollen neue Erkenntnisse in Bezug auf die Optimierung von zu Hause durchgeführten Trainingsabläufen gewonnen werden, wobei besonders Strategien der Motivationssteigerung durch zielgerichtete Human Robot Interaction (HRI), evaluiert werden.

A7 Example of informed consent document used within E3

Information und Einverständnis zur Projektteilnahme

Im Rahmen des Projekts „**physicAAL**“ wird die Eignung eines sozial unterstützenden Roboters für ein Training älterer Menschen im Hausbereich erforscht.

Es sollen neue Erkenntnisse im Bezug auf die Optimierung der Mensch-Roboter-Interaktion an Hand von zu Hause durchgeführtem physischem Training, gewonnen werden. Denn eine regelmäßige, effiziente und eigenständig durchgeführte Trainingseinhaltung führt zu einer Steigerung des Trainingserfolgs.

Ziel des Projekts und der Benutzertests ist es zu erforschen, wie Interaktion und damit verbundene Motivationsstrategien mit einem sozial unterstützendem Roboter optimal für die Zielgruppe um- und eingesetzt werden kann. Es gilt zu erforschen, ob ein sozial unterstützender Roboter für das physische Training älterer Menschen geeignet ist.

Im Rahmen des Projektes wird ein bereits eingesetztes Prototypensystem (der menschenähnliche Roboter **NAO** und ein zur Bewegungserfassung geeignetes Kinect-System) verwendet.

Die Testsitzungen, die gemeinsam mit Ihnen durchgeführt werden, werden nicht länger als 120 Minuten dauern.

Alle im Rahmen des Projekts erhobenen Daten und gestellten Diagnosen werden Ihnen in vertraulicher Weise offen gelegt. Ausgebildete Experten werden das Projekt beaufsichtigen und Ihnen bei Fragen und Problemen zur Verfügung stehen.

Persönliche Daten werden anonymisiert und sind nur für Projektmitarbeiter zugänglich. Alle Projektmitarbeiter unterliegen der Schweigepflicht. Die Ergebnisse der Forschung werden in wissenschaftlichen Studien und Berichten veröffentlicht. Zu keiner Zeit ist ein Rückschluss auf Ihre Person anhand der Veröffentlichungen möglich.

Ihre Teilnahme am Projekt ist freiwillig und Ihnen steht jederzeit frei, Ihre Teilnahme ohne Angabe von Gründen und ohne Konsequenzen für Sie wieder zurück zu ziehen.

EINVERSTÄNDNISERKLÄRUNG

1. Ich, _____, stimme freiwillig zu, an den Aktivitäten des Projekts "physicAAL" teilzunehmen, wie sie oben, bzw. in den mir zuvor ausgehändigten Informationsunterlagen beschrieben sind.
2. Ich habe die Projektinformationen gelesen und verstanden.
3. Darüber hinaus ist mir bewusst, dass ein Abbruch des Tests meinerseits keine nachteiligen Folgen für mich hat.
4. Ich wurde darüber informiert und bin damit einverstanden, dass ich für meine Teilnahme keine finanzielle Entschädigung erhalte. Die Teilnahme an der Studie ist für mich mit keinerlei Kosten verbunden.
5. Ich wurde in einem persönlichen Gespräch über die Aufgaben, Risiken und Ablauf der Testdurchführung aufgeklärt und meine Fragen wurden für mich befriedigend und umfassend beantwortet.
6. Ich bin damit einverstanden, dass Teile des Tests fotografisch oder mit Videokamera festgehalten werden. Die dabei gewonnenen Daten werden nur zu wissenschaftlichen Zwecken verwendet. Eine Verwendung des Materials zu anderen Zwecken bedarf meiner vorherigen Zustimmung.
7. Ich weiß, dass ich mich bei Fragen oder anderen Anliegen, jederzeit an den für die Benutzereinbindung verantwortlichen Mitarbeiterin (Daniela Krainer, CEIT Raltec) wenden kann. Anschrift und Telefonnummer habe ich in Form eines Informationsbriefes erhalten.

Unterschrift des Teilnehmers

Datum

Name des Teilnehmers

Unterschrift des Versuchsleiters

Datum



DI Mag. Franz-Lothar Werner

Kagranerplatz 43, Vienna 1220, AUSTRIA
franz.werner@gmail.com • +43 (0) 6503399100 • <http://www.linkedin.com/in/franzwerner>

WORK EXPERIENCE

FH Campus Wien, University of Applied Sciences

Head of Degree Program

Jun 2016 – Present

Management of the interdisciplinary master-programme "Health Assisting Engineering" including the associated section for interdisciplinary applied research on assistive technologies.

- Management of the master-programme including personal matters.
- Acquisition and management of cooperative research projects.
- Applied research in the fields of health and care technologies, focusing on technologies to support therapy and care.
- Teaching technical and interdisciplinary subjects in the field of "Active and Assisted Living".

raltec - Researchgroup for assistive living technologies

Senior Researcher

Oct 2014 – Present

Main research areas: user-centered design, smart homes, assistive robotics

- User-centered design and evaluation of a solution that supports the work of care personnel within an institutional setting (project SignAAL).
- Conducting a commissioned meta-study on potentials and risks of robotics as assistive technologies (project potenziAAL).

FH Campus Wien, University of Applied Sciences

Lecturer

Sep 2014 – Jan 2016

Teaching within the course "Quality of Life and Assistive Technologies" of the master-programme "Health Assisting Engineering".

Central European Institute of Technology

Senior Researcher

Jun 2006 – Aug 2014

Interdisciplinary applied research in the field of "Active Assisted Living" and rehabilitation technologies.

- Participation in international and national cooperative research projects, partly in a coordinating role.
- Acquisition of nationally and EU funded cooperative research projects including proposal coordination and project development.
- Definition of research goals and pursuing research questions for assistive technologies.
- Developing prototypes in the field of assistive technologies in close collaboration with industry: defining functional requirements, conducting risk assessments and quality assurance, usability engineering, user involvement, coordinating and performing technical implementation and integration.
- Performing dissemination tasks, including the editing of project- and company related media and representing the company / projects at conferences and exhibitions.

Theobroma Systems

Software engineer

Mar 2008 – Oct 2008

Software engineer developing security relevant embedded devices for medical technology.

Vienna University of Technology

Tutor

Sep 2006 – Feb 2007

Tutor at the Institute for Information Systems at the Database and Artificial Intelligence Group (DBAI) for the course "Data modelling"

EDUCATION

Vienna University of Technology

Doctoral programme in Computer Science Oct 2012 – Present
Thesis: "Potential of Socially Assistive Robotics for an Application within the Field of Active and Assisted Living". Planned finalization in 2020.

Adviser: Professor Dr. Wolfgang Zagler

Master of Science (MSc.) "Medical Software Science" Oct 2002 – May 2007
Focus areas of the studies were "Bio-signal Analysis and Pattern Recognition", "Simulation and Biometric Studies", "Clinical Medicine" and "Health Management". Obtained the MSc. degree summa cum laude.

Master of Science (MSc.) "Information Management" Oct 2005 – May 2007
Focus areas of the studies were "Economy & Law", "Teaching Methodology" and "Process Engineering". Obtained the MSc. degree summa cum laude.

KTH Royal Institute of Technology, Stockholm, Sweden

Semester abroad studying "Medical Software Science" Aug 2004 – Jan 2005

HTBLA Donaustadt

Studying "Electrical Engineering" Sep 1996 – Jun 2001

SKILLS

Technical skills

- Hardware design for research prototypes (embedded microprocessor boards)
- Software design for research prototypes (application level and embedded software)
- Several programming languages including: Matlab, C, C++, Java, SQL, HTML

Other skills

- Project management for research projects
- Reviewer for the scientific journals "Gerontechnology", "The Gerontologist" and several conferences.
- Writing of scientific research proposals

LANGUAGES

German: native language

English: business fluent

Swedish: basic skills

INTERESTS

Sports - Co-leading the Viennese volleyball sports club "BeachUnion Wien" with around 70 players.

Media - Video editing, Picture editing

Travelling (Round the World 2007, Borneo, Taiwan, Japan)