

DISSERTATION
zur Erlangung des akademischen Grades
Doktor rerum naturalium
(Dr. rer. nat.)

An Interdisciplinary Concept for Human-Centered Explainable Artificial Intelligence

Investigating the Impact of Explainable AI on End-Users

Katharina Blandina Weitz, M.Sc.

eingereicht an der Fakultät für Angewandte Informatik
Lehrstuhl für Menschzentrierte Künstliche Intelligenz
Universität Augsburg

2023

Erstgutachterin: Prof. Dr. Elisabeth André
Zweitgutachter: Prof. Dr. Bernhard Bauer
Drittgutachterin: Prof. Dr. Ute Schmid

Tag der mündlichen Prüfung: 28. Juni 2023

Katharina Blandina Weitz: An Interdisciplinary Concept for Human-Centered Explainable Artificial Intelligence - Investigating the Impact of Explainable AI on End-Users

Colophon

This dissertation was typeset with the help of **KOMA-Script** and **L^AT_EX** using the **kaobook** class.

The source code of this book is available at:

<https://github.com/fmarotta/kaobook>

“Sometimes it seems that
The going is just too rough
And things go wrong
No matter what I do
Now and then I feel
Like life is just too much
You’ve got the love
I need to see me through”
The Source - You Got the Love

Für meine Eltern

Abstract

Since the 1950s, Artificial Intelligence (AI) applications have captivated people. However, this fascination has always been accompanied by disillusionment about the limitations of this technology. Today, machine learning methods such as Deep Neural Networks (DNN) are successfully used in various tasks. However, these methods also have limitations: Their complexity makes their decisions no longer comprehensible to humans - they are black-boxes. The research branch of Explainable AI (XAI) has addressed this problem by investigating how to make AI decisions comprehensible. This desire is not new. In the 1970s, developers of intrinsic explainable AI approaches, so-called white-boxes (e.g., rule-based systems), were dealing with AI explanations. Nowadays, with the increased use of AI systems in all areas of life, the design of comprehensible systems has become increasingly important. Developing such systems is part of Human-Centred AI (HCAI) research, which integrates human needs and abilities in the design of AI interfaces. For this, an understanding is needed of how humans perceive XAI and how AI explanations influence the interaction between humans and AI. One of the open questions concerns the investigation of XAI for end-users, i.e., people who have no expertise in AI but interact with such systems or are impacted by the system's decisions.

This dissertation investigates the impact of different levels of interactive XAI of white- and black-box AI systems on end-users perceptions. Based on an interdisciplinary concept presented in this work, it is examined how the content, type, and interface of explanations of DNN (black box) and rule-based systems (white box) are perceived by end-users. How XAI influences end-users mental models, trust, self-efficacy, cognitive workload, and emotional state regarding the AI system is the centre of the investigation. At the beginning of the dissertation, general concepts regarding AI, explanations, and psychological constructs of mental models, trust, self-efficacy, cognitive load, and emotions are introduced. Subsequently, related work regarding the design and investigation of XAI for users is presented. This serves as a basis for the concept of a Human-Centered Explainable AI (HC-XAI) presented in this dissertation, which combines an XAI design approach with user evaluations. The author pursues an interdisciplinary approach that integrates knowledge from the research areas of (X)AI, Human-Computer Interaction, and Psychology.

Based on this interdisciplinary concept, a five-step approach is derived and applied to illustrative surveys and experiments in the empirical part of this dissertation. To illustrate the first two steps, a persona approach for HC-XAI is presented, and based on that, a template for designing personas is provided. To illustrate the usage of the template, three surveys are presented that ask end-users about their attitudes and expectations towards AI and XAI. The personas generated from the survey data indicate that end-users often lack knowledge of XAI and that their perception of it depends on demographic and personality-related characteristics. Steps three to five deal with the design of XAI for concrete applications. For this, different levels of interactive XAI are presented and investigated in experiments with end-users. For this purpose, two rule-based systems (i.e., white-box) and four systems based on DNN (i.e., black-box) are used. These are applied for three purposes: Cooperation & collaboration, education, and medical decision support. Six user studies were conducted for this purpose, which differed in the interactivity of the XAI system used. The results show that end-users trust and mental models of AI depend strongly on the context of use and the design of the explanation itself. For example, explanations that a virtual agent mediates are shown to promote trust. The content and type of explanations are also perceived differently by users. The studies also show that end-users in different application contexts of XAI feel the desire for interactive explanations.

The dissertation concludes with a summary of the scientific contribution, points out limitations of the presented work, and gives an outlook on possible future research topics to integrate explanations into everyday AI systems and thus enable the comprehensible handling of AI for all people.

Zusammenfassung

Seit den 1950er Jahren haben Anwendungen der Künstlichen Intelligenz (KI) die Menschen in ihren Bann gezogen. Diese Faszination wurde jedoch stets von Ernüchterung über die Grenzen dieser Technologie begleitet. Heute werden Methoden des maschinellen Lernens wie Deep Neural Networks (DNN) erfolgreich für verschiedene Aufgaben eingesetzt. Doch auch diese Methoden haben ihre Grenzen: Durch ihre Komplexität sind ihre Entscheidungen für den Menschen nicht mehr nachvollziehbar - sie sind Black-Boxes. Der Forschungszweig der Erklärbaren KI (engl. XAI) hat sich diesem Problem angenommen und untersucht, wie man KI-Entscheidungen nachvollziehbar machen kann. Dieser Wunsch ist nicht neu. In den 1970er Jahren beschäftigten sich die Entwickler von intrinsisch erklärbaren KI-Ansätzen, so genannten White-Boxes (z. B. regelbasierte Systeme), mit KI-Erklärungen. Heutzutage, mit dem zunehmenden Einsatz von KI-Systemen in allen Lebensbereichen, wird die Gestaltung nachvollziehbarer Systeme immer wichtiger. Die Entwicklung solcher Systeme ist Teil der Menschzentrierten KI (engl. HCAI) Forschung, die menschliche Bedürfnisse und Fähigkeiten in die Gestaltung von KI-Schnittstellen integriert. Dafür ist ein Verständnis darüber erforderlich, wie Menschen XAI wahrnehmen und wie KI-Erklärungen die Interaktion zwischen Mensch und KI beeinflussen. Eine der offenen Fragen betrifft die Untersuchung von XAI für Endnutzer, d.h. Menschen, die keine Expertise in KI haben, aber mit solchen Systemen interagieren oder von deren Entscheidungen betroffen sind.

In dieser Dissertation wird untersucht, wie sich verschiedene Stufen interaktiver XAI von White- und Black-Box-KI-Systemen auf die Wahrnehmung der Endnutzer auswirken. Basierend auf einem interdisziplinären Konzept, das in dieser Arbeit vorgestellt wird, wird untersucht, wie der Inhalt, die Art und die Schnittstelle von Erklärungen von DNN (Black-Box) und regelbasierten Systemen (White-Box) von Endnutzern wahrgenommen werden. Wie XAI die mentalen Modelle, das Vertrauen, die Selbstwirksamkeit, die kognitive Belastung und den emotionalen Zustand der Endnutzer in Bezug auf das KI-System beeinflusst, steht im Mittelpunkt der Untersuchung. Zu Beginn der Arbeit werden allgemeine Konzepte zu KI, Erklärungen und psychologische Konstrukte von mentalen Modellen, Vertrauen, Selbstwirksamkeit, kognitiver Belastung und Emotionen vorgestellt. Anschließend werden verwandte Arbeiten bezüglich dem Design und der Untersuchung von XAI für Nutzer präsentiert. Diese dienen als Grundlage für das in dieser Dissertation vorgestellte Konzept einer Menschzentrierten Erklärbaren KI (engl. HC-XAI), das einen XAI-Designansatz mit Nutzerevaluationen kombiniert. Die Autorin verfolgt einen interdisziplinären Ansatz, der Wissen aus den Forschungsbereichen (X)AI, Mensch-Computer-Interaktion und Psychologie integriert.

Auf der Grundlage dieses interdisziplinären Konzepts wird ein fünfstufiger Ansatz abgeleitet und im empirischen Teil dieser Arbeit auf exemplarische Umfragen und Experimente und angewendet. Zur Veranschaulichung der ersten beiden Schritte wird ein Persona-Ansatz für HC-XAI vorgestellt und darauf aufbauend eine Vorlage für den Entwurf von Personas bereitgestellt. Um die Verwendung der Vorlage zu veranschaulichen, werden drei Umfragen präsentiert, in denen Endnutzer zu ihren Einstellungen und Erwartungen gegenüber KI und XAI befragt werden. Die aus den Umfragedaten generierten Personas zeigen, dass es den Endnutzern oft an Wissen über XAI mangelt und dass ihre Wahrnehmung dessen von demografischen und persönlichkeitsbezogenen Merkmalen abhängt. Die Schritte drei bis fünf befassen sich mit der Gestaltung von XAI für konkrete Anwendungen. Hierzu werden verschiedene Stufen interaktiver XAI vorgestellt und in Experimenten mit Endanwendern untersucht. Zu diesem Zweck werden zwei regelbasierte Systeme (White-Box) und vier auf DNN basierende Systeme (Black-Box) verwendet. Diese werden für drei Zwecke eingesetzt: Kooperation & Kollaboration, Bildung und medizinische Entscheidungsunterstützung. Hierzu wurden sechs Nutzerstudien durchgeführt, die sich in der Interaktivität des verwendeten XAI-Systems unterschieden. Die Ergebnisse zeigen, dass das Vertrauen und die mentalen Modelle der Endnutzer in KI stark vom Nutzungskontext und der Gestaltung der Erklärung selbst abhängen. Es hat sich beispielsweise gezeigt, dass Erklärungen, die von einem virtuellen Agenten vermittelt werden, das Vertrauen fördern. Auch der Inhalt und die Art der Erklärungen werden von den Nutzern unterschiedlich wahrgenommen. Die Studien zeigen zudem, dass Endnutzer in unterschiedlichen Anwendungskontexten von XAI den Wunsch nach interaktiven Erklärungen verspüren.

Die Dissertation schließt mit einer Zusammenfassung des wissenschaftlichen Beitrags, weist auf Grenzen der vorgestellten Arbeit hin und gibt einen Ausblick auf mögliche zukünftige Forschungsthemen, um Erklärungen in alltägliche KI-Systeme zu integrieren und damit den verständlichen Umgang mit KI für alle Menschen zu ermöglichen.

Acknowledgements

This endeavour would not have been possible without the support and encouragement of Prof. Dr. Elisabeth André. Since the first day in her lab, she has inspired me with her questions, comments, and ideas. She has given me the opportunity to put my research interests into action and broaden my perspective. I have grown personally and professionally through the enriching work at her chair over the past years. Special thanks to Prof. Dr. Bernhard Bauer for spending time reading the many pages and reviewing my dissertation. I'm incredibly grateful to Prof. Dr. Ute Schmid, who has paved the way for my academic career. Thank you for the support, conversations, and joint work over the years.

My research and this dissertation would not have been possible without the help and discussions with my colleagues at the Chair of Human-Centered Artificial Intelligence, especially Tobias Huber, Stina Klein, Silvan Mertes, Dominik Schiller, and Stanislava Thull. Special thanks to Björn Petrak, the best project and teaching colleague you could wish for. Working without coffee breaks is possible but less enjoyable. Thanks to Simon Flutura, Kathrin Janowski, Pooja Prajod, and Hannes Ritschel for the entertaining breaks in the daily research routine. Thanks to my research colleagues at national and international institutions who enabled me to see beyond the horizon of my university, especially Kasper Hald and Lindsey Vanderlyn. Many thanks to all the participants in my studies and the students I worked with.

I am also grateful to my friends. Special thanks to Véronique Huffer, with whom I went through stressful and relaxed periods as a doctoral student. Thanks to Babsi Backert, Simone Hoffmann, Sabrina Lauermann, and Jessi Streber, who ensured my free time was full of joy and laughter.

Lastly, I am incredibly grateful for my parents, who have always supported and believed in me. Thanks to Mama, you taught me to be patient and structured in everything I do. Thanks to Papa, who showed me that it is important to question things and be critical and reflective. Thanks to Dirk and Michael, you bravely put up with your big sister's quirks. At last, a heartfelt thank you, Jens. Your unconditional love has always been a compass for me in the ups and downs of research and everyday life.

I will never forget the words of welcome from Prof. Dr. Udo Krieger at the beginning of my computer science studies, who said: "Always do what your heart burns for!"

I will!

Katharina

Contents

Abstract	v
Zusammenfassung	vi
Acknowledgements	viii
Contents	ix
Abbreviations	xxiv

I. OVERVIEW AND MOTIVATION

1

1 Motivation	2
1.1 Introduction	2
1.2 Developing Human-Centered Explainable AI	3
2 Research Objectives	5
2.1 What Are the Challenges of Human-Centered Explainable AI?	5
2.2 How Can HC-XAI Be Evaluated?	6
2.3 Why Combining Different Research Disciplines?	7
2.4 Why Investigating Three Purposes for XAI?	8
2.5 Why Investigating End-Users?	10
2.6 Contribution & Research Questions	11
2.6.1 Contribution	11
2.6.2 Research Questions	12
3 Structure of the Thesis	13

II. BACKGROUND

15

4 Artificial Intelligence	16
4.1 Definition	16
4.2 Knowledge-Based White-Box Approaches	17
4.3 Data-Driven Black-Box Approaches	18
5 Explanations	22
5.1 Definitions	22
5.1.1 Definition of Explanation	22
5.1.2 Definition of XAI	23
5.1.3 Other Definitions	23
5.2 Function of Explanations	24
5.2.1 Beneficial Functions of Explanations	25
5.2.2 Harmful Functions of Explanations	26
5.3 Explainable AI for Different AI Systems	27
5.3.1 XAI in Rule-Based Systems	28
5.3.2 XAI Methods for DNN	29

6	Human States	36
6.1	Theory of Mind and Mental Models	37
6.2	Trust	39
6.3	Self-Efficacy & Cognitive Workload	41
6.4	Emotions	43
7	Summary Background	44
 III. RELATED WORK		45
8	Overview	46
9	Concepts for Designing Human-Centered XAI	47
9.1	Concepts About the Structure of XAI	47
9.2	Concepts to Develop Interactive XAI	49
10	XAI for Different Purposes & Scenarios	52
10.1	XAI for Cooperation & Collaboration	54
10.1.1	XAI for (Re)building Trust in Human-Agent Interaction	54
10.1.2	Interactive XAI Interfaces for Cooperation & Collaboration	56
10.1.3	Impact of XAI on Users in Cooperation & Collaboration	57
10.1.4	Delimitation & Contribution of this Dissertation	58
10.2	XAI for Education	58
10.2.1	Analogue Teaching Techniques for XAI	59
10.2.2	Gamification	60
10.2.3	Intelligent Tutoring Systems	60
10.2.4	Delimitation & Contribution of this Dissertation	61
10.3	XAI for Medical Decision Support	61
10.3.1	Context-based XAI for Decision Support Systems in Healthcare	62
10.3.2	Stakeholder Needs Regarding XAI for Decision Support Systems	62
10.3.3	Delimitation & Contribution of this Dissertation	64
11	Summary Related Work	66
 IV. INTERDISCIPLINARY CONCEPT FOR HUMAN-CENTERED EXPLAINABLE AI		69
12	Overview	70
13	Presentation of the Interdisciplinary HC-XAI Concept	72
13.1	Outline	72
13.2	Explanation Design	73
13.2.1	Explanation Interface	74
13.2.2	Explainable Model	76
13.3	User Evaluation	78
13.4	Five-Step Approach to Develop HC-XAI	79
14	Summary Interdisciplinary Concept	83

V. END-USERS OF HUMAN-CENTERED EXPLAINABLE AI	84
15 Exploring End-Users of Explainable AI	85
15.1 Personas	85
15.2 Persona Approach for HC-XAI	86
15.2.1 Persona Template	86
15.2.2 Data-Driven Approach for Creating Personas	87
15.3 Overview of the Surveys	88
15.4 Scope and Objective of the Surveys	90
16 End-Users of Explainable AI in Education	92
16.1 Research Questions	92
16.2 Methodology	93
16.3 Participants	93
16.4 Results	93
16.4.1 RQ1: Knowledge & Attitude	93
16.4.2 RQ2: Personal Attributes	94
16.5 Lessons Learned	96
16.6 Personas of End-Users in Education	97
17 End-Users of Explainable AI in Companies	99
17.1 Research Questions	99
17.2 Methodology	100
17.3 Participants & Companies	101
17.4 Results	101
17.4.1 Results of Employees' Company Perspective	101
17.4.2 Results of Employees' Personal Perspective	103
17.5 Lessons Learned	105
17.6 Conclusion	106
17.7 Personas of End-Users in Companies	107
18 End-Users of Explainable AI in Mobile-Health	108
18.1 Research Questions	109
18.2 Survey Design	109
18.3 Methodology	110
18.4 Participants	111
18.5 Results	112
18.5.1 RQ1: Preferred App & Explanations	112
18.5.2 RQ2: Impact of Users' Attributes	115
18.6 Clustering of Results to Create Personas	116
18.7 Lessons Learned	119
18.8 Personas of End-Users in Mobile-Health	120
19 Summary End-Users	123
VI. TECHNICAL REALISATION	124
20 Rule-based System Implementations	125
20.1 Overview	125
20.2 Rule-based Industry Robot in VR	126
20.2.1 Virtual Reality Design	126

20.2.2	Explanation Design	127
20.3	Rule-based Conversational Dialog Partner	128
20.3.1	Game Design	128
20.3.2	Explanation Design	130
21	Convolutional Neural Network Implementations	132
21.1	Overview	132
21.2	CNN for Keyword Classification	133
21.2.1	CNN Setup	133
21.2.2	Explanation Design	134
21.2.3	Demonstrator Setup for the Participatory Machine Learning Show	135
21.3	CNN for Emotion Expression Classification	137
21.3.1	CNN Setup	137
21.3.2	Explanation Design	137
21.4	CNN for Pneumonia Classification	138
21.4.1	CNN Setup	138
21.4.2	Explanation Design	139
22	Summary Technical Realisation	141
VII.	METHODOLOGY FOR HUMAN-CENTERED EXPLAINABLE AI RESEARCH	142
23	Experimental Designs	143
23.1	Overview	143
23.2	Experiments in the Laboratory	144
23.3	Experiments in the Wild	145
23.4	Online Experiments	146
24	Measurements	147
24.1	Mental Models	149
24.1.1	Retrospection Questions	149
24.1.2	Prediction Task	149
24.1.3	Task Reflection	150
24.1.4	Users' Model of Self	150
24.2	Explanation Satisfaction	150
24.3	Trust	151
24.3.1	Trust in Robots	151
24.3.2	Trust in Automation - Jian et al. (2000)	151
24.3.3	Trust in Automation - Körber (2018)	151
24.3.4	Trust in Explanation - Hoffman, Mueller, et al. (2018)	152
24.4	Computer Self-Efficacy	152
24.5	Cognitive Workload	152
24.6	Emotions	153
25	Data Analyses	154
25.1	Quantitative Analyses	154
25.1.1	ANOVA	155
25.1.2	T-test	155
25.1.3	MANOVA	156
25.1.4	Correlation	156

25.2	Qualitative Analyses	156
25.2.1	Conventional Content Analysis	157
25.2.2	Summative Content Analysis	157
26	Summary Methodology	158
VIII.	EMPIRICAL INVESTIGATION OF HUMAN-CENTERED EXPLAINABLE AI	159
27	Explainable AI for Cooperation & Collaboration	160
27.1	Overview	160
27.2	Explainable AI in Industrial Human-Robot Cooperation - VR-Robot Study	161
27.2.1	Highlights	161
27.2.2	Introduction	161
27.2.3	Pilot Study	162
27.2.4	Methodology	164
27.2.5	Main Experiment	167
27.2.6	Discussion	171
27.2.7	Conclusion	173
27.3	Explainable AI in Human-Agent Collaboration - Conversational AI Study	174
27.3.1	Highlights	174
27.3.2	Introduction	174
27.3.3	Hypotheses	175
27.3.4	Pilot Studies	175
27.3.5	Main Study	176
27.3.6	Dataset	180
27.3.7	Results	181
27.3.8	Discussion	186
27.3.9	Future Work	188
27.3.10	Conclusion	188
27.4	Summary Cooperation & Collaboration Experiments	189
28	Explainable AI for Education	190
28.1	Overview	190
28.2	Explainable AI in Human-Virtual Agent Interaction - Gloria Study	191
28.2.1	Highlights	191
28.2.2	Introduction	191
28.2.3	Research Questions	191
28.2.4	Study Design	192
28.2.5	Methodology	193
28.2.6	Results	194
28.2.7	Discussion	196
28.2.8	Conclusion	200
28.3	Explainable AI with Virtual Agents in the Wild - Museum Study	200
28.3.1	Highlights	200
28.3.2	Introduction	200
28.3.3	Research Questions	201
28.3.4	Study Design	202
28.3.5	Methodology	203
28.3.6	Participants	203
28.3.7	Results	204

28.3.8	Discussion	206
28.3.9	Conclusion	208
28.4	Summary Education Experiments	209
29	Explainable AI for Medical Decision Support	210
29.1	Overview	210
29.2	Explainable AI in Facial Emotion Classification - NOVA Study	211
29.2.1	Highlights	211
29.2.2	Introduction	211
29.2.3	Research Questions	212
29.2.4	Study Design	213
29.2.5	Methodology	213
29.2.6	Participants	214
29.2.7	Results	214
29.2.8	Discussion	221
29.2.9	Conclusion	223
29.3	Explainable AI in Pulmonary Disease Classification - Pneumonia Study	225
29.3.1	Highlights	225
29.3.2	Introduction	225
29.3.3	Hypotheses	226
29.3.4	Study Design	226
29.3.5	Methodology	227
29.3.6	Evaluation Methods	228
29.3.7	Participants	229
29.3.8	Results	230
29.3.9	Discussion	234
29.3.10	Conclusion	237
29.4	Summary Medical Decision Support Experiments	237
IX.	CONCLUSION	239
30	Contribution	240
30.1	Conceptual Contribution	240
30.2	Empirical Contribution	241
30.2.1	XAI Design Related Research Questions	241
30.2.2	End-User Related Research Questions	244
30.3	Technical Contribution	248
31	Limitations	250
32	Future Work	252
32.1	Interactive HC-XAI	252
32.2	HC-XAI for Application-Grounded Scenarios	253
32.3	Ethical Considerations	254
X.	APPENDIX	256
33	Publications & Contribution	257
33.1	Publications Relevant for This Dissertation	257
33.2	Other Publications	259

34 Activities, Teaching, Awards & Roles	260
34.1 Peer-Review Activities & Editorials	260
34.2 Teaching	260
34.2.1 Masters Theses	260
34.2.2 Bachelors Theses	260
34.2.3 Student Projects	261
34.2.4 Lectures	261
34.3 Awards & Roles	261
34.4 Invited Talks	262
35 Surveys	264
35.1 Survey XAI in Companies - Translated Version	264
35.1.1 Personal Information	264
35.1.2 Information About the Company and Your Work Area	264
35.1.3 AI Applications in Your Company - General Information	265
35.1.4 AI Applications in Your Company - Concrete Applications	266
35.1.5 AI and XAI Knowledge	267
35.1.6 End	268
35.2 Survey XAI in Education - Translated Version	269
35.2.1 Personal Information	269
35.2.2 AI and XAI Knowledge	269
35.3 Survey XAI in Mobile Health	271
35.3.1 Personal Information	271
35.3.2 Presentation of Three Explanation Types	276
35.3.3 AI & XAI Knowledge	280
36 User Strategies to Resolve Miscommunication	282
36.1 Agrees With System	282
36.2 Corrects System	282
36.3 Asks for Clarification	283
36.4 Solves Without Help	283
36.5 Restarts	283
36.6 Tries Something New	283
36.7 Rephrases	284
Bibliography	285

List of Figures

2.1	This dissertation focuses on the <i>human-grounded XAI evaluation</i> , as defined by Doshi-Velez and Kim (2017). For this, six experiments are conducted. In addition, <i>specific stakeholder</i> needs are investigated using three surveys. From this, prototypical users (i.e., personas) are derived . . .	6
2.2	Components relevant for the interdisciplinary HC-XAI concept of this dissertation: <i>Artificial Intelligence</i> : Describes the AI systems used for an application. Based on the AI system, an explanation is designed. This design is chosen in line with the work in the field of <i>XAI</i> . Investigating the impact of the explanations of an AI system on users is done using theoretical concepts as well as empirical methods of <i>Psychology and HCI</i>	8
3.1	Structure of the chapters of this dissertation. Starting from the left (Background) to the right .	14
4.1	Illustration of a rule used in the MYCIN system. The IF-THEN rule leads to an intrinsic explainability (i.e., explainable by design). Example from Shortliffe et al. (1975, p. 305)	18
5.1	Illustration of different terms in the field of XAI. <i>Explanation</i> refers to an AI system that provides information on a “Why?” or “How?” question. <i>Interpretation</i> is mainly used in the ML domain. An XAI system can be <i>transparent</i> on different levels. <i>Understanding</i> refers to the goal of XAI: users should understand the decision or inner workings of an XAI system	24
5.2	Application of three XAI algorithms to explain a CNN’s classification of the emotion “happiness” of visual input (left image): LRP (Bach et al., 2015), Grad-CAM (Selvaraju et al., 2017), and LIME (Ribeiro et al., 2016)	30
5.3	Steps of the LIME algorithm (Ribeiro et al., 2016): (1) An original picture is (2) segmented into superpixel. (3) The segmented images are perturbed by greying superpixels out. (4) The resulting images are weighted according to their proximity to the instance (blue cross) which should be explained. With this, a sparse linear model (dashed line) is learned	31
6.1	Empirical findings on the impact of XAI on users’ <i>mental model</i> , <i>trust</i> , <i>self-efficacy</i> , and <i>cognitive load</i>	36
9.1	Involvement of different research areas to create XAI. Figure adapted from Loyola-Gonzalez (2019)	48
10.1	Taxonomy to cluster evaluation of XAI methods (illustration adapted from Doshi-Velez and Kim, 2017). The three evaluation steps include the presence or absence of human users and the complexity of the task. With the increasing complexity of the evaluation, the costs and specificity of the investigation and the application increase	53
10.2	Related research fields for the development of XAI for education. Figure adapted from Khosravi et al. (2022)	59
12.1	Three different levels of interactive XAI are examined in this dissertation: The effect of specific XAI methods (e.g., LIME, LRP, counterfactuals) on end-users perception is investigated in two experiments (left). For education, a hybrid version by combining a virtual agent with an XAI algorithm is investigated (middle). The impact of XAI in human-AI interaction (right) is investigated in two experiments where end-users have to solve a cooperative or collaborative task. For every level, different purposes are investigated (coloured in blue, rosé, and orange). For every purpose, surveys to investigate personas and user studies to investigate the impact of XAI were conducted	71

13.1	The interdisciplinary concept for HC-XAI: An AI system (i.e., white-box, black-box, hybrid approach) provides explanations. The explainable model (i.e., content and type of an explanation) and the explanation interface (i.e., level of interactivity) impact users' perceptions and actions. Knowledge about user attributes helps used to build personas to support an HC-XAI design . .	73
13.2	Three levels of interactive XAI are examined in this dissertation: Starting with a low level of interactivity by presenting XAI methods (e.g., different algorithms to provide XAI visualisations) to end-users. The second is more interactive. Here, a virtual agent presents XAI visualisations generated with LIME based on user input. The last one communicates explanations to end-users using during a cooperative/collaborative task using text	74
13.3	Difference between cooperation and collaboration in a human-robot task. While a human and a robot work in the same area on different objects during cooperation, a human and a robot work on the same object in a collaborative task. Figure is adapted from Malik and Bilberg (2019) . .	76
13.4	Four elements are central to creating human-centered AI explanations. Illustration adapted from Vilone and Longo (2021)	79
13.5	Process of the development of an HC-XAI design based on the interdisciplinary concept proposed in this dissertation	81
13.6	Different content and types of XAI were addressed in the six conducted experiments. In addition, the experiments investigate different levels of interactive XAI	82
15.1	The persona template for HC-XAI combines the PATHY 2.0 approach of Ferreira et al. (2018) and the relevant personalization features from XAI as described by Schneider and Handali (2019) .	87
15.2	A data-driven approach is used to fill out the HC-XAI persona template (Weitz, Zellner, & André, 2022; Zellner, 2021). Here, collected data are investigated based on pre-defined research objectives. Based on this investigation, similar categories of information are clustered in the next step. After this initial clustering, additional information about people that were clustered into the categories is added. Finally, the information of each cluster is filled into the persona template	88
15.3	Response of all 200 survey participants. Almost all participants have heard about the term 'AI'	89
15.4	<i>Outer circle</i> : response of all 200 survey participants whether they had heard about XAI. <i>Inner circle</i> : division of participants by application scenario. It can be seen that, especially in the mobile health and education scenario survey, participants did not know the term XAI	90
16.1	Stand at the museum, which was used to ask museum visitors about their attitudes towards AI and XAI. Figure from Weitz, Schlagowski, and André (2021)	92
16.2	Museum visitors' answers to the question "What future do you think we will have with AI?" .	94
16.3	Rating of museum visitors, whether XAI is important for different stakeholders (1=disagree; 7=fully agree). Error bars represent the 95% CI	95
16.4	Rating of museum visitors, whether the find AI relevant for different application areas. Results show a critical look at the use of AI in education, art, and leisure	95
16.5	Persona <i>Regina</i> , who is educated about AI but finds this topic only necessary in specific application areas	98
16.6	Persona <i>Dirk</i> represents a student who has little knowledge about AI and XAI	98
17.1	Participants rated (scale from 1 to 7) how important XAI is for stakeholder groups. Participants perceived XAI as important for all stakeholders, especially politicians and companies. Error bars represent the 95% CI	103
17.2	Rating of the AI technologies used in companies by employees. Employees perceived the AI technology significantly positively, compared to the mean of the rating scale, $*p < .05$, $**p < .001$. Error bars represent the 95% CI	104
17.3	<i>Wolfgang</i> represents a prototypical person in a leading role in a company. He is aware of the chances of AI for his company. To use AI efficiently, he has to support his employees with appropriate (X)AI training programs	107

18.1	Three types of interactive explanations when classifying stress/no stress were investigated in the survey. Images illustrate the explanations for the data-based app: <i>Live explanation</i> that allows users to change parameters, <i>Feature explanation</i> that displays the features used for the decision, where the size of the circle reports the importance of the feature for the classification, and <i>Ask-the-app</i> explanations where users can ask the app-specific questions regarding the decision. The hand icon symbolizes a user action. Figure adapted from Zellner (2021)	109
18.2	Rating of the Health Survey participants regarding the importance of XAI. Participants stated that XAI is important for different stakeholders, especially for end-users and companies	112
18.3	Participants' rating (ranging between 1 to 7) of the three different explanation types (i.e., ask-the-app, live explanation, and feature explanation) shows that ask-the-app invites users to try it out. Furthermore, ask-the-app and live explanations invite users to personalize them (i.e., selecting features that should be relevant for the classification). Error bars represent the 95% CI. * $p < .05$, ** $p < .001$	114
18.4	We generated word clouds for the different clusters of prototypical users based on the free-form feedback to the question "Why do you prefer this app over the other?". The overall sentiment for persona 1 and 2 was positive, while the sentiment for persona 3 was negative	117
18.5	Word clouds regarding the three types of interactive explanations (i.e., feature-based explanation, ask-the-app explanation, and live explanation) for the three personas. Bigger words indicate often-mentioned terms	119
18.6	Three personas were derived based on the empirical data: <i>Anni</i> , a power user, <i>Karl</i> , a casual user, and <i>Michael</i> , a sceptical user	122
20.1	Overview of the design of the explanation content and explanation type for the two conducted experiments in the industry and production application scenario	125
20.2	From the participant's perspective, the virtual environment features the robot, bottles, their platforms and indicators, and the display text on the desk surface. Figure from Hald, Weitz, et al. (2021)	127
20.3	Game interface with four puzzles: The <i>dials module</i> (dials; far left and right), the <i>button sequence module</i> (row of five buttons; bottom left), the <i>switches module</i> (stack of four toggle switches; centre-left), and the <i>button array module</i> (3 x 4 array of buttons; centre). The shield level and timer display users' remaining time. The warp drive indicator and indicators above each module show number of completed puzzles. Figure from Weitz, Vanderlyn, et al. (2021)	129
20.4	Overview of the rule-based dialog system ADVISOR (Li et al., 2020). The dialog system (orange) consists of a Natural Language Understanding module (NLU), a Belief State Tracker (BST), a handcrafted Policy, and a template-based Natural Language Generation module (NLG), which generates the textual output for the user. The user (blue) reacts to the output of the AI dialog partner by responding via text	130
21.1	Overview of the design of the explanation content and explanation type for the four conducted experiments in the education and healthcare application scenarios	132
21.2	A spectrogram of an audio sample (left), its segmentation into superpixels (centre) and the output for the user containing LIME visualisations and additional phoneme information (right). Figure adapted from Weitz, Schiller, et al. (2021)	133
21.3	Schematics of the used speech recognition system. 1) A spectrogram is generated from the raw audio waveform. 2) The spectrogram is used to calculate 20 MFCCs. 3) The MFCCs are fed into a convolutional neural network. 4) The learned features are then forwarded to the fully connected layers of the network. 5) Finally, the network's output is mapped to the corresponding target class. Figure adapted from Weitz, Schiller, et al. (2021)	134

21.4	The baseline group of the Gloria Study was presented with XAI visualisations of a spoken keyword (e.g., 'four'). With every XAI visualisation, the predicted label and the prediction accuracy of the speech recognition system were displayed. In the three experimental groups, an XAI visualisation was presented with one of the virtual agent modalities (i.e., text, voice, or visual presence) . . .	135
21.5	Demonstrator overview as used in the field study. (1) <i>Demonstration PC</i> , running the neural network for keyword recognition and calculating the XAI visualisations. (2) <i>Beamer & PA system</i> were used for agent displaying & sound, the XAI visualisations, and the presentation slides for the show. (3) <i>Smartphone</i> and microphone for recording the audio samples. Figure adapted from Weitz, Schiller, et al. (2021)	136
21.6	An example X-ray image classified as <i>Pneumonia</i> , as well as the different XAI visualisations used in our study. Figure adapted from Mertes et al. (2022)	139
21.7	Examples of counterfactual images are produced with the approach proposed in Mertes et al. (2022). The left image shows the original image in each pair, while the right image shows the corresponding counterfactual explanation. The red boxes were added manually to point the reader to the most altered regions. The original images in the top row were classified as <i>normal</i> , while the original images in the bottom row were classified as <i>pneumonia</i> . The shown counterfactual images were all classified as the opposite of their respective counterpart. Figure from Mertes et al. (2022)	140
23.1	Illustration of the Gloria Study setup for participants in the virtual embodied agent group. The virtual agent Gloria was displayed on a separate monitor on the left side of the participants. The XAI visualisations were displayed on a monitor in front of the participants. The keywords were spoken into a microphone (not in the picture). Figure adapted from Weitz, Schiller, et al. (2021)	145
23.2	Beginning of a public participatory machine learning show visited by non-experts in the German museum in Munich. Figure from Weitz, Schlagowski, and André (2021)	146
25.1	Overview of the different analysis methods used in this dissertation. Analysis methods were chosen depending on the data (qualitative or quantitative)	155
27.1	Textual explanation modality. Two robot failures were explained during the pilot study: a <i>calibration failure</i> (top) and a <i>computer vision failure</i> (bottom). Figure from Hald, Weitz, et al. (2021)	163
27.2	Rating of the robot in the no-failure and the two failure conditions of the pilot study. The ratings for the no-failure condition were significantly higher than for the two failure conditions. failure bars represent the 95% CI. $**p < .001$	165
27.3	The average trust scores for the first and second bottle-sorting task between the three explanation conditions. While the difference in the trust rating between the correct and the incorrect task were significant, the three explanation groups did not differ	170
27.4	The four puzzle modules that end-users had to solve collaboratively in the Conversational AI Study	178
27.5	Results of the Godspeed Questionnaire (Bartneck et al., 2009) revealed that participants perceived their AI dialog partner as intelligent, likeable, and safe but not anthropomorphic. <i>Mean</i> refers to the mean value of 3. failure bars represent the 95% CI, $**p < .001$	182
27.6	Strategies to resolve miscommunications and their frequency. <i>Blue</i> represents user strategies, <i>purple</i> represents system strategies, and <i>rosé</i> represents events that were not resolved	185

27.7	Example of a miscommunication situation during the study. The user misunderstood that grey buttons were disabled and, therefore, not important (highlighted in red). This misconception starts the miscommunication. The system had two strategies to handle such situations (purple): To repeat the question/task or to provide information about itself with the request to the user to reformulate the problem/question. The user tried several strategies to resolve the miscommunication (blue). In the end, the user agrees with the system without solving the task correctly	185
28.1	Results of the evaluation of five different aspects of the virtual agent Gloria. The rating was scaled between 1=disagree to 7=fully agree. Error bars represent a 95% CI	195
28.2	No significant differences were found in the participants' ratings on whether the displayed XAI visualisations were sufficient. The rating was scaled between 1=disagree to 7=fully agree. Error bars represent the 95% CI	197
28.3	Procedure during the ML-show: An interactive audio-corpus was collected and used to train a Neural Network during a lecture about Machine Learning. Afterwards, the model was tested by the participants. The virtual agent, Gloria, presented the results and XAI visualisations. At the end of the ML-show, participants had to answer a questionnaire. Figure from Weitz, Schlagowski, and André (2021)	202
28.4	Mean TA-EG ratings by category for the ML-show participants and non-participating museum visitors. Subcategories <i>competence</i> and <i>positivity</i> indicate significant differences between the two groups ($*p < .001$). Error bars represent the 95% CI	206
29.1	Overview of the experimental conditions of the NOVA Study. (0) Participants of the <i>baseline condition</i> received only images with facial emotion expressions with a classification result (e.g., happiness). In addition to this information, (1) participants of the <i>confidence value condition</i> received confidence value for each image, (2) participants of the <i>XAI visualisation condition</i> received images with highlighted superpixels relevant for the classification, (3) participants of the <i>XAI & confidence condition</i> got all available information	213
29.2	Rating of the participants to what extent they are confident in their description of the behaviour of the neural network model (1 = disagree to 7 = fully agree). 0=Baseline condition; 1 = Confidence values condition; 2 = LIME condition; 3 = LIME and confidence values condition. No significant differences between the conditions were found. Error bars represent the 95% CI	217
29.3	XAI visualisation generated by LIME for two images classified as happy by a neural network model. While in the left image the network focused on the mouth region, in the right picture the background seems to have had an impact on the model's decision. Figure from Heimerl et al. (2022)	218
29.4	Comparison between the average areas of interest according to the study participants and model agnostic explanations generated with LIME. The different faces show varying emotions. A: anger, B: neutral, C: disgust, D: sadness, E: happiness. Figure from Heimerl et al. (2022)	220
29.5	Rating of the participants to what extent they are confident in their classification of the emotional pictures (1=unsure to 8=sure). Error bars represent the 95% CI	220
29.6	Simplified illustration of our prediction task used in the Pneumonia Study. Figure from Mertes et al. (2022)	227
29.7	There are significant differences in explanation satisfaction, trust, and prediction accuracy between the counterfactual and saliency map conditions (LRP and LIME). Error bars represent the 95% CI., $*p < .05$	231
29.8	We found significant differences regarding self-efficacy and general confidence of the participants in their predictions of the AI between the counterfactual condition and the saliency map conditions (LRP and LIME). Error bars represent the 95% CI. $*p < .05$	232

29.9	Confidence rating of the participants for correct and false predictions. The significant difference between the counterfactual and saliency map conditions is based on the confidence in correct predictions, not incorrect ones. Error bars represent the 95% CI	232
29.10	Results of the task reflection questions. Error bars represent the 95% CI. * $p < .05$, ** $p < .001$.	233
29.11	Results of the emotion questionnaires. Participants in the counterfactual condition felt significantly less angry and more relaxed compared to the LRP saliency map condition. For LIME, no significant differences were found. Error bars represent the 95% CI. * $p < .05$, ** $p < .001$. .	234

List of Tables

5.1	Dark pattern definitions from Gray et al. (2018, p. 5) with examples in the context of explanations from Chromik et al. (2019, p. 3)	27
5.2	Suggested taxonomy from Gilpin et al. (2018) for DNN XAI methods and methods to evaluate them. <i>Processing</i> and <i>representation</i> stand for a more technical focus, <i>explanation producing</i> takes the interpretation and evaluation of humans into account, but are more for ML-experts than for end-users (Gilpin et al., 2019)	29
11.1	Overview of related work that tests XAI in user studies. The term “End-users” refers to people with no experience in AI/ML or in the investigated domain. In all other cases, the specific expertise of the evaluated user group is stated. Only the study of Gerlings et al. (2021) distinguishes between different stakeholders, their roles and needs regarding XAI (i.e., personas)	67
11.2	Listing of related work about human-centered XAI designs that discuss requirements, challenges, and possible designs of XAI. The work is clustered after whether they are related to one of the three purposes (cooperation & collaboration, education, and medical decision support) that are the focus of this dissertation or provide general ideas regarding XAI	68
13.1	Question categories that XAI should answer. For every question type, examples are given. Table adapted from Liao et al. (2021, p. 6)	77
15.1	Overview of the sample sizes of the conducted (online) surveys about (X)AI, which were used as the foundation for persona development	89
16.1	No significant correlations were found between gender, educational background and knowledge about AI and XAI. Only the age of participants correlated positively with knowledge about AI .	96
17.1	Rating of AI technology used in companies on five items. A one-sample t-test revealed that all items were perceived as significantly positive by employees	104
18.1	Cronbach’s α for the five personality items of the TIPI questionnaire (Gosling et al., 2003). Agreeableness and conscientiousness are below 0.7 and therefore indicating not sufficient reliability .	112
18.2	Rating of potential questions users would ask themselves when seeing a mobile health app stress classification. A one-sample t-test revealed that all question types except Why not?-questions were perceived as significantly important by participants	113
18.3	Participants answer the question, “How much time would you spend understanding an explanation in the mobile health app?”	115
18.4	XAI requirements and their description. Requirements were elicited from the user survey. Table adapted from Zellner (2021)	118
21.1	Example phrases of the virtual agent for group A (classifier personification) and group B (third person commentary)	135
23.1	Overview of the type of experiment, study focus and sample size of the conducted studies . . .	143
24.1	Overview of the measured variables and the corresponding method used in the experiments. In addition, standard variables (e.g., demographic information and attitude towards (X)AI) were asked in all experiments. More information about these variables can be found in Chapter 15 on page 85	148

25.1 Overview of the quantitative and qualitative analyses of the experiments. In all studies, basic statistical analyses of demographical data were conducted, not presented here. <i>Analysis</i> refers to additional inferential statistical analyses	154
27.1 Independent sample t-test. No significant differences in explanation quality (trustworthiness & understandability), performance and likeability between the two different explanation modalities (textual vs auditory) for both types of robot failure. Trustworthy refers to “help to trust or distrust the robot”, Understandable refers to “helps to understand how the robot works”	166
27.2 Paired t-tests. Significant differences in performance and trustworthiness between the two different failure types (calibration vs computer vision failure)	167
27.3 In the one-sample t-tests, */**significant differences were found regarding game engagement (GEQ) and all used scales of the Godspeed questionnaire	181
27.4 Main and sub-categories resulted from content analysis. For every sub-category (highlighted in bold), an example of participants’ free-form feedback is given. Every example response is from a different participant	183
28.1 Demographic information of the participants, distributed to the four experimental groups . . .	193
28.2 Evaluation of the LIME Explanations. Answers from participants to the question which further explanations they would have found helpful. Table adapted from Weitz, Schiller, et al. (2021) . .	196
28.3 Descriptive information about the participants of group A (classifier personification) and group B (third person commentary)	204
29.1 Demographic information of the participants. 0=Baseline condition; 1=Confidence values condition; 2=LIME condition; 3=LIME and confidence values condition	215
29.2 Rating of participants, if the confidence values and LIME visualisations are helpful and easy to understand (Conditions: 1=confidence values; 2=LIME visualisations; 3=LIME visualisations and confidence values)	216
29.3 Explanations were given by the participants about the behaviour of the neural network. Sentences in italics refer to the network’s behaviour when classifying images correctly, non-italic statements to incorrect classifications. Table from Heimerl et al. (2022)	218

Abbreviations

The list describes several abbreviations that will be later used often within the body of the document.

AI	Artificial Intelligence
AR	Augmented Reality
CML	Cooperative Machine Learning
CNN	Convolutional Neural Network
DNN	Deep Neural Network
HC-XAI	Human-Centered Explainable Artificial Intelligence
HCAI	Human-Centered Artificial Intelligence
HCI	Human-Computer Interaction
HRI	Human-Robot Interaction
ITS	Intelligent Tutoring Systems
ML	Machine Learning
MTurk	Amazon Mechanical Turk
NOVA	NOnVerbal behavior Analyzer
TiA	Trust in Automation
VR	Virtual Reality
XAI	Explainable Artificial Intelligence

I. OVERVIEW AND MOTIVATION

1 Motivation

1.1 Introduction	2
1.2 Developing Human-Centered Explainable AI	3

1.1 Introduction

“Oh, the AI seems not to be so smart as I thought!” told an amazed museums visitor while he tested a Deep Neural Network (DNN) for image recognition that I presented in the German Museum in Munich on the final event of the Science Year 2019, which revolved around the topic of Artificial Intelligence (AI).

Already at the beginning of the AI area, people had high expectations of such systems. In 1956, scientists met at the Dartmouth College summer workshop to work on seven major issues of AI, a term that was used in the workshop for the first time to unify labels such as *complex information processing* and *automata studies* (McCorduck, 2004). The scientists McCarthy, Minsky, Rochester, and Shannon, were highly motivated when they wrote the proposal for this workshop:

“We think that a significant advance can be made in one or more of these problems if a carefully selected group of scientists work on it together for a summer.” (McCarthy et al., 2006, p. 1).

This initial euphoria was followed in the next decades by the fact that problems in AI were not so easy to solve as McCarthy et al. (2006) thought. The disappointment reached its lowest point in two AI winters (Kaynak, 2021). The first winter was heralded in the 1970s. It was marked by general disappointment with the progress of AI systems, considered for example in the book “What Computers Can’t Do” (Dreyfus, 1992) and the “Lighthill report” (Lighthill, 1973). The second AI winter took place in the late 1980s. Back then, AI was dubbed impractical, which resulted in a stop of funding and attention (Hendler, 2008). Moreover, the expert systems used could not meet the expectations placed on them - especially the bad performance of these systems in realistic application scenarios disappointed people.

Today, we overcame this second winter due to the successful use of DNN, which allowed AI to achieve satisfactory results in various application scenarios. LeCun et al. (2015) describe the start of this development with the usage of DNN in object recognition. In the meantime, DNNs have been successfully used in a range of applications¹. But this success story is not untroubled: DNNs lack comprehensibility², which was already well-known before their breakthrough. For example, Derek Partridge already stated in the 1990s:

“I should make it clear that very little has actually been modelled with these networks to date. And, in addition, there are already a number of difficult and unsolved problems [...]. The overriding concern, in my opinion, is one of comprehensibility. In order to build and use complex computer systems, we

1: application areas of DNN include facial recognition, emotion expression, speech understanding, and many more
2: DNN are therefore called *black-boxes*

must be able to understand, at some level, how they are doing what they are observed to be doing” (Partridge, 1991, p. 64).

The complexity (e.g., handling a lot of data, huge model architectures) that Partridge describes leads to the problem that the learning process of DNN and the resulting decisions are no longer comprehensible to humans (Adadi & Berrada, 2018). This concern is still present and even increased with DNN-based systems in our everyday lives. National and international legal sides claim the transparency of AI-based systems to protect users from intransparent AI decision systems. For example, the European Commission (2018b) state in Art. 12 of the General Data Protection Regulation (GDPR):

“The controller shall take appropriate measures to provide any information [...] and any communication [...] relating to processing to the data subject in a concise, transparent, intelligible and easily accessible form, using clear and plain language [...].”

To prevent AI from a third AI winter due to missing transparency and the resulting absence of comprehensibility, the research area of Explainable AI (XAI) addresses this issue. XAI has set itself the goal of enabling users to understand and appropriately trust an AI system (Gunning & Aha, 2019). Research on XAI in the 1970s focused on explainability in white-box AI systems. White-box AI is intrinsically explainable due to its interpretable structure and design (Molnar, 2019). An example of such a system is a rule-based system. These systems use if-then rules, a form familiar to users and can easily be interpreted by them³. In comparison, developing XAI methods for DNN is a very young research area. At the beginning of XAI research for DNN, it was necessary to develop technical solutions to establish post-hoc explainability (i.e., without changing the DNN structure) (Molnar, 2019). Since about 2015, various XAI methods have been developed that make it possible to look inside the black-box (Adadi and Berrada, 2018 provide a broad overview). Methods specific to DNN were described, for example, by Bach et al. (2015). Methods applicable for different machine learning approaches (i.e., model-agnostic) were developed by researchers like Ribeiro et al. (2016).

3: e.g., MYCIN from Shortliffe and Buchanan (1975) and Shortliffe et al. (1975)

1.2 Developing Human-Centered Explainable AI

“Ehm...what did the red area in the picture mean again?” a participant asked me as they looked at a picture of their spoken word highlighted using the XAI method LIME.

AI systems have a problem known as the *accuracy-explainability problem*. While the older rule-based and expert systems of the 80’s are explainable by design but mostly not very accurate, the newer DNN approaches are characterized by very high accuracies for many tasks but lose explainability (Selvaraju et al., 2017). Both white and black-box approaches face *technical challenges* and *human-related demands* regarding XAI. For example, the issue of comprehensibility is about developing a system that uses XAI (technical-related issue) and providing AI explanations so that users can understand the system (human-related issue).

This dissertation considers the human perspective, which is subsumed in the term *Human-Centered XAI* (HC-XAI). HC-XAI focuses on creating XAI design so that humans can benefit from it. The term HC-XAI is inspired by Human-Centered AI (HCAI). HCAI provides a perspective on AI that includes technical challenges as well as human attributes (Riedl, 2019). HC-XAI has to take into account human-related demands that rely on the needs and goals of different stakeholders. Therefore, more is needed than implementing XAI from a technical point of view. When doing so, the impact of XAI on users is misjudged. Users could create incorrect mental models about the AI system's functioning or inappropriate trust in these systems. In addition, XAI could be misused and provide harmful explanations to users (e.g., forcing users to perform a specific action) (Chromik et al., 2019). Current AI systems' actual capacities and limits also needs to be considered. As a result, the effect of users' perception of AI and XAI is related to the capabilities of such systems.

The two AI winters have already shown that more than the mere existence of an AI approach is required to use it successfully in real application scenarios. Instead, it is necessary to adapt the methods so that they are capable of satisfying real-world requirements. For the usage of white- and black-box AI, it is not enough to achieve satisfactory results in different tasks; they must also be able to explain how they reached those goals, i.e., being comprehensible for users, especially end-users. End-users refer to people that interact with AI systems in their work or private life while having no expert knowledge⁴ about the functioning of white- or black-box AI approaches.

This dissertation investigates the impact of explanations (i.e., explainable model and explanation interface) on end-users in the context of HC-XAI. The peculiarity of the dissertation is that it presents and examines different levels of interactive XAI. For these interactive XAI, mental models and trust are especially investigated. Furthermore, self-efficacy, cognitive load, and users' emotions are evaluated in some of the conducted studies. Many authors (e.g., de Visser et al., 2020; Hancock et al., 2011; Lee and See, 2004) highlight the importance of trust in human-machine interaction. Especially establishing *appropriate trust*⁵ in users is a common goal in Human-Computer-Interaction (HCI) research. But how to foster appropriate trust? This dissertation outlines the idea that helping people to create accurate mental models⁶ of an AI system could be the key. With accurate mental models, end-users develop proper expectations about the behaviour of the AI system, its benefits, and its limitations. To create accurate mental models, giving them insights about the reasons why a system came to its decisions could be helpful. To gain these insides, XAI is used.

To summarize the general idea: In this dissertation, the impact of different levels of XAI on end-users trust, mental models, trust, self-efficacy, cognitive load, and emotions is investigated. In doing so, the helpfulness of XAI to support accurate mental models and thus enables users to appropriate trust in the XAI system is explored.

4: expert knowledge refers to knowledge about a particular subject. For example, engineers, computer scientists, and mathematicians often have detailed knowledge about the AI domain

5: i.e., trusting an AI system when it is reliable and correct and distrusting it otherwise (Lee & See, 2004)

6: i.e., an accurate cognitive representation of an AI system

This dissertation uses an interdisciplinary research approach and investigates HC-XAI in different AI application scenarios. In the following, the rationale for this design is elucidated. Subsequently, the research questions that are addressed in the dissertation are presented.

2.1 What Are the Challenges of Human-Centered Explainable AI?

“I have never heard about XAI!” one of my study participants said while filling out the questionnaire after an experiment.

This answer was common during all of my conducted studies and surveys. While many end-users are not aware of XAI, the other way around, researchers have little knowledge about the impact of XAI on end-users. To develop HC-XAI, it needs more than just developing and implementing XAI algorithms as previously outlined. Instead, HC-XAI must be including the investigation of the usability of XAI methods for end-users. The design of HC-XAI is struggling with two major challenges:

► Neglect of End-Users

Problem: Miller et al. (2017) criticize that the development of XAI focuses too much on the needs of developers themselves and takes too little account of the requirements of other stakeholders, such as end-users.

Contribution: The dissertation focuses on user needs and presents a persona template for HC-XAI design. To illustrate the usage of the template, three surveys of various complexity are conducted to outline user needs regarding the usage of XAI in companies, education, and mobile health.

► Unknown Impact of XAI

Problem: XAI aims to foster various goals like transparency, fairness, and understanding (Laato et al., 2022). But the development of XAI methods does not equate to the transparency, fairness, or understanding of these methods by humans (Lipton, 2018; Molnar, 2019), meaning that a technical working XAI algorithm does not have to be the best algorithm from the end-user perspective. From psychology research, it is known that when an explanation is generated, this changes the reasoning process of humans (Lombrozo, 2006). Whether and in which way XAI changes the reasoning process about an AI application is unknown.

Contribution: The dissertation investigates the impact of different explanation styles and contents in six AI application-based experiments on end-user regarding trust, mental models, self-efficacy, cognitive workload, and emotions. In addition, this dissertation pays special attention to the presentation and investigation of different levels of interactive XAI systems.

2.1	What Are the Challenges of Human-Centered Explainable AI?	5
2.2	How Can HC-XAI Be Evaluated?	6
2.3	Why Combining Different Research Disciplines?	7
2.4	Why Investigating Three Purposes for XAI?	8
2.5	Why Investigating End-Users?	10
2.6	Contribution & Research Questions	11
2.6.1	Contribution	11
2.6.2	Research Questions	12

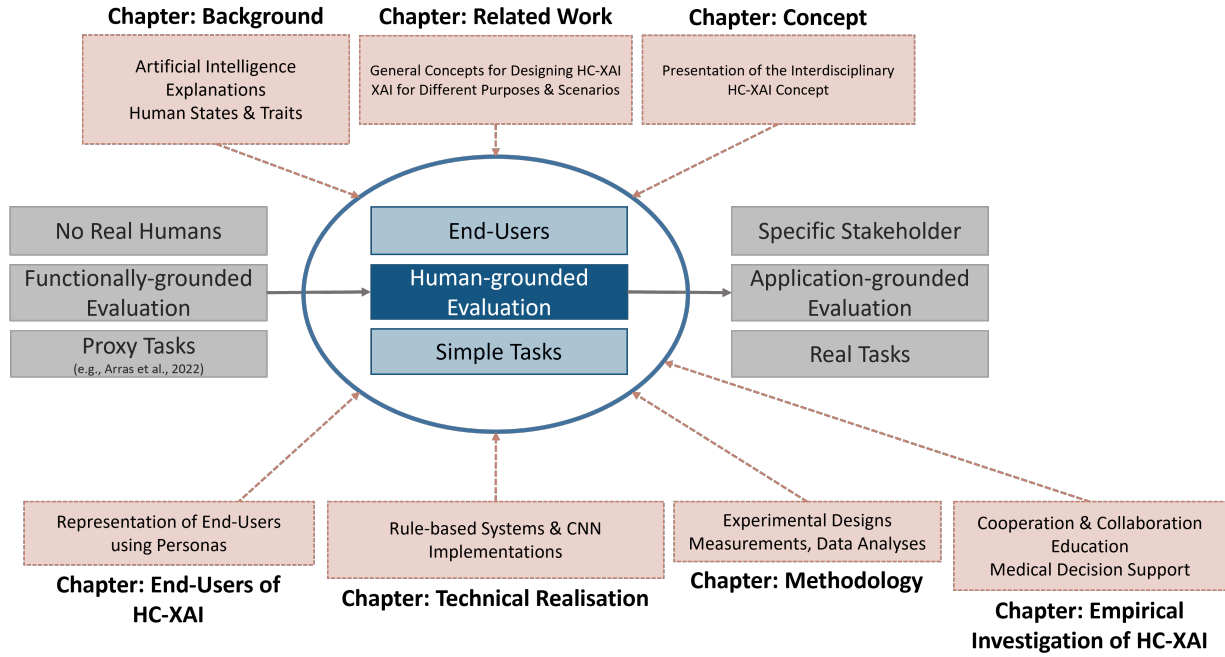


Figure 2.1: This dissertation focuses on the *human-grounded XAI evaluation*, as defined by Doshi-Velez and Kim (2017). For this, six experiments are conducted. In addition, *specific stakeholder* needs are investigated using three surveys. From this, prototypical users (i.e., personas) are derived

2.2 How Can HC-XAI Be Evaluated?

Doshi-Velez and Kim (2017) present an evaluation taxonomy divided into three steps to investigate XAI in a human-centered way. Each step in the taxonomy increases the investigation’s complexity and costs. Therefore, each step should be carried out carefully to build on the results of the previous steps for the next.

- ▶ **Step 1: Functionally-grounded Evaluation** In the beginning, experiments that require no humans are conducted to guarantee the proper working of the developed XAI algorithms. This work is done with a so-called proxy task. For example, Arras et al. (2022) and Tomsett et al. (2020) compared the fidelity of different XAI algorithms for DNN.
- ▶ **Step 2: Human-grounded Evaluation** In the next step, simple experiments with non-domain end-users (e.g., students, volunteering participants) are conducted. The goal of step 2 is to investigate the general impact of XAI algorithms on humans.
- ▶ **Step 3: Application-grounded Evaluation** Based on the results of step 2, step 3 focuses on the evaluation of a fully functioning XAI system in real scenarios (e.g., an explainable diagnosis AI tool in medicine) with domain experts (e.g., physicians).

The taxonomy of Doshi-Velez and Kim (2017) serves as a basis for this dissertation (see Figure 2.1). While technical evaluations (step 1) of different XAI algorithms are already available, user studies (steps 2 & 3), especially with end-users, are rare. Therefore, this dissertation investigates the impact of XAI on end-users in human-grounded evaluation (step 2). For

step 2, three purposes of AI (i.e., cooperation & collaboration, education, and medical decision support) are investigated. Six experiments with rule-based or DNN systems are designed, conducted, and analysed. For the six this experiments, the term end-user refers to laypeople with neither expert knowledge about AI nor the specific domain investigated. In addition, in three surveys for these application scenarios, end-user needs regarding XAI are explored and subsumed in prototypical end-users using the persona approach (see Figure 2.1 on the facing page). Overall, this dissertation presents data from a total of 683 end-users.

This dissertation provides a comprehensive look at end-user requirements and needs regarding XAI. For this, a persona template that can be used for HC-XAI design is presented. The usage of this template is illustrated with three user surveys of different complexity. In addition, the dissertation investigated different levels of interactive XAI in six experiments. The results provide an essential building block toward HC-XAI that consider both technical aspects and user needs.

2.3 Why Combining Different Research Disciplines?

“Build explanatory systems, not explanations.” is demanded by Mueller et al. (2021, p. 7).

This sentence lies in the assumption that explanations are more than only the sharing of information. This is in line with the statement of Miller (2019) saying that explanations are social. This means that explanations are more than just rational information exchange. Instead, explanations are part of a communication process that considers the counterpart’s needs, cognitive abilities, and expectations. Kim et al. (2021) present a theoretical framework that describes five different complexity levels of XAI:

- ▶ **Level 0:** No explanation is provided by the AI system
- ▶ **Level 1:** AI system presents one explanation type
- ▶ **Level 2:** AI system presents more than one explanation type
- ▶ **Level 3:** AI system considers user attributes (e.g., knowledge of the user)
- ▶ **Level 4:** AI system can create interactive explanations by communicating with the user

To reach level 4 (i.e., creating an interactive XAI), different explanation types and user attributes must be investigated to understand how XAI should be designed to be a “good and helpful explanation framework” for humans. This dissertation contributes to this goal by combining research of different disciplines: (1) AI, (2) XAI, and (3) Psychology & HCI (see Figure 2.2 on the next page). The interdisciplinary HC-XAI concept presented in this dissertation (for a detailed description, see Chapter 12 on page 70) provides demands on the design of the user interface (see *Explanation Design* in Figure 2.2 on the next page) but also take into account the technical capabilities of the used system (see *AI System* in Figure 2.2 on the following page) and the perception, action, and needs of

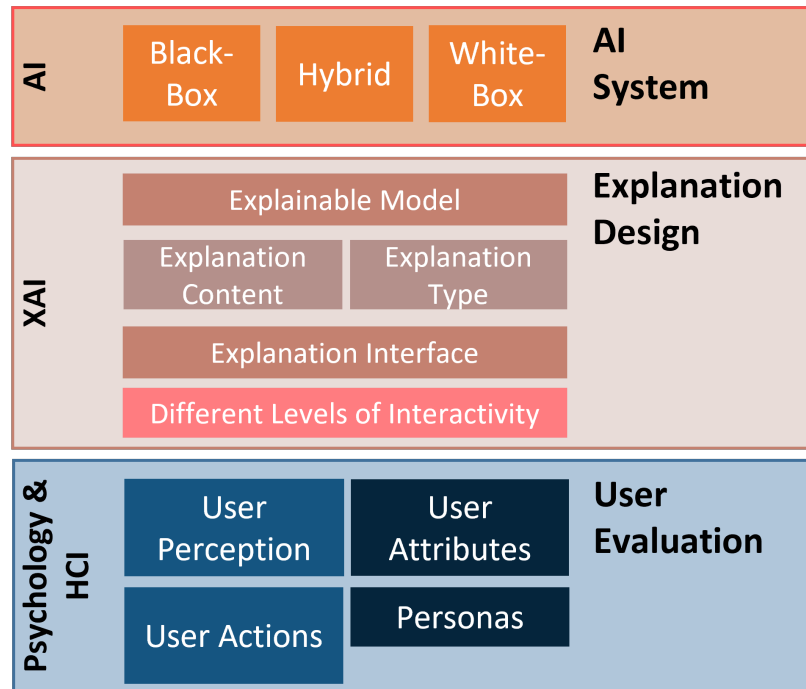


Figure 2.2: Components relevant for the interdisciplinary HC-XAI concept of this dissertation: *Artificial Intelligence*: Describes the AI systems used for an application. Based on the AI system, an explanation is designed. This design is chosen in line with the work in the field of *XAI*. Investigating the impact of the explanations of an AI system on users is done using theoretical concepts as well as empirical methods of *Psychology* and *HCI*

users (see *User Evaluation* in Figure 2.2). Therefore, the presented concept extends current XAI approaches by (1) providing an interdisciplinary perspective combining content from HCI, psychology and (X)AI research, (2) presenting a step-by-step approach based on the concept that serves researchers as a guide for a human-centered empirical investigation of XAI and (3) the application of the concept by using the step-by-step approach in the development of personas in three surveys as well as the investigation of different levels of interactive XAI in six experiments.

This dissertation creates more than isolated explanations but instead contributes towards HC-XAI that uses interactive explanation systems of different complexity. In doing so, it presents (1) a novel conceptual approach characterized by the integration of knowledge of interdisciplinary research areas and (2) a step-by-step approach based on the concept that serves researchers as a guide for the empirical investigation of XAI and (3) the exemplary application of the concept in the development of personas for three application contexts as well as the investigation of interactive XAI in six experiments.

2.4 Why Investigating Three Purposes for XAI?

Since AI is a broad topic, authors like Doshi-Velez and Kim (2017) and Wolf (2019) recommend using *scenarios* to investigate the impact of explanations on humans. Wolf (2019) present concrete but fictional scenarios to understand requirements regarding XAI. The description of possible scenarios should help developers and designers of such systems to think about the process (e.g., how to integrate explanations in an existing work

process) and the impact of content and style of an AI explanation (e.g., to decide which explanation supports uncertain users). This dissertation links these fictional scenarios to the application scenarios by investigating users' perceptions in the interaction with exemplary black-box and white-box AI systems. For this, XAI is investigated in scenarios for three purposes (i.e., cooperation & collaboration, education, and medical decision support) in six scenarios to examine the impact of XAI on end-users in different tasks. The scenarios represent typical situations where *users have little experience* in the field of (X)AI. In addition, the chosen scenarios are common in the XAI research community¹ as they fall into the *Reliable, Safe, and Trustworthy* (RST) design as described by Shneiderman (2020b). Shneiderman (2020b) describes applications characterized by different levels of human control and automation. A system is called *reliable, safe, and trustworthy* when it has high levels of human control and computer automation. An example in the field of a pain control device that meets these characteristics is a patient-guided and physician-monitored system (Shneiderman, 2020c). By taking into account patient needs (human control) and monitoring with the help of an AI system (computer automation), an RST system takes into account human needs and technical possibilities and thus enhances human well-being. The scenarios described in this dissertation have a similar claim: they are not intended to relinquish human control to the machine but to support humans in the decision-making process or in the activity to be carried out

1: for example, the use cases described in (1) Kraus et al. (2021) regarding XAI in healthcare and production & industry, and (2) Fiok et al. (2022) regarding XAI in education

For white- and black-box approaches, the effort of building such an HC-XAI is a branch of research that still contains many open questions regarding the design and impact of AI-generated explanations on humans. Especially the effect of XAI on humans' perception of the AI system needs to be adequately investigated. In addition, the impact of AI explanations in increasing dynamic human-machine interaction (e.g., in cooperation and collaboration tasks with agents or robots) has to be investigated to design interactive XAI that is valuable for users. To shed more light on this, six empirical studies in this dissertation examine the impact of XAI on end-users *mental models, trust, self-efficacy & cognitive load, and emotions*.

XAI for Cooperation & Collaboration To investigate the impact of XAI in cooperation and collaboration scenarios, two studies, including agents (i.e., a VR-robot arm and an AI dialog system), are used to examine the impact of verbal explanations in two industry-related scenarios. Both studies use rule-based approaches. The first study (i.e., **VR-Robot Study**) investigates the impact of robot failures in a virtual interaction setting, where a robot and end-users have to solve a sorting task. The second study (i.e., **Conversational AI Study**) uses a text-based AI dialog system in a cooperative puzzle game. End-users have to cooperate with this AI dialog system to solve the game.

XAI for Education Two studies with Convolutional Neural Networks (CNNs) are conducted, one in a laboratory (i.e., **Gloria Study**) and one in the wild - in a museum (i.e., **Museum Study**) to examine XAI for education purposes. The lab study investigates the impact of verbal explanations given by different representations of the virtual agent Gloria

(i.e., textual, voice, virtual presence) combined with XAI visualisations (i.e., LIME). The second study builds on the first and moves Gloria to the German Museum in Munich. During a participatory ML-show, she demonstrates the functionality of CNN and XAI to a broad audience of end-users in an edutainment-based setting.

2: for details about pain and emotion recognition via facial features, the work of Hassan et al., 2019 is recommended

XAI for Medical Decision Support Two studies using CNN in the field of healthcare applications were conducted to investigate the impact of XAI in decision support. The first study (i.e., **NOVA Study**) addressed facial emotion expression detection using a CNN. Automatic facial emotion detection supports clinical personnel or caring relatives to monitor patients' states². Especially when patients are not able to communicate verbally (e.g., babies, people with dementia), the investigation of the face helps infer the condition (e.g., whether the person is in pain) of the patient. The NOVA Study investigates end-users (i.e., laypeople) perception of three different XAI methods (confidence values, LIME visualisations, and a combination of both). The second study (i.e., **Pneumonia Study**) uses a CNN to predict whether X-ray images of lungs suffer from pneumonia. We investigate the impact of different XAI visualisations (i.e., LIME, LRP, counterfactuals) of these images on end-users perceptions.

For the investigation of the impact of XAI on end-users, three AI application purposes are investigated: cooperation & collaboration, education, and medical decision support. The unique feature of these experiments is that they use different levels of interactive XAI developed for the individual use case. All experiments presented in this dissertation are based on white-box or black-box AI systems, including their characteristics (e.g., differences due to the used architectures, accuracy) and limitations (e.g., mis- or nonunderstanding of participants during the interaction, misclassification).

2.5 Why Investigating End-Users?

End-users come into contact with AI systems in two ways: Either because they use them or because they are personally affected by the decision of an AI system (Laato et al., 2022). If AI systems have such a substantial impact on people's lives, they should, in a democratic understanding, be accessible to them. For example, the EU AI Act (European Commission, 2018a) addresses an AI strategy that ensures trustworthy AI that guarantees people's fundamental rights. The AI Act intends to empower individuals who derive the greatest possible benefit from AI systems.

In addition to these legal and societal imperatives to design HCAI for end-user, the fit of such systems to end-user is another challenge. Besides the fact that different AI scenarios demand different explanations, different types of end-users may demand different explanations. To integrate this in the design of explanatory AI systems, creating user models is necessary (Kass & Finin, 1988). In this dissertation, such user models are formed by developing *personas*. For this, a template for creating personas for HC-XAI is presented. An adaption of the PATHY 2.0 approach (Ferreira et al., 2018) combined with the suggestions for user-centered XAI design of Schneider

and Handali (2019) is used to design *personas* of XAI end-users. The usage of this template is illustrated by using three surveys investigating XAI in the context of companies, education, and mobile health. The surveys differ in their complexity and the research goal, indicating that the template can be used in different stages of the HC-XAI design process. The dissertation provides survey data from 200 participants.

This dissertation provides three survey-based descriptions of XAI end-users in the form of personas. The investigation of end-users includes different complexity levels. First, starting with a more general impression of XAI in an educational setting over a survey that addresses employees' experience of (X)AI in concrete applications in their companies, to the impression of end-users regarding a specific XAI use case for a mobile stress recognition app.

2.6 Contribution & Research Questions

In the following, the contribution of the theoretical part and the research questions for the empirical part of the thesis is addressed.

2.6.1 Contribution

Based on existing work in the field of HC-XAI, which is presented in the Related Work chapter (see Chapter 8 on page 46), the interdisciplinary HC-XAI concept of this dissertation is illustrated in Chapter 12 on page 70. In this concept, as already outlined, aspects from psychology (i.e., mental models, trust, self-efficacy, cognitive workload, and emotions), HCI (i.e., persona approach), AI (i.e., white- and black-box AI) and XAI (i.e., explainable model, explanation interface) are combined. The concept offers a generic perspective that can be applied to different AI systems. Furthermore, this dissertation aims to show that the presented concept can be filled with empirical life. The particular challenge of current HC-XAI is the design of interactive XAI systems. These are characterised by users not only receiving an explanation but can also react to it. In this dissertation, therefore, not only different types and contents (i.e. explainable model) of explanations are examined for their effect on end-user, but also the presentation for the user (i.e., explanation interface). The explanation interface is divided into different levels of interactivity, which are explained in more detail in Chapter 13 on page 72.

This dissertation contributes to scientific research by presenting an interdisciplinary HC-XAI concept based on previous research and applying it empirically to example applications. A five-step HC-XAI design approach based on the interdisciplinary concept is provided for this. In addition, a template for develop personas for HC-XAI investigations (see Chapter 15 on page 85 for details) is also noteworthy. Researchers can use the concept, the step-by-step approach, and the persona template as a guide for developing HC-XAI systems. To bring the concept into practical application, (1) three surveys investigate (X)AI attitudes of different end-user groups and develop personas from it and (2) six user studies investigate different levels of interactive HC-XAI are presented. While (1) is used as an illustration for the investigation of three different end-user groups

using the developed persona template, (2) presents an implementation of varying levels of interactive XAI systems (starting in Chapter 20 on page 125) for cooperation & collaboration (see Chapter 27 on page 160), education (see Chapter 28 on page 190), and medical decision support (see Chapter 29 on page 210). Before the six experiments, a detailed description of the study design, the methods used, and the evaluation is given so that future researchers can use these tools for preparing their own HC-XAI studies (starting with Chapter 23 on page 143).

2.6.2 Research Questions

To sum it up, for creating HC-XAI, the two main goals of the empirical part of this dissertation are (1) investigating the needs of end-users regarding XAI using surveys and (2) evaluating the impact of XAI in cooperation & collaboration, education, and medical decision making. Therefore, the research questions of this dissertation focus on two areas: *end-users* and *XAI design*.

End-Users

To investigate the needs and attitudes of end-users, the following research questions are investigated with the help of three surveys:

- ▶ **RQ-User-1:** What are end-users *knowledge, experiences, and attributes* towards AI and XAI? What do they expect from such systems?
- ▶ **RQ-User-2:** How are end-users demographic characteristics (e.g., age, educational background) related to the *knowledge, experiences, and attributes* toward AI and XAI?
- ▶ **RQ-User-3:** Which *personas* for human-centered XAI can be derived from empirical data about end-users? How do they *differ* regarding the application scenario?

XAI Design

The research questions related to the integration of XAI in three AI purposes (i.e., cooperation & collaboration, education, and medical decision support) that are investigated in six user studies explore the following:

- ▶ **RQ-XAI-1:** What are the *requirements and demands* for explanations in AI scenarios depending on the context of use?
- ▶ **RQ-XAI-2:** Which *aspects of an explanation* (i.e., type, content, interface) are helpful to end-users to *appropriate trust* and build *accurate mental models* about AI systems?
- ▶ **RQ-XAI-3:** How does XAI impact the *cognitive load, self-efficacy, and emotions* of end-users?

The dissertation is divided into eight parts (see the first row in Figure 3.1 on the next page). The contents covered in these parts are briefly explained in the following:

Background Here, basic concepts relevant to the dissertation are introduced. This includes the description of the topics *AI Systems* (see Chapter 4 on page 16), *Explanations* (see Chapter 5 on page 22), and *Human States* (see Chapter 6 on page 36). The section about AI systems gives an overview of AI in general, followed by the description of rule-based systems as a knowledge-based white-box approach and CNN as a data-driven black-box approach. These approaches are used in the experiments presented later in the dissertation. The section about explanations provides definitions for explanations and XAI and introduces the function of explanations for humans. In addition, XAI approaches for white- and black-box approaches will be described in detail. The human states section introduces five psychological concepts: mental models, trust, self-efficacy, cognitive load, and emotions. These are used for the empirical investigation of the interactive XAI in the six experiments of this dissertation.

Related Work These chapters give an overview of work in the field of explanations in HC-XAI. Here, *Concepts for Designing Human-Centered XAI* (see Chapter 9 on page 47) discusses work that indicates how explanations should be designed to fit human needs, expectations, and limitations. Next, in *XAI for Different Purposes & Scenarios* (see Chapter 10 on page 52), related work regarding XAI for cooperation & collaboration, education, and medical decision support is presented. For every section, the contribution of this dissertation and the delimitation regarding the related work are highlighted. Finally, the chapter closes with a summary of already done work and still open research gaps.

Interdisciplinary Concept for Human-Centered Explainable AI Based on the related work and the identified research gaps regarding the design of human-centered explanations, an interdisciplinary concept for HC-XAI is presented (see Chapter 12 on page 70). This concept connects the explanation design with a user evaluation and the specific requirements regarding the used AI system. It integrates work from Psychology, HCI, and (X)AI. After the individual components have been thoroughly introduced, based on the concept, a step-by-step approach for developing HC-XAI is presented that serves practitioners to design HC-XAI for their AI use cases.

End-Users of Human-Centered Explainable AI These chapters provide a template for developing personas in the context of HC-XAI (starting at Chapter 15 on page 85). The usage of this template is illustrated with

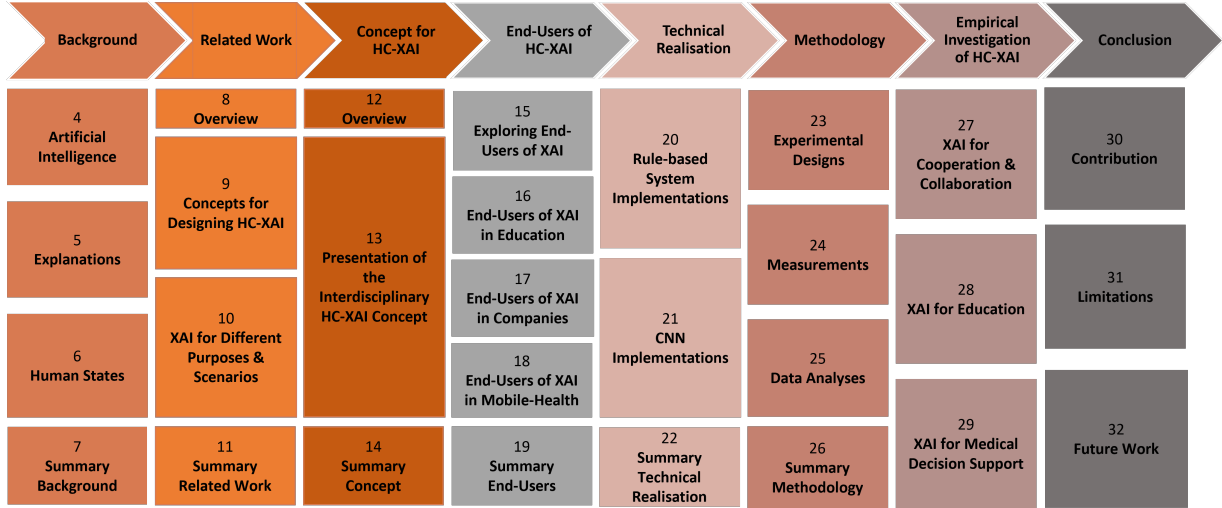


Figure 3.1: Structure of the chapters of this dissertation. Starting from the left (Background) to the right

three different surveys of various complexity and goals of the investigation. In doing so, attitudes and needs regarding AI explanations of 200 end-users are collected. Based on qualitative and quantitative analyses of the data, *Representations of End-Users* in the form of personas are presented.

Technical Realisation Since white- or black-box AI systems are used in all six conducted experiments, these chapters describe the technical setups used in more detail. In addition, the implementation of the used level of interactive XAI in the experiments is presented. The description is divided according to the type of AI system used: *Rule-based Systems* (see Chapter 20 on page 125) and *CNN* (see Chapter 21 on page 132).

Methodology for Human-Centered Explainable AI Research This part (starting in Chapter 23 on page 143) gives an overview of the *Experimental Designs*, *Measurements*, and *Data Analyses* used in the experiments of this dissertation. Furthermore, it highlights the scope and the limitations of the methods used.

Empirical Investigation of Human-Centered Explainable AI Here, six empirical studies in the field of *Cooperation & Collaboration* (see Chapter 27 on page 160), *Education* (see Chapter 28 on page 190), and *Medical Decision Support* (see Chapter 29 on page 210) with a total of 483 participants are presented. The report of the studies includes hypotheses, study design, evaluation methods, results, and a discussion of these.

Conclusion The dissertation ends with a conclusion of the theoretical and empirical findings in the field of HC-XAI. It highlights the *Contribution* (see Chapter 30 on page 240) of this dissertation, its *Limitations* (see Chapter 31 on page 250), and *Future Work* (see Chapter 32 on page 252) that has to be done to improve XAI in a way that humans benefit from it.

II. BACKGROUND

4

Artificial Intelligence

4.1 Definition	16
4.2 Knowledge-Based White-Box Approaches	17
4.3 Data-Driven Black-Box Approaches	18

In this dissertation, experiments with XAI in knowledge-based white-box and data-driven black-box systems are conducted. In the empirical part of this dissertation, rule-based systems are used as an example of a knowledge-based white-box approach, while CNNs are used for data-driven black-box approaches. The design and structure of CNN and rule-based systems entail specific capabilities and limitations. These, in turn, influence the generation of explanations. This chapter gives an introduction to such AI systems. It starts with a general definition of the term *Artificial Intelligence (AI)* and *Machine Learning (ML)*. Then, the concepts of rule-based systems and CNN are presented.

Parts of this chapter (highlighted in *italic*) are based on the work published in:

- *Description of CNN:*
Weitz, K. (2018). *Applying Explainable Artificial Intelligence for Deep Learning Networks to Decode Facial Expressions of Pain and Emotions* (Master's thesis), University of Bamberg.
- *Description of rule-based systems:*
 - Schmid, U., & **Weitz, K.** (2022). Künstliche Intelligenz und Psychologie - Von Kognitiver Modellierung bis Erklärbarkeit. In A. Schütz, M. Brand, & S. Steins-Löber (Eds.), *Psychologie: eine Einführung in ihre Grundlagen und Anwendungsfelder*, (pp. 219-231). Kohlhammer.
 - **Weitz, K.** (in publishing process). Erklärbare Künstliche Intelligenz im Kontext Bildung und Lernen. In C. deWitt, C. Gloerfeld, & S. Wrede (Eds.), *Künstliche Intelligenz in der Bildung*. Springer. doi: [10.1007/978-3-658-40079-8](https://doi.org/10.1007/978-3-658-40079-8)

4.1 Definition

McCarthy et al. (2006, p. 12) defined AI in their proposal about the summer research project in 1955 “[...] that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it”. This view of AI supports the branch of weak AI, indicating that AI is only able to *simulate thinking*, compared to strong AI, which assumes that AI *is thinking* (Russell & Norvig, 2016).

One of the subfields of AI covered in this dissertation are *rule-based systems*¹, another is *ML*. Rule-based systems mirror human expertise in a specific domain (e.g., diagnosis of diseases) using human-generated rules. Conversely, ML is a term to describe algorithms “[...]that improves its performance at some task through experience.” (Mitchell, 1997, p. 2). Here, rules or patterns are learned directly from the data instead of being pre-defined by humans. DNN, as a subfield of ML “allows computational

1: also referred to as *expert systems* and *knowledge-based systems*

models that are composed of multiple processing layers to learn representations of data with multiple levels of abstraction.”(LeCun et al., 2015, p. 436). A specific Neural Network is CNN.

4.2 Knowledge-Based White-Box Approaches

The term *white-box* for some AI approaches highlights the ante-hoc explainability of these. In the early days of AI², rule-based systems were seen as the key to developing intelligent systems. Rule-based systems are characterised by the modelling of human knowledge with the help of IF-THEN rules (Hayes-Roth, 1985). Rule-based systems have two essential components: “a knowledge base and an inference engine”(Hayes-Roth, 1985, p. 923). The knowledge base contains the rules and facts. Rules are expressions that specify actions. Facts represent true propositions (Hayes-Roth, 1985). To illustrate the idea of the definition of rules and facts, I present a simple example using the programming language PROLOG. In the following, I use a modified PROLOG example that Clocksin and Mellish (2003) presented in their book. The fact that *Jens likes coffee* is represented as:

2: in the 1960s and 1970s (Russell & Norvig, 2016)

```
1 | likes(jens, coffee).
```

While we could now list every caffeine-based drink that Jens likes as facts, we could instead define a rule that says *Jens likes drinks that include caffeine*:

```
1 | likes(jens, X):- includes(X, caffeine).
```

This code says that Jens likes *x* when *x* includes caffeine. Therefore, we can test for every drink, whether Jens likes it, without writing it as a list of facts. With PROLOG, it is possible to assess the knowledge base by asking questions (Clocksin & Mellish, 2003). For example, we could ask: “Is there an object that likes coffee?” and the PROLOG program would answer:

```
1 | ?-likes(X, coffee).
2 | X = jens;
3 | no
```

The answer includes the name of the object (i.e., Jens) that likes coffee. Since we do not have more objects stored in our facts, PROLOG prints no after that. If we had more people in our facts who like coffee, another name would appear here every time :enter is typed. Besides rules and facts, a rule-based system also includes the inference engine. The inference engine denotes the processing of the incoming input with the help of the rules and facts in the knowledge base and the output generated from it.

One of the first (and famous) rule-based systems was MYCIN³, which was developed in the 1970s (Shortliffe & Buchanan, 1975; Shortliffe et al., 1975). It was designed to assist physicians in making diagnoses and treatment recommendations. For this purpose, an extensive knowledge base about bacterial diseases, their symptoms, and the characteristics of the bacteria was collected. This database was the source for hundreds of IF-THEN rules which indicated a disease with a particular bacterium. For this, MYCIN required the answers to 45 questions and approximately 15-20 minutes to come to a decision and a corresponding explanation as well

3: implemented in the programming language LISP

as a therapy recommendation (Shortliffe et al., 1975) (see Figure 4.1 for an example rule of MYCIN).

Figure 4.1: Illustration of a rule used in the MYCIN system. The IF-THEN rule leads to an intrinsic explainability (i.e., explainable by design). Example from Shortliffe et al. (1975, p. 305)

```
IF
1) the stain of the organism is gramneg,
   and
2) the morphology of the organism is rod,
   and
3) the aerobicity of the organism is anaerobic
THEN
there is suggestive evidence (.6) that the
identity of the organism is bacteroid
```

4: this problem is called the *knowledge acquisition bottleneck* (Hanney & Keane, 1997)

However, these expert systems were not able to establish themselves in the long term (Brock, 2018), because (1) the collection of the required knowledge base was very time-consuming, (2) when the knowledge base was incomplete or incorrect, the expert systems were useless (Fogel et al., 1993), and (3) not every knowledge could be described explicitly (e.g., perception)⁴. Nevertheless, rule-based systems have, due to their explainability by design (i.e., intrinsic explainability) (Molnar, 2019), the advantage of setting up a system that behaves consistently for each user (P. R. Cohen, 2020). Currently, rule-based systems are being rediscovered to combine their strengths with ML to create hybrid approaches (Schmid et al., 2021).

For cooperation & collaboration experiments described later in this dissertation, rule-based systems are used. The usage of rule-based systems allowed us to fast prototyping, meaning that we could develop a functioning AI system for our studies that end-users could interact with in real-time. A detailed description of the two systems used for our studies can be found in Chapter 20 on page 125.

4.3 Data-Driven Black-Box Approaches

ML approaches are called *black-boxes* because of their lack of comprehensibility due to their complexity of calculations and the amount of data used. These approaches are data-driven (Molnar, 2019), meaning that they derive rules or patterns not from a knowledge basis but directly from data. While there are several approaches in ML, we focus on the following on CNN as an example of data-driven black-box approaches.

The architecture of CNN is inspired by the mechanisms of the visual perception of living beings (Gu et al., 2018). CNN was described for the first time in the late 1980s by LeCun et al. (1989). The idea of CNN is based on visual pattern recognition. Hubel and Wiesel (1959) discovered that neurons in a cat's cortex are specified to different stimuli. In their work three years later, they found the "most effective stimulus shapes are long narrow rectangles of light ('slits'), dark bars against a light background ('dark bars'), and straight-line borders separating areas of different brightness ('edges')" (Hubel & Wiesel, 1962, p. 559). In the 1980s, Fukushima (1988) described *neocognitron*, a neural network for visual pattern recognition that was inspired by the work of Hubel and Wiesel (1962). Fukushima (1988) described a hierarchical-ordered neural network separated into layers. While the first layers recognize only simple

patterns, the higher layers can integrate the simple patterns into more complicated ones. Fukushima (1988) also presented a use-case for their network: the recognition of handwritten digits.

Based on this work, LeCun, Bottou, Bengio, and Haffner (1998) introduced in the 1990s the idea of CNN. A CNN consists of convolutional layers, pooling layers, and fully-connected layers (LeCun et al., 2015). In contrast to classical (deep) neural networks, CNN uses convolutions (Goodfellow et al., 2016). The convolutional operation is achieved by using a filter, which scans over a given image. The through matrix multiplication calculated values are written into a feature map (LeCun et al., 2015). Convolutional layers can be seen as feature extractors (Lin et al., 2014), which makes it possible to have an end-to-end system that detects the features automatically and trains a classifier using these features. The resulting output of a convolutional layer is passed through an activation function (LeCun et al., 2015). Nowadays, the rectified linear unit (ReLU) activation function is applied chiefly (Goodfellow et al., 2016; Jarrett et al., 2009). Krizhevsky et al. (2017) describe this activation function as a non-linear, non-saturating function in the form of

$$f(x) = \max(0, x), \quad (4.1)$$

where $f(x)$ returns zero when $x < 0$ and $f(x)$ returns x when $x \geq 0$. DNNs using ReLUs have significantly shorter training times than saturating non-linearities (Krizhevsky et al., 2017). Another activation function is the softmax activation. It is often used in the last layer of a CNN (Goodfellow et al., 2016) to reflect the probability distribution of n classes. The softmax activation is formulated as

$$\text{softmax}(z)_i = \frac{\exp(z_i)}{\sum_j \exp(z_j)}, \quad (4.2)$$

where z is a vector of the inputs to the last layer (output layer) and i indexes the inputs of the vector z . A pooling layer follows a convolution layer. Pooling layers are used in CNNs to reduce the dimensionality and, therefore, the number of parameters in the network. This leads to shorter training time and helps to reduce overfitting. Overfitting can be seen as “memorizing the training cases” (Dreiseitl & Ohno-Machado, 2002, p. 254). One of the methods used is max pooling (Zhou & Chellappa, 1988). Here, the max values of different regions of the feature map are extracted and written into a max pooling map. Extracting the max values is done by a filter that does not overlap regions. To guarantee a non-overlapping filter of size $z \times z$, a stride defined as $s = z$ is used (Krizhevsky et al., 2017). For example, to scan the feature map with a filter size of 2×2 , a stride of 2 is necessary to ensure no overlapping regions. In a classical (deep) feedforward neural network, all layers are fully connected, which means that each neuron of the previous layer is connected to the following layer. In contrast, only a CNN’s last layers are fully connected. These layers are used to calculate the class score in a classification task.

Besides the architecture of a CNN, some relevant techniques are essential to make the network learn. According to Goodfellow et al. (2016), four things are essential to building a deep learning algorithm *DL*: specification of a dataset d , a cost function *cost_func*, an optimization procedure *opt*, and a network model m (e.g., CNN). In a semi-formalized description, it

can be said:

$$DL(x) = d(x) + m(x) + cost_func(m(x)) + opt(cost_func(m(x))), \quad (4.3)$$

5: sometimes referred to as ‘error function’ or ‘loss function’

where the cost function is applied to the model, and the optimisation minimises the cost function. A cost function⁵ is a function that quantifies the difference between the expected and actual outputs of a CNN. One cost function used in deep learning for multi-class classification tasks is cross-entropy, formulated as

$$H(P, Q) = -\mathbb{E}_{x \sim P} \log Q(x) = - \sum_x P(x) \log Q(x) \quad (4.4)$$

where P stands for the true distribution and Q stands for the distribution predicted by the model. The expectation of $f(x)$ with respect to $P(x)$ is denoted as $\mathbb{E}_{x \sim P}$. To minimise the output of the cost function, optimisation is needed. To optimise the layer weights during the training phase of a CNN and, therefore, to reduce the output of the cost function, backpropagation is used (Simonyan et al., 2014). The idea of backpropagation is not specific to CNNs and can also be used for calculating other functions (Goodfellow et al., 2016). The choice of the backpropagation method depends on the used cost function and the used network model (LeCun, Bottou, Orr, & Müller, 1998). Backpropagation is used to calculate a gradient (Goodfellow et al., 2016). The gradient represents the rate at which the costs C change with respect to weights and biases. The backpropagation algorithm is often misunderstood because backpropagation does not represent the entire learning algorithm for the CNN. Instead, the gradient-based optimisation method uses the gradient calculated by backpropagation for learning (Goodfellow et al., 2016). Therefore, the gradient represents the direction of the steepest change $\nabla C(x, y)$. The gradient is needed to apply a gradient-based optimisation function to update the weights. Optimisation methods can generally be categorised based on whether fixed or adaptive learning rates are used. As an optimisation method with fixed learning rates, stochastic gradient-descent (SGD) is nowadays often used (Goodfellow et al., 2016). A common approach using an adaptive learning rate is the adaptive moment estimation (Adam), introduced by Kingma and Ba (2015).

Besides using pooling layers to reduce the danger of overfitting, regularization techniques are used to prevent the network model from adapting itself over to the training set and performing poorly on unseen data. Goodfellow et al. (2016, p. 117) defined regularization as

“any modification we make to a learning algorithm that is intended to reduce its generalization error but not its training error.”

One of the practically used regularization techniques is dropout (Srivastava et al., 2014). The idea here is to randomly drop out neurons in a network, i.e., to remove the neuron with all its incoming and outgoing connections.

6: VGG is the research group’s name: Visual Geometry Group

A widely leveraged CNN architecture is VGG16⁶ (Simonyan & Zisserman, 2015). This architecture is used for the facial emotion recognition of the NOVA Study. For the Pneumonia Study, we implemented a CNN to detect pneumonia (i.e., binary task: pneumonia/no pneumonia) from X-ray images. In the two user studies of the education scenario, two

CNNs for speech recognition, based on the work of Sainath and Parada (2015), are used. A detailed description of all implemented CNNs can be found in Chapter 21 on page 132.

5 Explanations

5.1 Definitions	22
5.1.1 Definition of Explanation	22
5.1.2 Definition of XAI	23
5.1.3 Other Definitions	23
5.2 Function of Explanations	24
5.2.1 Beneficial Functions of Explanations	25
5.2.2 Harmful Functions of Explanations	26
5.3 Explainable AI for Different AI Systems	27
5.3.1 XAI in Rule-Based Systems	28
5.3.2 XAI Methods for DNN	29

To create XAI in a human-centered way, these systems need to be able to explain their decisions and inner workings so that humans can understand them. The following chapter will give a deeper look into the groundwork in the field of explanation research. After a definition of the terms *explanation* and *XAI*, psychological research will be presented to understand the impact of explanations on humans. After that, the focus shifts to the base work of XAI research in Computer Science.

Parts of this chapter are based on the following work:

- *Verbal explanations for rule-based systems:*
Weitz, K. (2021). Vertrauen und Vertrauenswürdigkeit bei sozialen Robotern - Stärkung von Mensch-Roboter-Vertrauensbeziehungen mithilfe Erklärbarer Künstlicher Intelligenz. In O. Bendel (Ed.), *Soziale Roboter* (pp. 309-323). Springer Gabler, Wiesbaden.
- *Description of the LRP and LIME algorithms:*
Weitz, K. (2018). *Applying Explainable Artificial Intelligence for Deep Learning Networks to Decode Facial Expressions of Pain and Emotions* (Master's thesis), University of Bamberg.

5.1 Definitions

While researchers largely agree that explanations are essential for humans to learn and understand their environment and the world, a consistent definition of XAI is missing. When reading papers on this topic, many different terms are used. For example, Vilone and Longo (2021) lists 36 terms related to XAI. Besides the term *XAI*, the terms *explanation* and *interpretation* are often used interchangeably (Adadi & Berrada, 2018). In addition, terms like *understanding* and *transparency* are used. In the following, these terms are defined, including the views of different researchers.

5.1.1 Definition of Explanation

The psychologists Siegler et al. (2002, p. 37) define (self-) explanations as “inferences about causal connections among objects and events.[...] In short, they are inferences concerning ‘how’ and ‘why’ events happen.” From a social scientist’s perspective, Miller (2019) sees an explanation as a cognitive and social process and a product. The cognitive process of an explanation means the process of “identifying the causes of a particular phenomenon” (Miller, 2019, p. 11). In social science, this identification is called attribution¹ (Miller, 2019). The result of this cognitive process is then the product - an explanation. The social process describes the transfer of knowledge between the explainer and explainee (Miller, 2019). In Miller’s work, the social aspect of explanations is important to consider when designing explanations for AI systems (Miller, 2019). Tomsett et al. (2018, p. 9) define explanation similar to (Lipton, 2018; Miller, 2019) as

1: attribution theory investigates the “perception of causality” (Weiner, 1972, p. 203)

“the information provided by a system to outline the cause and reason for a decision or output for a performed task.”

5.1.2 Definition of XAI

The term XAI was first mentioned in 2004 by van Lent et al. (2004) to describe the ability of their system to explain the behaviour of AI-controlled entities in a training system for the U.S. army. Adadi and Berrada (2018) highlight that the term XAI describes movements and endeavours more than a technical concept. Therefore, XAI can be seen as a research direction. The goal of XAI is to help users “to understand, appropriately trust, and effectively manage [...] artificially intelligent partners” (Gunning & Aha, 2019, p. 44). Gunning et al. (2019) describes in more detail the idea of XAI research: They see the motivation of XAI in providing comprehensible explanations to humans. They also name three helpful principles to create explanations in a more human-centered way: (1) explanation of AI capabilities and its understandings, (2) explanation of past, actual, and future behaviour, and (3) explanation of the information that is relevant for the action. Miller et al. (2017) highlight that XAI was already essential for expert systems developed in the 80s and 90s. Miller et al. (2017, p. 2) also focuses on the methods XAI uses by saying that XAI provides “[...] methods for automatically generating explanations of some type.”

5.1.3 Other Definitions

Interpretation & Interpretability

Interpretable AI or Interpretable ML is a term preferred by ML researchers (Adadi & Berrada, 2018). Here it refers “[...] to the capability of understanding the work logic in ML algorithms.” (Adadi & Berrada, 2018, p. 52141). Lipton (2018) criticizes that interpretability in the domain of interpretable AI research is ill-defined and leads to quasi-scientific research. To solve this problem, various authors have defined interpretability. Gilpin et al. (2018, p. 80) see interpretability as “the science of comprehending what a model did”. A similar view has Biran and Cotton (2017, p. 8) when calling systems interpretable, which they refer to as “if their operations can be understood by a human, either through introspection or through a produced explanation”. Choo and Liu (2018) gets more specific and refer to the interpretability of deep neural networks as the ability to identify features of the input (layer) that are relevant for classification in the output (layer). Montavon et al. (2018, p. 2) have a slightly different view as they see an interpretation as the “mapping of an abstract concept (e.g., a predicted class) into a domain that the human can make sense of.”

In the area of DNN, interpretability refers to *post-hoc interpretability* (Lipton, 2018; Montavon et al., 2018). Due to the interchangeability of the terms, some authors refer to *post-hoc explainability* as well (Adadi & Berrada, 2018). Except for the wording, the definition of post-hoc explanation is the same: it describes approaches used to shed light into black-box models (Adadi & Berrada, 2018) after a training (Molnar, 2019). Post-hoc interpretability is the counterpart of intrinsic explainable systems. These systems are, by definition, interpretable by themselves, meaning that they

are less complex (e.g., small decision trees) compared to systems that use post-hoc interpretability (Molnar, 2019). Gilpin et al. (2018) recommend differentiating between the terms interpretability and explainability because interpretability is just the first step to achieving explainability.

Transparency

Barredo Arrieta et al. (2020, p. 83) describe based on the work of Lipton (2018) transparency of AI systems as “the search for a direct understanding of the mechanism by which a model works.” Therefore, transparency is the counterpart of opacity (Lipton, 2018). Lipton (2018) differentiate between three levels of transparency: (1) transparency of the whole ML model, (2) transparency of specific components (e.g., parameters), and (3) transparency of the learning algorithm.

Understanding

Besides interpretation and explanation, Montavon et al. (2018) describes that a less algorithmic view of XAI is reflected in the term *understanding*. Understanding refers to a shallow description of a model without explaining the internal structure or algorithm (Montavon et al., 2018). Anjomshoae et al. (2019) refer to understanding in the context of robots and agents as *explainable agents*. Explainable agents explain their behaviour. In their work, they highlight the importance of not only using XAI to explain the decisions and inner workings from a technical point of view but using a *goal-driven XAI* by explaining the behaviour of robots and agents to users. Anjomshoae et al. (2019) conclude that this goal-driven XAI is more helpful for end-users to build an accurate mental model and, therefore, improve human-agent cooperation and collaboration. Gunning and Aha (2019) see understanding on the users’ side. With the explanation of an XAI system, users are (in the best case) able to understand the system’s pros and cons and know how the system will behave in the future.

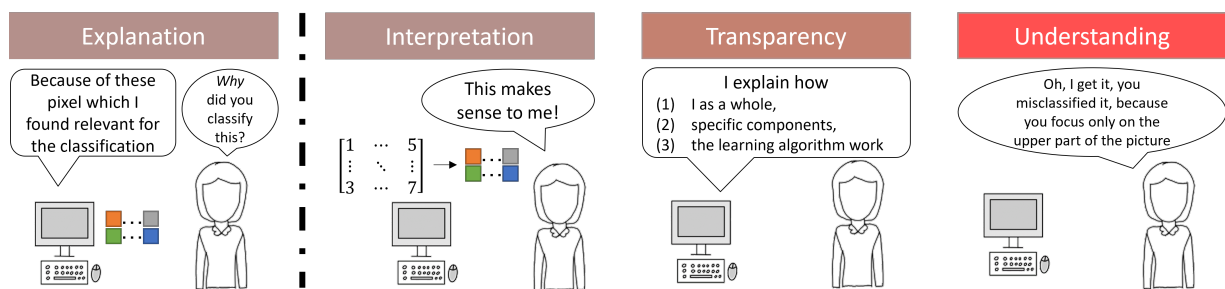


Figure 5.1: Illustration of different terms in the field of XAI. *Explanation* refers to an AI system that provides information on a “Why?” or “How?” question. *Interpretation* is mainly used in the ML domain. An XAI system can be *transparent* on different levels. *Understanding* refers to the goal of XAI: users should understand the decision or inner workings of an XAI system

5.2 Function of Explanations

An explanation answers a “Why?” or a “How?” question (Miller et al., 2017; Siegler et al., 2002; Wellman, 2011). Humans have a natural need

to acquire and provide explanations. For example, already children ask “why?” and try to gain explanations for phenomena they observe (Keil, 2006). There are various reasons why people ask for explanations (Miller, 2019). Lombrozo (2006) highlights that explanations help to infer a causal inference, generalize properties, and learn from examples. Learning from examples means that people want to understand or learn about something with the help of an explanation (Lombrozo, 2006). Here, explanations can be generated by the learner (i.e., self-explanations) or another person. Amsterlaw and Wellman (2006) investigated in their study with children between 3 and 4 years how performance in False Beliefs Tasks² could be improved with the help of self-generated explanations. For this purpose, they compared three groups of children (i.e., control, self-explanation, and comparison condition³): the examination of their explanations led the children to initiate developmental changes with their reflections, which enabled them to improve their performance and thus to solve False Belief tasks better. In addition, it was found that children who generated explanations also performed more successfully on transfer tasks than children who were not prompted to do so. These findings of self-explanations hold for third-person explanations; even these are not as effective as self-explanations.

These everyday experiences of explanations (Lombrozo, 2012), which are relevant for human-human interaction, are also vital in human-machine interaction. For example, users could ask in a medical context: “Why did the AI predict this disease?”⁴ or during a human-robot interaction: “Why did the robot failed in this task?”⁵. Therefore, in AI explanation research and XAI research, the explainer can also be a machine instead of a human (Molnar, 2019).

2: in False Beliefs Tasks, participants have to predict actions or thoughts of a person whose beliefs are incorrect (Amsterlaw & Wellman, 2006)

3: *self-explanation*: 24 false-belief problems over 12 sessions; *Comparison*: 24 false-belief problems, but fewer sessions and only half of the sessions with self-explanations; *Control*: Only pre- and post-tests

4: we investigated this question in our Pneumonia Study, detailed described in Chapter 29 on page 210

5: we investigated the impact of robot errors in our VR-Robot Study, as described in Chapter 27 on page 160

5.2.1 Beneficial Functions of Explanations

The study described earlier by Amsterlaw and Wellman (2006) and many other researchers highlight the importance and helpfulness of explanations for learning. From broad literature research about explanations in the social sciences, Miller (2019) concluded four attributes of explanations that are relevant to have in mind when designing XAI:

- ▶ **Explanations are contrastive:** People not only want to know why an event (e.g., a classification decision) occurs, but they want to know why this event occurs instead of another (possible) event.
- ▶ **Explanations are selected:** People do not expect all possible causes to be part of an explanation. Instead, 1-2 causes are sufficient for them as reasons. Cognitive biases influence the selection of these.
- ▶ **Probabilities probably don’t matter:** Even though AI research likes to work with probabilities, explaining probabilities or statistical correlations is not as catchy for people as describing the causes of an event.
- ▶ **Explanations are social:** Explanations are part of human communication and represent knowledge transfer from person to person.

Adadi and Berrada (2018) present four reasons why explanations for AI are needed:

6: for example, the Google image classifier which incorrectly classified People of Color as “Gorillas”: <https://www.bbc.com/news/technology-33347866> (last accessed on 26.07.2022)

- **Explain to justify:** Over the years, biased and therefore unfair AI systems were revealed and led to discussions among a broad public⁶. Therefore, explanations aim to present reasons for justifying an AI system’s decision to decide whether the decision is correct and unbiased.
- **Explain to control:** To control the outcome of an AI system is strongly connected to the explanations that help to justify a machine’s decision. Explanations help identify errors and help developers of such systems detect unknown vulnerabilities.
- **Explain to improve:** To detect errors in AI systems with the help of explanations helps in the next step to improve such systems by eliminating the errors.
- **Explain to discover:** Explanations can help gain new knowledge and insights.

Another function of AI explanations is the support of AI literacy and data literacy. *AI literacy* is described by Long and Magerko (2020, p. 2) “[...] as a set of competencies that enables individuals to critically evaluate AI technologies; communicate and collaborate effectively with AI; and use AI as a tool online, at home, and in the workplace.” *Data literacy* can be defined “[...] as the ability to understand and use data effectively to inform decisions.” (Mandinach & Gummer, 2013, p. 30). Especially in ML, where a huge amount of data are collected and processed, AI literacy is necessary to handle the data (e.g., by understanding statistical methods) and know about the problems and how to manage them. For example, the Google gorilla failure mentioned before was based on a training dataset that did not include images of People of Color. Therefore, the ML system tried to map the people into a similar (in terms of colour) but, of course, inappropriate and incorrect class.

5.2.2 Harmful Functions of Explanations

Chromik et al. (2019) describes yet another negative, *dark* function of explanations from which people other than the actual users benefit. In this context, Chromik et al. (2019) distinguishes five dark patterns that can be divided into two groups: (1) *phrasing of the explanation* and (2) *integration of the explanation into the user interface*. The five dark patterns they describe concerning explanations were inspired by the work on dark patterns of user experience design by Gray et al. (2018). The definitions of dark patterns presented below are based on Gray et al. (2018, p. 5), the related examples of explanations by Chromik et al. (2019, p. 3) (see Table 5.1 on the next page)

Table 5.1: Dark pattern definitions from Gray et al. (2018, p. 5) with examples in the context of explanations from Chromik et al. (2019, p. 3)

Dark Pattern	Definition	Transfer to the context of explanation
Nagging	Redirection of expected functionality that persists beyond one or more interactions	Interrupt users' desire for explanation and control
Obstruction	Making a process more difficult than it needs to be, with the intent of dissuading certain action(s)	Make users shun the effort to find and understand an explanation while interacting with explanation or control facilities
Sneaking	Attempting to hide, disguise, or delay the divulging of information that is relevant to the user	Gain from user's interaction with explanation/control facilities through hidden functions
Interface Interference	Manipulation of the user interface that privileges certain actions over others	Encourage explainability or control settings that are preferred by the system provider
Forced Action	Requiring the user to perform a certain action to access (or continue to access) certain functionality	Force users to perform an action before providing them with useful explanations or control options

5.3 Explainable AI for Different AI Systems

Different methods and approaches are used to explain knowledge-based white-box (e.g., rule-based) systems and data-driven black-box approaches (e.g., DNN). Rule-based systems are intrinsically explainable by design, meaning that XAI is part of the system's architecture. As presented in Chapter 4 on page 16, rule-based systems consist of IF-THEN rules. These rules exist in the simplest form of a condition (IF) and a prediction (THEN). With this structure, RBS are the most accessible and understandable AI systems as they provide human-readable rules in natural language (Molnar, 2019). In addition, their structure represents a familiar way humans think in everyday life (e.g., IF I feel hungry, THEN I eat something). Nevertheless, the explanation information has to be communicated to the user. Clancey (1983) criticizes rule-based systems, similar to Miller et al. (2017) did for DNN, that the rules are not easily understandable for non-experts. They argue that "rules are more than simple associations between data and hypotheses" (Clancey, 1983, p. 3). Therefore, even for an intrinsic explainable system like a rule-based system, one must think about communicating this information to end-users. In comparison to rule-based systems, DNNs are not explainable by design. Due to their structure including a multitude of neurons and parameters⁷, it is unclear to humans what they have learned during training. As a consequence, the decisions of DNN are not comprehensible to humans. Therefore, explanations must be included after the system's prediction, i.e., *post-hoc* (Adadi & Berrada, 2018). For this purpose, different methods have been developed in the last few years. These can be categorized into model-specific and model-agnostic approaches (Adadi & Berrada, 2018).

7: e.g., the famous image recognition CNN called *AlexNet* (Krizhevsky et al., 2017) consists of 650,000 neurons and 60 million parameters

In the following, I describe the unique features of XAI in the context of rule-based systems and DNN. The focus is on the procedures used in the experiments in this dissertation: (1) verbal explanations in rule-based

systems and (2) visual explanations using XAI algorithms (i.e., LIME, LRP, and counterfactuals).

5.3.1 XAI in Rule-Based Systems

Rule-based systems are often referred to as being interpretable by design (i.e., have an intrinsic explainability), meaning that the explanation is developed together with the decision system (Pedreschi et al., 2019). Barredo Arrieta et al. (2020) illustrate three different explainability domains for such transparent models: *Simulatability*, *decomposability*, and *algorithmic transparency*. Simulatability refers to a model that makes it possible for a human to think and reason about it. Decomposability describes the ability to explain each part of the model (i.e., input, parameter, calculation). For this, each model part must be understandable to a human without needing additional tools. Closely related to decomposability is algorithmic transparency. This means the possibility for people to understand the process that the model goes through. Barredo Arrieta et al. (2020, p. 90) highlight that rule-based models have three aspects that can be challenging regarding these three different explainability domains:

- ▶ **Simulatability:** Variables included in rules are readable, and the size of the rule set is manageable by a human user without external help
- ▶ **Decomposability:** The size of the rule set becomes too large to be analyzed without decomposing it into small rule chunks
- ▶ **Algorithmic Transparency:** Rules have become so complicated (and the rule set size has grown so much) that mathematical tools are needed for inspecting the model behaviour

In human-agent interactions (e.g., robots or virtual agents), the generation of verbal explanations seems to be a promising approach (Lyons, 2013; L. Zhu & Williams, 2020). A taxonomy of verbal explanations given by social robots is described by Stange et al. (2019). They illustrate four verbal explanations varying in content that a robot can provide about its behaviour: *perception-based*, *action-based*, *strategy-based*, and *need-based explanations*. They illustrate these different types in an example where a social robot moves towards a human and explains this behaviour. The perception-based explanation would reference the stimuli that led to the behaviour, “I moved to you because I saw you.” The action-based explanation would address the robot’s movement in this situation, “I wanted to move closer to you.” In contrast, the strategy-based explanation would result in a statement such as “I moved to you because I wanted to make contact with you.” Finally, an explanation based on the robot’s needs would read, “I approached you because I was lonely.” The effect of these four explanation contents was tested in the form of an online user study (Stange & Kopp, 2020), where users were shown six videos of the robot Pepper. In these, the different explanation behaviours of the robot were displayed. Users had to answer a questionnaire after each video. The results show that although all explanations led to an increase in the robot’s comprehensibility and desirability, the combination of multiple explanation contents (i.e., explanations with an intend x and a need y) showed the greatest improvements compared to the presentation of isolated explanations (Stange & Kopp, 2020).

Table 5.2: Suggested taxonomy from Gilpin et al. (2018) for DNN XAI methods and methods to evaluate them. *Processing* and *representation* stand for a more technical focus, *explanation producing* takes the interpretation and evaluation of humans into account, but are more for ML-experts than for end-users (Gilpin et al., 2019)

Purpose of explanation	Processing	Representation	Explanation Producing
XAI methods	Proxy methods Decision trees Saliency mapping Automatic-rule extraction	Role of layers Role of neurons Role of vectors	Scripted conversations Attention-based Disentangled representation
Evaluation methods	Completeness to model Completeness on task	Completeness on task Detect biases	Human evaluation Detect biases
Questions to answer	Why does this particular input lead to this particular output?	What information does the network contain?	Given a particular output, how can the network explain its behavior?

An approach of generating and using verbal explanations to communicate decisions for cooperative decision-making in medicine is presented by Schmid and Finzel (2020). They recommend using Inductive Logic Programming to generate rules which can be communicated to users. The speciality of their work is that they want to combine two *explanation types*: verbal explanations with visual explanations. In doing so, they intend to make it easier for end-users to understand the decisions of medical diagnoses made by black-box approaches. Verbal explanations provide a valuable basis for this, as they are a natural form of interaction for conveying information (recall Miller, 2019 “Explanations are social”).

5.3.2 XAI Methods for DNN

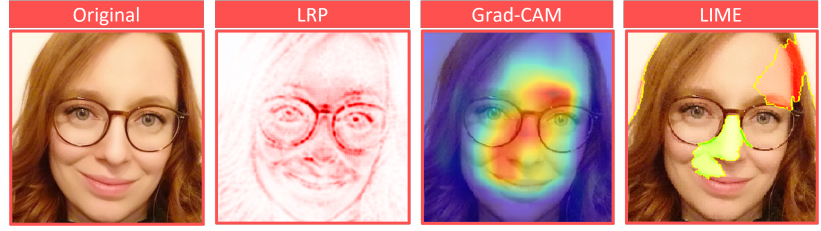
Explainable AI has been a relevant topic in the early days of AI, where researchers investigated the intrinsic explainability of rule-based systems. However, DNN brings new challenges to XAI due to its complexity, black-box characteristics, and widespread use in different domains. Gilpin et al. (2018) proposed a taxonomy for XAI in DNN applications (see 5.2). In their taxonomy, they differentiate between three different purposes for explanations.

- ▶ **Processing:** Explanations that focus on the processing of DNN aim to highlight the connections between the input and the output of a DNN. Common approaches are saliency methods like LIME (Ribeiro et al., 2016) and LRP (Bach et al., 2015), which are described in more detail in the following subsection.
- ▶ **Representation:** For explanations of representations of DNNs, the role of components of DNNs is investigated in more detail.
- ▶ **Explanation Producing:** Here, the focus lies on systems that produce explanations and present them to human users. Users refer here to ML experts with knowledge about ML and AI.

For all three purposes, they focus on *inside explanations*, that are tailored to ML-experts (Gilpin et al., 2019).

A variety of Saliency XAI methods has evolved since 2014 to address the problem of intransparency and incomprehensibility. Here, a distinction is made between *model-agnostic* and *model-specific* approaches. The former include methods such as LIME (Ribeiro et al., 2016) and SHAP (Lundberg

Figure 5.2: Application of three XAI algorithms to explain a CNN’s classification of the emotion “happiness” of visual input (left image): LRP (Bach et al., 2015), Grad-CAM (Selvaraju et al., 2017), and LIME (Ribeiro et al., 2016)



& Lee, 2017), which can not only be applied to DNN but can be used for various ML methods. Among others, model-specific methods are the LRP by Bach et al. (2015) and the Grad-CAM approach developed by Selvaraju et al. (2017). However, all these methods have in common that they highlight regions in images relevant for classification (see Figure 5.2).

LIME

The XAI approach of Ribeiro et al. (2016) was first presented in 2016 and belongs to the model-agnostic approaches, which means it can be used for different kinds of AI approaches (e.g., DNNs, decision trees, linear models). The idea of LIME (Local Interpretable Model-Agnostic Explanations) is to use local predictions to learn an interpretable model.

The key of the approach of Ribeiro et al. (2016) is *simplification*: For a simplification in image classification, the original representation of an instance which should be explained, denoted as $x \in \mathbb{R}$ is represented as an interpretable representation in the form of a binary vector $x' \in \{0, 1\}^{d'}$. Ribeiro et al. (2016) describe an explanation as a model $g \in G$, where G represents different kinds of interpretable models. In general, the explanation calculated by LIME looks like the following:

$$\xi(x) = \underset{g \in G}{\operatorname{argmin}} \{ \mathcal{L}(f, g, \pi_x) + \Omega(g) \}, \quad (5.1)$$

where $\Omega(g)$ stands for the complexity of the explanation $g \in G$. For example, using CNNs, the complexity measure is the number of non-zero weights. The complexity measure is the counterpart of interpretability, meaning the more complex an explanation is, the less interpretable it is by humans. The model being explained is denoted as $f : \mathbb{R}^d \rightarrow \mathbb{R}$. For multiple classification tasks, $f(x)$ represents the probability that x belongs to the relevant class. $\pi_x(z)$ serves as a proximity measure between a distance z and x and represents the locality. $\mathcal{L}(f, g, \pi_x)$ expresses the unfaithfulness of g in the approximation of f depending on the locality, given by π_x . The focus lies on two parts: to minimize $\mathcal{L}(f, g, \pi_x)$ to guarantee a local fidelity and to hold $\Omega(g)$ low to get a result that is still interpretable by humans. $\mathcal{L}(f, g, \pi_x)$ is approximated using samples which are weighted by π_x .

When using LIME for CNN image classifiers, the following steps of the LIME algorithm are passed:

1. The original image which is used for the prediction by the CNN is divided into super-pixels.

2. The original image is perturbed into sample instances by switching some super-pixels off. The L2 distance between the original and perturbed images is calculated and used later as weights π_x for the explanation model.
3. The created sample instances are used as input for the CNN image classifier. The classifier then calculates a prediction for each of the perturbed images.
4. Extraction of K features (super-pixels) of the CNN image classifier creates the maximum likelihood for the class the CNN predicted. The selection of the K features is made using a variant (Efron et al., 2004) of the Lasso algorithm from Tibshirani (1996). K stands for the number of features which should be extracted. Any number can be used here, but a higher value means more complexity (Efron et al., 2004). The weights for the K features are then learned using the least-squares method. The combination of Lasso with K features is named K-LASSO by Ribeiro et al. (2016).
5. The resulting relevant super-pixels can then be displayed on the image. The irrelevant super-pixels are greyed out.

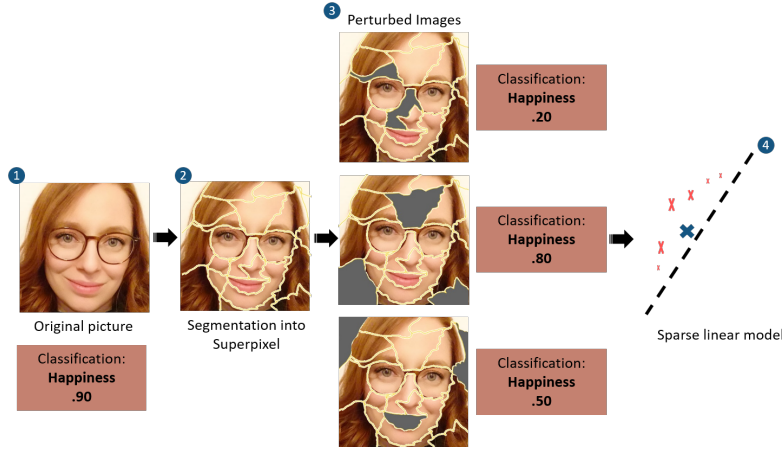


Figure 5.3: Steps of the LIME algorithm (Ribeiro et al., 2016): (1) An original picture is (2) segmented into superpixel. (3) The segmented images are perturbed by greying superpixels out. (4) The resulting images are weighted according to their proximity to the instance (blue cross) which should be explained. With this, a sparse linear model (dashed line) is learned

LRP

The LRP method, introduced by Bach et al. (2015), is optimized for DNN architectures. It uses pixel-wise decomposition as its central concept, combined with layer-wise relevance propagation (Bach et al., 2015). The general idea of pixel-wise decomposition is to look at the impact of each input pixel $x_{(d)}$ of an input image x to the prediction $f(x)$. One possibility to do that is to segment (=decompose) the prediction $f(x)$ is the sum of the terms of the input dimensions, notated as:

$$f(x) \approx \sum_{d=1}^V R_d. \quad (5.2)$$

$R_d < 0$ can be interpreted as evidence against the structure which should be classified, and $R_d > 0$ otherwise. The resulting Relevance R_d for each input pixel $x_{(d)}$ can be visualized in a heatmap by mapping every R_d to a colour space (Bach et al., 2015). LRP is an approach to achieve a pixel-wise decomposition as denoted in Equation 5.2. LRP defines

constraints that must be fulfilled when calculating the importance of pixels to a classification result of a neural network (Bach et al., 2015). These constraints are described in Equation 5.7 and Equation 5.8. Before the LRP approach for the entire network is explained, the function of LRP on a single neuron j is described (Lapuschkin et al., 2017): A neuron j gets a relevance score R_j from the higher layer. This relevance score is distributed proportionally to the contribution of the input neurons i of the neuron j . The distribution to i is based on the contribution of the i neurons in the forward pass:

$$R_{i \leftarrow j} = \frac{z_{ij}}{z_j} R_j. \quad (5.3)$$

z_{ij} is measuring the contribution of neuron i to the activation of neuron j . z_j represents the aggregation of all forward messages z_{ij} over i at j . The relevance value R_i is defined by all incoming relevance values, $R_{i \leftarrow j}$ of the neurons j in which i is involved:

$$R_i = \sum_j R_{i \leftarrow j}. \quad (5.4)$$

Therefore, the following local conservation property is given:

$$R_i = \sum_j R_{i \leftarrow j} \quad \text{and} \quad \sum_i R_{i \leftarrow j} = R_j. \quad (5.5)$$

With these formulas, it is possible to calculate the importance of pixels to a classification result of a neural network (Bach et al., 2015):

$$f(x) = \dots = \sum_{d \in l+1} R_d^{(l+1)} = \sum_{d \in l} R_d^{(l)} = \dots = \sum_{d \in 1} R_d^{(1)}, \quad (5.6)$$

where $R_d^{(l+1)}$ stands for the relevance score for each dimension $z_d^{(l+1)}$ of the layer $l + 1$, modelled by the vector z . The last layer is represented as $f(x)$ and the first layer of the network as $R_d^{(1)}$. The relevance of each neuron of the network except the last neurons (output neurons) is the first constraint of LRP. This first constraint is defined as:

$$R_i^{(l)} = \sum_{k: i \text{ is input for neuron } k} R_{i \leftarrow k}^{(l+1)}. \quad (5.7)$$

It should be noted that the term ‘input’ refers to the direction during classification, i.e., from a previous layer to a subsequent layer. The second constraint for LRP is defined as:

$$R_k^{(l+1)} = \sum_{i: i \text{ is input for neuron } k} R_{i \leftarrow k}^{(l)}, \quad (5.8)$$

which represents the sum over the sources at layer l for a fixed neuron k at layer $l + 1$. In comparison, Equation 5.7 represents the sum over the sinks at layer $l + 1$ for a fixed neuron i at a layer l . A visualization of the important components of LRP is displayed in Figure. The neuron activation of x_j represents a non-linear function of z_j . The pre-activations

z_{ij} measure the relative contribution of each neuron x_i to R_j . The relevance decomposition, based on the local and global pre-activations, is denoted as:

$$R_{i \leftarrow j}^{(l,l+1)} = \frac{z_{ij}}{z_j} \cdot R_j^{(l+1)} \quad (5.9)$$

A disadvantage of equation 5.9 is that for small z_j , relevance values $R_{i \leftarrow j}$ can take on unbounded values. A stabilizer $\varepsilon \geq 0$ can be used to counteract this. The Equation 5.9 can be adjusted as follows:

$$R_{i \leftarrow j}^{(l,l+1)} = \begin{cases} \frac{z_{ij}}{z_j + \varepsilon} \cdot R_j^{(l+1)} & z_j \geq 0 \\ \frac{z_{ij}}{z_j - \varepsilon} \cdot R_j^{(l+1)} & z_j < 0 \end{cases} \quad (5.10)$$

The LRP method is often stabilized using a $\alpha\beta$ -rule. Defining specific values for α and β is possible. The Equation 5.9 then changes to:

$$R_{i \leftarrow j}^{(l,l+1)} = R_j^{(l+1)} \cdot \left(\alpha \cdot \frac{z_{ij}^+}{z_j^+} + \beta \cdot \frac{z_{ij}^-}{z_j^-} \right) \quad (5.11)$$

The usage of α and β have the advantage through the stabilizing effect that they make it possible to visualize not only positive but also negative activations of pixels. The strength of the influence of negative and positive portions can be controlled with the choice of the respective α and β value (Bach et al., 2015; Montavon et al., 2018). Besides these parameters, Kohlbrenner (2017) showed that a ‘preset’ variant of the LRP algorithm achieves optimal results in calculating relevance maps. Using the preset approach, the relevance scores R_j for all neurons of the lowest (first) layer are uniformly distributed to the input neuron instead of using the $\alpha\beta$ values (Lapuschkin et al., 2017). To control the resolution of the heatmaps generated by LRP, Bach et al. (2016) describe an approach of a ‘mapping influence cut-off point’. This point defines the moment from which the forward mapping function of the classifier no longer influences relevance propagation since only the receptive field of the classifier is relevant. The cut-off at this point is called the ‘flat’ rule.

Counterfactuals

Wachter et al. (2018) introduced the concept of *unconditional counterfactual explanations*. They introduce their concept with the following example (Wachter et al., 2018, p. 844):

“You were denied a loan because your annual income was £30,000. **If your income had been £45,000, you would have been offered a loan.**”

The counterfactual in this sentence is highlighted in red and represents a statement about “[...] how the world would have to be different for a desirable outcome to occur.” (Wachter et al., 2018, p. 844). The authors highlight that various counterfactuals can exist because of various (desired) outcomes; therefore, different ways to reach them can exist. The concept of *the closest possible world* tries to reduce this complexity

by looking only at the smallest change in the world that is needed to reach the desired outcome. Nevertheless, Wachter et al. (2018) highlight that sometimes not the *closest possible world* but rather the *close possible worlds* can be more informative and relevant. In analytic philosophy, where Wachter et al. (2018) rooted their counterfactual explanation approach, the idea is that of *justified true belief*. This means that one does not simply believe that something is true, but one has to prepare a reason for believing it (Gettier, 1963). This justification can be used to answer the question: “Why do you believe in X?” (Wachter et al., 2018, p. 847). In addition, Wachter et al. (2018) include sensitivity in their counterfactual explanation approach. The sensitivity approach is defined by the following expression (Wachter et al., 2018, p. 847):

“If p were false, S would not believe p ”,

where p refers to a proposition that is known and S to the knowing subject. Counterfactual explanations are similar to this concept (Wachter et al., 2018, p. 847):

“If q was false, S would not believe p ”.

It is important to note that such a statement represents only the belief of a person and, therefore not need to be accurate. Going from such more general explanations to more of these that are relevant when explaining, for example, DNN, Wachter et al. (2018, p. 848) show a counterfactual explanation for this field:

“Score p was returned because variables V had values (v_1, v_2, \dots) associated with them. If V instead had values (v_1', v_2', \dots), and all other variables had remained constant, score p' would have been returned.”

Wachter et al. (2018) point out that the idea of a closest possible world is implicit in their definition reflected in the score p' .

To create counterfactual explanations as we did in our Pneumonia Study described in Chapter 29 on page 210, Generative Adversarial Networks (GANs) can be used. GAN is an architecture that consists of two multilayer perceptrons and was described by Goodfellow et al. (2014). The basic idea is to have a *generative model*, which is opposed to a *discriminative model*. This model must decide whether an example comes from the data distribution. Goodfellow et al. (2014) illustrate this idea using the example of money counterfeiters (i.e., generative model) and police officers (i.e., discriminative model). While the counterfeiters try to produce fake money and not get caught (i.e., by creating money that is very similar to the original), the police try to detect the counterfeits through the detection of the fake money. This approach is used in image-to-image translations. Here, one image is transferred into another image. The CycleGAN approach developed by J.-Y. Zhu et al. (2017) uses two GANs to translate one image (e.g., a zebra) into another domain (e.g., a horse). Their approach is based on the work of Johnson et al. (2016), which uses the VGG16 CNN architecture of Simonyan and Zisserman (2015). J.-Y. Zhu et al. (2017) denote the CycleGAN as:

$$\begin{aligned}\mathcal{L}(G, F, D_X, D_Y) = & \mathcal{L}_{GAN}(G, D_Y, X, Y) \\ & + \mathcal{L}_{GAN}(F, D_X, Y, X) \\ & + \lambda \mathcal{L}_{cyc}(G, F)\end{aligned}\quad (5.12)$$

The discriminators D_X and D_Y aim to maximize that objective function, while the generators G and F try to minimize it. More formally, it is denoted as (J.-Y. Zhu et al., 2017):

$$G^*, F^* = \arg \min_{G, F} \max_{D_X, D_Y} \mathcal{L}(G, F, D_X, D_Y). \quad (5.13)$$

For the Pneumonia Study⁸, an extension of the CycleGAN approach proposed by Mertes et al. (2022) is used to create counterfactual explanations. In their paper, they state that a counterfactual explanation for an image classifier, following the work of Wachter et al. (2018), should answer the question:

“What minimal changes to the input image would lead the classifier to make another decision?” (Mertes et al., 2022, p. 3)

This question defines two requirements for a counterfactual explanation (Mertes et al., 2022, p. 3): “The counterfactual image should look as similar to the original image as possible” and “The classifier should predict the counterfactual image as belonging to another class as the original image”. In a more formal notation, the CycleGAN Equation 5.12 is adjusted as follows:

$$\begin{aligned}\mathcal{L}(G, F, D_X, D_Y, C) = & \mathcal{L}_{GAN}(G, D_Y, X, Y) \\ & + \mathcal{L}_{GAN}(F, D_X, Y, X) \\ & + \lambda \mathcal{L}_{cycle}(G, F) \\ & + \mu \mathcal{L}_{identity}(G, F) \\ & + \gamma \mathcal{L}_{counter}(G, F, C)\end{aligned}\quad (5.14)$$

where μ represents an *Identity Loss Weight* and γ represents a *Counterfactual Loss Weight*. The identity loss ensures that input images belonging to the target class remain unchanged (J.-Y. Zhu et al., 2017). The counterfactual loss was proposed by Mertes et al. (2022) and is an extension to the CycleGAN approach of J.-Y. Zhu et al. (2017) to generate counterfactual images. Identical to the original CycleGAN approach, the discriminators D_X and D_Y try to maximize that objective function, while the generators G and F try to minimize it.

8: the study is described in Chapter 29 on page 210, the implementation in Chapter 21 on page 132

6

Human States

6.1 Theory of Mind and Mental Models	37
6.2 Trust	39
6.3 Self-Efficacy & Cognitive Workload	41
6.4 Emotions	43

1: emotions are not stated here since no empirical studies except the ones presented in this dissertation (i.e., VR-Robot Study & Pneumonia Study) were dealing with the impact of XAI on users' emotions during the writing of this dissertation

Explanations influence many facets of ourselves. Gunning and Aha (2019) state that psychological constructs like trust, mental models, and user satisfaction are relevant to measuring XAI's effectiveness. In the following, the concepts of *mental models*, *trust*, *self-efficacy*, *cognitive workload*, and *emotions*, which are central components of the experiments presented in this dissertation, will be introduced. Figure 6.1 gives an overview of the empirical findings of these constructs regarding XAI¹.

The presented concepts were also described in the following publications and served as a basis for this chapter:

- ▶ Heimerl, A., **Weitz, K.**, Baur, T., & André, E. (2022). Unraveling ML models of emotion with NOVA: multi-level explainable AI for non-experts. *IEEE Transactions on Affective Computing*, 13(3), 1155–1167. doi: [10.1109/TAFFC.2020.3043603](https://doi.org/10.1109/TAFFC.2020.3043603) *
- ▶ **Weitz, K.** (in publishing process). Erklärbare Künstliche Intelligenz im Kontext Bildung und Lernen. In C. deWitt, C. Gloerfeld, & S. Wrede (Eds.), *Künstliche Intelligenz in der Bildung*. Springer. doi: [10.1007/978-3-658-40079-8](https://doi.org/10.1007/978-3-658-40079-8)
- ▶ **Weitz, K.** (2021). Vertrauen und Vertrauenswürdigkeit bei sozialen Robotern - Stärkung von Mensch-Roboter-Vertrauensbeziehungen mithilfe Erklärbarer Künstlicher Intelligenz. In O. Bendel (Ed.), *Soziale Roboter* (pp. 309-323). Springer Gabler, Wiesbaden.
- ▶ **Weitz, K.** (2021). Towards Human-Centered AI: Psychological concepts as foundation for empirical XAI research. *it-Information Technology*, 64(1-2), 71-75. doi: [10.1515/itit-2021-0047](https://doi.org/10.1515/itit-2021-0047)
- ▶ Part of the section Emotions
Weitz, K. (2018). *Applying Explainable Artificial Intelligence for Deep Learning Networks to Decode Facial Expressions of Pain and Emotions* (Master's thesis), University of Bamberg.

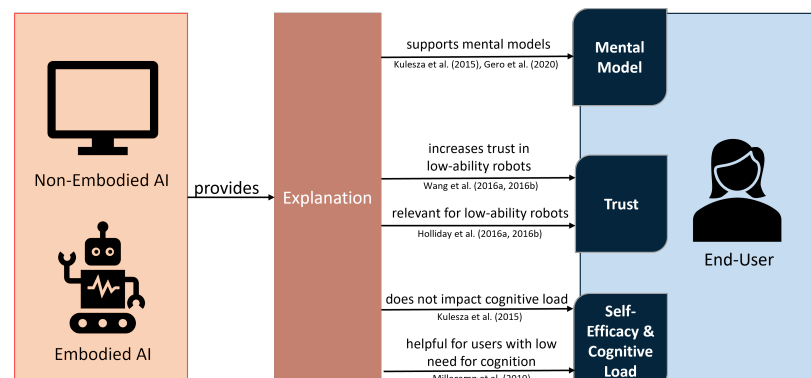


Figure 6.1: Empirical findings on the impact of XAI on users' *mental model*, *trust*, *self-efficacy*, and *cognitive load*

¹Licensed under the [Attribution 4.0 International License \(CC BY4.0\)](https://creativecommons.org/licenses/by/4.0/)

6.1 Theory of Mind and Mental Models

To design AI systems in a human-centered way, it is essential to understand people's beliefs and expectations of the technical system. The starting point for these considerations is the *Theory of Mind*, which was defined by Premack and Woodruff (1978, p. 515) in their work with chimpanzees: "An individual has a theory of mind if he imputes mental states to himself and others". Pylyshyn (1978) extended this assumption by saying that humans cognitively interpret the behaviours of others. This interpretation is the basis for explaining and evaluating the observed behaviour. In doing so, people form internal representations about what they perceive (Craik, 1967). These internal representations are also referred to as *mental models* (Johnson-Laird, 1983). In contrast, there is the so-called *conceptual model*. This describes an external representation (e.g. a mathematical formula, a formula) created by teachers, engineers or researchers (Greca & Moreira, 2000). A conceptual model is characterised by the fact that it is coherent with the knowledge of this group. Norman (1983) points out that mental and conceptual models are ideally related, but this need not be the case. For example, suppose people know nothing about the conceptual model of the earth. In that case, their mental model (e.g., earth as a disc) may be very different from the actual conceptual model (i.e., spherical earth model) (Vosniadou & Brewer, 1992).

Regarding the content of the mental models, Stangl (2012) separate between *cognitive mental models*, including thoughts and beliefs. In contrast, *emotional mental models* include emotions and feelings of a person towards things. Stangl (2012) criticise that most of the mental model research addresses cognitive mental models while neglecting the emotional part of it. Their study investigated the impact of the emotional mental model in the online environment of Second Life² on product and service presentations. Stangl (2012) asked unexperienced users before and after their visit of Second Life about their feelings towards service and product presentations in Second Life (i.e. "When I imagine a service presentation in Second Life (e.g., virtual massages) I feel:"). With their study, Stangl (2012) wanted to investigate whether there is a mismatch in the emotional mental model before and after visiting Second Life. Their results show a significant positive shift for the emotion "joy", indicating that users who perceived Second Life on their first visit as positive found even more positive feelings during the second visit. For the emotion "anger" they found precisely the opposite, a significantly negative shift, meaning that users who felt negative feelings during their first visit of Second Life perceived it even more negatively during their second visit. Both emotional shifts were seen in product and service presentations in Second Life. From these results, Stangl (2012) conclude that developers should also include the emotional mental models of users in the usability design of virtual environments to support a positive emotional feeling of users during a virtual experience and, more important, to change users' emotions positively, so that online communities grow due to the positive experience.

In the work of Khemlani and Johnson-Laird (2017), the authors highlight that people make systematic errors in their reasoning, which are caused due to *incomplete mental models*. Under *incomplete*, they understand that people have mental representations about "true" possibilities but not of

2: Second Life is a virtual reality where users can interact with each other by using avatars

“false” ones. They name this phenomena *illusory inferences* and highlight it in an example (Khemlani & Johnson-Laird, 2017, p. 19):

Imagine that you are in a restaurant, and suppose that only one of the following two assertions is true:

1. You have the bread.
2. You have the soup or the salad, but not both.

Also, suppose you have the bread. What, if anything, follows?
Is it possible that you also have either the soup or the salad?
Could you have both?

3: I, as the writer of this dissertation, also answered “no” on the first try

Based on the fact that one knows that they have the bread, the constructed mental model recommends choosing “no” to the question of whether you could have both soup and salad³. As Khemlani and Johnson-Laird (2009) highlight, this is the answer of most reasoners, but this conclusion is incorrect. The constructed mental model of most people is unable to recognize when (1) is true, (2) is incorrect, and therefore they could have both soup and bread or none of it. In more formal writing, it is expressed as the following (Khemlani & Johnson-Laird, 2009, p. 618):

Bread	¬Soup	¬Salad
Bread	Soup	Salad

The first row represents the premise that when one has bread, one could decide not to take the soup and the salad. The second row represents the premise that when one has bread, one can have both, soup and salad.

Early research such as that of Heider and Simmel (1944) shows that humans not only attribute properties, intentions, and needs to other living beings but also do so with objects (i.e., in their study, they used circles and triangles as objects). Similar attributions can also be observed in the interaction of humans with AI systems. The attributions and derived mental models can be *incomplete, unscientific, and unstable* (Norman, 1983), as the previous example of Khemlani and Johnson-Laird (2009) illustrates. Incorrect mental models also have an impact on the interaction of humans with AI systems. For example, Budiu (2019) shows that both new and experienced users often see personal assistants as an interface to another application. Still, their mental model does not correspond to reality (e.g., “Alexa reaches out into the magical world of the internet”). Alvarado et al. (2020) highlights the importance of taking these user beliefs into account. They used semi-structured interviews to gain insights into users’ beliefs about YouTube recommendations. They identified four groups of beliefs: previous actions, social media, recommender system, and company policy. They found that while users are aware of recommendation algorithms, their understanding of them is limited.

Nourani et al. (2021) demands not to underestimate the influence of *cognitive biases* on human mental models. Their user study highlights the importance of users’ first impressions of an AI system. Here, they identified an automation bias⁴ during an interaction with a video activity recognition tool. Participants produced significantly more errors in determining the correctness of AI-generated policies when in the AI model strengths condition (i.e., the AI system recognized the activities correctly). This result is due to an automation bias meaning that participants’ positive first impression of the AI system led to relying too much on the system

4: Cummings (2004) describe automation bias as a human tendency to hand over the responsibility to automation. It occurs in HCI when humans do not search or ignore contradictory information during a computer-based decision process

while participants who saw weaknesses did not. Regarding shaping the participants' mental models, XAI helped build confidence in users' mental models about the AI system.

From all this research, users' mental models should be considered when discussing HC-XAI. Including questions like "What do you understand by AI technology in the present application scenario?" should be pursued to address mental models for specific AI applications that can differ from application to application. This question can be fleshed out by assessing limitations ("What do you think AI technology *cannot do* in this application scenario?"). Through interaction with an AI system, users' mental models about it can be formed or changed (Rutjes et al., 2019). In this context, XAI can help users develop *accurate mental models*. Accurate mental models refer here to a mental model corresponding to the conceptual model. Kulesza et al. (2015) showed that when a text-classification system⁵ explains how it came to its decisions⁶, it helped end-users to build useful mental models. During the experiments, users had the task of improving the text-classification system. The authors found that users of the explanatory system improved their classifier up to twice as efficiently as users of the control group. In addition, this improvement correlated positively with users' mental model scores, indicating that an accurate mental model leads to better performance of the ML system. Similar results were found by Gero et al. (2020). When playing a word-guessing game with an AI agent, users win more often when they have an accurate mental model (i.e., correct estimation of the agent's abilities) about the AI agent.

5: they used a Naive Bayes Classifier

6: the authors refer here to *Explanatory Debugging*

6.2 Trust

Gilpin et al. (2018) sees explainability as a prerequisite for users to trust an AI system. In HCI, different definitions of trust exist. One of the most commonly used definitions is the one from Lee and See (2004). They define trust as the attitude of humans that an agent (e.g., a robot) will help them achieve goals in a situation characterized by uncertainty and vulnerability (Lee & See, 2004, p. 51). Authors like Lewis and Weigert (1985) and Madsen and Gregor (2000) separate trust into two components: *cognition-based trust* and *affect-based trust*: In the context of human-robot interaction, cognitive trust is described as a person's mental characteristics, reasons, and arguments toward an agent. On the other hand, affective trust describes a person's feeling towards an agent (Castelfranchi & Falcone, 2010).

Theoretical models of trust try to organize various relevant factors for trust in human-computer interaction. It can be distinguished between *types of trust* and *dimensions of trust*. Types of trust mean that trust in a technology (e.g., an AI system) can be based on different roots. Merritt and Ilgen (2008) distinguish between *dispositional* and *experience-based* trust. Dispositional trust here refers to a stable personality trait of a person. Experience-based trust refers to trust gained from prior experience with a machine. Hoff and Bashir (2015) used a literature review to develop a theoretical framework to cluster different types of trust. This framework includes three layers: *dispositional trust*, *situational trust*, and *learned trust*. Dispositional trust refers to personality traits based on biological

and environmental influences. Situational trust describes external aspects (e.g., the environment) and internal aspects (i.e., characteristics of the person) in a given situation. Finally, learned trust refers to the experiences that users have already gained with technology. Authors such as Hancock et al. (2011), de Visser et al. (2020), and Lee and See (2004) point out the importance of taking the dimensions of trust into account. Dimensions of trust highlight the relationship and interconnections of *trust*, *overtrust*, *distrust*, and *appropriate trust*. Lee and See (2004) see appropriate trust as the goal that interaction with technology should gain. Appropriate trust describes the fit between trust in the system and the system's actual capabilities. If users overestimate a system's capabilities, they overtrust it; if they underestimate it, they distrust it. Marsh and Dibben (2005) describe trust as a continuum between distrust (i.e., negative trust level) and untrust (i.e., positive trust level). Untrust is defined as "a measure of how much a person is trusted" (Marsh & Dibben, 2005, p. 21). Only when a *cooperation threshold* is reached do users trust the technology and are willing to cooperate with the system. Mistrust can be understood as misguided trust, which arises from betrayal (Marsh & Dibben, 2005).

Stanton and Jensen (2021) see the *Computers are Social Actors* (CASA) paradigm (Nass & Moon, 2000) helpful for researching human-machine trust. In the CASA paradigm, Nass and Moon (2000) state that humans attribute social rules and expectations to computers, even when computers do not have an anthropomorphic form like, for example, in social robots. They were inspired by the work about mindfulness and mindlessness of people from E. J. Langer (1992). Mindlessness and mindfulness describe the amount of social information a person is using (E. J. Langer, 1992). The usage of minimal structural cues characterizes mindlessness behaviour. In communication, structural cues refer to syntactical information, i.e., how the sentences in a dialogue are structured. For example, E. Langer et al. (1978, p. 637) describes that people that were asked for a favour expect a structure like

"FavorX + ReasonY → Comply".

In comparison, mindful behaviour incorporates content and context of social information (E. J. Langer, 1992). When interacting with computers, Nass and Moon (2000, p. 81) stated that people tend to "[...] mindlessly apply social rules and expectations to computers." even when users know that a computer is not a human. Stanton and Jensen (2021) highlights that the CASA paradigm shows that humans have the predisposition to show interactive behaviour to other humans as well as computers. Trust, and distrust help us to predict the actions of others during interactions (Stanton & Jensen, 2021). This behaviour can also be found in human-machine interactions (Stanton & Jensen, 2021).

Besides this general behaviour in interactions, individual attributions towards machines (e.g., learned through previous interactions) could be another relevant source for trust and distrust (Stanton & Jensen, 2021). For example, regarding trust in automated driving, Körber et al. (2018) investigated the impact of trust supporting vs trust decreasing introductory information on users' trust in an autonomous driving task. While the introductory information did not impact users' trust during the experimental drive, Körber et al. (2018) found that the individual trust level influences the users' perception of the environment while solving a non-driving-related task (NRDT). Users with higher trust in the automated

driving system looked longer at the NRDT task and less on the road and cockpit. Investigating human-robot trust, Gaudiello et al. (2016) found out that the kind of the task (i.e., social task vs functional task⁷) also has an impact of users' trust. User trust was in their experiment operationalized as the users' confirmation of the robot's answers in the tasks. In both kinds of tasks, first, the user was presented with two stimuli and had to answer a question (e.g., for the functional task: "Which is the most high-pitched sound: the first or the second?"). After that, the robot was asked the same question. Finally, the user was asked again whether they would change their answer. The social robot gave a contradiction for every answer from the user except for unambiguous answers. The results show that users had significantly more trust in the robot in the functional task than in the social task. Neither users' attributes, like attitude towards robots, nor their desire for control influenced their trust in the robot.

However, trust and mistrust also raise the question of whether and how *appropriate human-machine trust*⁸ to be promoted. Hoffman, Mueller, et al. (2018) highlight that with the help of appropriate trust, the *appropriate usage* of a machine or system can be achieved. An explanation could be a legit source to support appropriate trust between humans and machines (Hoffman, Mueller, et al., 2018). But for this, the impact of explanations needs to be investigated. For example, Wang et al. (2016a, 2016b) investigated how explanations about a robot's decisions affect user trust. Here, they compared explanations of robots with low and high capabilities. Low capabilities were reflected in frequent errors during decision-making, while for high capabilities, every decision made by the robot was correct. Wang et al. (2016a, 2016b) found that low-ability robots appeared more trustworthy when explaining their decision-making process, while this effect was not found for high-ability robots. These results suggest that explanations may be a helpful way to increase trust in robots. Holliday et al. (2016) investigated the impact of explanations on long-term users' trust towards AutoCoder⁹. They found that without explanations, user trust decreased over time. Explanations helped the users to understand (i.e., to create an accurate mental model) that the AutoCoder is a learning system that improved over time. This understanding was missing for the control group that did not receive explanations.

7: in the *functional task* users and a social robot had to compare physical stimuli regarding a perceptual aspect, e.g., the weight of two objects, while in the *social task*, users and the robot had to decide which objects are more suitable in a specific social context, e.g., wearing a bathing-cap or a pair of flip-flop in the social context of visiting a swimming pool (Gaudiello et al., 2016)

8: also referred to as *calibrated trust*

9: AutoCoder is an AI assistant that supports users in assigning qualitative data with codes

6.3 Self-Efficacy & Cognitive Workload

Cognitive load describes the limited capacities of humans' working memory (Sweller et al., 1998). Therefore, information should be presented in an *understandable* way that users can handle it (Sweller et al., 1998). Interaction with complex systems like AI technology demands many cognitive resources from users. In particular, dealing with technology and systems unfamiliar to users shows that much additional cognitive effort is required to enable access to and effective use of the technology. Besides the cognitive load during the actual interaction, self-efficacy aspects are relevant¹⁰ Bandura (2010, p. 1) describes perceived self-efficacy as "people's beliefs in their ability to influence events that affect their lives". Bandura (1977) points out that self-efficacy is the basis for all other factors (e.g., motivation) affecting users' actions. Only when a person is convinced that they can make a difference through their actions can they form the motivation to act based on this assumption. As Bandura

10: various studies in different domains found that cognitive load and self-efficacy are related (e.g., academics: Vasile et al., 2011, creativity: Redifer et al., 2021, online learning: Limakrisna et al., 2021).

describes, self-efficacy is a holistic construct that applies to all aspects of life. Self-efficacy regarding digital technologies is defined by Compeau and Higgins (1995) as *computer self-efficacy*. It describes the perceived self-efficacy of people regarding computers and related technologies. Numerous studies have found evidence that computer self-efficacy and user behaviour are related. For example, Hill et al. (1987) found a connection between perceived self-efficacy and the use of computers. People who were convinced that they had no control over the computer were less inclined to learn about or use the device. This assumption is consistent with Bandura (1977) suggestion that only direct interaction with a task or object helps minimize anxiety and induce users to change their behaviour (e.g., use technologies).

Information about the perceived self-efficacy of users can be used to design and adapt AI systems in a more user-centered way. For example, Wiggins et al. (2017) describes that the information about the perceived self-efficacy of users in human-human tutoring tasks can be used to adapt intelligent tutoring systems (ITS) to the abilities and preferences of the user. In their study, computer science students had to solve a JAVA programming task with the help of a human tutor. With this task, Wiggins et al. (2017) investigated the impact of self-efficacy on learning gain and frustrations of the students. The authors found a positive correlation between computer self-efficacy and computer self-efficacy, indicating that, for example, participants with high self-efficacy values also had high values of computer self-efficacy. Furthermore, they found that users with high and low self-efficacy values benefit differently from a tutor during their programming tasks. High self-efficacy persons especially benefit from the tutor's feedback so that their efficiency of learning increases. When a tutor had a comment on their programming code, high self-efficacy students stopped running the program. With this action, instead of thoroughly testing their program, they could immediately correct their code and discuss it with the tutor. Low self-efficacy students preferred to run the code and fix upcoming errors after the program finished. The authors conclude that a more explorative approach focusing on learning by doing could support low self-efficacy students. However, this approach can, at the same time, lead to higher frustration in the students. The work of Latikka et al. (2019) shows a connection between the self-efficacy of users and their acceptance of using robots. However, compared to Wiggins et al. (2017), the correlations could only be shown for the specific self-efficacy towards robots but not general self-efficacy.

Regarding explanations, Millecamp et al. (2019) investigated how user attributes like musical experience, tech-savviness¹¹, and need for cognition¹² impact the perception of explanations¹³ of a music recommender system. Users had to interact with a recommender system with and without explanations to create two-song playlists. They found that users' need for cognition impacts the interaction and perception of a recommender system and its explanations. Users with a low need for cognition tend to benefit more from the recommender's explanation than users with a high need for cognition. The authors reason that explanation help users with a low need for cognition to support their confidence in decisions (i.e., song selection), whereas, for users with a high need for cognition, explanations lower their confidence in decisions. Regarding the impact of cognitive load, Kulesza et al. (2015) found that the presentation of

11: Millecamp et al. (2019, p. 398) define *tech-savviness* as users "confidence in trying out new technology"

12: Cacioppo and Petty (1982, p.130) defined need for cognition as "individual differences in people's tendency to engage in and enjoy thinking"

13: in the form of bar charts and scatter plots

explanations leads not to a higher cognitive load of users, meaning they felt not overwhelmed by the additional information presented.

6.4 Emotions

Plutchik (1982) defined emotion as a complex inferring sequence of reactions to a stimulus, including cognitive evaluation, subjective changes, autonomous and neuronal arousal, and impulses for action and behaviour. This affects the stimulus that initiates the complex sequence. Emotions are one of the key characteristics of human experience (Vytal & Hamann, 2010). Emotional experiences permeate every area of (mental) life (Kassam et al., 2013). They influence the content and type of thoughts (Clore & Huntsinger, 2007), on decisions and actions (Damasio, 1994), and on memory and perception (Phelps, 2004; Phelps et al., 2006; Scott et al., 1997).

People show emotions not only towards other humans but also towards machines. The term user experience describes these subjective feelings (i.e., positive or negative) that people can have when interacting with technology (Hassenzahl, 2008). A BBC News online article describes negative user emotions towards personal computers with the following statistics:

“Almost a third of people had physically attacked a computer, 67% experienced frustration, exasperation and anger and more than 70% swore at their machines.” (BBC News, 2000)

Similar to cognitive and emotional mental models and cognitive and affective trust, researchers point out that emotions also have a cognitive and an emotional component (Cenefetelli, 2004). This raises the question of what influence technology has on users’ emotions. For example, the work of E. H. Park et al. (2022) shows that patients’ relatives rejected AI monitoring when they feared surveillance or relinquishing responsibility. However, the opposite effect (i.e., acceptance of AI monitoring) was found when fear of the quality of care and the health status of the person cared for prevailed. The authors also show that perceived controllability (i.e., self-efficacy) moderated the influence on the relationship between fear of surveillance and rejection of an AI monitoring system.

Regarding XAI, some research investigates the role of emotions in virtual agents. For example, Kaptein et al. (2017) provides approaches to use simulated emotions for a virtual agent to enhance agent explanations. However, empirical work in the field of XAI that investigates the other way around, i.e., examining the influence of XAI on user emotions, is rare. Therefore, two empirical studies exploring this aspect are presented in Chapter 27 on page 160 (VR-Robot Study) and in Chapter 29 on page 210 (Pneumonia Study).

7

Summary Background

Explanations are part of our everyday lives. In the background chapters, I overviewed the impact of AI-generated explanations on humans. Starting at the very beginning in Chapter 4 on page 16, defining what the term “AI” refers to and presenting rule-based systems as an example of white-box AI and CNNs as an example of black-box AI. The focus on these approaches is since these are used in the conducted experiments that this dissertation present. The implementation of the rule-based systems can be found in Chapter 20 on page 125, and the implementation of CNNs is described in more detail in Chapter 21 on page 132.

Chapter 5 on page 22 deals with the topic of Explanations. After defining the terms “explanation” and “XAI”, the broad field of XAI was illustrated by shed light on the differences and interconnections between the terms *interpretation*, *explanation*, *transparency*, and *understanding*. After that, I highlighted how explanations impact humans and the importance of explanations for human learning. Then, XAI approaches to generate explanations for rule-based systems, and DNN were presented. These approaches are used in the empirical part of this dissertation. After this overview of technical aspects of XAI, in Chapter 6 on page 36, an overview of the human perspective was given. For doing so, the psychological constructs of *mental models*, *trust*, *self-efficacy*, *cognitive workload*, and *emotions* were presented. The constructs are operationalised in the empirical part of this dissertation (see Chapter 23 on page 143 for a detailed description) and investigated in the six experiments presented in this dissertation (starting in Chapter 27 on page 160).

The background chapters build the foundation for understanding the following parts of the dissertation.

III. RELATED WORK

8

Overview

To design XAI in a human-centered way is not a novel idea. Many researchers have already put a lot of effort and thoughts into the design of XAI. In this chapter, first, this previous work on the design of XAI, which are of valuable knowledge for the design of HC-XAI, is given. Second, related work is presented that deals specifically with the three application purposes of (X)AI in this dissertation: cooperation & collaboration, education, and medical decision support.

The related work chapter ends with a summary of the already known XAI design insights and builds the basis for the interdisciplinary concept of this dissertation, described in detail in Chapter 12 on page 70.

This chapter deals with the question “**What concepts and ideas for the design of XAI already exist?**”. The following two sections are divided regarding the content of the concepts. First, ideas that dive into detail about XAI’s structure are presented. Here, the work focuses on precisely describing various components of XAI (e.g., the goal of the explanation). In the second section, the concepts focus on the evaluation and implementation of XAI. The objective of these approaches is to develop interactive XAI. Interactive XAI can be subsumed as the idea that an XAI system reacts to users’ needs, for example, by answering user questions.

9.1 Concepts About the Structure of XAI	47
9.2 Concepts to Develop Interactive XAI	49

9.1 Concepts About the Structure of XAI

Ribera and Lapedriza (2019) describe three aspects to archive explainability in AI systems:

- **WHY? - Goal of the explanation** They list different goals of XAI (e.g., verification of the system, improvement of the system, acceptance of technology) as reported by various researchers.
- **WHAT? - Content of the explanation** Here, the focus is on what content the explanation should include. Here, Lim et al. (2009, p. 2120) highlight to include information about the questions “*What* did the system do?”, “*Why* did the system do W?”, “*Why* did the system not do X?”, *What If*: What would the system do if Y happens?”, and *How To*: “How can I get the system to do Z, given the current context?”
- **HOW? - Types of the explanation** When the goal and the content of the explanation are clear, the question arises of how the explanation should look. This includes the method to generate the explanation (e.g., using an algorithm like LIME Ribeiro et al. (2016)) and the way to present it to the user (e.g., visualisation supported by textual explanations).

They highlight that the *Why*, *What*, and *How* of explanations have to fit user needs. They differentiate three types of users: developers & AI researchers, domain experts (e.g., clinicians), and lay users. Customising explanations to the target user group is the grand challenge of HC-XAI (Ribera & Lapedriza, 2019).

In the description of DARPA’s¹ planned XAI program, Gunning and Aha (2019) illustrate a three concepts approach to develop XAI. First, they present an explanation framework that divides XAI systems into an *explainable model* (e.g., DNN) and an *explanation interface* (e.g., interactive visualisations, diagrams, show-and-tell explanations). The explainable model represents a model that produces a classification, recommendation or decision that is explainable (e.g., post-hoc explanations for DNN). The explanation interface provides an explanation of the XAI system “[...] that justifies its recommendation, decision, or action” (Gunning & Aha, 2019, p. 50). Second, to describe these two components of the XAI system, the

1: Defense Advanced Research Projects
Agency

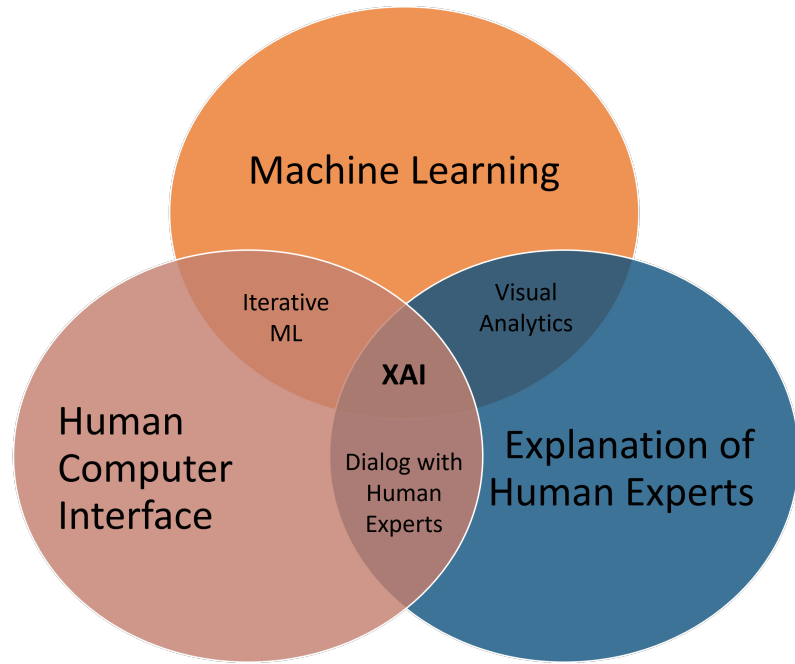


Figure 9.1: Involvement of different research areas to create XAI. Figure adapted from Loyola-Gonzalez (2019)

explanation framework also includes the user, who makes a decision based on the explanation presented. Finally, to assess whether the explanation was helpful for the decision, Gunning and Aha (2019) point out that the effectiveness of the explanation should be measured by evaluating different variables (e.g., user satisfaction, mental model, task performance, trust). In their work, they combine the technical development of an XAI system with a psychological model for explanations.

Liao et al. (2021) present a question-driven design process for XAI. They describe four steps to designing an XAI system: (1) Question elicitation, (2) Question analysis, (3) Mapping questions to modelling solutions, and (4) iterative design and evaluation. All these steps have in common that they focus on the users of the XAI system. In step 1, user needs regarding XAI are specified by taking a close look at the application area of the AI, which questions are asked by the users in this context, and which answers they expect. In step 2, the questions found in step 1 are evaluated, for example, by clustering similar questions and sorting the questions based on priority. This is done to identify essential user requirements. In step 3, the most important questions are used to find a technical solution that can answer these questions. In the last step, the developed technical solution is evaluated with the help of the users and iteratively improved until all gaps are closed.

In their overview paper on white- and black-box approaches for AI, Loyola-Gonzalez (2019) points to the need for interaction between different research domains to design XAI successfully. They point to the explanation of human experts (i.e., ML experts and experts from the application domain), human-computer interface, and machine learning as three essential components for the design of XAI (Figure 9.1).

Delimitation & Contribution of this Dissertation The work presented provides a valuable basis for the HC-XAI concept presented in

this dissertation (see Chapter 12 on page 70). This concept also uses the elements *content* and *types* of explanation, which is described by Ribera and Lapedriza (2019). Furthermore, inspired by Gunning and Aha (2019), the division of an XAI system into an *explainable model* and an *explainable interface* is also found in the HC-XAI concept of this dissertation. Liao et al. (2021) points to the involvement of users when designing XAI systems, a perspective also central to this dissertation. In addition, the work of Loyola-Gonzalez (2019) highlights the relevance of different domains in the development of XAI. With these related works as a basis, the HC-XAI concept presented in this dissertation provides an interdisciplinary approach that goes beyond the previous creations. The novelty includes, in particular, the integration of the persona approach as a contribution of HCI and the investigation of psychological constructs (e.g., mental models, trust) using both quantitative and qualitative data. Furthermore, the developed ideas are brought into practical use by presenting a step-by-step approach based on the concept that can serve the empirical investigation of HC-XAI. Possible practical investigations are illustrated in this dissertation by creating exemplary surveys for developing personas in three application contexts of varying complexity (see Chapter 15 on page 85). In addition to the surveys, six experiments investigated the effect of different levels of interactive XAI on user perception in six experiments (starting in Chapter 27 on page 160).

9.2 Concepts to Develop Interactive XAI

Kim et al. (2021) deal with how classification and evaluation of different XAI procedures can be accomplished. For this purpose, they present four foundational attributes which should evaluate the possible XAI procedures:

- ▶ **Explicit explanation representation:** This includes post-hoc and explainable-by-design approaches. Here it is less about the type of XAI system but rather about semantically representing the explanations of the XAI system and making them accessible to users.
- ▶ **Alternative explanations:** Kim et al. (2021) assume that alternative explanations for the same prediction increase user confidence in the system.
- ▶ **Knowledge of the explainee:** Explanations should be adapted to the explainee, i.e., the person receiving the explanation.
- ▶ **Interactivity of the explanatory system:** Here Kim et al. (2021) point out that already the Socratic Dialogues² used the interaction between the explainer and the explainee to gain knowledge through interaction.

With the help of these four attributes, explanations can be put into four levels, which I have already explained at the beginning of this dissertation in the chapter Research Objectives (see Chapter 2 on page 5).

- ▶ **Level 0:** No explanation is provided by the AI system
- ▶ **Level 1:** AI system presents one explanation type
- ▶ **Level 2:** AI system presents more than one explanation type

2: “The Socratic Dialogue is a philosophical group dialogue in which the participants guided by a facilitator and a number of ground rules strive to reach a consensus in answering a fundamental question on the basis of a real-life example or incident with the purpose of achieving new insights.” (Knezic et al., 2010, p. 1105)

- **Level 3:** AI system considers user attributes (e.g., knowledge of the user)
- **Level 4:** AI system can create interactive explanations by communicating with the user

Level 4 represents the most complex type of explanation, with interactive explanations conveyed using natural language communication. These are comparable to people's explanations to justify their actions and decisions. Kim et al. (2021) point out that developing such an interactive HC-XAI system is an enormous challenge but that the presented attributes can help to take a step in this direction.

Neerincx et al. (2018) stated that for the successful development of XAI in human-agent teams, an integrative XAI approach is required. They present their perceptual-cognitive explanation (PeCoX) framework, which is generic (i.e., model-agnostic and domain-agnostic) and, therefore, can be used for various applications in the domain of human-agent interaction. Their approach consists of three phases:

- **ε generation:** refers to the agent. An AI behaviour engine receives input from the cognitive and perceptual XAI modules. Cognitive XAI generates explanations that explain why an agent chose a particular action. Because this approach focuses on the agent's behaviour, which is related to reaching a specific goal, this is also called goal-based explanation. Perceptual XAI deals with the agent's perception and consists of sub-symbolic explanations (e.g., confidence values and counterfactuals).
- **ε communication:** Similar to content and types of explanations of Ribera and Lapedriza (2019), this phase deals with the presentation of the explanation to the user as well as the content of the explanation.
- **ε reception:** This phase deals with how well the user understands the explanation.

ε refers to the concept of explanation and is based on the work of Tididi et al. (2015). Tididi et al. (2015) emphasises that finding a formal definition of an explanation ε that is accepted by different research domains is a challenging work that has not yet produced a generally accepted result. Instead, their paper presents an ontology design pattern that supports the representation of an explanation in different research domains (e.g., philosophy, psychology).

Similar to Neerincx et al. (2018), the literature review of Anjomshoe et al. (2019) focuses on *explainable agency*. Explainable agency focus on goal-driven AI for robots and agents. Compared to data-driven AI (e.g., DNNs for image classification tasks), goal-driven AI explains the behaviour of robots and agents. This behaviour is driven by goals these entities want to achieve. Their review revealed, among others, that robot and agent explanations are primarily communicated through text (47%), followed by visualisations (21%). In addition, they found that multimodal explanation communication strategies were rare in current work. Regarding the categories of explanations (i.e., what is communicated through an explanation), 26 papers use introspective explanations that give insights into the reasoning process of the robot or agent. Another frequently identified category is teaching explanations. Here, the purpose of the explanation is to teach the user about the concepts that a robot/agent

has learned. Based on the work of Neerincx et al. (2018), Anjomshoae et al. (2019) integrate their literature review into the PeCoX framework. Anjomshoae et al. (2019) state that this should serve as a road map for researching robot and agent explanations:

- **Explanation generation:** The integration of an explanation generation module is necessary to enable robots and agents to generate and communicate explanations to users. Explanation generation should take into account elements of personalisation and context awareness.
- **Explanation communication:** This refers to the transmission of the explanation to a user or another robot/agent. Anjomshoae et al. (2019) see the communication of multimodal explanations (e.g., mimic expression, visual, audio) as a promising approach to use and communicate the explanations to users in different contexts.
- **Explanation reception:** One goal could be to use explanations to explain the state of the robot/agent and thus provide a better understanding for the user. To measure whether the explanation contributes to a better understanding of the user, it is necessary to measure how efficient the explanation is and how the user perceives it.

Delimitation & Contribution of this Dissertation The interactivity of explanations is an essential topic in research on HC-XAI. The related work indicates that interactivity could be achieved, mainly through communication between machines and humans. However, how this communication is designed depends on the purpose of the explanation (e.g., teaching the user vs giving insights into the machine's reasoning process). While researchers largely agree that interactive explanations can benefit users, two questions arise, which this dissertation addresses: (1) What might such interactive XAI look like? and (2) how do end-users perceive it? This dissertation provides studies on different levels of interactive XAI. The interactive XAI systems presented in this dissertation are illustrated using various application purposes: cooperation & collaboration, education, and medical decision support. In six experiments, end-users perception in terms of mental models, trust, self-efficacy, and emotions towards different levels of interactive XAI is investigated.

10

XAI for Different Purposes & Scenarios

10.1 XAI for Cooperation & Collaboration	54
10.1.1 XAI for (Re)building Trust in Human-Agent Interaction	54
10.1.2 Interactive XAI Interfaces for Cooperation & Collaboration	56
10.1.3 Impact of XAI on Users in Cooperation & Collaboration	57
10.1.4 Delimitation & Contribution of this Dissertation	58
10.2 XAI for Education . . .	58
10.2.1 Analogue Teaching Techniques for XAI . . .	59
10.2.2 Gamification	60
10.2.3 Intelligent Tutoring Systems	60
10.2.4 Delimitation & Contribution of this Dissertation	61
10.3 XAI for Medical Decision Support	61
10.3.1 Context-based XAI for Decision Support Systems in Healthcare	62
10.3.2 Stakeholder Needs Regarding XAI for Decision Support Systems	62
10.3.3 Delimitation & Contribution of this Dissertation	64

This chapter answers the question “**What challenges do specific AI applications pose for XAI?**”.

Thereby parts of the chapter are based on the following published work:

- *Cooperation & Collaboration*
Hald, K.*, **Weitz, K.***, Rehm, M., & André, E. (2021). “An error occurred!” - trust repair with virtual robot using levels of mistake explanation. In K. Ogawa, T. Yonezawa, G. M. Lucas, H. Osawa, W. Johal, & M. Shiomi (Eds.), *Proceedings of the 9th international conference on human-agent interaction* (pp. 218–226). ACM. doi: [10.1145/3472307.3484170](https://doi.org/10.1145/3472307.3484170)[†]
- *Education*
Weitz, K., Schlagowski, R., & André, E. (2021). Demystifying artificial intelligence for end-users: Findings from a participatory machine learning show. In S. Edelkamp, R. Möller, & E. Rueckert (Eds.), *Proceedings of KI 2021: Advances in artificial intelligence - 44th german conference on ai, virtual event, september 27 - october 1, 2021* (pp. 257–270). Springer. doi: [10.1007/978-3-030-87626-5_19](https://doi.org/10.1007/978-3-030-87626-5_19)[‡]
- *Medical Decision Support*
 - **Weitz, K.** (2021). Towards Human-Centered AI: Psychological concepts as foundation for empirical XAI research. *IT-Information Technology*, 64(1-2), 71–75. doi: [10.1515/itit-2021-0047](https://doi.org/10.1515/itit-2021-0047)
 - Heimerl, A., **Weitz, K.**, Baur, T., & André, E. (2022). Unraveling ML models of emotion with NOVA: multi-level explainable AI for non-experts. *IEEE Transactions on Affective Computing*, 13(3), 1155–1167. doi: [10.1109/TAFFC.2020.3043603](https://doi.org/10.1109/TAFFC.2020.3043603)[†]

Overview papers about XAI, like the one from Adadi and Berrada (2018) highlight the importance of XAI for different application scenarios (e.g., transportation, healthcare, finance, military). Wolf (2019) see in evaluating XAI with the help of fictional scenarios the chance (1) to investigate unique requirements of scenarios (2) to understand the impact of XAI in complex scenarios. A similar view have Doshi-Velez and Kim (2017), who also say scenario-based XAI studies should be conducted with people. Doshi-Velez and Kim (2017) present a taxonomy with three stages to cluster empirical work in XAI (see Figure 10.1 on the next page). Stage one represents the basis of empirical research: *proxy tasks*. Here, no real humans are needed for evaluation because the focus is *functionally-grounded* (i.e., formal definition of interpretability). In this proxy-tasks¹, XAI methods are analysed using different formal definitions of explainability (Doshi-Velez & Kim, 2017). While these kinds of tasks are also relevant to guarantee

1: see the work of Arras et al. (2022) and Kakogeorgiou and Karantzas (2021) to get an impression of such evaluations

*Both authors contributed equally to this work

[†]Licensed under the [Attribution 4.0 International License \(CC BY4.0\)](https://creativecommons.org/licenses/by/4.0/)

[‡]Adapted by permission from Springer Nature Customer Service Centre GmbH: Springer Nature Lecture Notes in Computer Science book series (Demystifying Artificial Intelligence for End-Users: Findings from a Participatory Machine Learning Show, Weitz, K., Schlagowski, R., & André E.), ©2021 Springer Nature Switzerland AG (2021)

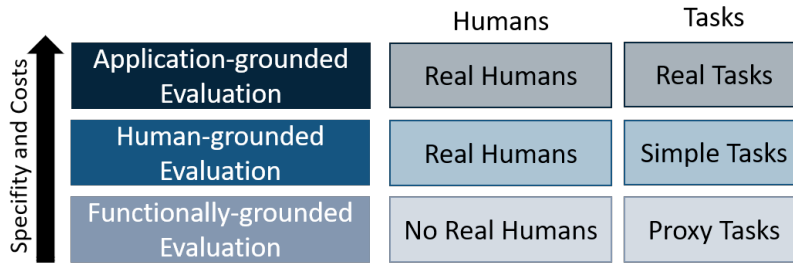


Figure 10.1: Taxonomy to cluster evaluation of XAI methods (illustration adapted from Doshi-Velez and Kim, 2017). The three evaluation steps include the presence or absence of human users and the complexity of the task. With the increasing complexity of the evaluation, the costs and specificity of the investigation and the application increase

that XAI methods work correctly, this dissertation focuses on evaluating XAI with human stakeholders (i.e., end-users) and their perception of XAI. Doshi-Velez and Kim (2017) refers to this type of evaluation that involves human as *human-grounded* and *application-grounded*. Human-grounded evaluation can be derived from the term; people participate in the evaluation in this step. The tasks they have to solve are simpler than in realistic scenarios. This kind of evaluation is chosen to test the general impression of XAI. For this, experiments with non-experts rather than domain experts are conducted. The evaluation is taken to a *application-grounded* level in the third stage. This includes the evaluation of XAI in real tasks with domain experts (e.g., clinicians for AI applications in healthcare).

Research indicates that humans tend to prefer a specific style of explanation. For example, Lombrozo (2007) conducted four experiments where participants had to rate explanations in a fictional story about alien illnesses. The results of these experiments revealed that people prefer simpler and more general explanations. Empirical research is needed to investigate whether such results, like from Lombrozo (2007), can be transferred into the field of XAI and whether XAI creates other or additional insights into human-machine interaction.

Nevertheless, empirical evidence of how XAI impacts users is rare. M. Langer et al. (2021) provides an overview of sources dealing with human stakeholders in the field of XAI. They show that most papers include work without empirical investigation with users (i.e., end-users, developers, deployers, and regulators).

Delimitation & Contribution of this Dissertation While researchers repeatedly emphasise the relevance of XAI for various application scenarios, many studies evaluating XAI tend to be technical (i.e., proxy tasks). Researchers demand the inclusion of users as a necessary next step, but so far, only a few studies are available that systematically investigate the impact of XAI on users. This dissertation addresses this problem and provides an evaluation of the perception of XAI using six experiments in a laboratory setting to investigate the general impact of XAI on end-user. Thus, this dissertation addresses the second step (i.e., human-grounded evaluation) in the taxonomy proposed by Doshi-Velez and Kim (2017) for evaluating XAI.

Since this dissertation deals with the empirical evaluation of XAI, the following related work for the three application purposes also focuses on empirical evaluations of XAI. These purposes address essential areas

2: for a definition of AI literacy, see subsection Beneficial Functions of Explanations in Chapter 5 on page 22

of AI usage: *Cooperation & collaboration* often deals with interactive AI in the form of robots or virtual agents. *Education* is a broad sector that has addressed computer science topics for many years. Here, XAI can be used to explain an AI system's inner workings and decisions to support AI literacy². Finally, a classical AI domain is the area of *decision support* where AI support humans in their decision-making process. I focus on the healthcare domain, where AI systems are already used.

10.1 XAI for Cooperation & Collaboration

3: agent mainly refers in these scenarios to autonomous acting systems like robots or virtual agents

Human-agent³ cooperation and collaboration is an expanding field of AI research. These interactions are characterised by highly dynamic, complex situations requiring a broad interaction skill set from the machine and knowledge of how to handle the machine from the human. The investigation of whether AI explanations are helpful in this context has been the subject of many research studies.

10.1.1 XAI for (Re)building Trust in Human-Agent Interaction

Failure, fault, or error are terms often used to describe a not intended behaviour of robots or agents (Honig & Oron-Gilad, 2018). A failure describes an "behavior or service being performed by the system to deviate from the ideal, normal, or correct functionality" (Brooks, 2017, p. 9). Failures are caused by errors in the system (e.g., mechanical errors) (Honig & Oron-Gilad, 2018). The causes of errors are faults. A fault is responsible for an error (e.g., a loose screw) (Honig & Oron-Gilad, 2018). Honig and Oron-Gilad (2018) underline that in human-robot interaction, robot failures are common due to complex and often unstructured situations. In studying robot reliability, Salem et al. (2015) found that a robot is perceived as less trustworthy after a failure. Despite the decrease in trust, participants followed the robot's instructions. When an error in human-robot interaction occurs, users often cannot understand how the error occurred, how to fix it, and how to avoid it in the future. This leads to performance degradation as well as mistrust (Holliday et al., 2016). Honig and Oron-Gilad (2018) present their Robot Failure Human Information Processing Model. Their model deals with the mitigation of robot failure. Their model approaches robots from a user-centred perspective by considering users' cognitive abilities (Honig & Oron-Gilad, 2018, p. 8):

- ▶ **Communicating Failures:** The focus here is on how a robot failure is communicated. This may involve visual indicators (e.g. using lights), screens presenting additional information, or audio and speech (e.g. sounds, up to verbal communication).
- ▶ **Perception & Comprehension of Failures:** This aspect deals with the user's perception of the failure. Only when the user perceives the failure can countermeasures be initiated by the user. It is important that users understand the robot error and react to it here. An influence on how users deal with a robot failure is also related to their mental model. If they have an incorrect model of the robot's behaviour, this can hinder problem-solving after an failure.

- **Solving Failures:** User motivation is needed to resolve a robot failure. If this is given, users must decide what they will do to solve the failure and translate it into action.

On this basis, mitigation strategies can be developed for the robot or for the user to deal with robot failures. Here, Honig and Oron-Gilad (2018) describes three different strategies: (1) setting expectations (i.e., the user can correctly assess the failure potential of the robot), (2) communicating properly (i.e., how and what is communicated in a failure situation), and (3) asking for help (i.e., robot request help from a user).

Besides these mitigation strategies, providing explanations about the robot's behaviour and decisions could also help increase user trust. de Visser et al. (2020) give an overview of trust in human-robot teams. They emphasize the importance of trust in successful human-robot tasks. They state two important directions for (re)building human-robot relationships through trust calibration: (1) predictive and preventative trust to minimize the risk of trust violation and (2) reactive and reparative trust when the trust is violated. One method for predictive and reparative trust that de Visser et al. (2020) propose is the usage of explanations. They suggest that explanations help calibrate trust by providing users with information about the robot. The work of de Visser et al. (2020) give a detailed overview of relevant concepts of trust in human-robot collaborations but misses empirical investigations. Similar is the work of Hellström and Bensch (2018) in which a theoretical model for human-robot communication and some possible example scenarios are proposed but not investigated in a user study.

XAI approaches are also discussed to gain insights into the behaviour and goals of the agent. For example, Sheh (2017) recommends the usage of decision trees in explanatory human-robot dialogues. They argue that decision trees have two major advantages: (1) the intrinsic explainability of if-then-else rules and (2) attribute-centric⁴ type of explanations. They aim to use decision trees in a human-robot dialogue and enrich it with visual explanation modalities like histograms. The work of Das et al. (2021) used hand-scripted explanations to investigate three different failure types (i.e., navigation, arm motion planning, and object detection) of an industrial robot arm. They differentiate four different contents of explanation (Das et al., 2021, p. 354):

- **Action Based:** e.g., "Robot could not find the object."
- **Context Based:** e.g., "Robot could not find the object because the object is hidden from view."
- **Action Based History:** e.g., "The robot finished scanning objects at its current location, but could not find the desired object."
- **Context Based History:** e.g., "The robot finished scanning objects at its current location, but could not find the desired object because the desired object is hidden from view."

In an online user study, they presented videos of the robot executing a task. After three videos where the robot solved the task, they saw six failure simulations without explanations. Afterwards, they asked participants to identify the cause of the error and possible solutions. Finally, additional 12 failure videos with explanations were presented. Again, participants had to determine the cause of the error and possible solutions. The results

4: explanation is not for one specific example but rather for attributes that are most relevant for different actions

indicate that context-based explanations are most appropriate for end-users to identify the reason for a robot error and come up with a helpful solution for this situation.

10.1.2 Interactive XAI Interfaces for Cooperation & Collaboration

I already mentioned that research like the work of Anjomshoae et al. (2019) and Neerincx et al. (2018) highlight the chance of using interactive XAI, especially in the context of agents and robots. In the field of human-agent interaction, Tenhundfeld et al. (2021) investigated the impact of virtual personal assistants (VPA) (e.g., Apple's Siri, Amazon's Alexa, Google Assistant) on users' mental models and trust. In particular, the researchers investigated the perception of interconnected VPA systems. Only participants who stated they used such assistants were allowed to participate in their online survey. In this survey, they were asked 13 questions about their experience with virtual assistants (e.g., "How often do you use [VPA]?", Tenhundfeld et al., 2021, p. 5) and their mental models about these systems (e.g., "Does the [VPA] you talk to know everything that the [VPA] others talk to, know?", Tenhundfeld et al., 2021, p. 5) as well as their trust in the system (e.g., "How much do you trust [VPA] based on your last interaction?", Tenhundfeld et al., 2021, p. 5). These users' self-reports revealed no consensus on whether VPAs are systems that everyone can interact with in the same way or whether their VPA is a unique system. Regarding the functionality and features of the VPAs, users could describe them with confidence. However, users did not have a common understanding of how VPAs worked, indicating that users lacked an understanding of how these systems function. Furthermore, the frequency of VPA usage does not change this lack of understanding. These results indicate the need to support users to gain a better understanding. One way to achieve this is to present explanations during human-machine interaction. Frameworks are needed to provide these explanations.

Rehse et al. (2019) introduced a framework to provide local and global explanations for a Smart-Lego-Factory in an Industry 4.0 setting. In their showcase example, they provide the following explanation types:

- ▶ **Feature-based explanations:** e.g., a diagram representing the impact of different features on the prediction
- ▶ **Rule-based explanations:** e.g., "If Crane position = TRUE AND Elapsed time >150 THEN "Positive"" (Rehse et al., 2019, p. 186)
- ▶ **Textual explanations:** e.g., "Crane position is the most important factor. Changing its value from True to False decreases the probability of the positive outcome by 30%." (Rehse et al., 2019, p. 186)
- ▶ **Standard ML-metrics:** e.g., accuracy, F1 score, AUC

Not only seeing the prediction of the DNN but also understanding it helps users interact adequately with the system and, for example, improve the system's predictions (Rehse et al., 2019). Sakai et al. (2021) proposed a framework for generating explanations in human-agent collaboration with the help of a Markov decision process (MDP). They included methods to gain knowledge about the agents' goals and their current understanding

of a specific situation. Sakai et al. (2021) evaluated their framework with two user experiments where users had to understand the agents behaviour and its goals⁵. In the first one, they investigated whether explanations that highlight the subgoals of an agent are sufficient for users (e.g., “Could you predict the route that the agent took?”, Sakai et al., 2021, p. 9). They found that presenting the agent’s subgoals or all agent states helped the users correctly predict the agent’s route compared to showing only random states. In the second experiment, they investigated users’ understanding of the agent’s actions (e.g., “Did you understand the reason for the next action taken by the agent?”, Sakai et al., 2021, p. 11). Here, they found that users wanted to know details of the route taken by the agent, resulting in non-significant results between showing all key points or presenting only the first key point. In their work, the authors point out that their approach needs to be evaluated in studies with real-world scenarios. The main challenge for these studies will be that the world model is designed so humans can interpret it. In addition, they highlight that a real-world evaluation would provide new insights into whether user understanding and the resulting acceptance of an agent explanation can be achieved with their framework.

5: the agent had to find and take a key, walk to a door and open it to get a reward

Garcia et al. (2018a) proposed MIRIAM⁶, a multimodal interface to explain the actions of an autonomous vehicle robot using a decision tree. The interface includes a chat interface where users can ask *why* and *why not* questions.

6: short for *Multimodal Intelligent inteR-action for Autonomous systems*

10.1.3 Impact of XAI on Users in Cooperation & Collaboration

In follow-up work, Garcia et al. (2018b) used the MIRIAM interface for a user study⁷. In three experimental conditions, they investigated varying levels of soundness (i.e., depth of details of an explanation) and completeness (i.e., amount of relevant information for the explanation). They found that providing highly complete and sound explanations leads to the highest user trust. Users’ mental model is built over time and varies between the conditions. The online study in human-robot interaction of Wang et al. (2016a) shows that explanations are beneficial for users to understand the decisions of a low-ability robot. In addition, explanations in human-robot interaction are only helpful when they help users in the decision-making process (i.e., explanation helps users be sure about what decision to make). In human-robot collaboration, Nikolaidis et al. (2018) conducted a collaborative experiment where a robot and a human had to solve a table-carrying task together. They found that verbal commands from the robot were most effective in forcing users to adapt to the robot (100% vs 60% in the non-verbal condition). When a robot uses verbal commands, *why* actions (e.g., “it thought it knew the best way of doing the task”) lead users to question the robot’s trustworthiness. However, when explanations were included (e.g., “I need to be able to see the door with my forward-facing camera”), 95% of the users adapted to the robot.

7: all participants were familiar with autonomous vehicles, e.g., software developers, meaning no end-users

10.1.4 Delimitation & Contribution of this Dissertation

Cooperation and collaboration are exciting topics for HC-XAI research in many respects: these are usually complex situations in which humans and machines have to solve a task together. However, errors can disrupt the interaction and, in the worst case, lead to a (permanent) loss of trust. The authors highlight the importance of developing robot error mitigation strategies for successful interaction and appropriate trust. Explanations can help to foster trust in interactions between humans and machines. However, it needs to be clarified whether explanations help to restore trust when errors occur during a user's interaction with a robot or an agent. Therefore, this dissertation looks at how explanations can be communicated in interactive settings where AI errors occur and how users perceive these explanations in terms of mental models, trust, self-efficacy, and emotions. For this, two interactive XAI systems are evaluated in two experiments (i.e., VR-Robot Study and Conversational AI Study) with end-users (starting in Chapter 27 on page 160) to investigate the general impact of verbal explanations on end-users. The unique feature of these experiments is that study participants do not evaluate explanations of a system presented to them (e.g. by watching videos of the system). Instead, participants have to solve a task with their AI counterparts, meaning that the explanations are presented during interaction with the robot or the agent, thus enabling investigation of the impact of explanations when conducting a cooperative and collaborative task.

10.2 XAI for Education

Using AI explanations in an educational setting could be a promising approach to (1) help to gain knowledge about AI, referred to as *AI literacy* and to (2) support learning about a topic in general. Long and Magerko (2020) claim that educators need to focus more on AI and XAI education. This is necessary since AI is impacting more and more every area of our lives and forces us more and more to train competencies and skills to handle these systems successfully (Long & Magerko, 2020). Khosravi et al. (2022) emphasise the relevance of different research disciplines (i.e., AI, HCI, and cognitive and learning sciences) for the development of XAI for education (see Figure 10.2 on the facing page). In addition, Khosravi et al. (2022) present a conceptual model for XAI in education. Their XAI-ED framework considers six dimensions regarding explainability for developing educational AI tools (Khosravi et al., 2022, p. 4):

- ▶ **Stakeholders:** Who are the main stakeholders? (e.g., parents, teachers, policy makers)
- ▶ **Benefits:** What are the main benefits? (e.g., AI literacy, accountability & trust)
- ▶ **Pitfalls:** What potential pitfalls need to be considered? (e.g., needless use of complex models, misconceptions)
- ▶ **Approaches:** What approaches are used for presenting explanations? (e.g., global vs local explanations)
- ▶ **Models:** What AI models are commonly used? (e.g., rule-based models, decision trees)

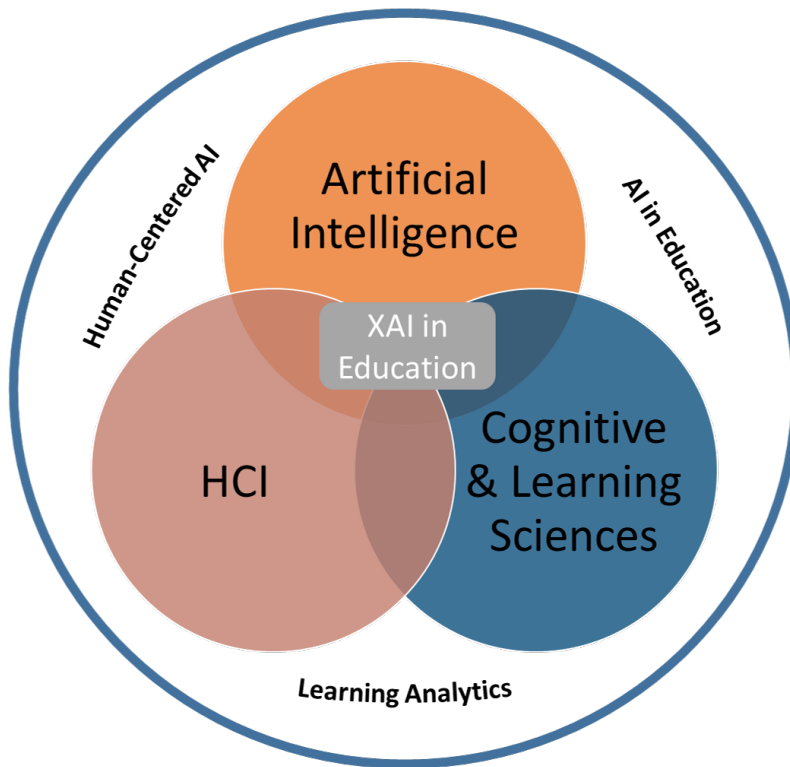


Figure 10.2: Related research fields for the development of XAI for education. Figure adapted from Khosravi et al. (2022)

- **Design:** How can educational AI tools be effectively designed? (e.g., user experiences, theory-driven design)

Khosravi et al. (2022) integrate users and the educational AI system, which consists of AI interfaces and AI models, into their framework, resulting in a very similar XAI framework to that of Gunning and Aha (2019). Gunning and Aha (2019) differentiate between an explainable model and an explanation interface (see Chapter 9 on page 47 for a more detailed description of these components). The following presents different approaches to integrating (X)AI in educational settings.

10.2.1 Analogue Teaching Techniques for XAI

Analogue teaching refers to imparting computer science concepts in an analogue, real-world-based way to learners (Kapaniari, 2020). Bell et al. (2009) adapted this idea in creating analogue material to teach introductory Computer Science principles like binary counting, search algorithms, and image representation⁸. The idea of creating analogue, *unplugged* material to teach people about science topics was extended by Lindner et al. (2019) for AI topics. They developed five AI-related unplugged activities to help students understand the basic concepts of classical and modern machine learning topics (i.e., decision trees vs neural networks/deep learning). For GANs, Virtue (2021) also presented an unplugged approach by challenging three user groups (i.e., Real, Generator, and Discriminator). While the Real and the Generator group have to draw animals on cards (only the Real group knows about the relevant attributes), the Discriminator group has to decide which cards came from the Real or Generator group. In addition, the Discriminator group can give one

8: the Computer Science Unplugged material can be found at <https://www.csunplugged.org> (last accessed on 29.09.2022)

suggestion on each card on how the drawing could be more realistic. This information is then given back to the group who draw the card. Bueff et al. (2022) presented an approach for teaching about XAI with the goal that participants can understand XAI methods and their respective challenges. They used classical didactic methods (e.g., presentation of the topic by the teacher, discussion groups, and assignments) in ten lectures and four tutorials. They collected feedback from 16 participants during the course. Over 87% of the students were not familiar with the topic of XAI. The participants' feedback shows that the designed course helps teach XAI to students without experience with XAI. While this work demonstrates the usefulness of pedagogical methods for teaching about XAI, its use is more limited to a school setting.

10.2.2 Gamification

Gamification is defined as “the use of game design elements in non-game contexts” (Deterding et al., 2011, p. 10). Ibáñez-Espiga et al. (2014) presented a Q-learning game platform to engage students to learn the programming Language C. Their game-based approach improved students' engagement and learning outcome. Including XAI methods in game-based evaluation is used by Fulton et al. (2020). They reported the idea of using *Games with a purpose* (GWAP)⁹ combined with XAI. Their study presents the idea of a multiplayer GWAP for XAI for Image Recognition. Here, one player, the *explainer* chooses a plain image and the top visual explanations¹⁰ that they assume which are helpful for two other players, the *guessers* to identify the image. The guessers are presented with the selected XAI visualisations and have to guess the plain image. In addition, the guessers provide information on how they interpret the visualisations and which images helped them with their guess (Fulton et al., 2020). Collecting quantitative and qualitative data should help gain deeper insights into how users select and interpret XAI visualisation. As the paper by Fulton et al. (2020) is a late-breaking report, a detailed evaluation of the study was not yet available. The usage of virtual agents in such gamification environments helps to increase enjoyment and self-efficacy in students (Jin, 2010). This effect, known as *persona effect* was already reported in the 1990s by Lester et al. (1997). The results of their study show the positive impact of virtual characters in learning environments. Students found well-designed virtual agents helpful, credible, and entertaining (Lester et al., 1997). The positive attributes of virtual agents could support XAI in educational settings in a similar way.

9: this type of games generate user data as part of the gameplay (von Ahn & Dabish, 2008). Example of a Google GWAP for image labelling: <https://crowdsourcing.google.com/> (last accessed on 29.09.2022)

10: The authors used the Lucid approach to generate XAI visualisations: <https://github.com/tensorflow/lucid> (last accessed on 29.09.2022)

10.2.3 Intelligent Tutoring Systems

Teaching about the usage of computers is quite an old domain starting with computer-based training and computer-aided instructions (Beck et al., 1996). They were the first digital attempt to help users in learning scenarios. The shortcoming of these systems was that they could not adapt to the user and their needs (Beck et al., 1996). *Intelligent Tutoring Systems* (ITS) address this problem by adapting to user needs. ITS has been used since the 1960s to support learners and teachers in independent learning tasks (Hoffman & Clancey, 2021). ITS uses pedagogical aspects (i.e., how to teach) and information about the learner (Beck et al., 1996).

Hoffman and Clancey (2021) propose to use insights from the research domain of ITS for XAI. They highlight that ITS could be a helpful tool for users to understand the capabilities and limitations of an AI system. Conati et al. (2021) presented an empirical study combining ITS with XAI. Their goal was to investigate *when* and *how* explanations should be presented to users during an educational algorithmic learning task. They used an interactive simulation to help students to understand the working of the constraint satisfaction algorithm using an ITS that provide hints to the students (e.g. “Use Reset less frequently”). Regarding this ITS, they offered three explanation categories, answering to the users (1) Why they delivered a hint, (2) why they were predicted to be lower learning, and (3) Why are the rules used for classification. For these answers, they used text and illustrations. Their results indicate that explanations are helpful in ITS, but an *one explanation fits all* approach is not the best learning strategy. Instead, user needs are essential to include in educational XAI designs. They found that participants with a low need for cognition¹¹ paid significantly less attention¹² to the presented explanations than participants with a high need for cognition.

11: Cacioppo and Petty (1982, p.130) defined need for cognition as “individual differences in people’s tendency to engage in and enjoy thinking”

12: attention was measured in seconds participants looked at the explanations

10.2.4 Delimitation & Contribution of this Dissertation

In education, the use of AI is an extensively researched topic. The idea of using AI to adapt learning processes to users is being taken up in ITS, for example. Other approaches deal with the transfer of AI knowledge to users. An XAI system that can communicate with users is a valuable approach here. This dissertation addresses interactive XAI in the context of education (starting in Chapter 28 on page 190) by presenting a system that makes use of visual explanations generated by the algorithm LIME to explain the decision of an AI-based speech recognition system to end-users (Gloria Study). The novelty of this approach compared to the previous work regarding XAI is that a virtual agent is used as a communication bridge between the visual explanations of a DNN and the end-user. The Gloria Study investigates the impact of a virtual agent in combination with XAI visualisations on end-users trust, XAI and agent perception. Building on this, a participatory ML-show is presented in this dissertation, designed to teach large groups of end-users about the functionality, capabilities, and limitations of DNNs (Museum Study). Our approach in the Museum Study is unique because, compared to previous work focusing on individuals, we address DNN education for large groups of end-users. Moreover, this study represents one of the few works that go out of the classroom into a broader educational sector (i.e., a museum).

10.3 XAI for Medical Decision Support

Decision support systems have a long tradition in AI research. Especially in the healthcare domain, these systems have a long research history, as one of the early rule-based decision support systems like MYCIN (Shortliffe & Buchanan, 1975; Shortliffe et al., 1975) show (more details about MYCIN can be found in section Rule-based Systems in Chapter 4 on page 16). The idea for developing MYCIN and subsequent clinical

decision support systems is to assist clinicians in medical decision-making (Berner, 2007). In addition, medical decision-making can be helpful for patients (i.e., patient decision support using personal health records) and for administrative staff (e.g., documentation, cost reduction) (Sutton et al., 2020). While these systems work successfully in experimental settings in the lab, they mostly fail in practice due to the poor contextual fit (Yang et al., 2019) and emerging ethical and legal issues (Lucieri et al., 2020). Today, DNN achieve remarkable success in different medical image classification tasks, for example, in dermatology (Esteva et al., 2017) and radiology (Nam et al., 2019). In these examples, the DNN performance was comparable to or even better than human medical experts. Nowadays, the challenge is integrating these highly performant DNNs into an XAI design so that they can be beneficially used by medical personnel (e.g., for diagnosis of diseases) and their patients (e.g., to understand the diagnosis) in various contexts.

10.3.1 Context-based XAI for Decision Support Systems in Healthcare

An example of context-based decision-making in the medical domain is detecting emotional states. Here, AI-based systems promise to support physicians and nurses and identify the emotional state of patients to adjust their treatment in the best possible way. The detection of pain using facial information of patients (for an overview, see Hassan et al., 2019) is a pursued approach in ML research. Another health-related application scenario is the explanation of diagnoses (e.g., by using X-ray images) (Tjoa & Guan, 2020). The usage of DNN for those image analysis tasks in the medical domain seems to be beneficial to support clinicians (Balaji & Lavanya, 2019). To generate explanations for those image analysis tasks, XAI relies on visual highlighting of important areas of the input data using XAI visualisation methods like LIME (Ribeiro et al., 2016) or LRP (Bach et al., 2015) (see a detailed description of these methods in Chapter 5 on page 22). Tjoa and Guan (2020) give an overview of a broad set of model agnostic and model-specific XAI methods. They show that there are almost no empirical studies with humans investigating the effect of these XAI visualisations. They are urgently needed in the medical domain to ensure the efficient and safe usability of AI in this context.

10.3.2 Stakeholder Needs Regarding XAI for Decision Support Systems

While various XAI algorithms deliver information about a DNN, these methods are often not easily interpretable for a non-expert. They can not be used out-of-the-box as an explanation (Stieler et al., 2021). To investigate the needs of different stakeholders in healthcare scenarios, Gerlings et al. (2021) conducted a case study in a start-up that developed AI decision tools in the healthcare domain. The authors' investigation was done on the X-ray AI decision tool the start-up is developing. Their investigation used a combination of online workshops and semi-structured interviews with seven company employees and collected written documents (e.g., PowerPoint presentations and business proposals). The results show that there is not a *prototypical user* for XAI. Instead, the need for XAI is related

to the different concerns of different stakeholders. However, the authors noted that they did not interview the application's users (i.e., clinicians). This was different in the semi-structured interviews of Xie et al. (2019). Here, six physicians from various fields of application (i.e., pathology, orthopedy, neurology, family physician, general physician, and cardiology) participated. With these interviews, Xie et al. (2019) wanted to investigate the research question of how medical professionals use their patients' data to make a diagnosis or treatment. Behind this lies the authors' assumption that clinicians find XAI more comprehensible when the system thinks and talks like them. Their interview results provide insights into which aspects play a role in clinicians' use of data (e.g., theoretical validity: to what extent symptoms can be confirmed with the known theories from medical literature). In addition, the physicians stated that they use a range of computer-based systems in their daily practice, especially for data collection and storage. To make greater use of such systems for diagnosis, the doctors stated two central prerequisites: the systems must provide a reason (i.e., why? explanation) for their diagnosis, and they must be able to make such decisions in a personalised way, i.e., take into account individual patient information.

Holzinger et al. (2017) highlight that medical professionals have to deal with a broad spectrum of heterogeneous data (e.g., images, biological data, text). They emphasise the urge for *explainable medicine*. When medical professionals work with AI systems that support them in medical decisions, they must have the opportunity to understand the AI's decision and decision process. The authors also point out that it is assumed that medical professionals always explain their own decisions. This is not the case: sometimes, they are unwilling or unable to express the reasons for their decisions. Asking clinicians about their impression of XAI was the idea of Tonekaboni et al. (2019). They conducted qualitative interviews to explore the clinicians' understanding of explainability and what they expect from ML in their daily work. In addition, they showed the doctors two hypothetical, interactive scenarios tailored to the doctors' specialities. In the Intensive Care Unit (ICU) task, an AI tool was presented that predicts the risk of cardiac arrest. In the Emergency Department (ED) task, an AI tool was presented that performs patient classification based on triage reports. The results show that physicians understand XAI as a confirming (i.e., justifying) tool for their diagnosis. In addition, physicians indicated that they would use an AI repeatedly if the AI predictions frequently matched their personal experience.

Holzinger et al. (2017) promote the benefits of such a *hybrid human-in-the-loop approach*. Here, the medical AI systems gain support from the knowledge of human experts. Bruckert et al. (2020) picks up on this idea in using *interactive machine learning* for medical AI systems. Their roadmap paper presents their Comprehensible AI (cAI)-transition-framework. Here they refer to mutual explanations of the LearnWithMe system (Schmid & Finzel, 2020), where medical experts can investigate a medical AI system's decisions by displaying explanations and being able to correct them. While the theoretical concept of the LearnWithMe system sounds promising, a broader empirical investigation involving users is missing.

Cai et al. (2019) presented an XAI system in the medical domain that was empirically evaluated. They presented a DNN for prostate cancer

13: here, pictures from the TCGA Research Network's Pathology image repository were used: <https://tcga-data.nci.nih.gov/docs/publications/tcga> (last accessed on 29.09.2022)

14: i.e., nucleus annotations in the image and percentage of these annotations, and classification result

diagnosis to 21 domain experts (i.e., pathologists). The evaluation took part in three parts: (1) *pre-test*: Semi-structured interviews were conducted to gain insights into the pathologists' expectations and previous experience regarding an AI diagnosis system. (2) *test*: Pathologists had to diagnose prostate cases with the help of an AI diagnosis system¹³. (3) *post-test*: Impression of the pathologists about further information on the need to work with the AI system. The qualitative evaluation of the results showed that clinicians' needs regarding XAI must be considered to use medical AI systems successfully. For example, participants stated that they are not only interested in the accuracy of the system but also the competencies of the system (e.g., how conservative/liberal the system decides) and knowledge about the basis for the implementation (e.g., whether strengths/weaknesses reflecting the human counterpart have been incorporated). In the user study of Evans et al. (2022), the impact of XAI in the context of diagnostic pathology was also addressed. Evans et al. (2022) used a two-source approach: (1) they conducted an online user study and (2) had semi-structured expert interviews with board-certified pathologists. In both, a Ki-67 quantification task was conducted where images of an AI output and the Ki-67 quantification¹⁴ were shown to participants. In addition, seven XAI methods (i.e., Saliency Maps (local), Saliency Maps (global), concept attribution, prototypes, counterfactuals (one-axis and two-axis), and confidence scores) were presented in the online survey and the interviews. Saliency maps were perceived as confusing or not useful, and concept attribution examples were perceived as intuitively understandable but not helpful. On the other hand, counterfactuals, confidence values, and prototypes were perceived as valuable and easy to understand. The results show that clinicians attach great importance to explanations that are easy to understand. The results show that clinicians attach great importance to explanations that are easy to understand. Evans et al. (2022) explain this result with the face-paced work in clinics. The doctors' statements that the support of an AI would help them save time strengthen this interpretation. Regarding the usefulness of the explanations, the doctors stated that they serve as *sanity checks* for the AI model, for example, when an explanation highlights an irrelevant part of an image.

10.3.3 Delimitation & Contribution of this Dissertation

XAI, in the context of medical decision support, is a vast research field currently receiving a lot of attention. Therefore, focusing on the design of HC-XAI here becomes necessary if these systems are to be used successfully by doctors or if the decisions of such systems impact patients and should be communicated to them. Previous work has focused on the needs of professionals (e.g. physicians) in this area. End-users are less in the focus of the present work. In addition, much of the related work shown is qualitative, using interview studies' methodology. This dissertation differentiates itself from the works mentioned here in that it focuses on investigating end-users (i.e., laypeople) in this context. End-users are a group affected by medical decision support systems and are the minor user group having insights into AI or the medical domain. Therefore, this dissertation addresses two important aspects that have received less attention in research so far: (1) The effect of XAI in a medical context on

trust, mental model, explanation satisfaction, cognitive load, and emotions on end-users. (2) The implementation of experimental studies under controlled conditions, whose data are analysed qualitatively and quantitatively (details see Chapter 25 on page 154). In Chapter 18 on page 108, end-users are interviewed exemplarily on their attitudes towards a mobile health app to gain insights about their fears, needs, and desires regarding XAI. In addition, this dissertation (starting in Chapter 29 on page 210) investigates which types of explanations end-users find helpful when using the NOVA software to explore a medical decision support system (i.e., CNN) for facial emotion recognition (NOVA Study) and how different types of visual explanations are perceived by end-users when presented with X-rays (Pneumonia Study).

Table 11.2 on page 68 gives an overview of the papers that deal with the human-centered *design of XAI* and was presented in the Related Work chapter. In Table 11.1 on the next page an overview of *user studies* presented in the Related Work chapters is given. It should be noted that these user studies listed are from the areas of cooperation & collaboration, education, and medical decision support, as these topics are central to this dissertation.

So, what are the insights of the related work for this dissertation?

A review of the papers shows that current work in the field of XAI can be grouped into one of two directions: (1) **XAI design concepts** including ideas and challenges regarding XAI in general or for different purposes but lack empirical evaluation in the form of user studies or (2) **XAI user studies** that investigating the impact of XAI on human users in specific applications but lack a conceptual framework integrating the various relevant components for HC-XAI design. This dissertation addresses both aspects by (1) proposing an *interdisciplinary concept for HC-XAI* and (2) using this framework for *AI application-based experiments* for cooperation & collaboration, education, and medical decision support in *human-grounded* evaluations. For these experiments, a *broad methodological spectrum* (i.e., a combination of quantitative and qualitative data) on the evaluation of *AI systems* (i.e., CNN and rule-based systems) with *end-users* is presented.

A broad community of researchers proposes the importance of a human-centered view in the development of XAI. Evaluating application-based scenarios seem a promising approach to creating XAI with a focus on user needs. The Related Work chapters presented research about XAI for three purposes where AI is used (i.e., cooperation & collaboration, education, and medical decision support). Investigating how a *good explanation* should look like, authors like Ribera and Lapedriza (2019) recommend focusing on the *goal*, *content*, and *types* of explanations. Gunning and Aha (2019) focus on distinguishing between the *explanation interface* and the *explainable model*. This work is the basis for the interdisciplinary HC-XAI concept presented in this dissertation. The concept extends these approaches in three aspects: (1) providing an interdisciplinary perspective combining content from HCI, psychology and (X)AI research, (2) presenting a step-by-step approach based on the concept that serves researchers as a guide for the empirical investigation of XAI and (3) the exemplary application of the concept in the development of personas for three application contexts as well as the investigation of interactive XAI in six experiments.

To create XAI with a focus on user needs, work on XAI for three different purposes (i.e., cooperation & collaboration, education, and medical decision support) was presented in the Related Work. It shows that empirical studies with end-users are rare. Instead, empirical research in this domain focus on domain experts (e.g., clinicians) or ML/AI experts. Therefore,

this dissertation evaluates the impact of XAI on end-users with neither ML expertise nor domain knowledge (e.g., patients). This dissertation not only investigates the effect of XAI on this user group in six experiments but also collects information about the potential stakeholders using surveys and generates prototypical users, so-called personas, based on the survey results.

Table 11.1: Overview of related work that tests XAI in user studies. The term “End-users” refers to people with no experience in AI/ML or in the investigated domain. In all other cases, the specific expertise of the evaluated user group is stated. Only the study of Gerlings et al. (2021) distinguishes between different stakeholders, their roles and needs regarding XAI (i.e., personas)

Paper	User Group	Dependent Variable	Study Design
Cooperation & Collaboration			
Das et al. (2021)	End-users	Task performance	▲
Garcia et al. (2018b)	ML-experts	Trust, Mental Model,	▲
Nikolaïdis et al. (2018)	End-users	Trust, mental model, adaptability	▲
Sakai et al. (2021)	End-users	Task performance, mental model, expl. satisfaction	?
Tenhundfeld et al. (2021)	End-users	Mental model, trust	▲
Wang et al. (2016a)	End-users	Cognitive load, trust, performance, understanding	▲
Education			
Bueff et al. (2022)	Students*	Experience, task performance	□
Conati et al. (2021)	Students*	Expl. satisfaction, attention, learning gains, perception of hints	△
Fulton et al. (2020)	AI interested people	Task performance, engagement	▲
Medical Decision Support			
Cai et al. (2019)	Pathologists	Mental model	◇ △
Evans et al. (2022)	Pathologists	Mental model	◇ ▲
Gerlings et al. (2021)	ML experts	Mental model	◇
Tonekaboni et al. (2019)	Clinicians ^o	Mental model	◇
Xie et al. (2019)	Clinicians	Mental model	◇

* = with Computer Science knowledge, ^o = familiar with ML

◇ = interview, □ = field, △ = lab, ▲ = online, ? = not explicitly stated

Table 11.2: Listing of related work about human-centered XAI designs that discuss requirements, challenges, and possible designs of XAI. The work is clustered after whether they are related to one of the three purposes (cooperation & collaboration, education, and medical decision support) that are the focus of this dissertation or provide general ideas regarding XAI

Paper	Key Message
General	
Anjomshoe et al. (2019)	Literature review and resulting roadmap for goal-driven XAI in human-agent interaction
Doshi-Velez and Kim (2017)	Presentation of a three-step approach for clustering investigations of XAI
Gunning and Aha (2019)	Presentation of an XAI framework, including an explainable model, explanation interface and the user
Kim et al. (2021)	Sorting XAI into four levels with the help of foundational attributes
Liao et al. (2021)	A four-step question-driven design process with a focus on user needs
Loyola-Gonzalez (2019)	Presentation of three research domains (i.e., ML, human-computer interface, and explanation of human experts) that are relevant for XAI design
Neerincx et al. (2018)	Presentation of the generic PeCoX framework to generate, communicate, and present explanations to users
Ribera and Lapedriza (2019)	Three relevant aspects of XAI: goal, content, and types of explanation
Cooperation & Collaboration	
de Visser et al. (2020)	Presentation of an integrative model for human-robot trust calibration with methods for explanation (i.e., confidence values & contrastive explanations)
Hellström and Bensch (2018)	Presentation of a goal-driven model for human-robot interaction that helps users to understand the robot's state of mind
Rehse et al. (2019)	Presentation of a full automated factory prototype using feature-, rule-, text-based XAI methods
Sheh (2017)	Overview of techniques for explainable robots with a strong focus on expert systems
Education	
Hoffman and Clancey (2021)	Suggestion to combine ITS with XAI to benefit from both approaches
Khosravi et al. (2022)	Presentation of the XAI-ED framework considering attributes of users and educational AI systems and consisting of six key questions for XAI in education
Medical Decision Support	
Bruckert et al. (2020)	Presentation of a research roadmap for cAI-transition-framework with a focus on mutual explanations (i.e., cooperative & interactive explanations)
Holzinger et al. (2017)	Overview of current state-of-the-art XAI methods for images, *omics data, and text in medicine
Schmid and Finzel (2020)	Introduction of the concept of mutual explanations for medical decision support by combining DNN and interpretable models by design (i.e., Inductive Logic Programming)
Tjoa and Guan (2020)	Overview and categorisation of different XAI methods for medicine

IV. INTERDISCIPLINARY CONCEPT FOR HUMAN-CENTERED EXPLAINABLE AI

12

Overview

From the Related Work chapters, it becomes visible that XAI research mainly focuses on one of two things: (1) presenting design concepts that deal with ideas or recommendations on how to design HC-XAI or (2) evaluating the impact that XAI has on users, focusing on the evaluation of XAI or the evaluation of the needs of different users. The literature search also revealed that studies on XAI rarely emphasised end-users. This dissertation addresses these gaps by creating an interdisciplinary concept for HC-XAI that is then be used (1) to investigate the needs of end-users of XAI systems and (2) to investigate the impact of different levels of interactive XAI on end-users mental model, trust, self-efficacy, cognitive workload, and emotions.

Creating XAI systems is one of the main goals of HCAI. Riedl (2019, p. 3) defines HCAI as “a perspective on AI and ML that intelligent systems must be designed with an awareness that they are part of a larger system consisting of human stakeholders, such as users, operators, clients, and other people in close proximity.” Therefore, it is necessary to take this human stakeholder’s abilities, beliefs, and perceptions into the design of explanations of AI. Riedl (2019) distinguishes two directions of HCAI:

- ▶ **AI-systems understanding humans** To successfully interact with humans, AI must be able to understand human behaviour, intentions, and beliefs. To achieve this, Riedl (2019) says that AI must acquire *commonsense knowledge*. This can be *declarative* (e.g., cars drive in the right or left lane - depending on the country) or *procedural* (e.g., a waitperson will not bring the bill until you ask for it).
- ▶ **Humans understanding AI-systems** To guarantee a successful usage of AI and to prevent misuse, users have to understand the AI system they are interacting with. Unfortunately, current machine learning approaches (e.g., DNN) used in AI systems are not interpretable out of the box. In particular, Riedl (2019) points out that already existing XAI methods (e.g., LIME, LRP) alone are insufficient to make DNN decisions understandable to non-experts.

In this dissertation, the second aspect, **humans understanding AI-systems**, will be addressed in particular. In doing so, I consider the challenging task of using existing approaches to design different levels of interactive XAI. These will then be presented to end-users in six experiments. Here, my investigations are threefold (see Figure 12.1 on the facing page): First, focusing on the interaction between humans and AI by investigating a robot and a virtual agent in one cooperative and one collaborative task¹. Second, examine the impact of specific XAI algorithms on end-users perception. Third, combining a virtual agent and XAI visualisations. Here, visualisations generated by the XAI algorithm LIME benefit from a natural interaction through an agent representation and communication to users using speech. In addition to the experiments, a

1: *cooperative*: human and machine working on different objects to reach a goal; *collaborative*: human and machine working on the same object (Roschelle & Teasley, 1995)

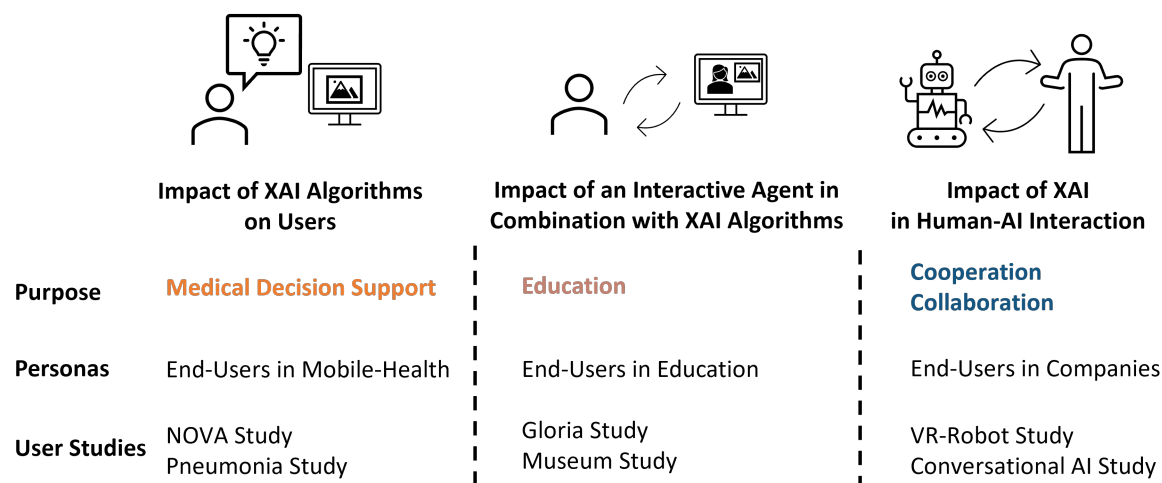


Figure 12.1: Three different levels of interactive XAI are examined in this dissertation: The effect of specific XAI methods (e.g., LIME, LRP, counterfactuals) on end-users perception is investigated in two experiments (left). For education, a hybrid version by combining a virtual agent with an XAI algorithm is investigated (middle). The impact of XAI in human-AI interaction (right) is investigated in two experiments where end-users have to solve a cooperative or collaborative task. For every level, different purposes are investigated (coloured in blue, rosé, and orange). For every purpose, surveys to investigate personas and user studies to investigate the impact of XAI were conducted

template for creating personas in the context of HC-XAI will be presented. The usage of the template is illustrated in three surveys.

13

Presentation of the Interdisciplinary HC-XAI Concept

13.1 Outline	72
13.2 Explanation Design . .	73
13.2.1 Explanation Interface . .	74
13.2.2 Explainable Model	76
13.3 User Evaluation	78
13.4 Five-Step Approach to Develop HC-XAI	79

13.1 Outline

Why is the HC-XAI concept presented in this dissertation interdisciplinary? Because it takes into account different perspectives from different research domains and uses their strengths as a basis for HC-XAI. The importance of investigating psychological elements (e.g., trust, mental models) and the design of explanations for XAI in white- and black-box approaches have already been shown in the Background and Related Work chapters. In addition to these components, HCI, with its human-centered user design and the design of prototypical personas, also has a place in the interdisciplinary concept (see Chapter 15 on page 85 for more details). Likewise, the interdisciplinary HC-XAI concept reflects the division into *explainable model* and *explanation interface* already introduced by Gunning and Aha (2019). The concept of this dissertation further divides the explainable model into content and type of explanation to distinguish between the content component (i.e., what is explained) and the type of explanation (i.e., how it is explained). In addition, the concept includes the four central elements (i.e., end-user, goals, contents, and language) as stated in the work of Vilone and Longo (2021) (see Figure 13.4 on page 79 for more details). The question of “Who is the receiver of the explanation?” is represented in the concept with the component *User Evaluation*. The investigation of this question with a focus on end-user can be found in the part End-Users of Human-Centered Explainable AI (see Chapter 15 on page 85). Focusing the goals of an explanation (i.e., “What questions should be answered?”) is related to the content of the explanation (i.e., “What information should be contained in?”). Both are addressed in the concept within the component *Explanation Design*. The investigation of different explanation types and contents in interactive XAI is presented in Chapter 27 on page 160 (cooperation & collaboration), Chapter 28 on page 190 (education), and Chapter 29 on page 210 (medical decision support). In detail, the interdisciplinary concept consists of the following three components:

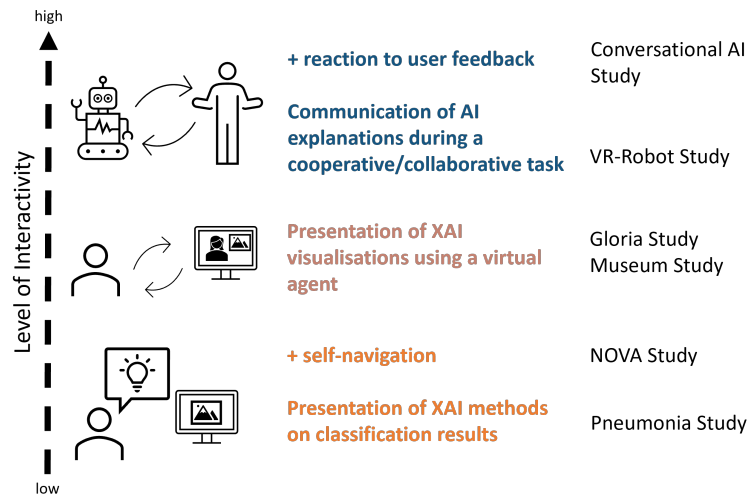
- ▶ **AI System** Describes the system for which the explanation is generated. Here, classical white-box approaches, the newer black-box approaches, or a hybrid approach combining both can be used. In this dissertation, as a representative of knowledge-based white-box techniques, rule-based systems are used. A representative of data-driven black-box approaches, CNNs are used. For these CNNs, a variety of different XAI methods are examined in the experiments of the dissertation. Rule-based systems offer the advantage that they can be quickly deployed in complex scenarios, such as human-AI cooperation and collaboration, to gain initial empirical data that can be transferred and extended (when needed) to DNN at a later stage.
- ▶ **Explanation Design** As we saw in the Related Work chapter, there are many possible ways to design HC-XAI. Different types and

13.2.1 Explanation Interface

Why should empirical user studies on HC-XAI be conducted? The Background chapter about explanations (see Chapter 5 on page 22) shows that there are already extensive findings in human-human explanations that could be transferred to the area of XAI. However, Shneiderman (2000) advise against such transfers, as interactions with machines are different and more limited than with humans. Nevertheless, the findings from psychological research are an essential guide for the design of XAI (Miller, 2019). Still, they cannot replace user studies in which the effect of AI explanations on users is examined.

As stated in the Related Work chapters, I focus in this dissertation on the investigation of XAI for cooperation & collaboration, education, and medical decision support by presenting six different user studies. These user studies investigate different levels of interactivity and analyse the impact of (1) the presentation of different XAI algorithms, (2) virtual agents in combination with XAI visualisations, and (3) XAI in human-AI cooperation & collaboration (see Figure 13.2).

Figure 13.2: Three levels of interactive XAI are examined in this dissertation: Starting with a low level of interactivity by presenting XAI methods (e.g., different algorithms to provide XAI visualisations) to end-users. The second is more interactive. Here, a virtual agent presents XAI visualisations generated with LIME based on user input. The last one communicates explanations to end-users using during a cooperative/collaborative task using text



Investigation of the Impact of XAI Algorithms on End-Users When focusing on XAI in a human-centered way, more is needed than to develop XAI algorithms like LIME, LRP, or counterfactual explanations (these approaches were presented in Chapter 5 on page 22). In addition, these explanations' impact on users must be investigated to understand their usability, benefits, and downsides. For this, a broad research community agrees that *human-stakeholders* (e.g., their needs, mental models, experiences), as well as the *purposes of different AI application scenarios* (e.g., healthcare, military, sales, finance), have to be taken into account when creating XAI (e.g., Doshi-Velez and Kim, 2017; Gerlings et al., 2021; Gilpin et al., 2019; Hoffman, Klein, and Mueller, 2018; Miller, 2019). I investigate the impact of the outcome (e.g., XAI visualisation) of specific XAI algorithms (i.e., LIME, LRP, counterfactual, confidence values) on users' perception in the NOVA Study and the Pneumonia Study. These experiments investigate mental models, trust, self-efficacy, cognitive workload, and users' emotions. In both studies, participants were *presented different*

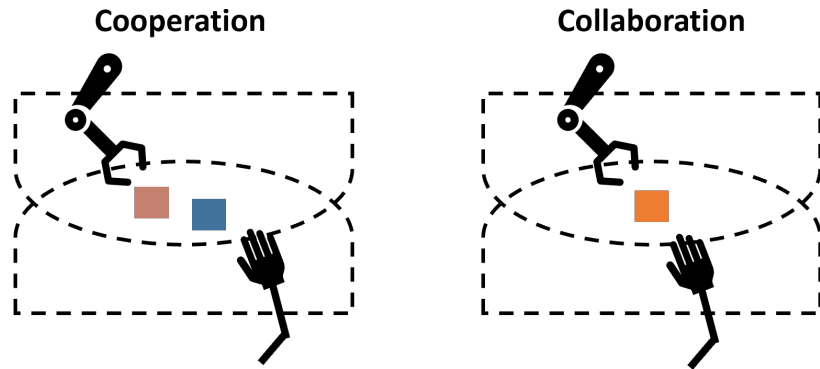
XAI methods on classification results of a neural network (see the upper part of Figure 13.2 on the preceding page). While in the Pneumonia Study, participants had to rate XAI visualisations on X-ray images that were presented to them one another, in the NOVA Study, people could freely navigate through the NOVA software to get an impression of the CNN predictions. Therefore, while both studies represent two XAI systems with a low level of interactivity, the NOVA Study allowed participants to more freely explore the results of the XAI methods used CNN through exploratory use of the NOVA software. In contrast, in the Pneumonia Study, the XAI visualisations to be evaluated were automatically presented to participants.

Investigation of the Impact of Virtual Agents Combined with XAI Algorithms on End-Users Another part of the experimental investigations of this dissertation deals with a hybrid approach that combines the two previously explained areas of inquiry. Therefore, in two of the studies presented in this dissertation (i.e., Gloria Study and Museum Study), a *virtual agent shown XAI visualisations* generated by LIME (see the middle part of Figure 13.2 on the facing page). Both studies represent a medium level of interactivity since the participants had to produce the data predicted by the CNN by themselves (i.e., speaking words in a microphone that the CNN classified). Still, the explanations were shown to them without the capacity to react to them accordingly. In the educational context, combinations of different teaching methods are common (e.g., for teaching programming (Mohorovicic & Strcic, 2011)). In our Museum Study, we offered the possibility for end-users to train and test an AI system themselves and to have explanations presented to them by a virtual agent based on the trained system. Therefore, they were involved in the whole process, beginning with data generation, training, and testing the CNN with self-chosen example words afterwards.

Investigation of XAI in Human-AI Cooperation & Collaboration Both studies in cooperation (i.e., VR-Robot Study) and collaboration (i.e., Conversational AI Study) investigate the interaction between a user and an agent. Nevertheless, they differ in interaction: In the VR-Robot study, robots and users have to solve a cooperative task, whereas, in the Conversational AI Study, both have to work on a collaborative task. Although the terms sound similar, they differ significantly in their meaning. Roschelle and Teasley (1995, p. 70) defines *collaboration* as “[...] a coordinated, synchronous activity that is the result of a continued attempt to construct and maintain a shared conception of a problem.”. The authors highlight the difference between cooperation and collaboration in people’s roles during problem-solving. Therefore, *cooperation* is described “[...] as an activity where each person is responsible for a portion of the problem-solving.” (Roschelle & Teasley, 1995, p. 70) while *collaboration* means the “[...] as the mutual engagement of participants in a coordinated effort to solve the problem together.” (Roschelle & Teasley, 1995, p. 70). Malik and Bilberg (2019) have illustrated the difference between collaboration and cooperation for an example task of direct robot-human interaction (see Figure 13.3 on the next page). Both studies in this dissertation represent a high level of interactivity since the used systems provide the possibility to *communicate AI explanations* during the task and to react to the feedback

of users (see the lower part in Figure 13.2 on page 74). Explanations were part of the interaction and communicated to the end-user utilising text.

Figure 13.3: Difference between cooperation and collaboration in a human-robot task. While a human and a robot work in the same area on different objects during cooperation, a human and a robot work on the same object in a collaborative task. Figure is adapted from Malik and Bilberg (2019)



13.2.2 Explainable Model

The explainable model is divided into the content and the type of explanation. Both are described in more detail now.

Explanation Content

Liao et al. (2021) explored which questions XAI can or should answer. They proposed nine question categories that can evolve for an XAI application (see Table 13.1 on the next page). Ribera and Lapedriza (2019) indicate that besides these theoretical ingredients to design explanations in a human-centered way, the *technical development* and the *empirical evaluation* of these ideas are necessary. Conati et al. (2021) provide a user study investigation of “Why?” and “How?” questions in the context of an ITS for learning an algorithm for constraint satisfaction problems. In their work, users mainly accessed the “Why?” explanation types. In the work of Castelli et al. (2017) user preferences and demands regarding visualisations of a smart home system. They found eight questions that users wished to ask their smart home system, including “What?” questions (i.e., “What (has) happened in my home?”, “What is the current status of my home?”). Cotter et al. (2017) investigated “Why?” and “How?” questions of Facebook’s News Feed algorithm that is used to personalise the News Feed of Facebook users. They found that the News Feed posts focus more on “Why?” than “How?” questions. Therefore, more of the decision of the algorithm than its inner working is explained to users. Lim et al. (2009) used beside “Why?” and “How?” questions also “Why not?” and “What if?” questions in their experiment. Here, users were presented with different explanations of a physical activity recognition system. “Why?” and “Why not?” questions helped users gain a better understanding and increased their task performance and trust in the system. The “Why?” questions were beneficial to create a precise understanding of the AI system. for “How?” and “What if?” no benefit was found (Lim et al., 2009).

Similar to previous work, this dissertation investigates “Why?” and “How?” questions in explanations. These are then complemented by

Question	Examples
Global "how"	What is the system's overall logic? What kind of algorithm is used?
Why	Why/How is this instance given this prediction?
Why not	Why is this instance NOT predicted to be [a different outcome Q]?
How to be that (different prediction)	How should this instance change to get a different prediction Q?
How to still be this (the current prediction)	What is the scope of change permitted for this instance to still get the same prediction?
What if	What would the system predict if this instance changes to ...?
Performance	How accurate/precise/reliable are the predictions?
Data	What kind of data was the system trained on?
Output	What kind of output does the system give?

Table 13.1: Question categories that XAI should answer. For every question type, examples are given. Table adapted from Liao et al. (2021, p. 6)

"What?", "Why not?", and "What if?" questions. "What if?" questions are used within counterfactual explanations, which we investigated in the Pneumonia Study. For XAI visualisations created with LIME, "Why?" and "What?" questions are used in the Gloria Study. While areas that are coloured green are used to answer "Why?" questions (i.e., "Why did the DNN classify this?"), areas coloured red are meant to answer "What?" questions (i.e., "What part in the image speaks against the classification?"). Explanations that provide an answer to "Why not?" questions play an essential role in our cooperation studies. Here we explain to the end-users why a system has made a mistake (VR-Robot Study) or a wrong assumption (Conversational AI Study).

Explanation Type

In addition to the content of the explanation, the type of explanation plays a role in the presentation to end-users. According to Ribera and Lapedriza (2019), the type of explanation includes the method of how the explanation is generated (e.g., the LRP algorithm) and the form of how the explanation is presented to the user (e.g., as a visual explanation). Vermeire et al. (2021) offer an approach to help data scientists find the appropriate XAI method for their use case. In addition to user needs and application context, the technical requirements of the explanation methods (e.g., practical usage of an XAI method or the explanation properties) should also be taken into account (Vermeire et al., 2021). For example, it makes little sense to choose an explanation method like LRP if you want to explain the inner workings of a rule-based system. Similarly, a technique like LIME does not lend itself to time-critical operations since LIME takes some time to generate explanations. The types of explanations presented in this thesis include visual (i.e., LIME, LRP, counterfactuals), verbal (i.e., text and speech), and numerical (i.e., confidence values) representations. To choose the type of explanation, as recommended by Vermeire et al. (2021), we first considered the particular conditions of the purpose of AI in the studies (i.e., cooperation & collaboration, education, and medical decision support). In addition, we used pilot studies to consider the end-user needs and adapt the explanations accordingly for the main studies. We followed

this approach in the two studies on cooperation and collaboration (i.e., VR-Robot Study and Conversational AI Study). Because of the focus on investigating the influence of different levels of XAI in the interaction between humans and machines, we selected linguistic explanations in four studies. Linguistic explanations are natural interactions humans use intuitively with other human and machine entities. Therefore, we presented textual explanations to end-users in the VR-Robot Study and the Conversational AI Study. In contrast, in the Gloria Study and Museum Study, verbal explanations of a virtual agent were given.

These different types of explanations can also be combined. Finzel et al. (2021) present a technical approach to create multi-modal explanations (i.e., textual and visual explanations) for end-users. For this, they use Inductive Logic Programming. However, while Finzel et al. (2021) propose a first proof-of-concept implementation and present an exemplary use-case for medical decision making (Schmid & Finzel, 2020), an empirical evaluation with real users is missing. This dissertation's two studies for educational purposes use multi-modal explanations (i.e., verbal and visual explanations). Compared to Finzel et al. (2021), we investigate the impact of such combinations of visual and verbal explanations on end-users in the Gloria Study and Museum Study. For this, we used LIME to generate XAI visualisations for a CNN. We combined these with the virtual agent Gloria to communicate as naturally as possible with end-users. For conducting such HC-XAI user studies, the targeted user group has to be taken into account. I describe how the interdisciplinary concept integrates users in the following.

13.3 User Evaluation

One of the challenges of current XAI research is to consider different user groups with different requirements regarding AI explanations. Gilpin et al. (2019) distinguish between *inside explanations*, providing a technical explanation and therefore addressing technical expert-users like programmers and developers and *outside explanations* that answers “why” questions and are consequently helpful for society, e.g., non-expert end-users. Already Wick and Thompson (1989) emphasise in their research on rule-based systems that explanations for developers of AI systems (i.e., traced-based explanations) are fundamentally different from explanations for end-users (i.e., reconstructive explanations). While in traced-based explanations, the reasoning process of the AI system is considered, reconstructive explanations are actively reconstructed during the problem-solving process of the AI system. This means that the explanation directly provides information on the solution and not the whole reasoning process of the AI. Trace-based explanations instead present all possible information that the system can generate during the reasoning process to show developers possible errors in the reasoning process. This dissertation explicitly investigates HC-XAI for end-users (i.e., reconstructive explanations) and compares different kinds of explanations that differ in content and type. Vilone and Longo (2021) describe four main factors that are relevant to create such explanations for AI: end-user, goals, contents, and language (see Figure 13.4 on the facing page).

Structure of an explanation

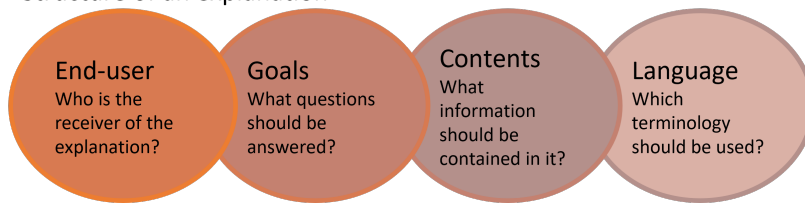


Figure 13.4: Four elements are central to creating human-centered AI explanations. Illustration adapted from Vilone and Longo (2021)

Ribera and Lapedriza (2019) distinguish between *ML-experts* (i.e., developers and AI researchers), *domain experts* (e.g., physicists, lawyers), and *lay users* (i.e., recipients of the decision like patients or loan applicants). Already Kass and Finin (1988) highlight the importance and the practical need of personalised user-models for creating good explanations. Junior and Filgueiras (2005) present one possibility to create such user-models through the development of *personas*¹. Mueller et al. (2021) criticise that such user-models or personas are neglected in current XAI research.

1: personas are fictitious users created from real user observations (Junior & Filgueiras, 2005)

Including end-users in the HC-XAI concept enables researchers and developers to test their XAI application with users and design it human-centered, i.e., adapted to users' needs. On the one hand, I include the aspect of personas, which capture users' needs concerning XAI. But my concept also integrates users in the evaluation of XAI. For this purpose, various experiments are described in this dissertation. First, collected survey data will be analysed to gain insights into end-users of XAI and create *personas* of typical XAI end-users. Second, different *AI systems* (i.e., white-box and black-box approaches) are used in experiments to investigate variations of the *explanation design* regarding the content, type, and interface of explanations on end-users (see Figure 13.6 on page 82).

13.4 Five-Step Approach to Develop HC-XAI

For researchers in classical HCI, the components of the HC-XAI approach presented in the following will be familiar. The process I present is very similar to user experience (UX) designs used in HCI to develop human-centred software solutions². In experience-based design, presented by Wright and McCarthy (2010), the user's experience is placed in the foreground when developing technology. The *responsible design process* by Peters et al. (2020) present five phases explicitly addressing the design of AI technologies. Peters et al. (2020) take into account ethical aspects in the design of AI systems as well as a design that considers the well-being of users in every phase of development. However, the approach of Peters et al. (2020) does not tackle XAI directly. The topic of integrating XAI in the design process of AI systems is addressed by Ehsan et al. (2022). Their *seamful XAI design process* is intended to enable AI practitioners to identify and counteract possible pitfalls in AI systems as early as the design stage. To this end, they present three phases that address so-called seams (i.e., mismatches, gaps): (1) Envisioning Breakdowns: led by the question to stakeholders, "what could go wrong?" concerning the AI technology; (2) Anticipating & Crafting Seams from Breakdowns: In this step, stakeholders develop actionable steps for handling seams; (3) Designing with Seams: In the last step, stakeholders have to select seams

2: see an example for the human-centred UX design process: <https://www.interaction-design.org/literature/article/what-is-a-usability-engineer> (last accessed on 06.12.2022)

that should be displayed and those that should be hidden. The goal is to empower users to interpret AI results and take informed actions.

The interdisciplinary concept in this dissertation is similar to the UX designs used in HCI. It also focus on the needs of users and integrate the user actively in different steps of the development process of HC-XAI. Compared to the seamful XAI design process (Ehsan et al., 2022), my proposed 5-step approach focuses less on only the seams that might occur when using HC-XAI. Instead, both positive and negative aspects of HC-XAI are considered. While Ehsan et al. (2022) presents a proof-of-concept study for their approach, this only includes a general survey on possible pitfalls during an AI lifecycle. The development and testing of a corresponding application are not part of the study. In contrast, besides investigating users using the persona approach, I also use my five-step approach to conduct user studies to give practitioners an impression of the impact of XAI on users and what such evaluation could look like. In their study, Ehsan et al. (2022) interviewed AI experts and domain experts (i.e., researchers, data scientists, and UX designers). In contrast, I focus explicitly on end-users.

The five steps are presented in more detail below. It should be noted that for creating an HC-XAI design for a specific application, the step-by-step approach has to be based on a use case that provides a meaningful use of AI for a particular purpose and the relevance of HC-XAI in this usage of AI.

- **Step 1 - Information Collection:** The first step of the user-centred approach is to gain insights about the user group. This includes collecting information about user needs, opportunities and risks, and existing approaches/procedures/routines that users already use in the domain. This information gathering helps researchers gain insight into the domain for which HC-XAI is to be developed. In addition, by consulting users from the beginning, they are brought on board, which reflects an appreciation of users' opinions and experiences, while at the same time taking into account difficulties and bottlenecks that could hamper the design of XAI from the very beginning.
- **Step 2 - Definition of Personas:** Based on the collected information, prototypical users (i.e., personas) of the future HC-XAI system are defined. This approach makes it possible to structure the collected information about the user group. A template for the design of such personas in the context of HC-XAI is presented in Chapter 15 on page 85. The personas are intended to support developers in the HC-XAI design by addressing important information about motivation, needs, and possible problems from the user's point of view.
- **Step 3 - Development of XAI Design:** The development of the XAI for the specific use case integrates the identified user needs, which are illustrated with the created personas. The explanations are designed in terms of their content, type (e.g. visual, textual) and interface so that they meet the requirements of the personas. For this, the goal for deploying the HC-XAI system must be well defined. This phase is usually an iterative process, starting with a low-fidelity prototype that is improved through several user evaluations towards a high-fidelity prototype (Rudd et al., 1996).

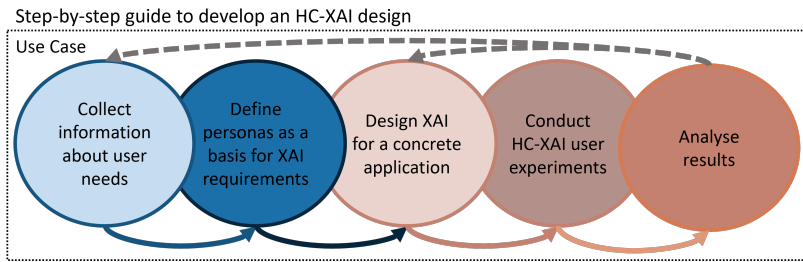


Figure 13.5: Process of the development of an HC-XAI design based on the interdisciplinary concept proposed in this dissertation

- **Step 4 - Conducting HC-XAI Experiments:** To test the application's XAI design, studies are conducted with users. The aim is to investigate whether the explanations meet the requirements of the users of the application.
- **Step 5 - Analysis of Results:** The results will be analysed after the experiments have been completed. The findings provide insights into whether the HC-XAI system has met the previously defined objectives. Depending on the results, the XAI design is adapted, or, if necessary, users are interviewed again.

Components of the step-by-step approach can be found in this dissertation: Steps 1 & 2 are examined in Chapter 15 on page 85 exemplarily for three different groups of end-user. Regarding steps 3-5, this dissertation differs from the 5-step approach presented. Although applications for different purposes of AI are introduced and evaluated with end-users, they do not directly refer to the previously generated personas. Instead, controlled experiments will be presented that deal with the general influence of explanations on end-users (e.g., trust, mental models). Thus, as already explained in the Introduction chapter (see Chapter 1 on page 2), this dissertation follows up with the Doshi-Velez and Kim (2017) postulated by *human-grounded evaluation*. Thus, this dissertation provides fundamental insights into the perception of XAI by end-users in different scenarios. In future studies, the presented 5-step approach can be applied to one throughout the design for a specific use case. Steps 3-5, as taken up in this dissertation, will now be explained: For Step 3, the technical implementation is explained in Chapter 20 on page 125 (for rule-based systems) and in Chapter 21 on page 132 (for CNN). Furthermore, in Chapter 23 on page 143, the experimental design for the conducted studies is presented. Steps 4 & 5 deal with the implementation and evaluation of exemplary studies on the topic of HC-XAI in the areas of cooperation & collaboration (Chapter 27 on page 160), education (Chapter 28 on page 190), and medical decision support (Chapter 29 on page 210). Here, the type, content, and interface of the explanation are investigated (see Figure 13.6 on the next page). As stated before, the interface of explanation has different complexity, starting from simple interaction where XAI visualisations are presented to participants (i.e., Pneumonia Study, NOVA Study) to a more interactive one that combines verbal communication of a virtual agent with XAI visualisations (i.e., Gloria Study). Finally, an even more interactive approach where an AI could communicate an explanation via text during the interaction is presented (i.e., Conversational AI Study).

Cooperation & Collaboration		Education		Medical Decision Support	
VR-Robot Study		Gloria Study		NOVA Study	
Content of explanation	Type of explanation	Content of explanation	Type of explanation	Content of explanation	Type of explanation
Why not? How?	Verbal - Text (explanation vs. explanation & solution)	Why? What?	Verbal - Text Verbal - Speech Visual (LIME)	Why?	Numerical (Confidence values) Visual (LIME)
Cooperation with an industrial robot		Interaction with a virtual agent		Interaction with the NOVA software	
Conversational AI Study		Museum Study		Pneumonia Study	
Content of explanation	Type of explanation	Content of explanation	Type of explanation	Content of explanation	Type of explanation
Why? Why not?	Verbal - Text (personal vs. impersonal)	Why?	Verbal - Speech (first vs. third person) Visual (LIME)	Why? What if?	Visual (LIME vs. LRP vs. counterfactuals)
Collaboration with an AI dialog partner		Interaction with a virtual agent		Presentation of XAI visualisations	

Figure 13.6: Different content and types of XAI were addressed in the six conducted experiments. In addition, the experiments investigate different levels of interactive XAI

In this chapter, the elements of the interdisciplinary concept were introduced, which serve as the basis for this dissertation's empirical investigations (i.e., surveys and experiments). The concept comprises three central components: an AI system (i.e., white-box, black-box, hybrid approaches), the explanation design (i.e., explainable model and explanation interface), and a user evaluation. The design of the explanations depends on the application scenario and the user group. Regarding the explanation design, this dissertation investigates (1) specific XAI algorithms (i.e., LIME, LRP, counterfactuals) when presenting explanations, (2) the influence of XAI in human-AI interactions where explanations are provided in natural language, and (3) a hybrid variant of (1) and (2) where a virtual agent presents XAI visualisations using natural language. The studies are characterised by various levels of complexity of interactive XAI, ranging from simple presentations of XAI visualisations (i.e., Pneumonia Study) to approaches using different XAI modalities (i.e., a virtual agent that communicate in natural language and XAI visualisations), towards communication between AI and user using natural language (i.e., Collaboration Study). The concept's practical implementation is illustrated in the following chapters, starting in the next chapter with the investigation of end-users needs regarding XAI.

V. END-USERS OF HUMAN-CENTERED EXPLAINABLE AI

Exploring End-Users of Explainable AI

15

In the previous chapter, I presented the interdisciplinary concept of HC-XAI. In this dissertation, this concept is applied to end-users. To create an HC-XAI system for end-users, we need (1) information about end-users regarding (X)AI and (2) a concrete XAI system for a specific purpose or application. While we explore (2) in six experiments for three different purposes (i.e., cooperation & collaboration, education, medical decision support) in the chapters 29, 28, and 29, this chapter will investigate (1). The following provides insights into end-users general impressions of AI and XAI. Based on three conducted surveys, I will create personas to summarize the empirical findings in prototypical (X)AI end-users. An adaption of the PATHY 2.0 approach of Ferreira et al. (2018) will be used.

For this chapter, the following work served as a basis:

► *Persona Approach for HC-XAI*

- **Weitz, K., Zellner, A., & André, E. (2022).** What Do End-Users Really Want? Investigation of Human-Centered XAI for Mobile Health Apps. *arXiv preprint arXiv: 2210.03506* <https://doi.org/10.48550/arXiv.2210.03506>
- Zellner, A. (2021). *Towards personalized explanations in digital health - user-centered explanations for mobile health applications* (Master's thesis). University of Augsburg.*

15.1 Personas

The term persona refers to a technique in the HCI research community, especially in usability research (Castro et al., 2008). The purpose of creating personas is to create prototypical users of a software system (Castro et al., 2008). In the persona approach called PATHY 2.0¹ of Ferreira et al. (2018), empirical data of users are analysed to extract information about user needs, expectations and problems regarding a software product. These information are structured into six fields (Ferreira et al., 2018, p. 283):

- *Who*: Description of the persona who will be using the application
- *Context*: Characteristics of the persona's routine, aspects of the environment in which the persona lives, and people with whom the persona has contact.
- *Technology experience*: Experiences that the persona has had with other technologies or applications, and information regarding application characteristics that the persona likes and does not like
- *Problems*: Problems faced by the persona and which can be solved by the application to be designed. The goal here is to increase the understanding of the users' issues.
- *Needs*: Needs to be met to solve the problems described in the problems field.

15.1	Personas	85
15.2	Persona Approach for HC-XAI	86
15.2.1	Persona Template	86
15.2.2	Data-Driven Approach for Creating Personas	87
15.3	Overview of the Surveys	88
15.4	Scope and Objective of the Surveys	90

1: PATHY refers to emPATHY

*Co-supervised by the author of this dissertation

- ▶ *Existing solution*: Existing solutions related to the ideas and interfaces to be improved or included in the application to be designed for solving the identified problems.

Compared to other persona approaches like the one of Acuña et al. (2012) that mostly describe the application requirements from a more technical view, the PATHY 2.0 approach focus on the empathy of the developer towards the user. To raise empathy, Ferreira et al. (2018) focus on conducting studies to collect and understand user needs. In doing so, the PATHY 2.0 approach focus on the personal attitudes of users and their previous experiences towards technology to elaborate the problems and issues for a new application (see “Who” and “Context” in Figure 15.1 on the facing page). When designing personas, Ferreira et al. (2018) describe the challenge that personas should have information about the potential end-user, but the amount of detail can be overwhelming. Guo et al. (2011) recommend designing personas in a way that they are:

- ▶ *accurate* and report only *relevant information* instead of including irrelevant personal details (e.g., place of residence, family situation)
- ▶ developed for a *specific purpose*
- ▶ evaluating *attitude* and *behaviour*, meaning not to focus exclusively on user features (e.g., personality, age) but also include what they want from a system

15.2 Persona Approach for HC-XAI

The previously described principles and recommendations serve as a basis for creating personas in this dissertation. The persona template developed and a data-driven approach to fill out the template using user surveys are presented in the following.

15.2.1 Persona Template

Schneider and Handali (2019, p. 7) describe that when XAI is being personalized, four relevant categories emerge:

- ▶ **Prior knowledge**: What does a user know?
- ▶ **Decision information**: What information does the user want for the decision?
- ▶ **Preferences**: What does the user like/prefer?
- ▶ **Purpose**: What should the explanation be used for?

To address Guo et al. (2011) recommendation to focus the personas less on non-relevant and often too detailed information, the PATHY 2.0 approach (Ferreira et al., 2018) is used in a slightly adapted form for this dissertation. In addition, the personalization features of Schneider and Handali (2019) are included in the template (see Figure 15.1 on the next page).

The following content is part of the adapted template and a corresponding summary of each section:

- ▶ **Who**: Description of who will use or be impacted by the AI application. This field introduces the user, their traits, frustration, and concerns.

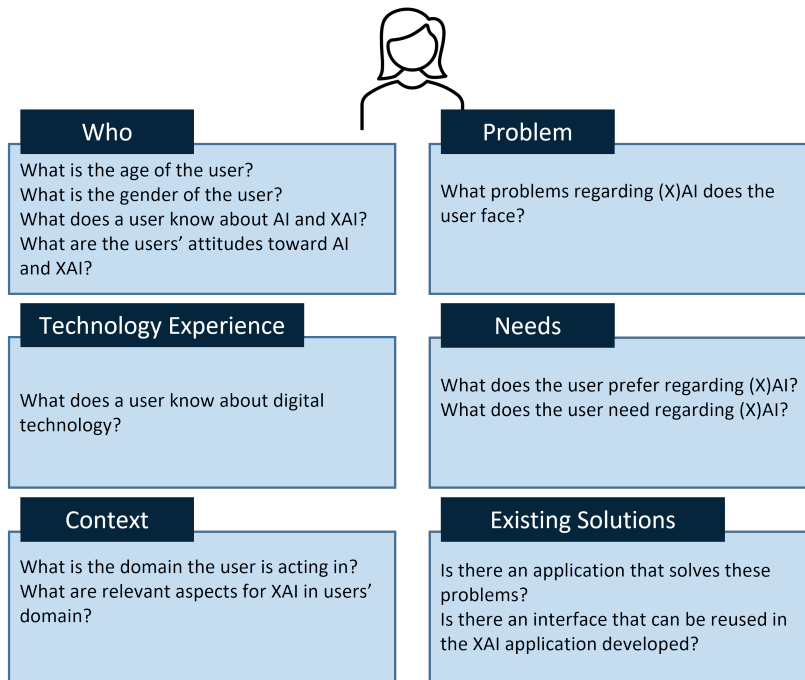


Figure 15.1: The persona template for HC-XAI combines the PATHY 2.0 approach of Ferreira et al. (2018) and the relevant personalization features from XAI as described by Schneider and Handali (2019)

- **Technology experience:** describes user experiences with other AI applications from this domain or in general. Additionally, the users' attitudes regarding XAI and AI are stated.
- **Context:** refers to the domain where the person utilizes the AI application. In addition, relevant information about the environment the user is acting in is pictured.
- **Problems:** Problems and issues regarding missing or insufficient HC-XAI are stated here.
- **Needs:** states user needs to be considered when solving the problems stated.
- **Existing solutions:** A listing of existing solutions trying to solve a similar problem or similar ideas and approaches to improve HC-XAI in this context.

For the development of personas, we used an approach similar to Holzinger et al. (2022). They present a 5-step process for developing personas for AI, starting in step 1 with identifying user groups. In this dissertation, we focus on end-users. In step 2, information about users is collected. We did this using surveys. In steps 4 and 5, the personas were created and visualized at the end. We do this at the end of every survey evaluation.

15.2.2 Data-Driven Approach for Creating Personas

How do you create a persona from the data collected about users? We used a data-driven approach (Weitz, Zellner, & André, 2022; Zellner, 2021). Here, the idea is to find information in the data based on a pre-defined research scope and objective (e.g., investigating end-users general attitude towards XAI as we did in the Museum Visitors Survey)² that can be divided into clusters (see Figure 15.2 on the following page). For this purpose, the questions/statements in the questionnaire relevant to the object of investigation are considered. Then, similar opinions or

²: more information about the conducted surveys can be found in the next section

Figure 15.2: A data-driven approach is used to fill out the HC-XAI persona template (Weitz, Zellner, & André, 2022; Zellner, 2021). Here, collected data are investigated based on pre-defined research objectives. Based on this investigation, similar categories of information are clustered in the next step. After this initial clustering, additional information about people that were clustered into the categories is added. Finally, the information of each cluster is filled into the persona template



comments are grouped into categories. This can be done by investigating the free-form feedback of the participants, as we did in the Mobile Health Survey by using conventional content analysis (more about this approach in Chapter 25 on page 154) or using the answers of closed answer formats (e.g., rating scales). Based on this initial clustering, other variables of the persons in the cluster are added (e.g., age or gender). Finally, this information about users is filled in the persona template. This way, a picture of a prototypical person, who is representative of users in a particular cluster, is obtained based on the object of investigation.

15.3 Overview of the Surveys

3: people who use voice assistants, social media, fitness trackers, and more (Alizadeh et al., 2020)

Alizadeh et al. (2020) investigated users' understanding of AI by interviewing 50 AI-technology users³. The authors found seven clusters of users' understanding of AI, including concepts like ML, expert systems, or Neural Networks. But their findings also suggest that users mix up AI-based systems with automation systems (e.g., a door that automatically closes). To investigate end-users needs and (maybe incorrect) understanding of XAI, we conducted three surveys for the AI application purposes of cooperation and collaboration, education, and medical decision support. Holzinger et al. (2022) emphasize the importance of asking stakeholders about their attitudes towards AI. This information is essential, as it influences whether and how users will utilize an AI system later. Therefore, the surveys address end-users in a company (cooperation & collaboration), a museum (education), and a mobile health app (medical decision support).

Survey	Content	N
Education Museum Visitors Survey	(X)AI attitudes of end-users interested in gaining knowledge by visiting a museum	58
Companies AI in Companies Survey	(X)AI impressions in the workplace of end-users working in German-based companies	50
Mobile Health Stress App Survey	End-users rating XAI in the context of stress detection on their mobile device	92

Table 15.1: Overview of the sample sizes of the conducted (online) surveys about (X)AI, which were used as the foundation for persona development

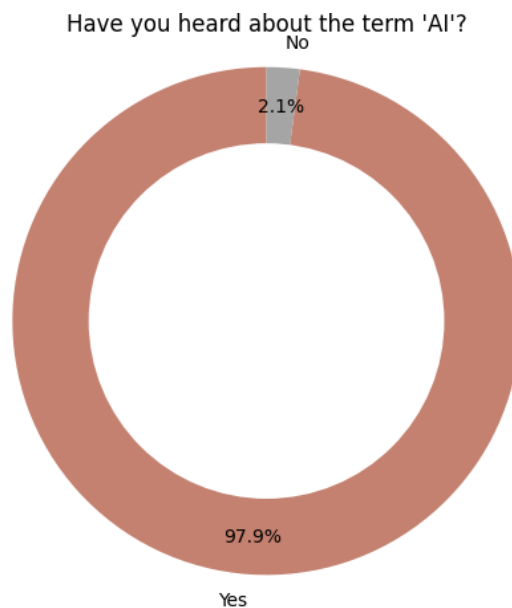


Figure 15.3: Response of all 200 survey participants. Almost all participants have heard about the term 'AI'

The data presented were collected in different surveys in 2019-2021 (see Table 15.1). The participant size N refers to the amount of data used for the analyses (i.e., cleaned dataset).

From all our surveys, we collected data of 200 end-users between 8 and 70 years ($M = 39.3$, $SD = 13.8$). Eighty-seven participants were female, 112 were male, and one identified themselves as divers. We asked all participants about their knowledge of AI⁴ and XAI⁵. This question was adapted from the Eurobarometer 2017 study of the European Commission (European Commission, 2017). In general, almost all of the participants knew the term AI (see Figure 15.3).

Regarding XAI, a lot more did not know the term (see Figure 15.4 on the following page). When separated by the AI application's purpose, it becomes aware that users in a company setting are more familiar with XAI than education and mobile health scenarios.

4: the AI definition presented in all surveys was: "The term 'Artificial Intelligence' is often used to describe machines (or computers) that mimic 'cognitive' functions that humans associate with the human mind, such as "learning" and "problem solving" and is oriented on the definitions given by Russell and Norvig (2016)

5: XAI was described to all participants as: "With the help of explainable Artificial Intelligence, it should be possible to have a better understanding of artificial intelligence". In this definition, the purpose of XAI was in focus

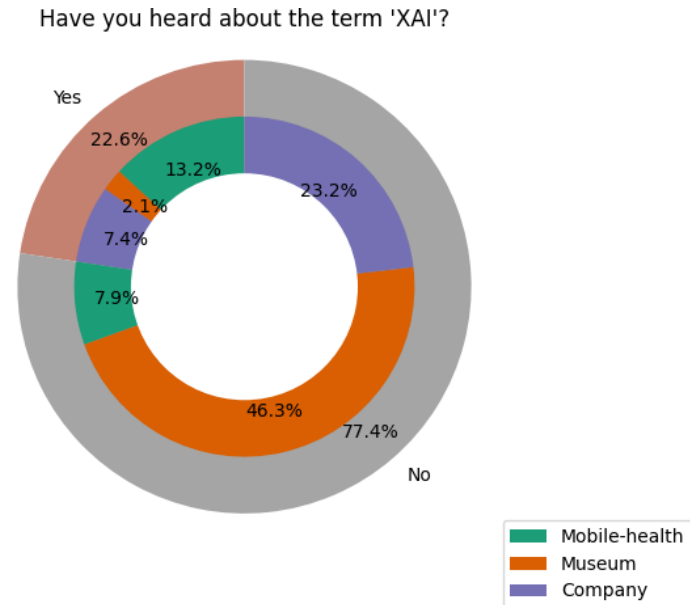


Figure 15.4: *Outer circle:* response of all 200 survey participants whether they had heard about XAI. *Inner circle:* division of participants by application scenario. It can be seen that, especially in the mobile health and education scenario survey, participants did not know the term XAI

15.4 Scope and Objective of the Surveys

The three surveys presented in this dissertation (i.e., education, company, mobile health) reflect different complexities and objectives when collecting information about user groups. These will be briefly explained and motivated in the following.

End-users in Education The survey I conducted at the Deutsches Museum in Munich is a general survey of end-user attitudes towards (X)AI. Here, the general attitude of the people was surveyed without asking about a concrete application scenario. This type of survey represents the most basic survey, which makes it possible to get a first impression of a domain about which one has no other information. As explained in the Related Work chapters, educational offers on XAI are hardly available. Surveys of this user group are also non-existent. The questionnaire was distributed to museum visitors at the Deutsches Museum in Munich to derive initial general attitudes about XAI. The personas developed from the questionnaire serve as a basis for developing initial offers on XAI for end-user in this area.

End-users in Companies This survey focuses more on concrete (X)AI applications than the survey in the German Museum. It addresses employees in German companies. The survey asked employees about the current and future use of AI in their company and the required competencies to use it. Furthermore, the survey asked about specific AI applications but did not examine specific XAI methods for these applications. Instead, XAI represented a sub-part in the survey, which was also asked to analyse the possible potential for XAI in this context.

End-users in Mobile Health The last survey presented in this dissertation is also the most detailed survey regarding XAI. Based on a concrete

application that detects a person's stress level, the application was fictional but oriented on actual technical possibilities, such as the ability of wearables to measure the pulse. Concrete needs regarding explanations in this context were asked among users. For this purpose, different types, contents, and interfaces of explanations were presented, which were evaluated by the users. In addition, users' explanatory behaviour was evaluated when they were asked to categorise a person as stressed/not stressed. All this information provides a detailed picture that should give developers of XAI applications concrete recommendations to design such systems.

In the following, detailed insights into these studies regarding end-users of AI in education, companies, and mobile health are presented. At the end of each chapter, the survey findings are used to create personas.

16

End-Users of Explainable AI in Education

16.1	Research Questions . . .	92
16.2	Methodology	93
16.3	Participants	93
16.4	Results	93
16.4.1	RQ1: Knowledge & Attitude	93
16.4.2	RQ2: Personal Attributes	94
16.5	Lessons Learned	96
16.6	Personas of End-Users in Education	97

A museum is an educational environment where people want to learn more about past and present issues through the exhibits on display. In addition, museums use special exhibitions to introduce visitors to specific topics. For example, the German Museum Munich has many exhibits in the field of technology (e.g. robotics, electronics). Moreover, in 2019, the museum maintained a special exhibition for the “Year of Science” on AI. In this scope, we developed a participatory ML-show which is presented later in this dissertation (see Museum Study in Chapter 28 on page 190). In addition, it seemed to be the appropriate place for us to ask users about their knowledge and attitude towards (X)AI in education. Thus, during the ‘focus week AI’ at Deutsches Museum Munich, we asked museum visitors their attitudes towards technology in general and AI and XAI in particular (see Figure 16.1). This evaluation and its results are reported in the following. We published parts of this work in:

- **Weitz, K., Schlagowski, R., & André, E. (2021).** Demystifying artificial intelligence for end-users: Findings from a participatory machine learning show. In S. Edelkamp, R. Möller, & E. Rueckert (Eds.), *Proceedings of KI 2021: Advances in artificial intelligence - 44th german conference on AI, virtual event, september 27 - october 1, 2021* (pp. 257-270). Springer. doi: [10.1007/978-3-030-87626-5_19](https://doi.org/10.1007/978-3-030-87626-5_19)*



Figure 16.1: Stand at the museum, which was used to ask museum visitors about their attitudes towards AI and XAI. Figure from Weitz, Schlagowski, and André (2021)

16.1 Research Questions

The research questions we set ourselves here are as follows:

*Adapted by permission from Springer Nature Customer Service Centre GmbH: Springer Nature Lecture Notes in Computer Science book series (Demystifying Artificial Intelligence for End-Users: Findings from a Participatory Machine Learning Show, Weitz, K., Schlagowski, R., & André E.), ©2021 Springer Nature Switzerland AG (2021)

- **RQ1: Knowledge & Attitude** Which knowledge and attitudes do end-users in a museum have about AI and XAI and their application areas?
- **RQ2: Personal Attributes** How are the personal attributes (e.g., age, gender) of end-users related to their impressions of (X)AI?

16.2 Methodology

We used a paper and pen questionnaire to ask museum visitors about their impressions of (X)AI¹.

1: the complete questionnaire can be found in the Appendix

Demographic information We asked museum visitors about their age, gender, and educational background. Regarding (X)AI, we asked about their knowledge and attitude towards AI and their knowledge about XAI. In addition, we queried in which areas AI should be used (e.g., education, security, care-work) and which future humans will have with AI (i.e., life will be better/worse, negative and positive effects will be balanced, I don't know). Finally, for XAI, we investigated for whom XAI would be essential (i.e., end-users, politicians, researchers, and companies).

Technical Affinity. To measure the technical affinity of participants using the TA-EG questionnaire (Karrer et al., 2009) was queried. The 19 items of the questionnaire had to be rated on a 5-point Likert scale (1 = fully applies to 5 = does not apply at all).

16.3 Participants

Fifty-nine museum visitors took part in our field survey. Unfortunately, we had to remove the answers of one visitor due to too many unanswered questions. Therefore, for the following analyses, answers from 58 museum visitors (29 female, 29 male) between 8 and 66 years ($M = 30.3$, $SD = 16.5$) are considered. The educational background of the visitors was mixed and ranged from school students to secondary school graduates to university graduates. None of the participants took part in the participatory ML-show, which will be reported later in this dissertation.

16.4 Results

16.4.1 RQ1: Knowledge & Attitude

Museum visitors showed a significant weak technology affinity in a one-sample t-test, $t(57) = -8.55$, $p < .001$, $d = 1.12$ (large effect) compared to the mean of 3 (5-point Likert scale). When taking a look closer, we found a significant positive correlation² between technical affinity and age ($r_{sp} = .0313$, $p = .020$) and a significant negative correlation with gender ($r_{sp} = -.374$, $p = .004$). This indicates that female or older visitors had a higher technical affinity.

2: we calculated Spearmans' rang correlations

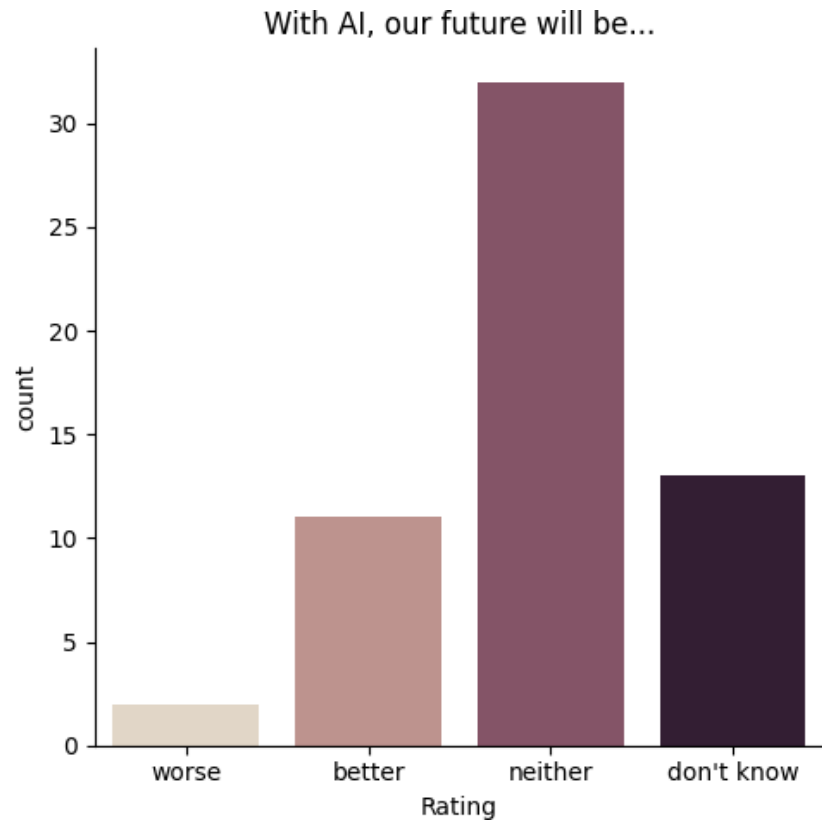


Figure 16.2: Museum visitors' answers to the question "What future do you think we will have with AI?"

Most of the visitors had heard about AI in the last 12 months (97%), but only 24% of them had heard about XAI. 60.3% of the museum visitors stated that they could roughly explain the term AI, while 25.9% of visitors said they could explain it in detail. 13.8% just had heard the term but could not explain it.

When asked about the future of AI, a large proportion ($n = 32$) of visitors were confident that it would be neither particularly good nor bad. Only two visitors had a rather poor outlook on AI in the future (see Figure 16.2).

Regarding XAI, visitors found it an important topic, especially for researchers and companies, followed by end-users and politicians (see Figure 16.3 on the facing page).

Although visitors have a non-negative image of AI, it is apparent that they do not endorse AI for use in all application areas. In particular, they reject the use of AI in the areas of education, art, and leisure, while they are more in favor of its use in the household, transportation, safety, and care sector.

16.4.2 RQ2: Personal Attributes

The following results regarding possible correlations of demographic characteristics and technical affinity on the perception of (X)AI are reported.

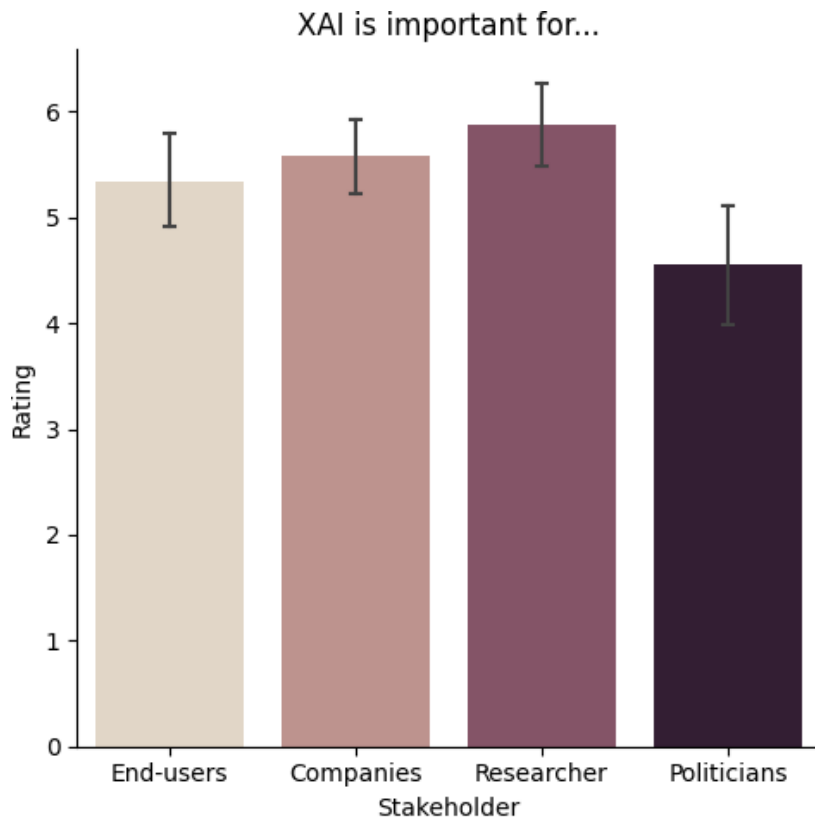


Figure 16.3: Rating of museum visitors, whether XAI is important for different stakeholders (1=disagree; 7=fully agree). Error bars represent the 95% CI

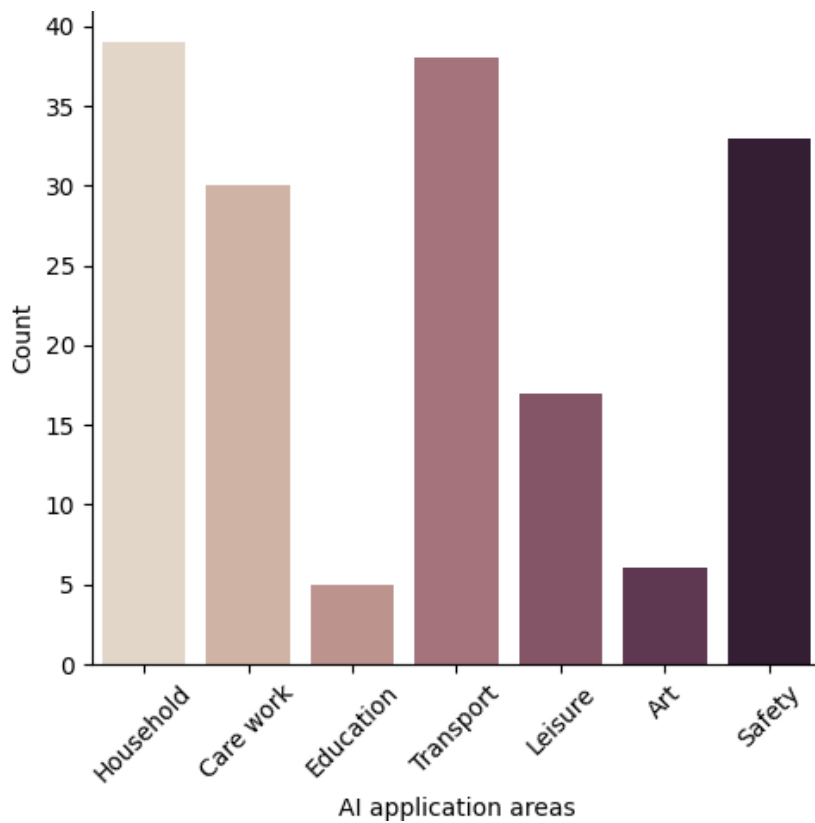


Figure 16.4: Rating of museum visitors, whether the find AI relevant for different application areas. Results show a critical look at the use of AI in education, art, and leisure

Demographic Information Regarding gender and educational background, we found no significant correlation with knowledge about AI or XAI (see Table 16.1). Only age correlates positively with AI knowledge, meaning that the older the museum visitors are, the more likely they will have heard of AI.

Table 16.1: No significant correlations were found between gender, educational background and knowledge about AI and XAI. Only the age of participants correlated positively with knowledge about AI

Demographic information	Knowledge about	
	AI	XAI
Age	.272*	.005
Gender	.189	.000
Educational background	.232	.094

* $p < .05$

Technical Affinity Regarding the technical affinity of museum visitors, we found no correlations with knowledge about AI ($r_{sp} = -.062$, $p = .643$) and XAI ($r_{sp} = .047$, $p = .726$).

16.5 Lessons Learned

Most End-Users in Education have Rough an Idea of AI A large part of users has heard about the term AI. Most users have heard of it and could explain it roughly. This shows that many users encounter the topic in their everyday lives but do not yet have a deeper understanding of AI and how it works. This makes the promotion of AI literacy³, as Long and Magerko (2020) demand, very important.

3: “We define AI literacy as a set of competencies that enables individuals to critically evaluate AI technologies; communicate and collaborate effectively with AI; and use AI as a tool online, at home, and in the workplace.” (Long & Magerko, 2020, p. 2)

The situation is different with XAI, a term unknown to most. Nevertheless, users perceive this topic as relevant for various stakeholders. In general, it can be seen that there is no generally negative attitude toward AI and no opposing view of the future, but also not a positive one. This lack of clarity about the effects of AI should be addressed and discussed in educational programs so that end-users can understand the advantages and disadvantages of AI and use this knowledge to develop a nuanced attitude toward AI. XAI can be utilized here to present the pros and cons in an understandable form (e.g., to illustrate the influence of biases).

Education is not Seen as an AI Application Area AI is not seen as an application area in the field of education. This is in line with research expressing scepticism about using digital technology in the context of education. Authors like Luckin et al. (2016) highlight the benefits of AI in an educational setting. They differentiate between benefits that can currently be achieved with the help of AI in education (e.g., personal tutoring systems, learning in VR) and the possibilities that can be achieved in the future with the help of AI in the education sector (e.g., AI as a lifelong learning partner). But to leverage these benefits, they point out that educators need to be aware of the opportunities with AI. Weitz et al. (2017) already found that educators’ experiences with digital technologies strongly influence whether educational opportunities for children in this area are supported and considered necessary. Focusing only on the training of teachers would be too short-sighted: the critical view of AI in

education among museum visitors shows that educational offerings about AI and XAI need to be developed and offered to a broad population.

Demographic Information Have An Impact on Technical Affinity, but not on AI Knowledge While we found a significant relationship between age and gender on technical affinity, we found no relationship between gender and educational background on the attitude towards (X)AI. Only age seems relevant for AI knowledge, which makes sense since our sample of museum visitors also includes very young persons. Since computer science classes are not yet compulsory in most grades in Germany, most students have not yet come into contact with AI-related topics. While the increasingly urgent demand for the integration of information technology topics into the school curricula is discussed in politics and society⁴, AI education is not anchored in the school curriculum as a subject.

4: see, for example, the introduction of a mandatory computer science subject in grades 5 to 10 in high schools and community colleges in Schleswig Holstein/Germany: <https://www.heise.de/news/Informatik-wird-an-Schleswig-Holsteins-Schulen-Pflichtfach-6037135.html> (last accessed on 02.03.2022)

16.6 Personas of End-Users in Education


This part aims to explore how prototypical end-users could benefit from (X)AI in education. Based on the insights gained from the questionnaire during the focus weeks AI of the Deutsches Museum München, two prototypical user personas are now presented. The first one is *Regina*, who represents the survey findings in a person who knows AI but is critical of AI in the educational sector. The reason for this could be that while she has experience with AI in her household using a robot vacuum cleaner, she has no experience with AI in her educational history as a student. Regina aims to get a broader overview of the benefits and limitations of AI in different application areas (see Figure 16.5 on the following page).

The second one is *Dirk*, who has less AI knowledge than Regina. In his school curricula, AI education is not a subject. In his free time, he is not interested in technology. Because he is not technical affine in general, the educational concepts which address AI should be (1) in an application area that he is interested in and (2) not require a lot of technological knowledge (see Figure 16.6 on the next page). Non-digital approaches like the CS-Unplugged concept (Bell et al., 2009) could be a starting point for teaching end-users like Dirk about computer science and (X)AI.

Figure 16.5: Persona *Regina*, who is educated about AI but finds this topic only necessary in specific application areas

Regina

Critical Stakeholder




Age: 50 years
Gender: Female
Technology Experience: Has heard about AI and rates her knowledge as experienced
Context: Uses AI components in her household (e.g., robot vacuum cleaner). XAI is an unknown term for her
Problem: She interacts with AI in her household and sees it as a helpful tool. For education, she does not see a benefit of AI usage
Needs: Has a good impression of AI in general. How she could benefit from XAI is not clear to her.
Existing Solutions: Online courses that give overviews about the usage of AI in different application areas are available on online platforms
Goal of (X)AI: Give her an overview of the benefits and limitations of AI in different areas and how XAI addresses this issue

Figure 16.6: Persona *Dirk* represents a student who has little knowledge about AI and XAI

Dirk

Uninformed Stakeholder



Age: 12 years
Gender: Male
Technology Experience: He is not technical affine. He has little knowledge of AI and no knowledge of XAI.
Context: Goes to school. Here, AI is not a topic to be taught.
Problem: He is confronted with AI in his private life but lacks knowledge about it, e.g., did not know how this technology works.
Needs: Understanding the AI he is using so that he can assess its risks and benefits.
Existing Solutions: Non-digital education concepts about computer science (e.g., CS-Unplugged) could be a starting point.
Goal of (X)AI: Supports him in gaining knowledge about AI and XAI.

End-Users of Explainable AI in Companies

17

The online survey focused on identifying the actual state of (X)AI-related issues and potential in companies. To achieve this, employees of companies of different sizes and sectors were asked about AI technology's current and future development in their company by means of an online survey. The chapter is based on the work published in:

- **Weitz, K.**, Dang, C. T., André, E. (2022). Do We Need Explainable AI in Companies? Investigation of Challenges, Expectations, and Chances from Employees' Perspective. *arXiv preprint arXiv: 2210.03527* <https://doi.org/10.48550/arXiv.2210.03527>
- André, E., Bauer, W., Braun, M., Dang, C. T., Peissner, M., **Weitz, K.** (2021). Projektbericht: KI-Kompetenzentwicklung bei Sach-und Produktionsarbeit. *Lernende Systeme - Die Plattform für Künstliche Intelligenz*

17.1	Research Questions	99
17.2	Methodology	100
17.3	Participants & Companies	101
17.4	Results	101
17.4.1	Results of Employees' Company Perspective	101
17.4.2	Results of Employees' Personal Perspective	103
17.5	Lessons Learned	105
17.6	Conclusion	106
17.7	Personas of End-Users in Companies	107

17.1 Research Questions

We asked each employee about the current status as well as the strategic planning of using AI systems in their company. For this, we formulated the questions of the survey from two perspectives: (1) a broader *company perspective* and (2) a *perspective of employees* working in these companies¹. Since we want to investigate employees' perspectives as they are interacting with a (future) AI system in the company, it is important to note that the company perspective reflects the employees' subjective perception and not the company's slogan.

Employees' Company Perspective The company perspective may generally serve companies that do not yet, hardly or already use AI technologies for further strategic orientation and planning. For example, what are company motivations, usage areas, or issues? Here, the experiences and decisions gained from the current state help assess the individual potential by introducing or using (X)AI technologies. To inquire about the company perspective, that is, the view of employees about AI in their company, including a look at the existing AI applications and those planned for the future, we formulated the following research questions:

- **RQ-C1: Motivation & Risks** What motivations and risks for their company do employees see in using AI technologies?
- **RQ-C2: Usage of AI Technology** Do companies use AI technology, and if so, which applications already exist in companies?
- **RQ-C3: Future Plans** What are companies' plans regarding AI technologies?
- **RQ-C4: AI Training** What are companies' structures and plans for (X)AI employee-trainings?

1: abbreviation interpretation of the research questions: RQ = research question, E = Employee, C = Company

Employees' Personal Perspective Insights from the general attitude, knowledge, or acceptance of (X)AI technologies, including demographics, from employees' perspectives show the state of practical implementation in companies. For example, this guides improvements or what to look for when implementing AI technologies. To investigate the personal perception of employees regarding (X)AI and their experiences with AI technologies in their companies, we formulated the following research questions:

► **RQ-E1: Knowledge & Attitude**

- **RQ-E1a** How do employees rate their (X)AI knowledge and attitudes towards (X)AI?
- **RQ-E1b** How is the knowledge and attitude towards XAI related to demographic data?

► **RQ-E2: Rating of AI Technology**

- **RQ-E2a** How do employees rate the AI technologies used in their company?
- **RQ-E2b** Is there a correlation between personal AI knowledge/attitude and the rating of AI technologies in the company?
- **RQ-E2c** How does the perception of the AI technology used in their company differ depending on demographic data (e.g., age, educational attainment, company position)?

17.2 Methodology

2: the complete questionnaire can be found in the Appendix

We derived a questionnaire² with groups of questions addressing each of our formulated research questions and distributed the questionnaire as an online survey through multipliers of the *Plattform Lernende Systeme/accatech* (e.g., chambers, competence centres, corporate leaders) to cover a broad portfolio of companies and their employees. The questionnaire was in German and addressed employees of German-based companies. In this work, we focused mainly on employees with experience or knowledge of AI technologies in their companies to obtain valid results. For the evaluation, we used the following questions:

Demographic Data We collected information from participants about their age, gender, educational background, knowledge and attitude about (X)AI, their rating of the importance of XAI for different stakeholders (7-point Likert scale), and their role in the company.

Company Information To get an overview of the size and domain of the company, we asked questions about the sector and in which area (i.e., production or office work) the participants work. Here we used a combination of predefined answers and free-form answers.

AI Technology - Strategy To investigate the strategic plans towards AI for the company, we asked about plans for the usage of AI (e.g., “In which areas do your company plan to make changes with the help of Artificial Intelligence in the next years?”). Furthermore, we addressed chances (i.e., “What is driving AI development in your company?”) and risks (i.e., “What are challenges, obstacles, or problems for your company in implementing AI?”). We gave predefined answer options for each question and the possibility of writing free-text answers.

AI Technology - Usage Here, we requested detailed information about the AI technologies used (i.e., the task/goal of the AI, the field of application, the autonomy of the AI, and the duration of use). In addition, we asked, inspired by the overview of XAI metrics of Hoffman, Mueller, et al. (2018), five items on a 7-point Likert scale (1 = not at all, 7 = extremely), regarding the AI technology’s reliability, usefulness, transparency, operability, and comprehensibility.

Training Offers We investigated companies’ general and AI-specific training offers. In addition, we investigated which group of employees would benefit from AI-related training and how this training should be conducted.

17.3 Participants & Companies

We collected data from 50 participants between 25 and 66 years ($M = 45.0$, $SD = 11.3$). Thirty-four of the participants were male, and 16 were female. 80% of them had an academic educational background (i.e., bachelor/-master’s degree or higher). 24% were employed in medium-sized, 56% in big-sized companies. Here, 84% had a domain expert role, scientific expert role, or leading position. Workers and temporary staff were 16% barely represented. The companies’ sector was broadly distributed, with technical services (20%) and manufacturing (20%) as the most mentioned sectors.

17.4 Results

In the following, the results are separated according to our research questions into a subsection including analyses of the employees’ company perspective and a subsection presenting the results of employees’ personal perspective.

17.4.1 Results of Employees’ Company Perspective

RQ-C1 to RQ-C3: Motivation, Usage, & Future Plans for AI The strongest *motivation of companies* for using AI technology is an increase in productivity ($n = 23$), followed by an increase in flexibility ($n = 21$), customer requirements ($n = 18$), and adjustment of business models ($n = 18$). *Risks* by using AI are financial aspects ($n = 24$), qualification of employees ($n = 21$), and acceptance by employees ($n = 18$). 56.8% of the participants

stated that their company uses AI technology in prototypes (12 companies) or applications daily (13 companies). Furthermore, over two years, AI technology has been used in 54.2% of companies. More details were revealed by the free-form answers about the *application areas of AI*. Here, we found four clusters:

- ▶ **Quality Assurance:** Mostly, participants stated that the AI technologies help monitor and predict production quality (e.g., by predictive maintenance using image classification), which assures the quality of the produced goods or the functioning of the machines used.
- ▶ **Process Optimization:** Due to streamlining of processes (e.g., by automatically evaluating and clustering Big Data), processes are optimized. This leads to a cost reduction due to shorter and more efficient processes.
- ▶ **Support Employees:** AI is also used to support employees in fulfilling their tasks successfully, especially in office work. The usage of AI here covers a broad spectrum, from a simplification of bookkeeping to support of office-based work processes (e.g., software as a service³)
- ▶ **Interaction & Communication:** This includes communication with customers or employees by means of chatbots (e.g., check-in process of a guest in a hotel) as well as interaction in the form of robots within a physical environment (e.g., intelligent positioning where a robot pick up goods).

3: Buxmann et al. (2008, p. 500) describe the usage of the software as a service: “customers are provided with a standard software solution as a service via the Internet.”

For the *future*, participants stated that their companies focus on the usage of AI to change processes within the organization ($n = 29$), followed by the goal of developing new technologies ($n = 25$), and changes in the organization of the company ($n = 21$).

Although we have a small sample of employees, the clusters found for AI applications (e.g., process optimization), as well as the reported risks (e.g., qualification of employees) seen in the use of AI in companies, are very similar to the results of larger surveys from over 500 industry companies in Germany (Bitcom Research, 2019, 2020).

RQ-C4: AI Training Companies expecting a change in the working field. 75.7% expect an upgrading (i.e., demands on employees will grow), while 8.1% expect a downsizing (i.e., Demands on employees will reduce). 13.5% stated they were unsure or expected that the demands would not change. 48.6% of the companies stated that their employees already have the qualifications to handle the new demands. In contrast, 37.8% indicated that their employees do not have the needed qualifications, and 13.5% of the companies were not sure.

86.5% of the companies answered that they have a general qualification program (e.g., for training courses), while 5.4% stated that they do not have such an infrastructure, and 8.1% were not sure.

AI training was seen for 71% of the companies as relevant for employees with a leading position or domain experts. Only 26% stated that AI training was seen as suitable for workers. In-house training was often named as a method to educate employees regarding AI.

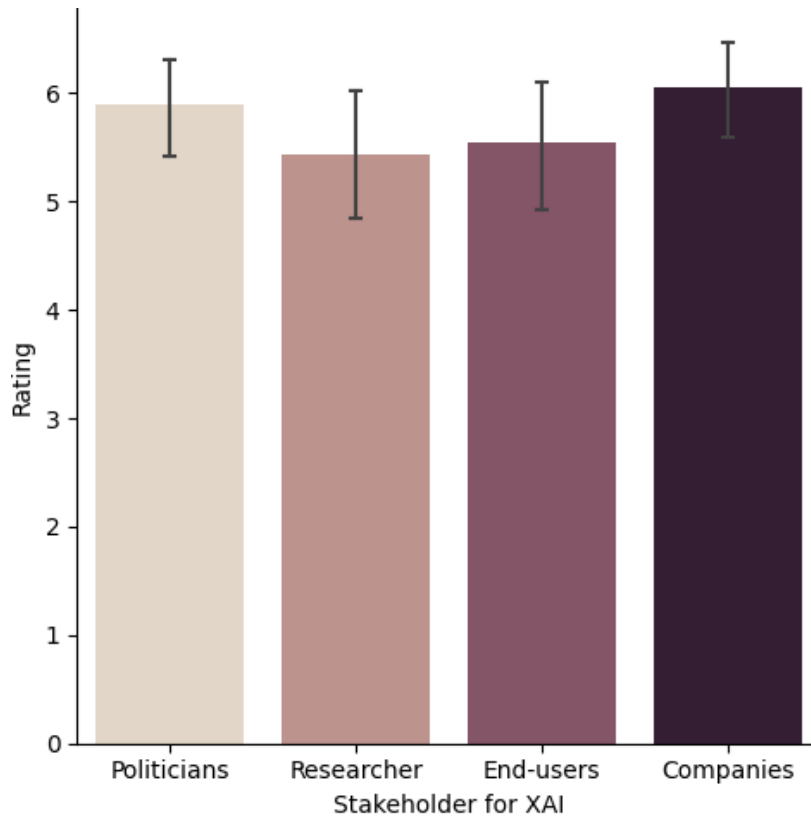


Figure 17.1: Participants rated (scale from 1 to 7) how important XAI is for stakeholder groups. Participants perceived XAI as important for all stakeholders, especially politicians and companies. Error bars represent the 95% CI

17.4.2 Results of Employees' Personal Perspective

RQ-E1a: AI Knowledge & Attitude All of the participants have heard of the term “Artificial Intelligence”. 87% of them agreed with our given definition of AI. Five participants had a different definition of AI in mind, especially focusing on “the cloud” rather than on physical machines or indicating that the term “AI imitates human behaviour” is not correct to them. 62% of the participants had heard about the term XAI, while 37.5% did not. Participants found XAI relevant for all the interest groups queried (items ranged from 1 = not important to 7 = very important), especially for companies ($M = 6.05$, $SD = 1.41$) and politicians ($M = 5.90$, $SD = 1.50$).

Conducting a one-sample t-test, we found that participants had a significantly positive view towards AI compared to the mean value of the rating scale (i.e., $M = 4$, 7-point Likert scale), $t(39) = 7.92$, $p < .001$, $d = 1.25$ (large effect)⁴ (see Figure 17.2).

RQ-E1b: XAI Knowledge & Attitude Employees rated their experience with AI technology in the company significantly positive compared to the mean of the rating scale (i.e., $M = 4$, 7-point Likert scale) (see Table 17.1) for the items comprehensibility, transparency, reliability, usefulness, and operability (see Figure 17.2).

RQ-E2a: Rating of AI Technology - (X)AI Attitude We found a significant positive correlation⁵ between employees' attitude towards AI and their rating of the AI technology in their company, $r_{sp} = .71$,

4: interpretation of the effect size d according to J. Cohen (1988) is:

$d < .05$: small effect;

$d = 0.5-0.8$: medium effect;

$d > 0.8$: large effect

5: we calculated Spearmans' Rang correlations

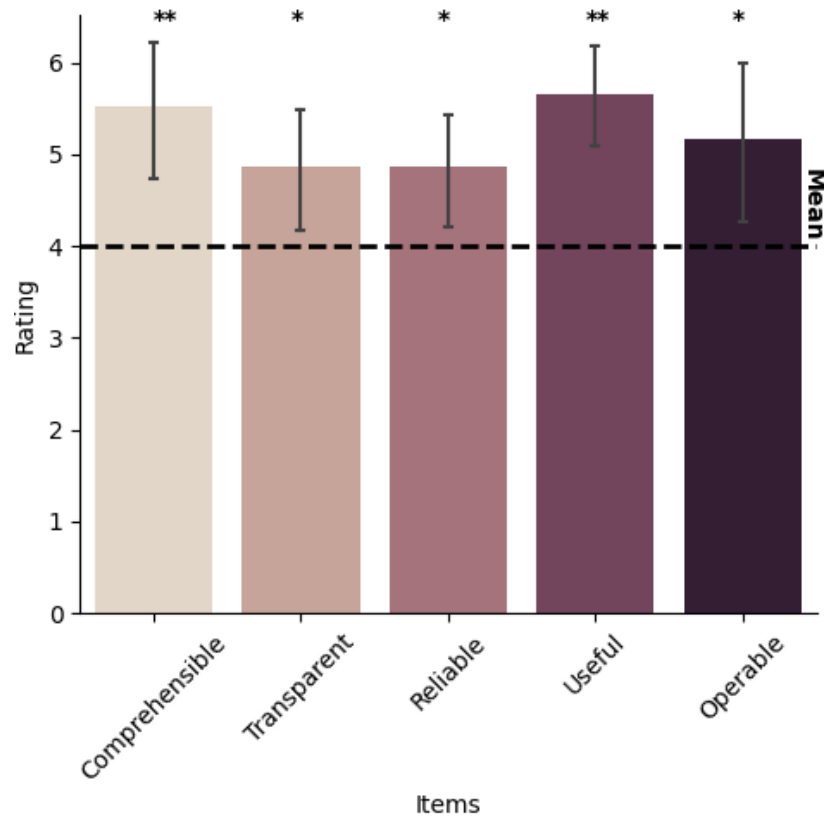


Figure 17.2: Rating of the AI technologies used in companies by employees. Employees perceived the AI technology significantly positively, compared to the mean of the rating scale, $*p < .05$, $**p < .001$. Error bars represent the 95% CI

Table 17.1: Rating of AI technology used in companies on five items. A one-sample t-test revealed that all items were perceived as significantly positive by employees

Rating item	<i>t</i> (22)	<i>p</i>	<i>d</i>
useful	5.79	< .001**	1.21
reliable	2.87	.009*	0.60
operable	2.60	.016*	0.54
comprehensible	3.88	< .001**	0.81
transparent	2.43	.024*	0.51

* $p < .05$, ** $p < .001$

$p < .001$, meaning that the higher the personal attitude towards AI of the employees, the higher is their positive perception of the AI technology in their company. The same significant positive relationship was found for the employees' attitude towards XAI and their rating on their company's AI technology, $r_{sp} = .56$, $p = .007$.

RQ-E2b and RQ-E2c: Rating of AI Technology - Demographic

Demographic attributes such as age, gender, and educational background of employees did not correlate with the perception of AI technology in their companies, but with their *role in the company*, $r_{sp} = -.61$, $p = .003$. This indicates that participants with a higher position in the company perceived the AI technology as less favourable.

The knowledge about XAI correlates positively with educational background, $r_{sp} = .53$, $p < .001$. Regarding XAI attitude, the demographic attributes *company role*, $r_{sp} = -.42$, $p = .008$, and *educational background*, $r_{sp} = .38$, $p = .015$ showed a significant correlation. These correlations indicate, similar to the perception of AI in the company, that a higher position leads to a less positive attitude towards AI. However, at the same

time, the educational background positively impacts the knowledge and attitude towards XAI.

17.5 Lessons Learned

Based on the results of our online survey, we report lessons learned that should be taken into account when designing XAI for companies:

Convince Management and Promote (X)AI Education Companies see increasing working demands for employees. To cope with these arising demand, AI-specific training is necessary. However, companies stated that they see AI training as important for employees in leading positions, neglecting workers. Experts highlight that the training of all employees, including workers, is necessary to guarantee an efficient application of AI (André et al., 2021, p. 28):

“Companies need AI competencies on the one hand from experts who develop AI and on the other hand from skilled workers who apply AI. Not all employees who work with AI have to be data scientists. Rather, everyone should be given a task in the field to which they can contribute, and they must be qualified for it.” *Andrea Sticht (Infineon Technologies AG)*

We found a correlation between employees’ attitude towards AI and their rating of AI technology used in the company, but no correlation of this rating with demographic data except for company role. For XAI, we found that the knowledge and the attitude about XAI depend on the educational background and the company position of the employees. These findings are similar to the ones we found in Weitz, Schlagowski, and André (2021) (see also Chapter 28 on page 190). Here we found that demographic information such as age and gender have no impact on users’ perception of (X)AI in an educational setting, but the educational background has an impact on the trustworthiness of the AI system. Hence, it is highly worthwhile to create and foster a positive attitude towards AI from the very beginning, especially in the leading management, to achieve appropriate trust⁶ and successful usage of deployed AI technologies later on. Since it is one of the goals of XAI to support appropriate trust in AI technology (Hoffman, Mueller, et al., 2018), training that includes XAI techniques could be the key.

6: appropriate trust refers to trust in a technical system that matches the true capabilities of it (Lee & See, 2004)

XAI is Known and an Important Issue XAI is already a known term for many employees, which is contrary to our findings in the other surveys (see also the overview Figure in Chapter 15 on page 85). This indicates that company employees are more in touch with the problem of explainability and are aware that this is an important topic. As reflected in the ratings, XAI is considered necessary, especially for companies. This awareness represents a fruitful basis for developing XAI for real-world applications.

Companies Should Address the Goal of XAI Our results also suggest that our respondents perceive AI technology as already comprehensible and transparent. Nevertheless, as our results suggest, training employees regarding AI is seen as a challenge for companies. Furthermore, we found that AI usage leads to increasing requirements of employees. With these results, whether and when XAI should be used in companies arises. We can imagine two possibilities for XAI usage: (1) XAI, especially in supporting employees in their actual tasks. Here, XAI's goal is to provide good explanations supporting people in their work (e.g., diagnosis of malfunctioning parts). (2) XAI can be used to train employees to explain the inner workings of AI (in training) to help employees work successfully with it by understanding it better, i.e., gaining AI competence. In addition, XAI can help reduce fears towards AI technology that employees may have. Overall, to identify concrete goals of XAI in companies, further studies have to investigate in more detail which and to what extent XAI methods are used in the company.

It is Necessary to Address All Employee Groups In general, we found that employees perceive the company's AI technologies as comprehensible, transparent, reliable, useful, and operable. While these results are encouraging, it is essential to note that we have responses almost exclusively from employees with academic backgrounds who are leaders or have domain expertise. Therefore, it remains unclear whether employees with other backgrounds have similar impressions. Thus, for further studies on XAI in companies, special attention should be paid to reaching other target groups, such as workers and untrained staff who operate with AI. In addition, by recruiting participants via the Plattform Lernende Systeme, a selection bias (Heckman, 1979), leading to responses, especially from people interested in the topic and therefore having a more positive view towards (X)AI.

17.6 Conclusion


The elevation of company success and innovation through AI is one reason why companies address AI in their strategic plans. Legal regulations force them to have comprehensible AI systems. XAI refers to methods that address this issue. While research is just starting to investigate the impact of XAI on end-users in lab experiments, real-world applications are not the focus of investigation right now. To design and evaluate XAI for companies, the perceptions and needs of employees should be given attention to using AI in a human-centered way. Our online survey has moved research closer to this goal by investigating employees' perspectives towards X(AI). Our findings in this project report suggest that fostering a positive attitude toward AI on the management level is essential for successfully integrating AI technologies in companies. XAI is already a known topic for employees and is perceived as an important issue. With our insights, we encourage researchers to include employees' attitudes towards (X)AI in their design to create a more HC-XAI.

17.7 Personas of End-Users in Companies

Based on our online survey of companies, I present a prototypical end-user in this area. Since most survey participants had a domain expert role or a leading position, *Wolfgang* represents a leading person in a German-based company who is already aware of the importance of using AI for the future of his company. He is also aware that his employees need skills to handle the new requirements that emerge with the usage of AI. Therefore, he has to train his employees to support them with upcoming new challenges in AI. He can use his company's existing internal training program (see Figure 17.3).

Wolfgang

Company Leader



Age: 40 years
Gender: Male
Technology Experience: Has heard about AI and has a positive attitude towards AI and XAI.
Context: He works in a leading position in a German-based company and is confronted with the topic of AI in his work.
Problem: AI is used in his company as a prototype, but he sees new educational demands of his employees when interacting with AI.
Needs: Support employees to work successfully with an AI system.
Existing Solutions: His company has a general training program, which includes training on various topics. This could be used for (X)AI training as well.
Goal of (X)AI: Support him in offering appropriate (X)AI in-house training courses for his employees in the company.

Figure 17.3: *Wolfgang* represents a prototypical person in a leading role in a company. He is aware of the chances of AI for his company. To use AI efficiently, he has to support his employees with appropriate (X)AI training programs

18.1	Research Questions	109
18.2	Survey Design	109
18.3	Methodology	110
18.4	Participants	111
18.5	Results	112
18.5.1	RQ1: Preferred App & Explanations	112
18.5.2	RQ2: Impact of Users' Attributes	115
18.6	Clustering of Results to Create Personas	116
18.7	Lessons Learned	119
18.8	Personas of End-Users in Mobile-Health	120

The Mobile Health Survey addresses end-users views about stress classification with the help of a mobile phone. For this, two representations of an AI-based classification system were investigated: (1) A data-based app and (2) a photo-based app. Data-based refers to an application that uses sensor data (e.g., heart rate) and context data (i.e., calendar entries). Photo-based refers to an application that uses features of a user's image (e.g., eye region). For these two apps, a variety of explanatory styles and explanation contents were investigated. The presented work was published in:

- ▶ **Weitz, K., Zellner, A., & André, E. (2022).** What Do End-Users Really Want? Investigation of Human-Centered XAI for Mobile Health Apps. *arXiv preprint arXiv: 2210.03506* <https://doi.org/10.48550/arXiv.2210.03506>
- ▶ **Zellner, A. (2021).** *Towards personalized explanations in digital health - user-centered explanations for mobile health applications* (Master's thesis). University of Augsburg.*

We used three different kinds of explanation designs to investigate these different explanation styles and contents: *live explanation*, *feature-based explanation*, and *ask-the-app explanations* (see Figure 18.1 on the next page). All three explanation types have in common that the user can actively influence the explanation. This makes the explanation system interactive (Weld & Bansal, 2019). However, since we were interested in user preferences and less in concrete AI models, the explanations shown are not based on real AI models.

- ▶ **Live explanation** The live explanation implements an exploratory explanation paradigm (Shneiderman, 2020a). The idea is that the user finds out how input variables (e.g., resting pulse) affect output (e.g., stress classification) by trial and error.
- ▶ **Feature explanation** A feature tag approach was chosen for this type of explanation¹. In our explanations, different features of a (fictional) ML model are used. The size of the displayed feature represents the influence of the feature on the prediction. More detailed information about a feature is displayed to the user by clicking on it.
- ▶ **Ask-the-app explanation** This explanation is a dialogue-inspired explanation. This type of explanation allows the user to ask questions about the prediction of the AI system or to choose from several predefined questions, the one whose answer interests him.

1: Here, data can be displayed in different sizes and colours indicating, for example, different degrees of importance (Halvey & Keane, 2007)

*Co-supervised by the author of this dissertation

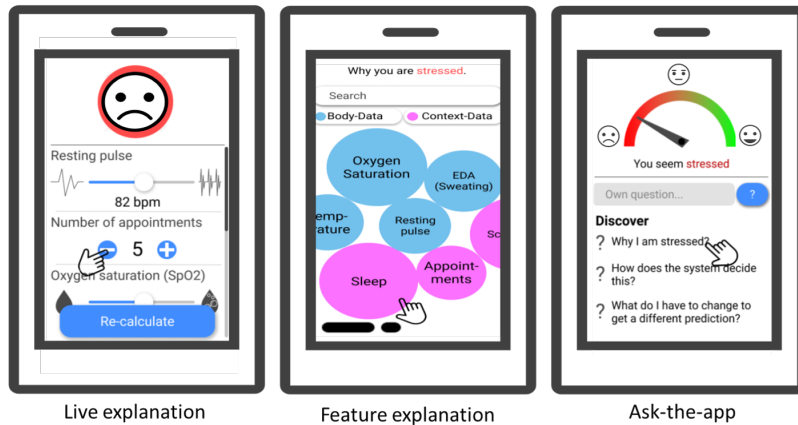


Figure 18.1: Three types of interactive explanations when classifying stress/no stress were investigated in the survey. Images illustrate the explanations for the data-based app: *Live explanation* that allows users to change parameters, *Feature explanation* that displays the features used for the decision, where the size of the circle reports the importance of the feature for the classification, and *Ask-the-app* explanations where users can ask the app-specific questions regarding the decision. The hand icon symbolizes a user action. Figure adapted from Zellner (2021)

18.1 Research Questions

The research questions (RQ) of the survey include:

- ▶ **RQ1: App & Explanation Preferences**
 - **RQ1a:** Which representation of an AI classification system (i.e., data-based or photo-based) do end-users prefer?
 - **RQ1b:** Which content and type of explanation are preferred by end-users?
- ▶ **RQ2: Impact of Users' Attributes** Are end-users attitudes towards explanations related to personality, technical affinity, or demographic attributes (e.g., gender, age)?

18.2 Survey Design

We conducted an Amazon Mechanical Turk (MTurk) online survey to investigate the research questions. Here, we addressed English and German-speaking people. On average, the survey took 30 minutes and participants were compensated \$4. The online survey comprised three phases:

- ▶ **User information & Preferences** At the beginning of the survey, we collected demographic information, personality traits, and users' attitudes regarding technical affinity and their attitudes and usage of health applications. We asked about participants' attitudes towards (X)AI only at the end of the survey to not bias users.
- ▶ **Mobile Health Application** Next, we presented the users with an example of a photo-based app and a data-based app and asked them about their preferences for one of the two apps. We then asked how much they would like an explanation of the app shown and what questions they would ask of the app.
- ▶ **Explainability** Based on their preference for one of the two mobile health apps, participants saw the input data of an potential person, which differed regarding of the type of the app². Participants had the task to classify this person as stressed/not stressed and how they would explain their decision (i.e., "Assume that you are supposed to explain this decision to someone else. You may use all information the app provides.").

2: For the data-based app, heart rate, blood oxygen, EDA, calendar entries and sleeping time of the imaginary person has been displayed. For the photo-based app, a picture of a person was shown

- **Personalised Explanations** In the last section of the survey, participants rated three different types of explanations and provided information about the time they would spend understanding the explanation.

18.3 Methodology

3: the complete questionnaire can be found in the Appendix

The survey included the following questions³:

Personality To assess the personality of the participants, we used the Ten-Item Personality Inventory (TIPI) questionnaire developed by Gosling et al. (2003). TIPI investigates the Big-Five personality construct (i.e., extraversion, agreeableness, conscientiousness, emotional stability, and openness to experience). For this, ten items (e.g., “I see myself as sympathetic, warm” for agreeableness) on a 7-point Likert scale (1 = Disagree strongly to 7 = Agree Strongly) were asked.

Technical Affinity To measure the technical affinity of the participants, we used the Affinity for Technology Interaction Short scale (ATI-S) (Wessel et al., 2019), which includes five items less than the original ATI scale (Franke et al., 2019). Each of the four ATI-S items (e.g., “I like testing the functions of new technical systems”) was rated on a 6-point Likert scale (1 = completely disagree to 6 = completely agree).

Health App We investigated participants’ usage of mobile devices for healthcare applications. After that, participants were asked which of the two apps presented (i.e., data-based app or photo-based app) they would use and then had to justify their preferred app via free-form feedback.

Explanations - Type Participants had to rate three different types of explanations (see Figure 18.1 on the preceding page). For this, three visualisations were evaluated: *live explanation*, *feature-cloud explanation*, and *ask-the-app explanation*. Here, we used five items of the Explanation Satisfaction Scale (ESS) proposed by Hoffman, Mueller, et al. (2018): understanding, satisfaction, sufficient detail, useful for users’ goals, and precision (rating from 1 = I disagree strongly to 5 = I agree strongly). In addition, we asked participants about their willingness to try the explanation and their intention to personalise the explanation (e.g., “I would like to determine for myself what factors are considered for the ask-the-app explanation.”) on a 5-point Likert scale (1 = I disagree strongly to 5 = I agree strongly). Afterwards, participants could state what they liked or disliked about each explanation type using free-form feedback.

Explanations - Content Regarding the content of explanations, users were asked to rate which of the four presented questions they would ask themselves (i.e., “How likely would you be to ask yourself any of the following questions while using the app?”) on a 5-point Likert scale (1 = extremely likely to 5 = extremely unlikely). The questions presented are inspired by Hoffman, Mueller, et al. (2018, p. 4) and Lim et al. (2009, p.

2120) and are intended to map user needs to the app and to the explanations required:

- ▶ Why do I get this prediction?
- ▶ How does the system come up with this prediction?
- ▶ Why did I not get another prediction?
- ▶ What do I have to change to make the system change its prediction?

In addition, we investigated participants' preferred content of explanation (i.e., level of detail, comparison with the average of users) by contrasting two explanations (e.g., explanation with few details, explanation with many details) and having participants rate which of the explanations they would prefer.

Explanation - Time To investigate the amount of time participants are willing to spend for explanations, we asked two questions: (1) How much time would you invest in understanding the app's explanation?" (less than 1 minute; 1-2 minutes, 2-5 minutes, more than 5 minutes) and (2) "If you had the opportunity to ask questions to the app or interact with the app, would you be willing to put more time into understanding an explanation?" (5-point Likert scale; 1 = totally disagree, 5 = totally agree)

18.4 Participants

A total of 92 participants between 24 and 70 years ($M = 42.10$, $SD = 10.4$) finished the online survey. Forty-two participants were female, 49 were male, and 1 divers. All participants stated that they are currently in the United States. In addition, 35.8% of the participants indicated that they had a secondary degree or apprenticeship degree, and 64.1% noted that they have a university degree.

68.6% of the participants use a mobile phone in their daily life, and 28.3% use a mobile phone in combination with a wearable. 50% of the participants reported using health-related apps at least once a week. Especially the usage of fitness apps ($n = 44$), followed by wellbeing apps ($n = 23$), and nutrition apps ($n = 23$) was reported.

97.8% of the participants stated that they had heard the term AI and 96.7% of them agreed with the definition of AI given. Furthermore, 95.7% of the participants indicated that they had a slightly or strong positive attitude towards AI (rated 4 or more on a 7-point Likert scale). Regarding XAI, 95.7% of the participants were not familiar with the term. Nevertheless, after giving a definition of XAI, 93.5% had a slightly or strong positive attitude towards XAI (rated 4 or more on a 7-point Likert scale). Participants stated that XAI is important for different stakeholder, especially end-users and companies (see Figure 18.2 on the next page).

Technical Affinity Participants had an technical affinity of $M = 4.03$, $SD = 1.14$ with a Cronbach's α of .88⁴. The results of a one-sample t-test show that participants had significantly high values in technology affinity, $t(91) = 4.47$, $p < .001$, $d = .47$ (medium effect) compared to the mean of

4: Wessel et al. (2019) suggest to report Cronbach's α as an indicator of reliability. Here, values between 0.7 to 0.8 indicate a satisfactory reliability (Bland & Altman, 1997)

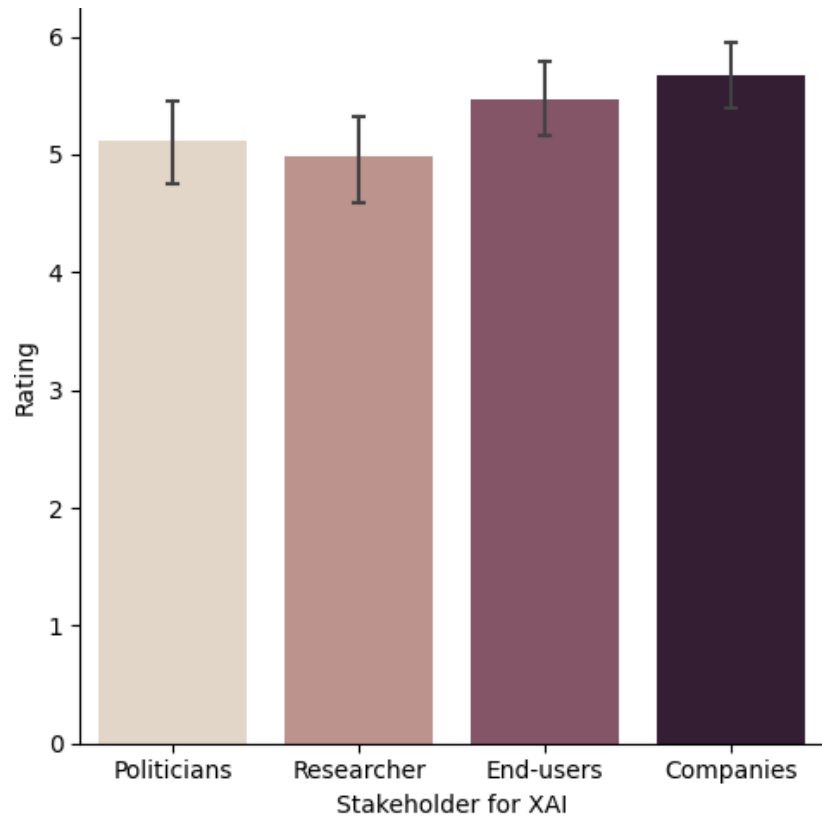


Figure 18.2: Rating of the Health Survey participants regarding the importance of XAI. Participants stated that XAI is important for different stakeholders, especially for end-users and companies

Table 18.1: Cronbach's α for the five personality items of the TIPI questionnaire (Gosling et al., 2003). Agreeableness and conscientiousness are below 0.7 and therefore indicating not sufficient reliability

TIPI item	Cronbach's α
Extraversion	.81
Agreeableness	.42
Conscientiousness	.67
Emotional Stability	.84
Openness to Experiences	.71

5: we calculated Spearmans' rang correlations

3.5 (6-point Likert scale). The technical affinity did not significantly correlate⁵ with age nor gender.

Personality Regarding the TIPI's Big-5 items, the agreeableness and conscientiousness scales had a very low Cronbach's α , indicating the low reliability of these scales and were therefore excluded from further analyses.

18.5 Results

18.5.1 RQ1: Preferred App & Explanations

RQ1a: Preferred App

With 88%, most of the participants preferred the data-based app. Reasons for this could be investigated in the free-form answers. We found three general reasons why people prefer one of the two apps:

Question type	<i>t</i> (91)	<i>p</i>	<i>d</i>
Why?	9.26	< .001**	0.97
Why not?	0.19	.849	0.02
How?	9.16	< .001**	0.95
What?	3.75	< .001**	0.39

***p* < .001

Table 18.2: Rating of potential questions users would ask themselves when seeing a mobile health app stress classification. A one-sample *t*-test revealed that all question types except Why not?-questions were perceived as significantly important by participants

- **Dependability:** Thirty-eight participants (3 of them preferred the photo-based app) stated several things regarding dependability. Here it was frequently mentioned that the data-based app is more reliable since it uses much more data for evaluation than the photo-based app (e.g., “It would be able to collect data about me that would not be visible in a photo.”). In general, participants were very critical of the reliability and general functioning of the photo-based app (e.g., “The other one [photo-based app] seems like pseudoscience.”), whereas the data-based app was perceived as more objective (e.g., “Further, I think the data-based application is more likely to be objective, whereas deducting characteristics from a photo seems more subjective and therefore less valuable to me.”).
- **Privacy:** Twenty-six participants (1 of them preferred the photo-based app) raised privacy concerns. In particular, people did not feel comfortable providing a photo of themselves and therefore tended to use the data-based app (e.g., “I don’t like the idea of taking my photograph in an app, because I may not have control over how that photograph may be used.”). Regarding the data-based app, participants tend to assume that the shared information was not so personal compared to the photo-based app (e.g., “I suspect that more personal information about myself could be obtained via the photo than via the data used in the data-based application.”).
- **Usability:** Twenty-two participants (5 of them preferred the photo-based app) stated reasons related to the app’s usability. While the photo-based app was described as easier and quicker to use (e.g., “It is quicker to understand and look over), the data-based app was described as more understandable (e.g., “I feel like it gives me a better understanding of my health than my face”).

RQ1b: Preferred Explanations

Preferred Content of Explanation Based on the triggers *why*, *how*, *why not*, and *what* proposed by Hoffman, Mueller, et al. (2018), we investigated which information participants wanted to know from a stress-classification app. Results of a one-sample *t*-test show that all questions except the Why not?-question were significantly positively rated and can therefore be assumed to be relevant for users in mobile health scenarios (see Table 18.2). In addition, users tend to prefer more detailed explanations (72.8%) as well as explanations that include a comparison to the average (68.5%).

Preferred Type of Explanation Overall, 56.8% of the participants prefer to use the ask-the-app explanation in their daily life, followed by the feature explanation (23.5%) and the live explanation (19.8%).

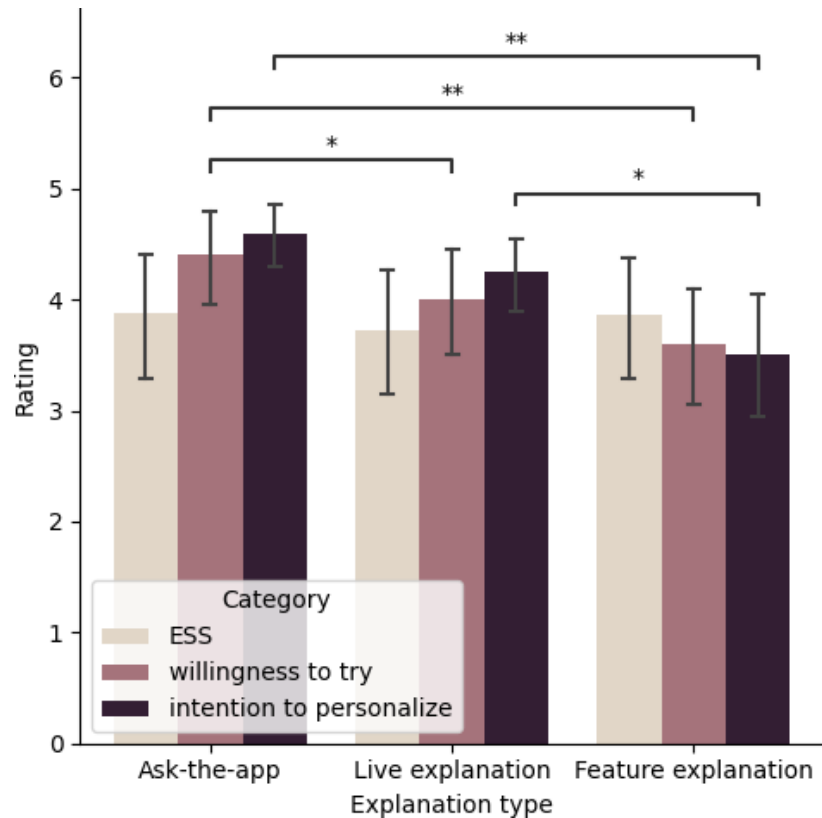


Figure 18.3: Participants' rating (ranging between 1 to 7) of the three different explanation types (i.e., ask-the-app, live explanation, and feature explanation) shows that ask-the-app invites users to try it out. Furthermore, ask-the-app and live explanations invite users to personalize them (i.e., selecting features that should be relevant for the classification). Error bars represent the 95% CI. * $p < .05$, ** $p < .001$

Comparing the three *types* of explanation (i.e., live explanation, feature-cloud explanation, and ask-the-app explanation), we found that the explanation satisfaction did not differ significantly between the types (see Figure 18.3).

Regarding the willingness to try the explanation type, the ask-the-app explanation was rated significantly higher than the live explanation, $t(91) = -2.42$, $p = .018$, $d = .25$ (small effect) and than the feature explanation, $t(91) = -3.53$, $p < .001$, $d = .37$ (medium effect). While the explanation satisfaction seems similar to each explanation type, the ask-the-app invites users to try it out - maybe because the app forces an interaction through asking questions (see Figure 18.3).

Regarding the intention to personalise the explanation, the live explanation was rated significantly higher than the feature explanation, $t(91) = 2.97$, $p = .004$, $d = .31$ (medium effect). The ask-the-app explanation was rated higher than the feature explanation, $t(91) = 4.29$, $p < .001$, $d = .45$ (medium effect). The feature explanation seems not so interesting for users to personalise (see Figure 18.3).

Time Spend for Explanations Regarding the interaction time, participants' answers were evenly distributed between 1 to 2 minutes up to more than 5 minutes (see Table 18.3 on the facing page). Interestingly, the results indicate that users are willing to spend some time understanding an explanation.

When asking whether participants are willing to spend more time with an explanation when asking questions or interactively interact with the app,

Time	Percent
less than 1 minute	6.5%
1-2 minutes	30.4%
2-5 minutes	32.6%
more than 5 minutes	30.4%

Table 18.3: Participants answer the question, “How much time would you spend understanding an explanation in the mobile health app?”

we found a statistically significant difference to the mean of 3 (5-point Likert scale) conducting a one-sample t-test, $t(91) = 11.5$, $p < .001$, $d = 1.19$ (large effect), indicating that users tend to spend more time with an app that provides an interactive interface for explanations.

18.5.2 RQ2: Impact of Users’ Attributes

We will report correlations regarding user attributes and the explanations presented in the following. However, the reported values only reflect correlations, not causal relationships.

Demographic Information Regarding the time participants would spend to understand an explanation, we found a significant positive correlation with the participants’ age ($r_{sp} = .24$, $p = .019$), indicating that the older the users, the more willing they are to spend time to understand an AI’s explanation. On the other hand, no significant correlations were found for gender and educational background. Regarding time participants would spend on an explanation when they have the option to ask questions during the explanation, we found a significant positive correlation with the participants’ attitude towards AI ($r_{sp} = .30$, $p = .004$) and XAI ($r_{sp} = .28$, $p = .008$). A more positive attitude towards AI as well as XAI leads to higher time investments to understand an explanation.

Regarding attitude towards AI and XAI, we only found a significant positive correlation between gender and the attitude towards AI ($r_{sp} = .23$, $p = .026$), indicating that men have a more positive attitude towards AI. On the other hand, for attitude towards XAI, no significant impact on age and gender was found. Regarding educational background, we found no relationship with participants’ attitudes toward (X)AI nor their willingness to spend more time with an explanation or to ask questions about an explanatory system.

These results indicate that the attitude towards (X)AI is an essential driver for users to take the time to interact with an explanatory system. Gender and age seem to have an impact on the attitude towards AI but not on XAI. Also, the educational background does not indicate the attitude towards (X)AI in general nor the interaction with an explanatory system.

Personality To evaluate the impact of personality on the different explanation ratings, due to Cronbach’s α , only the traits *Extraversion*, *Emotional Stability*, and *Openness to Experiences* are used.

Regarding time participants would spend on an explanation when they have the option to ask questions during the explanation, we found a significant positive correlation with the personality trait Extraversion ($r_{sp} = .22$, $p = .036$) and Openness to Experiences ($r_{sp} = .23$, $p = .030$).

This indicates that users with higher values in these two personality traits are more willing to spend time with an interactive explanation.

Regarding the time participants would spend understanding an explanation, we found no significant relationship with any personality traits.

Regarding the content of the explanation (i.e., comparison to the average user and degree of detail of the explanation), we found no significant correlation with any of the personality traits.

Technology Affinity Regarding the time participants would spend to understand an explanation, we found a significant positive correlation with the technical affinity of the participants ($r_{sp} = .27, p = .010$), indicating that users with higher values in technical affinity tend to invest more time in understanding an explanation.

Regarding the content of the explanation (i.e., comparison to the average user and degree of detail of the explanation), we found no significant correlation with technology affinity.

18.6 Clustering of Results to Create Personas

Building on the findings collected during the survey and the requirements derived, we designed user personas in the next step. The different argumentative identified during the application selection have indicated first divergences in the motivation and attitude of the users. We found three different views of users: Dependable and accuracy focused (Persona 1 - Anni), perception and usability focused (Persona 2 - Karl), and privacy and commitment focalised (Persona 3 - Michael) (see Figure 18.6 on page 122). With these three clusters as a basis for three personas, additional answers were investigated for possible patterns and affirmations of the personas. Due to the low number of respondents who chose the photo option, the results are based on the proportion of the data-based app. The positive and negative impressions for the different explanation types were now separated according to the three initial clusters (see Figure 18.4 on the facing page). Analysing users' free-form feedback content, we found an overall positive sentiment for persona 1 with 99.4% and persona 2 with 98.5%. However, for persona 3, we found 96.1% an overall negative sentiment, reflecting the aversion against the photo-based app.

Next, we summed up general requirements found during the analysis of the results into Table 18.4 on page 118. The requirements were elicited from the participants' preferences, intentions, and feedback for the three interactive explanations shown.

After that, we investigated how users of the three clusters responded to the three interactive explanation types (i.e., ask-the-app, live-, and feature explanation). A summary of the free-form feedback for the three explanation types is displayed in Figure 18.5 on page 119. All users mention similar terms when comparing the different explanation types at a glance.

Our results for *RQ1b: Preferred Explanations* paint a more precise picture. Based on the closed-ended questions, the *ask-the-app explanation* was generally perceived as the most satisfying. However, the personas express



Figure 18.4: We generated word clouds for the different clusters of prototypical users based on the free-form feedback to the question “Why do you prefer this app over the other?”. The overall sentiment for persona 1 and 2 was positive, while the sentiment for persona 3 was negative

different tendencies of satisfaction. Persona 1 and Persona 3 mentioned the improved depth and amount of details the most. Thus, both positively perceived the feature of asking questions. Conversely, Persona 2 focused on engaging design and understandability while criticizing the ask feature, as it is indirect and time-consuming. Persona 3 also criticizes the high head complexity that the app entails. For the *feature-based explanation*, Persona 2 appreciated the ability to see the features’ impacts directly the most. Persona 1 is fond of the overview it offered, but at the same time, with 75%, it was also the one who criticized the lack of comprehensibility and usefulness the most. Persona 3 paints a balanced picture with the highest liking for the overall design, while superficiality and lack of depth are the most criticized.

From this more qualitative analysis, we further explored the persona types quantitatively. 27 users reflecting Persona 2 report a significantly higher emotional stability ($M = 5.80$, $SD = 1.44$) than the 34 of Persona 1 ($M = 5.25$, $SD = 1.36$) and the 26 of Persona 3 ($M = 4.54$, $SD = 2.09$), $p = .045$. Furthermore, 34 participants identified as Persona 1 rated why, how, and why not question significantly higher in the probability of self-questioning than others assigned to different personas.

Table 18.4: XAI requirements and their description. Requirements were elicited from the user survey. Table adapted from Zellner (2021)

Requirement	Description
Simplicity	The interface should be easy and fast to understand
Forecast	To see the impact of a parameter if its value would be changed
Quality	Explanations should provide sufficient details
Queryable	It should be possible to ask follow-up or personal questions towards the AI
Personalized	The contents and the explanation should be personalised towards the recipient. Personalised explanations can take more time
Context provision	Users should be able to provide additional context for the measured features
Feature explanations	Features used in the prediction should be explained

The question of what would have to change did not indicate any significant differences dependent on different personas as Persona 1 ($M = 3.79$, $SD = 1.20$), Persona 2 ($M = 3.19$, $SD = 1.24$), and Persona 3 ($M = 3.77$, $SD = 1.31$) demonstrated similar interests.

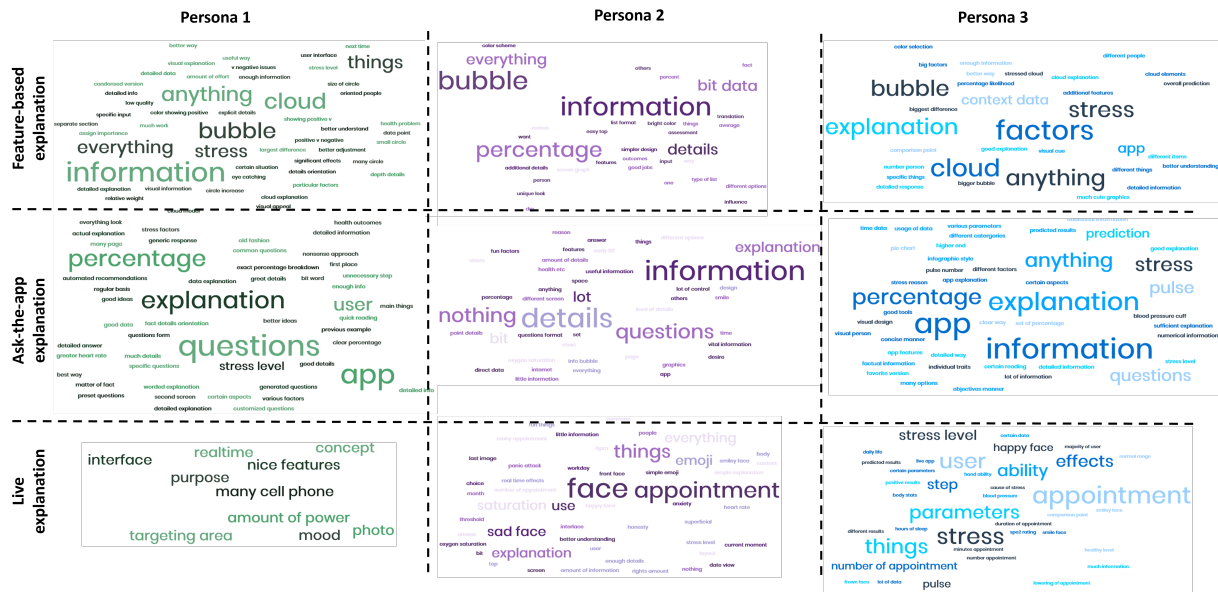


Figure 18.5: Word clouds regarding the three types of interactive explanations (i.e., feature-based explanation, ask-the-app explanation, and live explanation) for the three personas. Bigger words indicate often-mentioned terms

18.7 Lessons Learned

Data-based XAI Applications are Preferred The results of the survey show that participants preferred the data-based health app. Reasons for this were the *dependability* and *usability* of the system. In addition, *privacy* was mentioned by participants as an important reason for choosing the data-based app. The reason for this was mainly the argument that pictures are very personal information and that they do not want to share pictures that show their vulnerability (i.e., stressed) with a system where they do not know if these pictures will be used only for this purpose. In these statements, it can be seen that XAI designers should be aware of users' subjective feelings since they are powerful drivers of decisions. Barredo Arrieta et al. (2020) indicate that privacy awareness should be created with the help of XAI. Further research should investigate whether privacy concerns about a healthcare application persist even though these applications provide explanations.

Users Want Detailed Explanations Similar to the results in the pilot studies we conducted in Hald, Weitz, et al. (2021) and Weitz, Vanderlyn, et al. (2021) (presented in the Chapter 27 on page 160), users stated in the survey that they want very extensive information within an explanation. However, this statement must be taken with a grain of salt because it makes a difference whether users are expressing their desires for explanations for a potential app or whether they are being asked to evaluate the explanations of an app they are actually using. Whereas users in the pilot study of Weitz, Vanderlyn, et al. (2021) indicated that they wanted extensive and detailed information in an explanation, the main study showed that detailed explanations were quickly found to be annoying because (1) information was repeated and (2) the extent of the explanation was often disruptive to actual task completion because too much information

impacted users. Similar were the results we found in Hald, Weitz, et al. (2021) (see Chapter 27 on page 160 for details). While participants stated they wished for explanations with a suggestion of solutions, in the main study, these explanations did not improve participants' impression of the VR robot. In future studies, the desire for detailed explanations should always be evaluated directly with the application to be used.

Interaction and Personalisation are Important for Users The survey results show that users prefer the personalisation of explanations and the possibility of using an interactive explanatory system. Users would also invest more time in dealing with the explanatory system for these aspects. While explanation satisfaction is similarly high for all types of explanations, there are differences in the willingness to try out such a system and the personalisation of users. The explanation types ask-the-app and live explanation show the highest approval by users. These types are also the ones that allow users to interact with the system, such as asking questions or changing parameters that are relevant to stress. In addition, participants would ask themselves a lot of questions (i.e., Why?, Why not?, How? What?) when receiving a stress classification decision. These empirical results support the claim of Shneiderman (2020a), who uses the example of an interface for mortgage loan explanations, which allows users to interact with the system and to try out and experience for themselves the results that arise from changes to parameters. The interactivity in an explanatory system is desired by users and must be considered in the XAI design of future studies.

18.8 Personas of End-Users in Mobile-Health

Our online survey showed that the application area healthcare triggers security concerns among users due to the collection of sensitive health data. The persona *Michael* is prototypical for this critical user group (see Figure 18.6 on page 122). The concerns affect the use of the app. Persona 1 (Anni) and Persona 3 (Michael) placed the most value on the changeability and exploratory character of the interactive explanations. In contrast, Persona 2 (Karl) valued the simplicity and understandability of the design more and the ability to provide a good overview. Persona 1 criticized the lack of details and inability to be more specific about the feature measured the most. Persona 2 added the criticism of not showing enough information simultaneously.

Persona 1: Power User Based on the qualitative and quantitative analysis of the data collected following statement was derived for the first persona: **Power user Anni, who enjoys details and technology, willing to put significant effort into applications found beneficial.** According to the Oxford Learners' Dictionary, the term power user refers to someone able to use more advanced features and engage in more complex topics than other users (Oxford Learner's Dictionary, 2022). Anni is intended to mirror users in our survey who actively use mobile phones to track their health, e.g., to improve their fitness. An interactive XAI design is the basis to satisfy Anni's request for queries to the app that encapsulates these traits. Anni is described as a persona with high

precision standards and who is unsatisfied with uncertainties or vague statements. According to the data, her dominant personality traits are diligence, tolerance, and emotional stability. Further, the data shows Persona 1 as engaging in technology and an active user of mobile health apps. Therefore, practical explanations should extend her knowledge and allow her to deepen her understanding of the respective field. In contrast, unsuitable explanations lack these features.

Persona 2: Casual User The following persona statement can be derived from the previous results for the second persona: **Casual user appreciates easy and fast consumable information and does not want to spend too much time retracing AI.** Data on Persona 2, *Karl*, displayed temperate interest in details while understanding the core principles of the prediction, why, and how it was derived are essential. A deeper immersion is often not perceived as necessary by casual users. Hence, simplicity, understandability, as well as general appearance are valued. Due to the distribution of Persona 2 in the personality traits, his most pronounced traits are emotional stability, diligence, and openness to new experiences. He uses a smartphone and monitoring apps but is less intensive than Anni (Persona 1). Persona 2 is more usability-focused. Hence, an intuitive design is vital for Karl. This type displays reduced interest in intentions besides why and how. Existing solutions that focus on providing overviews can serve as suitable orientations.

Persona 3: Sceptical User Our online survey showed that the application area of healthcare triggers security concerns among users due to the collection of sensitive health data. The persona *Michael* is prototypical for this critical user group. This last type is influenced by hesitancy and general usage concerns. Therefore, the subsequent persona statement was constructed: **Sceptical user, who is reserved about sharing too much information with new applications, commitment increases with increased trust.** Michael is described as somewhat sceptical and reserved. Privacy and nontransparent applications agitate him. Thus, he uses trusted applications and limits monitoring apps on non-invasive systems. Since Persona 3 is initially very reserved, too forceful or manipulative explanations are rejected. The commitment could gradually increase with Michael gaining more trust in the application. Personalisation turns out to be more a means for incremental learning than for instant adaptation of explanations. Apps implementing such a learning curve can serve as valuable inspirations. This persona stood out for its concern and reluctance to divulge personal data. For example, many users chose the data-based application over the photo-based approach since a photo of one's face was seen as too deep an invasion of privacy. The concerns affect the use of a mobile health app. XAI design could address these concerns and aim to increase transparency and promote user trust in the app.

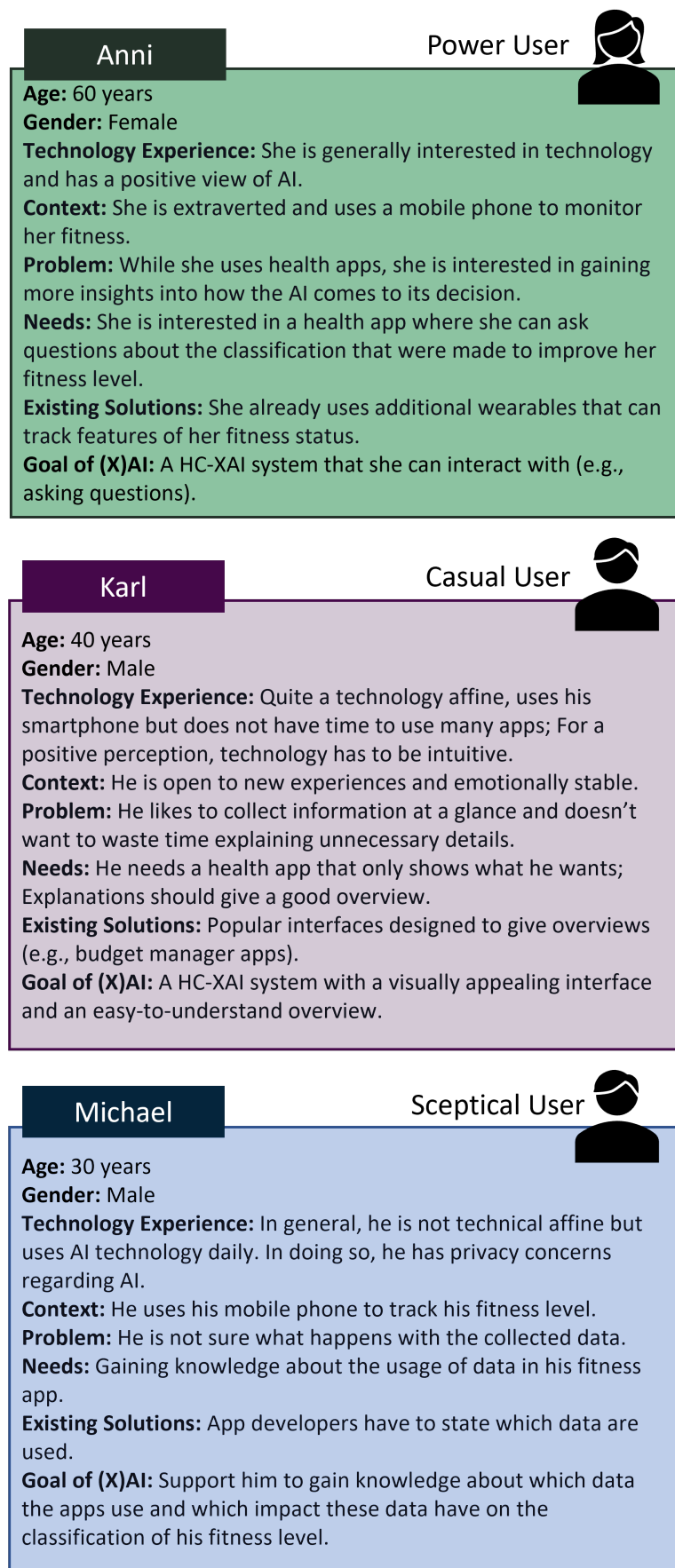


Figure 18.6: Three personas were derived based on the empirical data: *Anni*, a power user, *Karl*, a casual user, and *Michael*, a sceptical user

The evaluation of end-user in different scenarios shows that the opinions about (X)AI vary. While museum visitors are generally critical of the use of AI in education, employees in companies see great potential in the use of AI for their company to increase their productivity, work more flexibly, and meet customer requirements better. Employees consider XAI essential, and they are satisfied with the transparency of AI in their companies. However, companies' employees also note that using AI brings more job requirements that necessitate training workers who have to interact with AI. In healthcare, users are sceptical about using stress recognition apps that use images to predict stress. In particular, they express privacy concerns. Regarding XAI, end-users prefer interactive explanations to ask questions to the app.

By evaluating three surveys in different contexts of use for XAI, this chapter has also illustrated that different scopes can be addressed by such surveys, depending on the survey objectives. The persona approach presented here can be applied to these surveys of various capacities and enables XAI designers and developers to integrate user needs from the initial needs assessment to the design of concrete XAI applications.

VI. TECHNICAL REALISATION

20.1 Overview

In the previous chapters, I discussed user needs in different application scenarios. The following chapters will explain the technical foundations for implementing the AI systems we used in the six experiments presented in this dissertation. We used rule-based systems as an example of a knowledge-based white-box approach in the VR-Robot Study and the Conversational AI Study on a technical basis. These are suitable for creating prototypes that can be replaced with other architectures (e.g., DNN, or learned rules) later. The technical implementation for both studies is clarified in more detail in the following sections. In addition, the design of the XAI components (i.e., explanation content and explanation type) is described (see Figure 20.1).

Cooperation & Collaboration	
VR-Robot Study	
Content of explanation	Type of explanation
<i>Why not? How?</i>	<i>Verbal - Text (explanation vs. explanation & solution)</i>
Cooperation with an industrial robot	
Conversational AI Study	
Content of explanation	Type of explanation
<i>Why? Why not?</i>	<i>Verbal - Text (personal vs. impersonal)</i>
Collaboration with an AI dialog partner	

20.1 Overview	125
20.2 Rule-based Industry Robot in VR	126
20.2.1 Virtual Reality Design	126
20.2.2 Explanation Design	127
20.3 Rule-based Conversational Dialog Partner	128
20.3.1 Game Design	128
20.3.2 Explanation Design	130

Figure 20.1: Overview of the design of the explanation content and explanation type for the two conducted experiments in the industry and production application scenario

The technical implementations described in this chapter are based on the following publications:

- *VR Robot Study*
Hald, K.*, **Weitz, K.***, Rehm, M., & André, E. (2021). “An error occurred!” - trust repair with virtual robot using levels of mistake explanation. In K. Ogawa, T. Yonezawa, G. M. Lucas, H. Osawa, W. Johal, & M. Shiomi (Eds.), *Proceedings of the 9th international conference on human-agent interaction* (pp. 218-226). ACM. doi: 10.1145/3472307.3484170[†]

*Both authors contributed equally to this work

[†]Attribution 4.0 International License (CC BY4.0)

► *Conversational AI Study*

Weitz, K.*, Vanderlyn, L.*, Ngoc, T. V., & André, E. (2021). “It’s our fault!”: Insights into users’ understanding and interaction with an explanatory collaborative dialog system. In A. Bisazza & O. Abend (Eds.), *Proceedings of the 25th conference on computational natural language learning, conll 2021, online, november 10-11, 2021* (pp. 1-16). Association for Computational Linguistics. doi: [10.18653/v1/2021.conll-1.1](https://doi.org/10.18653/v1/2021.conll-1.1)[‡]

20.2 Rule-based Industry Robot in VR

In the VR-Robot Study, we investigated the impact of explanations on trust repair in a close-proxemic cooperation task between an industrial robot and a human. For this, we developed an experimental setup in a virtual environment and two different contents of explanation: (1) explanation about a robot mistake, answering the question “**Why** did the robot **not** solve the task?” and (2) explanation with solution after a robot mistake, answering the question “**How** can I (i.e., the user) prevent the mistake the next time?”.

20.2.1 Virtual Reality Design

We conducted our pilot and main study in a VR setup. The following describes the virtual environment and the robot arm used in detail.

Virtual Environment The setup for the virtual environment was inspired by a real-world robot scenario, described by Hald, Rehm, and Moeslund (2021). Here, an industrial robot arm and a user had to sort cones from one side of the table to another in close proxemics. We transferred this setup in a slightly adapted way into VR for our experiment. For this, we used a VR environment that was implemented using the Unity 3D game engine with a SteamVR plugin (Hald, 2021). We used the HTC Vive VR headsets and Vive Wand 6 degrees-of-freedom controllers to interact with this environment. The virtual environment consisted of an office environment with desks and office chairs, with participants being situated in an isolated corner of the room. Within reach of the participant was a desk with the robot mounted on top (see Figure 20.2 on the next page). On the table was also a white square platform at either side of the robot with a little copy of the bottles involved in the test shown next to them, indicating which shapes of bottles have to be put where. The task involved sorting bottles by whether they had a round or square base. There were four bottles on each platform at startup, two red and two blue each, and both had one bottle of each colour that did not match the shape. When the test started, the participant and the robot had to switch two bottles between the platforms to complete the shared objective. Between the two platforms was room to display text to convey instructions and explanations to the participants. The text was displayed on the surface, similar to a projected AR overlay. The participants picked up the red

*Both authors contributed equally to this work

[‡]Attribution 4.0 International License (CC BY4.0)

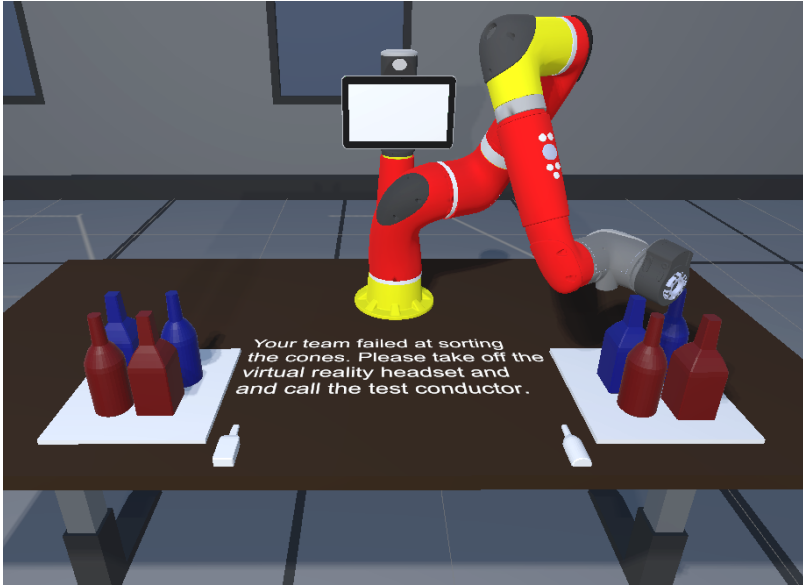


Figure 20.2: From the participant's perspective, the virtual environment features the robot, bottles, their platforms and indicators, and the display text on the desk surface. Figure from Hald, Weitz, et al. (2021)

bottles by moving a controller within 20 cm of their centre and pressing the trigger. Letting go of the trigger released the bottle, and they dropped straight down as they could not be thrown. If a bottle were dropped on the floor, rather than requiring the participant to pick it back up, it would be moved back to its initial position.

Virtual Robot The robot was modelled after the Rethink Robotics Sawyer robot arm ¹. The robot is developed for collaborative tasks in industry and production (Lawrence, 2019). The robot's behaviour is adapted from the physical robot arm used in Hald, Rehm, and Moeslund (2021). Here, the movements of the robot were pre-programmed. The movements were also pre-programmed on the real robot and then transferred and recorded in the Unity environment for the VR robot. Therefore, the VR-robot's motions were based on inverse kinematics which considered the real robot's height and weight (Hald, 2021). In addition to the robot's movements, the sound of the actuators and springs was also recorded to make the VR robot seem as realistic as possible (Hald, 2021).

1: More information about this robot can be found on <https://www.rethinkrobotics.com/de/sawyer> (last accessed on 29.09.2022)

20.2.2 Explanation Design

We conducted a pilot study to investigate different explanation contents and types for our main study.

Explanation Content We investigated the impact of explanations on trust repair after a robot mistake occurred. Therefore, the content of the explanations answered the question *Why* did the robot *not* succeed? We created two different robot errors:

- *Colour vision error:* To illustrate the colour vision error, the robot shown is moving a bottle of an incorrect shape. The explanation given was: "A computer vision error occurred. The system did not

successfully distinguish the shapes in the current lighting conditions.”

- *Calibration error*: Here, the robot knocked over one of the bottles while moving them. The explanation given was: “A calibration error occurred. The motion planner did not properly compensate for the robot’s momentum.”

Explanation Type In addition, we tested which of the two different *explanation types* (i.e., *textual* or *auditory*) are preferred by participants. We presented the participants with videos of a virtual robot sorting bottles of different shapes on either side of a table. The setup is illustrated in Figure 27.1 on page 163.

Based on the results of our pilot study, for the main study, we used the textual colour vision error and the explanation content of the colour vision error, respectively. In addition, based on the feedback of the pilot study, we investigated a third condition (i.e., explanation with solution) where in addition to the explanation, a solution on how to prevent such errors in the future was presented to the participants.

20.3 Rule-based Conversational Dialog Partner

The goal of the Conversational AI Study was to investigate the influence of explanations in a collaboration situation between humans and AI. For this purpose, we developed a cooperative game in which a user and a machine could only achieve a common goal and thus won the game by working together. The design of the game was inspired by a collaborative control room simulator from the German Aerospace Center (Schulze Kissing & Bruder, 2016), which is used to investigate the impact of a complex workspace on human behaviour. In their scenario, a team of three people has the task of monitoring production processes in various technical facilities. The team must watch the displays on their screen (i.e., their control centre). Each team member only sees the information for their facility location. Communication between the participants is required to identify and correct any anomalies in the energy supply of the production process. The simulation environment presented by Schulze Kissing and Bruder (2016) relies on the joint task of a team using verbal communication. We have used this idea for our Conversational AI Study. Here, a human has to collaborate with a rule-based dialog partner to solve four different puzzles and share their knowledge to solve the task. For the puzzle design, we were inspired by the human-AI collaborative games developed by Polyak et al. (2017) and van Waveren et al. (2019), and by the virtual reality game *Keep Talking and Nobody Explodes* from Steel Crate Games².

2: the game can be found on https://keep_talking_game.com (last accessed on 29.09.2022)

20.3.1 Game Design

We created a communication-based puzzle game and a dialog system based on this work. Each user is shown four puzzles in this game but is not given instructions to solve them. Conversely, the dialog system

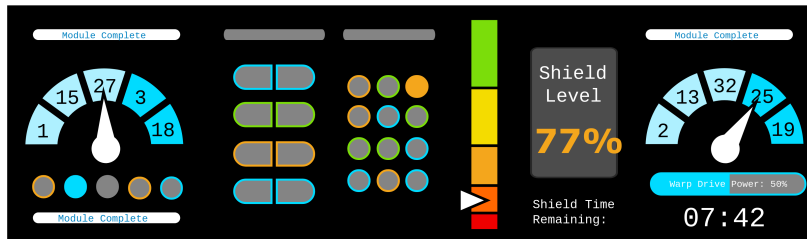


Figure 20.3: Game interface with four puzzles: The *dials module* (dials; far left and right), the *button sequence module* (row of five buttons; bottom left), the *switches module* (stack of four toggle switches; centre-left), and the *button array module* (3 x 4 array of buttons; centre). The shield level and timer display users' remaining time. The warp drive indicator and indicators above each module show number of completed puzzles. Figure from Weitz, Vanderlyn, et al. (2021)

has instructions for solving the puzzles but cannot “see” or interact with them. Therefore the user must communicate with the dialog system via a text-chat interface to decide which actions to take (i.e., where to click to solve the puzzle). The game’s theme was inspired by space travel, where increasing human-AI collaboration (Bluethmann et al., 2003) is taking place. The game is comprised of a control panel (see Figure 20.3) and a chat interface (not shown) to communicate with the dialog system.

Puzzle Modules

Three of the four puzzle modules rely on the user to accurately describe the board (e.g., number of dials or colour/sequence of buttons); the fourth, a memory game, requires the user to remember their previous actions. This choice was inspired by van Waveren et al. (2019), who stress the importance of human players having a role beyond just following instructions. Only the most recent dialog turn is displayed to the user to simulate a spoken interaction. Additionally, there is a time limit with a time penalty for every wrong action to encourage users to communicate with the system. To learn the game, we provided a mandatory tutorial introducing each type of puzzle³.

3: tutorial and game puzzle configurations were different

Dialog System

We used the open-source ADVISER toolkit (Li et al., 2020) to implement a rules-based conversational agent. Following the work of P. R. Cohen (2020), a rule-based system was preferred for this study as we wanted to ensure the conversational agent would behave consistently for each user. As previous work (Salem et al., 2015; Yin et al., 2019; Yu et al., 2016, 2017) has shown that trust is most damaged where mistakes are made, we hard-coded a scenario where the system made an incorrect assumption into three of the four puzzles. The dialog system consists of a regex-based natural language understanding module, which matches user utterances against a series of hand-crafted regexes to determine user intent (see Figure 20.4 on the following page). The dialog history is then stored in a Beliefstate Tracker, which the rules-based policy uses to decide on the following system response. Finally, the system response is communicated to the user through a template-based Natural Language Generation module, which chooses a set of pre-defined templates to convert the system intent into natural language output.

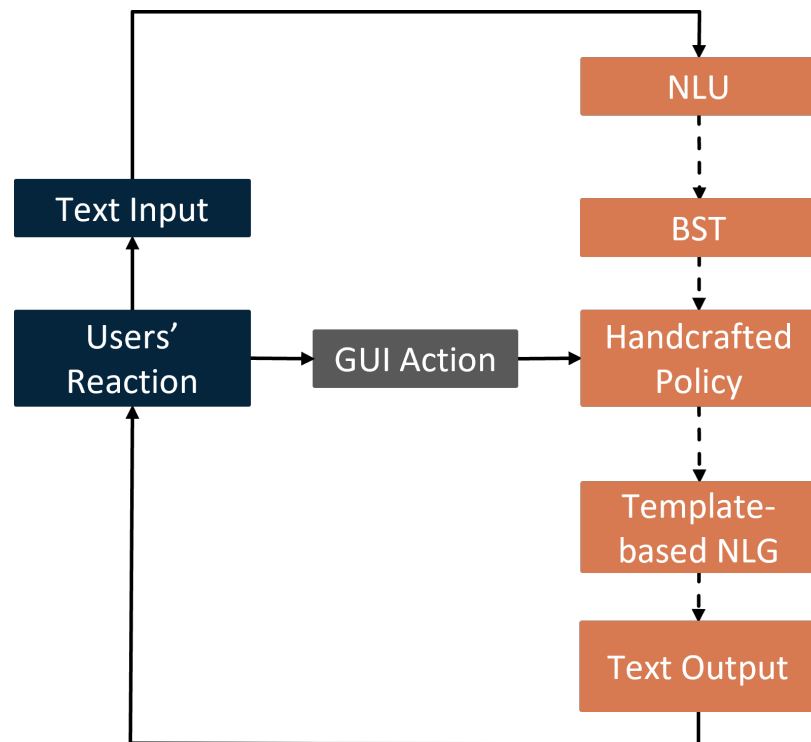


Figure 20.4: Overview of the rule-based dialog system ADVISOR (Li et al., 2020). The dialog system (orange) consists of a Natural Language Understanding module (NLU), a Belief State Tracker (BST), a handcrafted Policy, and a template-based Natural Language Generation module (NLG), which generates the textual output for the user. The user (blue) reacts to the output of the AI dialog partner by responding via text

20.3.2 Explanation Design

We created a new ‘explain’ dialog act for the explanations and extended the dialog policy to present an explanation with every instruction to the user.

Explanation Content Explanations took the form of the relevant evidence from the belief state which supported the given instruction. For example: “since the previous switch is blue, you should activate the left side”. The explanation content answered the question of *Why?* when the belief state of the dialog partner was correct and *Why not?* when the assumption of the dialog partner was incorrect. This approach was chosen as previous work (Garcia et al., 2018b; Kulesza et al., 2013) has found users best understand explanations that provide evidence rather than try to explain a process.

Explanation Type As part of the design of the dialog system, we implemented two different styles for language output. Our work here was inspired by Kunkel et al. (2019), who found that more *personal* explanations increase trust in movie recommendations. In this previous work, the authors conclude that more personal utterances should increase trust in the system. However, “personal” is a complex term to define. In Kunkel et al. (2019), the authors’ approach involved getting recommendations and explanations from actual humans, and the participants knew which recommendations came from humans vs from a recommender system, blurring the line between a preference for “personal” and the known phenomenon that humans prefer advice from other humans (Dietvorst et al., 2015). Nevertheless, the work of Kunkel et al. (2019) uses more

companion-like interactions in the human condition (e.g., by addressing the user as “buddy”). In our design, we interpret *personal/impersonal* similarly. Regarding the conversational style of the AI dialog partner, we modelled two versions: The AI dialog partner as a companion or assistant. While an assistant represents an “useful machine”(Sundar et al., 2017, p. 89), a companion is characterized by a less severe behaviour.

We modelled the more personal, companion-like AI dialog partner using the following features:

- ▶ second person direct address (e.g., “you should activate the left button”),
- ▶ first person plural pronouns (e.g., “we are in the third round”),
- ▶ less formal speech such as contractions (e.g., “Thanks for playing, don’t forget your user id...”)
- ▶ variation of word order/word choice (synonyms) in templates,
- ▶ using backchannels (e.g., “Okay, what color is the second switch?”)

In contrast, a more impersonal, assistive AI dialog partner was modelled using passive voice (e.g., “the left button should be activated”), an absence of first-person plural person pronouns, more formal speech without contractions, no variation in templates, and an absence of backchannels.

21.1 Overview	132
21.2 CNN for Keyword Classification	133
21.2.1 CNN Setup	133
21.2.2 Explanation Design	134
21.2.3 Demonstrator Setup for the Participatory Machine Learning Show	135
21.3 CNN for Emotion Expression Classification	137
21.3.1 CNN Setup	137
21.3.2 Explanation Design	137
21.4 CNN for Pneumonia Classification	138
21.4.1 CNN Setup	138
21.4.2 Explanation Design	139

21.1 Overview

We used CNN implementations as an example of data-driven black-box approaches for the remaining four experiments. In the NOVA Study, Gloria Study, and Museum Study, we trained CNN for emotion expression classification on images, Pneumonia classification on X-rays, or keyword classification on speech, respectively. In addition to the description of the CNN implementations, the design of the explanations for each system is described in detail (see Figure 21.1).

Education		Medical Decision Support	
Gloria Study		NOVA Study	
Content of explanation	Type of explanation	Content of explanation	Type of explanation
Why? What?	Verbal - Text Verbal - Speech Visual (LIME)	Why?	Numerical (Confidence values) Visual (LIME)
Interaction with a virtual agent		Interaction with the NOVA software	
Museum Study		Pneumonia Study	
Content of explanation	Type of explanation	Content of explanation	Type of explanation
Why?	Verbal - Speech (first vs. third person) Visual (LIME)	Why? What if?	Visual (LIME vs. LRP vs. counterfactuals)
Interaction with a virtual agent		Presentation of XAI visualisations	

Figure 21.1: Overview of the design of the explanation content and explanation type for the four conducted experiments in the education and healthcare application scenarios

This chapter is related to parts of the following publications:

► *Gloria Study*

Weitz, K., Schiller, D., Schlagowski, R., Huber, T., & André, E. (2019). “Do you trust me?”: Increasing user-trust by integrating virtual agents in explainable ai interaction design. In C. Pelachaud, J. Martin, H. Buschmeier, G. M. Lucas, & S. Kopp (Eds.), *Proceedings of the 19th ACM international conference on intelligent virtual agents, IVA 2019, paris, france, july 2-5, 2019* (pp. 7-9). ACM. doi: [10.1145/3308532.3329441](https://doi.org/10.1145/3308532.3329441)*

► *Museum Study*

Weitz, K., Schlagowski, R., & André, E. (2021). Demystifying artificial intelligence for end-users: Findings from a participatory machine learning show. In S. Edelkamp, R. Möller, & E. Rueckert (Eds.), *Proceedings of KI 2021: Advances in artificial intelligence -*

*Licensed under the [Attribution 4.0 International License \(CC BY4.0\)](https://creativecommons.org/licenses/by/4.0/)

44th german conference on AI, virtual event, september 27 - october 1, 2021 (pp. 257-270). Springer. doi: [10.1007/978-3-030-87626-5_19](https://doi.org/10.1007/978-3-030-87626-5_19)[†]

► *NOVA Study*

Heimerl, A., **Weitz, K.**, Baur, T., & André, E. (2020). Unraveling ML models of emotion with NOVA: multi-level explainable AI for non-experts. *IEEE Transactions on Affective Computing*.

► *Pneumonia Study*

Mertes, S., Huber, T., **Weitz, K.**, Heimerl, A., & André, E. (2022). GANterfactual-Counterfactual Explanations for Medical Non-experts Using Generative Adversarial Learning. *Frontiers in Artificial Intelligence*, 5, 825565. doi: [10.3389/frai.2022.825565](https://doi.org/10.3389/frai.2022.825565)[‡]

21.2 CNN for Keyword Classification

We used a CNN for spoken keyword classification in the Gloria Study and the Museum Study. Based on this system, we generated explanations presented by the virtual agent Gloria. In the following, the implementation of the keyword classification, the explanation design, and the setup for the participatory ML-show in the German Museum Munich are described.

21.2.1 CNN Setup

For the training of the CNN, we used spectrograms. These were calculated from the respective audio signal (i.e., spoken keyword) and used as input for our classification model. Spectrograms are visual representations (i.e., images) of audio samples and display sound pressure levels as pixel values over the dimensions of time (x-axis) and frequency (y-axis). Figure 21.2 illustrates the spectrogram for the spoken input word 'house' on the left.

As a prediction model, we used the neural network architecture proposed by Sainath and Parada (2015). This classification model uses a CNN to

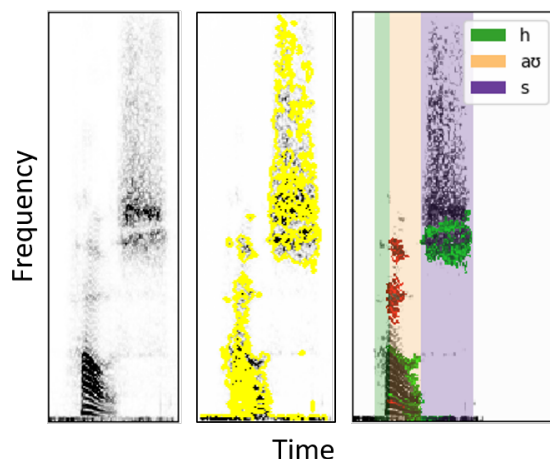


Figure 21.2: A spectrogram of an audio sample (left), its segmentation into superpixels (centre) and the output for the user containing LIME visualisations and additional phoneme information (right). Figure adapted from Weitz, Schiller, et al. (2021)

[†]Adapted by permission from Springer Nature Customer Service Centre GmbH: Springer Nature Lecture Notes in Computer Science book series (Demystifying Artificial Intelligence for End-Users: Findings from a Participatory Machine Learning Show, Weitz, K., Schlagowski, R., & André E.), ©2021 Springer Nature Switzerland AG (2021)

[‡]Licensed under the [Attribution 4.0 International License \(CC BY4.0\)](https://creativecommons.org/licenses/by/4.0/)

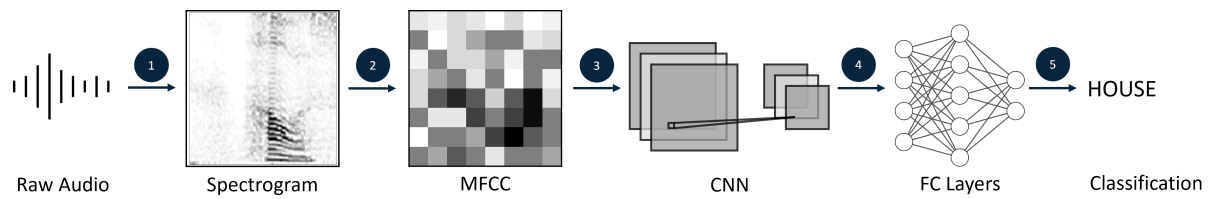


Figure 21.3: Schematics of the used speech recognition system. 1) A spectrogram is generated from the raw audio waveform. 2) The spectrogram is used to calculate 20 MFCCs. 3) The MFCCs are fed into a convolutional neural network. 4) The learned features are then forwarded to the fully connected layers of the network. 5) Finally, the network’s output is mapped to the corresponding target class. Figure adapted from Weitz, Schiller, et al. (2021)

generate abstract features based on Mel-frequency cepstrum coefficients (MFCCs) which are derived from the spectrograms of the raw audio waveforms. These features are then fed into a fully-connected layer which finally predicts the target class, which is one of the keywords (labels) of the training dataset (see Figure 21.3).

We trained our model on the speech command dataset provided by Warden (2018). This dataset consisted of instances from 35 different spoken words and was specifically designed to train and evaluate audio keyword classification systems. The comparably high ratio of samples per class to the overall number of classes and the high variance for speakers and sound quality enabled us to train a reasonably robust model for our specific use case.

21.2.2 Explanation Design

For our studies, we chose the LIME framework by Ribeiro et al. (2016) to explain the automatic recognition of spoken keywords. In the Gloria Study, this explanation was presented by three different representations (i.e., text, voice, visual presence) of the virtual agent Gloria¹. In the Museum Study, we varied Gloria’s personification (i.e., first-person or third-person).

1: developed by Charamel: <https://vupp-etmaster.de/> (last accessed on 18.10.2022)

2: “What are superpixels that speak against the classification?”

3: phonemes are small units of sound that can be used to distinguish one word from another

Explanation Content To generate a visual explanation for a specific prediction (keyword) of our classification model, the input spectrogram was first segmented by the Felzenszwalb algorithm for image segmentation (Felzenszwalb & Huttenlocher, 2004) (see Figure 21.2, centred image) into superpixel. Using LIME, superpixels that are found to have a significant impact in favour of a specific label are highlighted green for the user, whereas red highlighted segments speak against the predicted label (see Figure 21.2, right image). The LIME visualisations represent the answers of *Why?* (i.e., green superpixel) and *What?* (i.e., red superpixel)². To further enhance the explainability of the LIME visualisations, we presented a phoneme-based³ segmentation of the input word to the user. Therefore they are particularly well suited to assist with establishing a relation between how humans understand spoken language and the visualisations our system provides. The phoneme segmentation of the spectrogram is generated through the WebMAUS tool developed by Kisler et al. (2017). An example of this segmentation for the spoken word ‘house’ can be seen on the right side in Figure 21.2 on the preceding page.

Explanation Type In the Gloria Study, we investigated three different representations of the virtual agent Gloria (see Figure 21.4). While the explanation content using LIME was the same, Gloria repeated the displayed information of the spectrogram using natural language via text, speech, or speech in combination with a visual presentation of the agent.

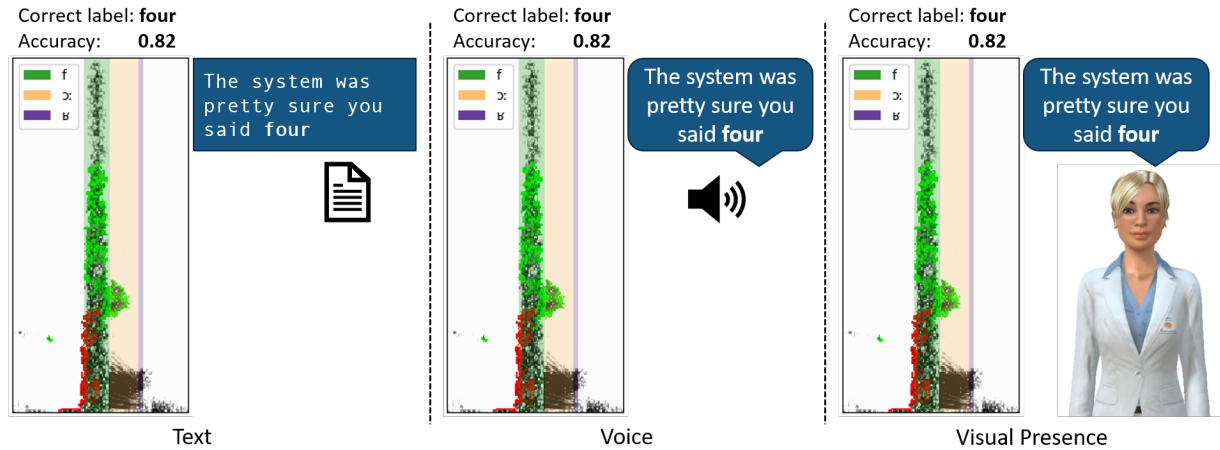


Figure 21.4: The baseline group of the Gloria Study was presented with XAI visualisations of a spoken keyword (e.g., 'four'). With every XAI visualisation, the predicted label and the prediction accuracy of the speech recognition system were displayed. In the three experimental groups, an XAI visualisation was presented with one of the virtual agent modalities (i.e., text, voice, or visual presence)

In the Museum Study, we varied the personification of Gloria:

- ▶ Personifying the classifier by speaking in first person (e.g., “I am sure you said <keyword>”) or
- ▶ Commenting on the classifier’s processes in the third person (e.g., “The neural network was sure you said <keyword>”).

For both situations, different versions of the virtual agent’s phrases (see examples in Table 21.1) were hard coded in advance.

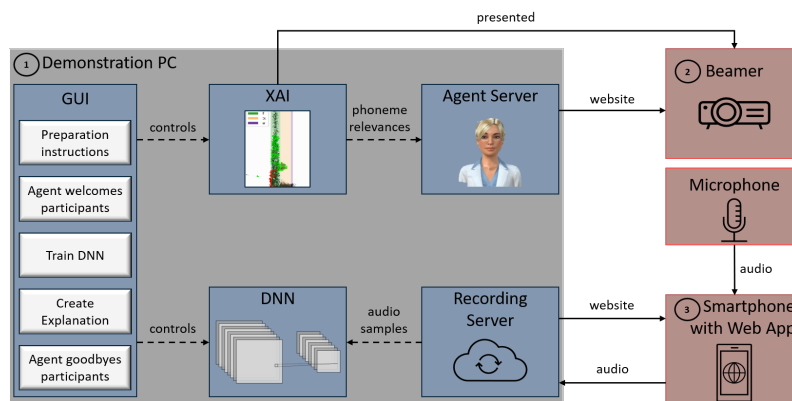
Table 21.1: Example phrases of the virtual agent for group A (classifier personification) and group B (third person commentary)

Situation	Group A (classifier personification)	Group B (third person commentary)
Welcome	"Hello my name is Gloria. You can train me to recognize a keyword of your choice. Please tell me which word I should recognize!"	"Hello my name is Gloria. I will help you to train a neural network for a keyword of your choice. Please tell me which word the system should recognize!"
Start of Training	"Thank you for your recordings. I will now start to train myself so that I will be able to understand you later."	"Thank you for your recordings. The neural network will now be trained so that it will be able to understand you later."
Communication of Prediction Results	"I am <sure/unsure> that you said <keyword>. Should i explain why?"	"The system was <sure/unsure> that you said <keyword>. Should i explain why?"
Explanation of XAI Visualisation	"The system found phoneme number <phoneme no.> to be particularly important for its prediction."	"I found phoneme number <phoneme no.> to be particularly important for my prediction."

21.2.3 Demonstrator Setup for the Participatory Machine Learning Show

After we conducted the Gloria Study experiment in the lab, we extended our approach of explaining keywords with the help of LIME and the virtual agent Gloria by developing a participatory ML-show in the German Museum Munich. The demonstrator for the show in the Museum Study (see Figure 21.5) mainly consisted of a *demonstration PC* including a high

Figure 21.5: Demonstrator overview as used in the field study. (1) *Demonstration PC*, running the neural network for key-word recognition and calculating the XAI visualisations. (2) *Beamer & PA system* were used for agent displaying & sound, the XAI visualisations, and the presentation slides for the show. (3) *Smartphone* and microphone for recording the audio samples. Figure adapted from Weitz, Schiller, et al. (2021)



performance GPU (Nvidia GTX 1060) for improved training performance and a smartphone which was used to record and transmit audio samples for training and prediction of the neural network over WLAN. The demonstration PC was connected to a *beamer* which displayed the virtual agent, the generated spectrograms, and the XAI visualisations generated by the XAI framework. In parallel, the demonstration PC hosted a website providing audio recording and transmission functionalities on a server in the local network. An android app containing a browser window was used on the *smartphone* to access the site when the audience recorded the audio samples. As in our lab study, we used for the recognition of audio keywords the neural network architecture proposed by Sainath and Parada (2015) and generated XAI visualisations with LIME (Ribeiro et al., 2016). To make the resulting XAI visualisations better readable, we again used the webMAUS API (Kisler et al., 2017) to highlight areas within the spectrograms that contain the phonemes of the actual spoken word (ground truth).

The moderator of the show, who was instructed in advance, operated the main application by using a step-by-step structured GUI that enabled them to (1) start and stop the training process of the neural network, (2) start prediction for a recorded audio sample, (3) review transmitted audio files, and (4) calculate XAI visualisations after prediction.

The virtual agent Gloria was integrated into a website hosted locally and displayed with a browser on the demonstration PC. Communication between the virtual agent and the main application was implemented with WebSockets.

As soon as about 80 audio samples were recorded and transmitted to the demonstration PC, the moderator used pre-programmed software functionalities to label the samples and merge them with a subset of the speech command dataset provided by Warden (2018) (we used data for 11 classes/keywords, 80 samples each) to create the training corpus. Then, the moderator started the training process for the prediction model. To give the participants a feeling of how good the classifier was after this relatively short time (the typical validation accuracy was about 80%), we decided not to use any pre-trained networks and instead train the network from scratch for each show.

21.3 CNN for Emotion Expression Classification

In the NOVA Study, we investigated the impact of explanations on participants interacting with the software tool NOVA. NOVA⁴ is an annotation tool that helps humans to label social signals (e.g., gestures, facial expressions) in data (Baur et al., 2013). Interacting with NOVA, people are interactively involved in the workflow and can thus already improve the system with their expertise during the training of ML-systems. Baur, Heimerl, et al. (2020) describe that NOVA can also be used in the context of Explainable AI. By integrating model-agnostic and model-specific methods and confidence values, users can gain detailed insight into the machine learning model and assess whether the system they are training is already working correctly or needs further training data.

4: abbreviation for NOnVerbal behavior Analyzer

21.3.1 CNN Setup

Our study used NOVA to improve a neural network model that recognizes emotional facial expressions based on image data. As a neural network architecture, we chose a CNN and applied transfer learning to improve the performance of our model. Transfer learning is based on taking already learned knowledge about one domain and transferring it to another to improve generalization (Goodfellow et al., 2016). In our case, we took advantage of the learned knowledge of the VGG16 (Simonyan & Zisserman, 2015). Here we stripped the fully connected layers of the network responsible for the mapping onto the domain-specific classes. We then added our fully connected layers corresponding to recognizing emotional facial expressions. Finally, we trained our network on the AffectNet facial expression corpus (Mollahosseini et al., 2019). The corpus provides data annotations for Neutral, Happy, Sad, Surprise, Fear, Anger, Disgust, Contempt, None, Uncertain, and Non-face. Out of those categories, we chose a subset (anger, disgust, happiness, sadness, and neutral) to train our neural network model. This subset consists of four of Ekman's six basic emotions (happiness, sadness, anger, surprise, disgust, fear) (Ekman & Friesen, 1971). We chose not to consider surprise and fear to reduce the complexity of the classification task. In the user study, our trained model was used to predict visible emotions in images of facial expressions. Those images have not been part of the training set and therefore unknown to the model.

21.3.2 Explanation Design

We generated two different types and contents of explanations: a visual explanation using the LIME algorithm (Ribeiro et al., 2016) and a value-based explanation using confidence values (Baur, Heimerl, et al., 2020).

Explanation Content The content of the explanations answered the question “Why did the model classified this image as *<emotion>?*”. The LIME-based explanation answered this question by highlighting the relevant superpixel for the classification. The confidence-based explanation

showed the percentage of confidence in labelling the image with emotion.

Explanation Type The type of explanation is related to the content. While LIME presents a visual explanation type by highlighting superpixel in the image, confidence values are a numerical explanation type showing the model's confidence in the prediction for a single image made. Both were displayed to the user as part of the NOVA interface.

21.4 CNN for Pneumonia Classification

In the Pneumonia Study, we used a trained CNN in the first step to predict binary classifications (i.e., pneumonia vs no pneumonia) for X-ray images of lungs and a GAN in the second step to generate counterfactuals to explain these classifications to the study participants.

21.4.1 CNN Setup

5: link to the RSNA Pneumonia Detection Challenge: <https://www.kaggle.com/c/rsna-pneumonia-detection-challenge/> (last accessed on 29.09.2022)

To evaluate our system, we trained a CNN to decide whether or not X-ray images showed lungs that suffer from pneumonia. As a dataset, we used the data published for the RSNA Pneumonia Detection Challenge⁵ by the Radiological Society of North America. The original dataset contains 29,700 frontal-view x-ray images of 26,600 patients. The training data is split into three classes: *Normal*, *Lung Opacity* and *No Lung Opacity/Not Normal*. We took only the classes *Normal* and *Lung Opacity*, as Franquet (2018) argue that opacity of lungs is a crucial indicator of lungs suffering from pneumonia, and we only wanted to learn the classifier to distinct between lungs suffering from pneumonia and healthy lungs. Therefore, other anomalies that do not result in opacities in the lungs are excluded from the training task to keep it a binary classification problem. All duplicates from the same patients were removed as well. The resolution of the images was reduced to 512x512 pixels. Subsequently, we randomly split the remaining 14,863 images into three subsets: *train*, *validation*, and *test*.

We trained an AlexNET architecture (Krizhevsky et al., 2017) to solve the described task. We slightly modified the architecture to fit our needs. These modifications primarily include L2 regularization to avoid overfitting. Further, we replaced the loss function with an MSE loss, which worked well for our classification task. After training the classifier on the *train* partition for 1000 epochs, it achieved an accuracy of 91,7% on the test set. State-of-the-Art classifiers achieve much better performance values than our classifier does. However, our work aims to *explain* the decisions of a classifier. Explaining an AI model does not only include explaining decisions where the AI was right but also cases where the AI was wrong, as a complete understanding of an AI also covers both cases. Thus, we found that a *perfect* classification model would not be an appropriate tool to measure the performance of an XAI system, resulting in our decision not to improve the classifier performance further (i.e., we did not conduct any hyperparameter tuning or model optimization).

21.4.2 Explanation Design

We investigated three visual explanation types (i.e., LIME, LRP, and counterfactuals), including two different explanation contents (i.e., Why? and What if?)

Explanation Content The explanation content was related to the explanation type. While LIME and LRP highlight relevant areas of the X-ray image that were relevant for the classification of the image (i.e., *Why?*), counterfactuals represent an X-ray image where the input image is modified in a way such that the classifier would have made a different prediction. By doing so, the users of counterfactual explanation systems are equipped with a completely different kind of explanatory information (i.e., *What if?*). Figure 21.6 shows an example of the three different XAI visualisations for one of the X-ray images used in our experiment.

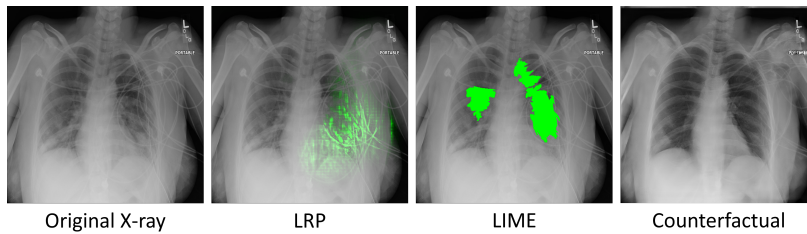
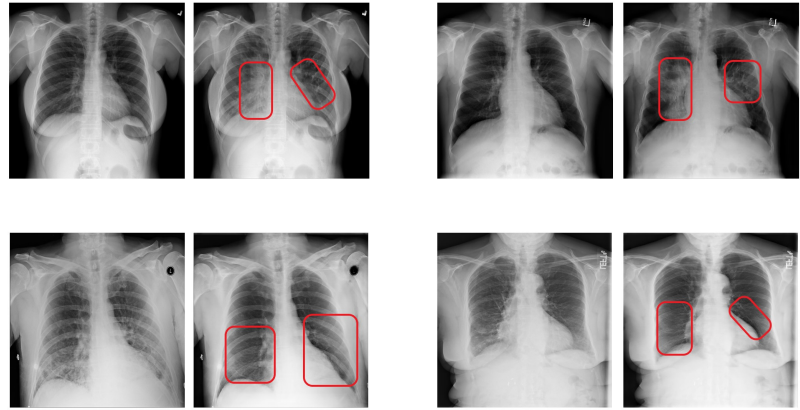


Figure 21.6: An example X-ray image classified as *Pneumonia*, as well as the different XAI visualisations used in our study. Figure adapted from Mertes et al. (2022)

Explanation Types Participants in the LRP condition were assisted by heatmaps generated through Layer-wise Relevance Propagation using the z -rule for fully connected layers and the $\alpha 1\beta 0$ -rule for convolutional layers, as recommended by Montavon et al. (2019). The LIME condition contained highlighted superpixel, which LIME generated. Here, we chose the *SLIC* segmentation algorithm, which Schallner et al. (2019) found to perform well in a similar medical use case. We used the default values for the remaining hyperparameters and showed the five most important superpixels. For both LIME and LRP, we omitted the negative importance values since those were highly confusing to participants in our pilot study. Participants in the counterfactual condition were shown counterfactual images generated by the proposed approach of Mertes et al. (2022). As a training dataset for the counterfactual explanations, we used the *train* partition of the same dataset that we used for our CNN classifier. The architecture of both the generators as well as both the discriminators was adopted from J.-Y. Zhu et al. (2017). Examples of counterfactual images produced by feeding images from the test partition into our trained generative model are shown in Figure 21.7. Here, the main structure and appearance of the lungs are maintained during the translation process, while the opacity of the lungs is altered. This was expected due to the pneumonia class of the used dataset being defined by lungs that show a certain degree of opacity.

To avoid cherry-picking while still ensuring variety in the images, we randomly chose 12 images for our user study based on the following constraints: To ensure that the classifier equally makes false and correct predictions for each class, we wanted 3 true positives, 3 false positives, 3 true negatives, and 3 false negatives. Furthermore, inspired by Alqaraawi

Figure 21.7: Examples of counterfactual images are produced with the approach proposed in Mertes et al. (2022). The left image shows the original image in each pair, while the right image shows the corresponding counterfactual explanation. The red boxes were added manually to point the reader to the most altered regions. The original images in the top row were classified as *normal*, while the original images in the bottom row were classified as *pneumonia*. The shown counterfactual images were all classified as the opposite of their respective counterpart. Figure from Mertes et al. (2022)



et al. (2020), we used the AI model's confidence to ensure diversity in the images. Decisions where the model is certain are often easier to interpret than decisions where the AI model struggles. Since our prediction classifier mainly had probability values between 0.8 and 1, we randomly chose one X-ray image with values of 0.8, 0.9 and 1 (rounded) out of each of the sets of true positives, false positives, true negatives, and false negatives.

The six studies on Explainable AI conducted in this dissertation were done with real, i.e., working AI systems. This section describes the implementation of these systems in detail. Rule-based systems were used for two studies (VR-Robot Study and Conversational AI Study), while CNN was used for the remaining four. In addition, we explained how the explanations of these systems were generated. Here, a distinction was made between the explanation content (i.e., what kind of information is part of the explanation) and the explanation type (i.e., how the explanation is presented). The explanation content answered (depending on the implemented system) the questions: *Why?*, *Why not?*, *How?*, *What?* and *What if?*. The explanation type included *verbal* (i.e., text and speech), *visual* (i.e., LIME, LRP, counterfactuals), and *numerical* (i.e., confidence values) representations. After presenting the technical part of the experiment in this dissertation in the last chapters, we will now have a closer look at the study designs and analysis methods of the experiments.

VII. METHODOLOGY FOR HUMAN-CENTERED EXPLAINABLE AI RESEARCH

23.1 Overview

In the following chapters, I will go into the considerations under which the experiments in this dissertation were designed, which method we used to measure the psychological constructs (e.g., trust, mental models), and which procedures I used for the evaluation. Doshi-Velez and Kim (2017) present a taxonomy of evaluation approaches for HC-XAI. They differentiate between three evaluation categories that are increasing in specificity and costs: (1) proxy tasks, (2) simple tasks, and (3) real tasks. We focus in all experiments on the second category, the *simple tasks*. Our experiments addressed specific application scenarios (i.e., industry, education, and healthcare) but were conducted with end-users not working in these domains. This had two reasons: (1) before developing *real tasks* which inherently have a low *internal validity*¹, the conducting of *simple tasks* in a more controlled experimental setup is needed. The investigations of our work can be used to evaluate in the next step (that is not part of this dissertation) in real task to ensure *external validity*.

In the following, the *designs* used in the experiments of this dissertation are presented (for an overview, see Table 23.1).

23.1 Overview	143
23.2 Experiments in the Laboratory	144
23.3 Experiments in the Wild	145
23.4 Online Experiments	146

1: internal validity describes whether a conducted study is free from biases while external validity describes whether a study and its results can be generalized to other contexts (Campbell & Stanley, 1963)

Study Title	Type	Study Focus	N
Cooperation & Collaboration			
VR-Robot Study	Lab	Verbal XAI in a cooperative task	30
Conversational AI Study	Online	Verbal XAI in a collaborative game	117
Education			
Gloria Study	Lab	Virtual agent modalities & LIME	60
Museum Study	Field	Virtual agent modalities & LIME	105
Medical Decision Support			
NOVA Study	Lab	LIME & Confidence values in facial emotion recognition	53
Pneumonia Study	Online	LIME, LRP, and counterfactuals in Pneumonia detection	118

Table 23.1: Overview of the type of experiment, study focus and sample size of the conducted studies

We decided to use three different designs: experiments in the laboratory, experiments in the wild, and experiments that were conducted online (see the following sections for more details). The goal was to investigate the end-users perception of XAI. Regarding XAI, we investigated the *explanation type*, the *explanation content*, and the *explanation interface*.

Each explanation interface represents a different level of interactivity as described in Chapter 13 on page 72.

The data protection officer of the University of Augsburg approved all conducted studies in this dissertation. At the beginning of each experiment, all participants were informed about the goals, duration, their GDPR rights, and the possibility of quitting during the experiment. All participants participating in our online studies via recruiting platforms were financially compensated at the minimum wage level. At the local studies, financial compensation or another kind of compensation was also paid (e.g., a drink voucher). Only in our experiment in the wild, the Museum Study, we offered no reward since the ML-show was part of the museums' visit and, therefore, an offer of the museum to their visitors.

23.2 Experiments in the Laboratory

Controlled environments (e.g., laboratories) are the first step for investigating the impact of XAI on users since existing biases can be prevented or controlled systematically. Laboratory experiments support internal validity (Campbell & Stanley, 1963), which is the basis for later experiments in the wild, i.e., in a less controlled setup.

We conducted three of our experiments in a laboratory setup: (1) VR-Robot Study, (2) Gloria Study, and (3) NOVA Study. We used rooms at the University of Aalborg (VR-Robot Study) and the University of Augsburg (Gloria Study, NOVA Study). During the experiments, there was always an experimenter in the room. This person led through the experiment, was the contact person for questions and took over the introduction and de-briefing at the beginning and end of the experiment.

We chose an in-situ implementation due to the hardware and software requirements for the experiments. We needed the NOVA software system in the *NOVA Study*. Therefore, we prepared a computer for the study, with all required software packages for NOVA installed. In addition, as part of the study, a drawing task was performed using a specific paint program installed and adjusted beforehand. For the *Gloria Study*, in addition to two monitors, an audio recording device was required to record spoken keywords from participants (see Figure 23.1 on the facing page). For the *VR-Robot Study*, a test room was specially prepared for the requirements of a VR setting. In addition to plenty of space for the test subjects to move around in VR, there was sufficient room for the extensive VR equipment.

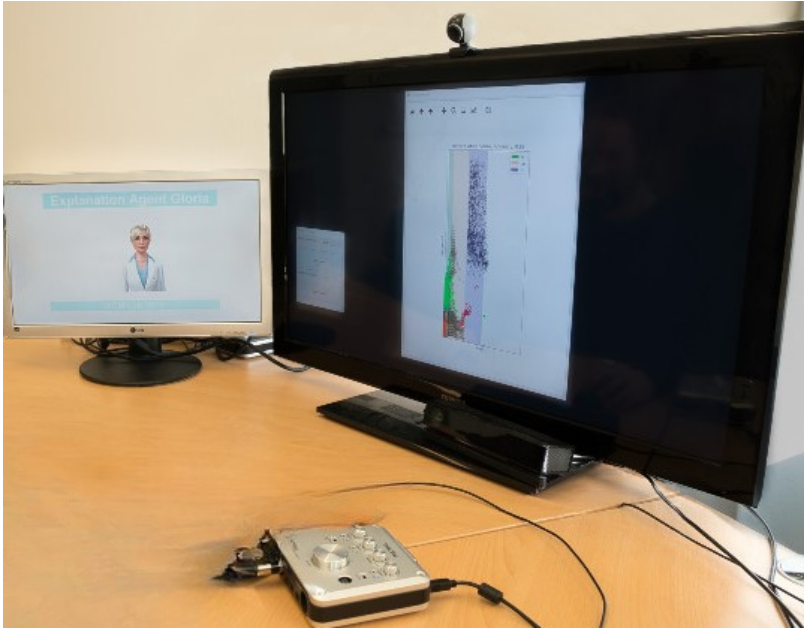


Figure 23.1: Illustration of the Gloria Study setup for participants in the virtual embodied agent group. The virtual agent Gloria was displayed on a separate monitor on the left side of the participants. The XAI visualisations were displayed on a monitor in front of the participants. The keywords were spoken into a microphone (not in the picture). Figure adapted from Weitz, Schiller, et al. (2021)

23.3 Experiments in the Wild

We conducted one experiment in the wild, the *Museum Study* (see Figure 23.2 on the next page). While experiments in the wild help to foster external validity (Campbell & Stanley, 1963), it brings a lot of uncertainty with them. Detailed planning is necessary for all experiments, especially those outside the laboratory. In a less controlled setting, possible biases and influencing factors are already considered in the planning phase. In our case, we had the additional challenge that the experimenters were not part of our scientific team, but instead, the museum personnel (1) had no experience in working with an AI system and (2) had only a little knowledge about our study goals. The second point was intentional, as we wanted to ensure not to introduce any biases based on the investigators' knowledge about the study goals. Besides these challenges regarding the experimenters, we had challenges regarding the organisational and spatial nature of the museum. The participatory ML-show should take place for a whole year in the museum. Therefore, we designed a system that can be used by non-AI experts and is stable to run for an entire year in the museum. The ML-show was integrated into the museum's daily routine. This allowed people to participate in the show at fixed times and continue their museum visits after participation. The evaluation, which provided the study data, took place at the show's end in either digital or paper form. It was found that people wanted to continue their visit to the museum and that there was little response to the digital evaluation. More people completed the paper-based version, but with over 2500 ML-show visitors, we received only 47 usable responses. This clearly shows the particular demands of recruiting and motivating participants in an experiment in the wild.



Figure 23.2: Beginning of a public participatory machine learning show visited by non-experts in the German museum in Munich. Figure from Weitz, Schlagowski, and André (2021)

23.4 Online Experiments

During the COVID-19 pandemic, two of our experimental studies had to be shifted to an online setup to eliminate the risk of infection from in-situ studies. These offered the advantage that many participants in a short time could process them. Therefore, we conducted two online experiments: (1) Conversational AI Study and (2) Pneumonia Study. To acquire participants, we use the MTurk platform²(Conversational AI Study), and the Clickworker platform³(Pneumonia Study).

2: Amazon Mechanical Turk: <https://www.mturk.com/> (last accessed on 29.09.2022)

3: Clickworker: <https://www.clickworker.de/> (last accessed on 29.09.2022)

Besides the advantages of online studies, it must be emphasized that the research and the associated questionnaire must be prepared very carefully. Unlike studies in the laboratory or the field, the participants can ask no questions to an experimenter during the study. Therefore, misunderstandings or incorrect processing of tasks during the experiment can only detect (if at all) afterwards. We used pilot experiments in all online studies to ensure that questions and study designs were adequately chosen. In the pilot experiments, we tested the general study design and the quizzes and tutorials we used to introduce the task of the respective experiment to the participants. In addition, we included pre-study quizzes in the main experiments of both studies to ensure acceptable quality of responses. Only participants who successfully solved the quiz, tailored to the study's content, were admitted to the main experiment. Furthermore, we used the MTurk and Clickworker platform options (e.g., selecting only participants with a particular performance score) to recruit participants who worked particularly carefully.

Our six experiments aimed to investigate users' perception of XAI for three application purposes. The psychological constructs of mental models, trust, self-efficacy, cognitive load, and emotions were investigated. A broad set of instruments is available to make these constructs operationalisable (i.e., measurable). However, selecting the right ones depends on the research questions, the experimental design, and the statistical analysis. Only when the measurements fit these criteria can studies gain new knowledge about whether and how XAI influences users' mental models, trust, self-efficacy, cognitive load, or emotions.

In every study conducted for this dissertation, common questions about personal information were asked, similar to the survey questions presented in Chapter 15 on page 85. These questions included age, gender, experience with AI in general, experience in the specific application used in the study (e.g., NOVA framework, chatbots, personal assistants), and their knowledge about AI and XAI. In addition, study-specific questions were addressed (e.g., rating of the virtual agent in the Gloria Study), which are directly described in the respective chapter of the experiment. Besides demographic and study-specific questions, measurement tools to elicit mental models, trust, self-efficacy, cognitive load, and emotional state of users were used. These are presented in the following. Since there exists a variety of methods¹, the description is limited only to methods used in the empirical studies of this dissertation. An overview of the measured variables and the methods used for each study is given in Table 24.1 on the next page.

24.1 Mental Models	149
24.1.1 Retrospection Questions	149
24.1.2 Prediction Task	149
24.1.3 Task Reflection	150
24.1.4 Users' Model of Self	150
24.2 Explanation Satisfaction	150
24.3 Trust	151
24.3.1 Trust in Robots	151
24.3.2 Trust in Automation - Jian et al. (2000)	151
24.3.3 Trust in Automation - Körber (2018)	151
24.3.4 Trust in Explanation - Hoffman, Mueller, et al. (2018)	152
24.4 Computer Self-Efficacy	152
24.5 Cognitive Workload .	152
24.6 Emotions	153

1: an overview about methods for XAI research can be found in Hoffman, Mueller, et al. (2018)

Table 24.1: Overview of the measured variables and the corresponding method used in the experiments. In addition, standard variables (e.g., demographic information and attitude towards (X)AI) were asked in all experiments. More information about these variables can be found in Chapter 15 on page 85

	Measured Variable	Used Method
Cooperation & Collaboration	VR-Robot Study Trust Emotions Trust Explanation Satisfaction Self-efficacy	TiR (Schaefer, 2013) DEQ (Harmon-Jones et al., 2016) TiE (Hoffman, Mueller, et al., 2018) ESS (Hoffman, Mueller, et al., 2018) CSE-B (Bernacki et al., 2015)
	Conversational AI Study Trust Game Experience Agent perception Mental model AI	TiA-K (Körber, 2018) GEQ (IJsselsteijn et al., 2013) Godspeed (Bartneck et al., 2009) Retrospection technique (Hoffman, Mueller, et al., 2018)
Education	Gloria Study Trust XAI perception Agent perception	TiA-J (Jian et al., 2000) Likert-scale rating & free-form Likert-scale ratings & free-form
	Museum Study Technical affinity Trust Agent perception XAI perception	TA-EG (Karrer et al., 2009) TiA-J (Jian et al., 2000) Likert-scale ratings & free-form Likert-scale rating & free-form
Medical Decision Support	NOVA Study Trust Mental model AI Mental model Self XAI perception Cognitive Load Self-efficacy	TiA-J (Jian et al., 2000) Task reflection (Hoffman, Mueller, et al., 2018) & confidence Likert-scale rating (Hoffman, Mueller, et al., 2018) Helpfulness & easy to understand using a Likert-scale rating NASA-TLX (Hart & Staveland, 1988) CSE (Compeau & Higgins, 1995)
	Pneumonia Study Mental model AI Mental model Self Explanation Satisfaction Trust Emotions Self-efficacy	Task prediction, task reflection (Hoffman, Mueller, et al., 2018) Likert-scale rating (Hoffman, Mueller, et al., 2018) ESS (Hoffman, Mueller, et al., 2018) TiA-K (Körber, 2018) DEQ (Harmon-Jones et al., 2016) CSE-B (Bernacki et al., 2015)

24.1 Mental Models

One method to make the theory of mind development measurable is false-belief tasks. These were first described by Wimmer and Perner (1983). Here, children were asked questions about their beliefs about an observed third person (e.g., “What do you think, where will Peter search for the rabbit?”). For XAI research, similar questions are asked to investigate users’ mental models. However, the third person they are to judge does not represent another human but the AI system. For example, Kulesza et al. (2015) investigated a user’s mental models of an AI for email classification (i.e., “What do you think, which topic would the AI assign the text?”). Here, an AI sorted text into a category based on the content of the mail. For the AI, two variables were essential for this classification: class ratios (i.e., size of messages in the category) and feature presence (i.e., words that are important for a specific category). After an introduction, participants could explore the AI system on their own. Afterwards, participants had to improve AI predictions by labelling messages. Participants in the control group received only the accuracy value. In contrast, participants of the experimental group were presented with “Why?” explanations for the classification of text into a category (e.g., Why was this text classified into the category football?). After that, Kulesza et al. (2015) evaluated the participants’ mental models by letting them rate messages regarding (1) which topic would the AI system classify the message and (2) why it would classify that. For the empirical studies in this dissertation, we used methods for mental model elicitation as described by Hoffman, Mueller, et al. (2018). Each method is presented with a short description and examples of how the methods were used in the six experiments of this dissertation.

24.1.1 Retrospection Questions

Inspired by Friedman et al. (2018) who asked students about their beliefs about the changing of seasons, Hoffman, Mueller, et al. (2018) suggests using open questions for getting a holistic view of *users’ belief about an AI system*. In the *Conversational AI Study*, we asked all participants four retrospective questions about their perception of their AI-based dialog partner: (1) “Describe your dialog partner”, (2) “What information was important for your dialog partner?”, and (3) + (4) “What was easy/difficult for your dialog partner?” The answers to these different questions are intended to help create a complete picture of the user’s mental model of the AI system. In addition, the various questions also help to make visible any contradictory beliefs that users may have. For example, regarding question (2), the responses of some participants revealed that they perceived the AI dialog partner to be very limited in its capabilities (e.g., unable to solve complex problems). But in question (2), they described the AI as having human-like characteristics that require higher cognitive abilities (e.g., the dialog partner intentionally tried to mislead them).

24.1.2 Prediction Task

A prediction task aims to investigate whether users can predict the *future actions of an AI system* (Hoffman, Mueller, et al., 2018) correctly. This

task can be combined with a confidence rating to examine how sure participants are in their predictions. For example, in the *Pneumonia Study*, the participants had to predict what the AI model would decide for a given X-ray image (i.e., “What do you think will the AI decide? Base your prediction on the explanation”) with a dichotomous answer format (i.e., “suffering from pneumonia” or “not suffering from pneumonia”). In addition, we asked them about their confidence in their answer (i.e., “How confident are you that you predicted the decision of the AI correctly?”) on a 7-point Likert scale.

24.1.3 Task Reflection

This method allows users to describe their *reasoning about an AI system*. The task reflection approach is more of a qualitative approach, where participants are asked about their understanding of how the AI system works. This can be combined with a quantitative rating of their confidence in understanding the AI. For example, in the *NOVA Study*, after each five-minute interaction with NOVA, we asked users about their assumptions about why the model recognized pictures with emotional expressions of persons as wrong or correct, respectively. This free-form feedback was combined with a 7-point Likert scale that allowed users to evaluate their confidence in their statements. In the *Pneumonia Study*, the task was similar. After completing the prediction task, the participants were asked to describe their understanding of the AI’s reasoning. For this, the participants were asked two questions about their mental model of the AI: “What do you think the AI pays attention to when it predicts pneumonia?” and “What do you think the AI pays attention to when it predicts healthy lungs?”

24.1.4 Users’ Model of Self

Besides investigating the users’ model of the AI system, we investigated the model of the user themselves. This was done to understand how their *perception of the AI system differs from users’ self-perception*. The self-perception can also indicate whether users transfer their conceptions and mental models to the AI system. For example, at the end of the *NOVA Study*, we showed all participants five images with people expressing emotions (i.e., anger, neutral, disgust, sadness, happiness) and asked them to (1) classify the emotion themselves and rate their confidence in the decision (8-point Likert scale) and (2) draw the areas on the face that was relevant for them to identify the emotion. In the *Pneumonia Study*, we asked the participants how they would predict the shown X-ray image of a lung (i.e., “Do you think the original X-ray shows a person suffering from pneumonia or not?”) combined with a confidence rating, (i.e., “How confident are you that your diagnosis is right?”) on a 7-point Likert scale, (1 = not at all confident, 7 = very confident).

24.2 Explanation Satisfaction

To measure users’ subjective satisfaction with the presented explanations, we used the Explanation Satisfaction Scale (ESS) proposed by Hoffman,

Mueller, et al. (2018). The ESS consists of eight items (e.g., “This explanation of how the [software, algorithm, tool] works is satisfying.”), each rated on a 5-point Likert scale (1 = I disagree strongly to 5 = I agree strongly). We used the ESS in the *VR-Robot Study* and the *Pneumonia Study* to investigate users’ impressions of the presented explanations.

24.3 Trust

In particular, two methods are used to measure trust: evaluation using questionnaires and behavioural observation. In our experiments, trust was assessed with the help of validated questionnaires.

24.3.1 Trust in Robots

The 14-item version of the Schaefer Human-Robot-Trust questionnaire (TiR) (Schaefer, 2013) was used to investigate the trust of participants in the virtual robot during the experiment in our *VR-Robot Study*. Each of the 14 items had to be rated on an 11-point Likert scale ranging between 0% and 100% with the sentence “What % of the time will this robot [item]”.

24.3.2 Trust in Automation - Jian et al. (2000)

Trust is regarded as a user trait in the Trust in Automation Scale (TiA-J) from Jian et al. (2000). The TiA scale is one of the most commonly used trust scales in HCI (Hoffman, Mueller, et al., 2018). With 11 items, the TiA measures six subscales of Trust: Fidelity, loyalty, reliability, security, integrity, and familiarity on a 7-point Likert scale (1 = not at all, 7 = extremely). We used the TiA-J in three of our studies to investigate the trust of the participants: *Gloria Study*, *Museum Study*, and *NOVA Study*.

24.3.3 Trust in Automation - Körber (2018)

Another questionnaire that investigates Trust in Automation (TiA-K) is proposed by Körber (2018). TiA-K consists of 19 items on six subscales (i.e., reliability/competence, understanding/predictability, familiarity, the intention of developers, the propensity to trust, and trust in automation). Compared to the questionnaire of Jian et al. (2000), in the TiA-K, the two-items subscale (i.e., trust in automation subscale) can be used to measure trust in automation. Körber (2018) points out that the two-items subscale is sufficient to measure trust if people have no previous experience with the system. For the *Collaboration Study*, we used the whole TiA-K to investigate the trust of the participants in their AI dialog partner. In the *Pneumonia Study*, we used the two-items subscale (i.e., “I trust the system” and “I can rely on the AI system”) to evaluate trust in the presented AI system.

24.3.4 Trust in Explanation - Hoffman, Mueller, et al. (2018)

In the *VR-Robot Study*, we used one item of the ESS to investigate the participants' trust regarding the explanations given (i.e., "This explanation lets me judge when I should trust and not trust the robot", 5-point Likert scale).

24.4 Computer Self-Efficacy

To measure the computer self-efficacy of the participants, we used the Computer Self-efficacy scale (CSE) (Compeau & Higgins, 1995). This scale consists of 10 items that ask the user to estimate their perceived self-efficacy when using an AI system. Another possibility to measure computer self-efficacy is inspired by Bernacki et al. (2015) (CSE-B), who used one question (i.e., "How confident are you that you could solve a math question like this one in the future?") to investigate users' self-efficacy towards a technical system when evaluating a cognitive tutor for algebra. In the *NOVA Study*, we investigated the computer self-efficacy of participants when interacting with the software NOVA (e.g., "I could complete the job using the software package if I had only the software manuals for reference"). The CSE items were initially answered with "Yes" or "No". If a user answered "Yes", they were then asked on a 10-point Likert scale how confident they would be with this item (1= not confident at all, 10 = totally confident). In the *Pneumonia Study*, we used the question motivated by Bernacki et al. (2015). We adapted this item so that the wording fits our experiment (i.e., "How confident are you that you could detect pneumonia using the presented explanations in the future?"; 10-point Likert-scale rating). We used two items to measure the self-efficacy toward the robot in the *VR-Robot Study*. For this, we used two variations of the item of Bernacki et al. (2015) to address users' perception of future successful interaction with the robot and their error-solving competencies (i.e., "How confident are you that you would successfully interact with a robot like this one in the study in the future" and "How confident are you that you could solve a robot error like this one in the study in the future?").

24.5 Cognitive Workload

In one of our studies, we also collected data about participants' subjective cognitive workload using the NASA-TLX questionnaire (Hart & Staveland, 1988). The NASA-TLX consists of 6 scales (i.e., mental demand, physical demand, temporal demand, performance, effort, and frustration level) with 21 gradations that can be clustered in "low", "medium", and "high". We used the NASA-TLX in the *NOVA Study* to investigate participants' cognitive load of the previously performed task using the NOVA software.

24.6 Emotions

Besides the subjective rating of the explanations, we investigate the emotional feelings of participants in the *Pneumonia Study* and the *VR-Robot Study*. For this purpose, we used the subscales *anger*, *happiness*, and *relaxation* of the Discrete Emotions Questionnaire (DEQ)(Harmon-Jones et al., 2016). Here we asked participants: “While doing [specific task of the experiment], to what extent did you experience these emotions?” Each subscale consists of four items (e.g., “chilled out” for relaxation) ranging from 1 (not at all) to 7 (an extreme amount). To the best of our knowledge, our experiments were the first to investigate XAI’s impact on users’ emotional states.

25.1 Quantitative Analyses	154
25.1.1 ANOVA	155
25.1.2 T-test	155
25.1.3 MANOVA	156
25.1.4 Correlation	156
25.2 Qualitative Analyses	156
25.2.1 Conventional Content Analysis	157
25.2.2 Summative Content Analysis	157

Table 25.1: Overview of the quantitative and qualitative analyses of the experiments. In all studies, basic statistical analyses of demographical data were conducted, not presented here. *Analysis* refers to additional inferential statistical analyses

Olds et al. (2005) give an overview of different qualitative and quantitative evaluation techniques that are used in engineering education. They highlight that the selection of the techniques should be made depending on the study design. Ibáñez-Espiga et al. (2014) referred to these techniques and pointed out that a combination of qualitative and quantitative evaluation in a so-called “mixed-method” approach gains more profound insights into the investigated topic. In their case, they examined the impact of gamified learning activities on learning the programming language C by using qualitative data (i.e., log files) and quantitative data (i.e., answers in questionnaires). In our six studies, we also collected qualitative and quantitative data (see Table 24.1 on page 148 for an overview of the methods used for data collection). To draw interpretations and conclusions about the impact of XAI on end-users, we used different inferential statistical analysis techniques presented in the following.

Study Title	Analysis
Cooperation & Collaboration	
VR-Robot Study	MANOVA, ANOVA, t-tests (paired, independent, post-hoc)
Conversational AI Study	MANOVA, one-sample t-test, correlation*, content analysis
Education	
Gloria Study	ANOVA, contrast analysis
Museum Study	MANOVA, ANOVA, post-hoc tests, correlation*
Medical Decision Support	
NOVA Study	MANOVA, one sample t-test
Pneumonia Study	MANOVA, ANOVA, content analysis

*= Spearman’s correlation

25.1 Quantitative Analyses

The quantitative data were collected using validated questionnaires or were generated by using the qualitative analysis approach presented later. These data were analysed using inferential statistical methods such as MANOVA¹, ANOVA², contrast analysis, t-tests, as well as correlations (see Table 25.1 and the *Quantitative Data* branch in Figure 25.1 on the next page). The data analysis strategy was planned during the experiment design phase. For this, statistical hypotheses and the analysis methods to evaluate them were defined before experiments started.

1: multivariate analysis of variance

2: analysis of variance

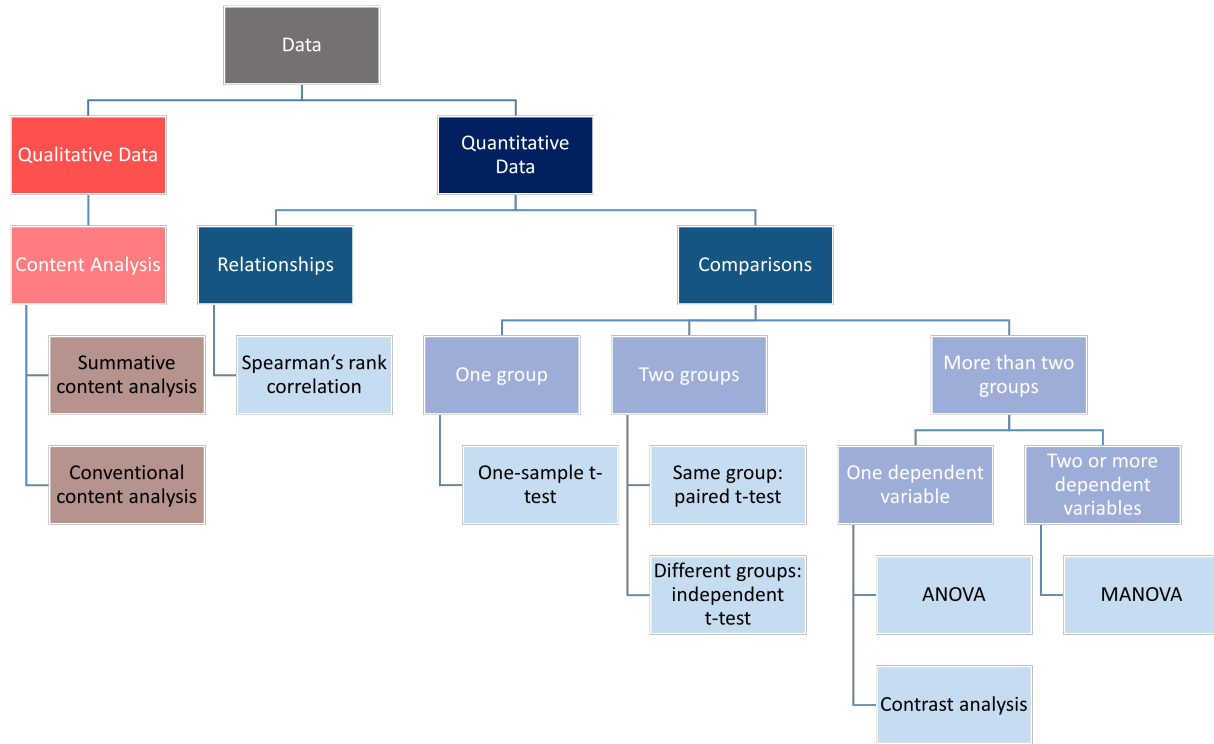


Figure 25.1: Overview of the different analysis methods used in this dissertation. Analysis methods were chosen depending on the data (qualitative or quantitative)

25.1.1 ANOVA

ANOVAs were conducted when more than two experimental groups were compared, which was, for example, the case in the *Gloria Study*. We had four experimental groups (i.e., visual explanation only and three explanatory virtual agent conditions). A particular case of ANOVA is the contrast analysis. Contrast analysis is a specific way of testing directional hypotheses (planned contrasts) that uses linear contrast coefficients to weight the means of the groups that are compared (Field et al., 2012). This method offers insights into group differences as it allows for specific and more precise comparisons between groups. Contrast analysis can be used when a particular effect is expected. As with all inferential statistical procedures, appropriate directional hypotheses must be formulated prior to the experiment (Field et al., 2012). Based on these requirements, we calculated a contrast analysis in the *Gloria Study*. We benefited from the advantages of contrast analysis since it leads to a higher power, makes post-hoc testing obsolete, and the effect sizes are easier to interpret.

25.1.2 T-test

T-tests were conducted when only two groups were compared. In the *VR-Robot Study*, we compared participants before and after a robot mistake occurred. To investigate the impact of the error on participants' trust, we used a paired t-test. Paired t-tests compare the mean of the same participants in a pre-post test design (Field et al., 2012). Another version

of t-tests is the independent t-test, where the means of two different (i.e., independent) groups of participants (e.g., different XAI groups) are compared. We used this kind of test when conducting post-hoc testing after an ANOVA. In post-hoc tests, pairwise comparisons are performed to discover the direction of the detected significant difference that the ANOVA revealed (Field et al., 2012). When the assumptions of the t-test are violated (e.g., homogeneity of variance or normal distribution), we calculated the non-parametric equivalent, the Mann-Whitney U-test (e.g., for the post-hoc comparisons in the *Museum Study*) (Field et al., 2012). To control for an increasing α -error due to multiple post-hoc tests, we used the Holm correction to adjust the p-values. The last type of t-test we used was the one-sample t-test. We used this test to compare the mean of a variable to the mean value of a rating scale (e.g., $M = 3$ on a 5-point Likert scale) to determine whether the participants rated a variable higher or lower than the mean of the rating scale. This kind of test was used, for example, in the *NOVA Study*, where we compared participants' rating of NOVA to the mean of the rating scale (i.e., $M = 4$ on a 7-point Likert scale).

25.1.3 MANOVA

3: a dependent variable represents the effect caused by an independent variable, e.g., different explanation conditions or pre-post tests

ANOVA and t-test are applicable when only one dependent variable³ is investigated. MANOVA was conducted when two or more dependent variables were investigated. Instead of comparing single variables in separate ANOVA, one MANOVA can be conducted. Only when the results of the MANOVA are significant, an ANOVA and post-hoc tests calculated (Field et al., 2012). For example, we ran a MANOVA, followed by ANOVA and post-hoc tests in the *NOVA Study*. We investigated trust, self-efficacy, and cognitive workload as dependent variables.

25.1.4 Correlation

We used Spearman's rank correlations to investigate relationships between variables in the *Museum Study* and the *Conversational AI Study*. Spearman's rank correlation belongs to the non-parametric procedures and is used to detect monotonic (e.g., linear) relationships of discrete ordinal or continuous variables (Field et al., 2012). For example, we found in the *Conversational AI Study* a significant negative relationship indicating that participants with more experience in cooperative games were misunderstood more often by the AI dialog system.

It should be noted that although these correlations show interesting relationships, they should not be interpreted causally. This means that one variable does not have to cause a connection with another variable. To verify this, further experiments need to be conducted in which these variables are changed in a controlled manner.

25.2 Qualitative Analyses

For qualitative analyses, we used the free-form feedback of the participants. The feedback was analysed using content analysis (Hsieh &

Shannon, 2005) (see *Qualitative Data* branch in Figure 25.1 on page 155). Content analysis annotates users' expressions using predefined labels or labels generated from the collected data. In this way, user trends can be identified and quantified while giving users greater freedom of expression compared to closed-ended questions⁴. Hsieh and Shannon (2005) distinguish between conventional content analysis, directed content analysis, and summative content analysis. In our studies, we used the *conventional content analysis* and the *summative content analysis*. To the best of our knowledge, we were the first to use these qualitative analysis techniques to investigate XAI's impact on users' mental models. The results of summative as well as conventional content analysis can be directly reported, as we did in the *Conversational AI Study*, or they can be used for further statistical analyses, as we did in the *Pneumonia Study*.

4: i.e., questions that use predefined answer options

25.2.1 Conventional Content Analysis

The conventional content analysis aims to cluster the answers to free-form questions into categories. The categories are not defined in advance but are derived from the data (Hsieh & Shannon, 2005). We used this approach in the *Conversational AI Study*. Here, we had two independent annotators who labelled the data of our free-form answers depending on the content of the answers. From this, they created categories. For example, answers to the question "What was difficult for your dialog partner?" that included content that the AI dialog partner had a hard time handling misunderstandings were collected in the category "mistakes and misunderstandings".

25.2.2 Summative Content Analysis

For summative content analysis, certain words or content are determined in advance. The occurrence of these words is subsequently examined in the data and counted. Until this point, the procedure is purely quantitative (Kondracki et al., 2002). In the next step, a qualitative part is added. Based on the words defined in advance, alternative terms in the text that describe the same content are added and included in the count. We used this procedure in the *Pneumonia Study*. Here, correct words were defined in advance in answering the question "What do you think the AI pays attention to when it predicts pneumonia?". These included, for example, "opacity". These occurrences were counted. In the next step, alternative terms such as "white colour" or "lung shadows" found in the data were also identified as correct descriptions and counted.

Careful preparation of empirical studies goes a long way toward ensuring that their results are valid and replicable. For this, it is necessary to consider the study design, measurement techniques, and data analyses. This section explained which variables were collected using various methods in our XAI studies. We used validated questionnaires and supplemented them with evaluation methods such as task reflection or prediction tasks, as suggested by Hoffman, Mueller, et al. (2018). In doing so, our conducted experiments used a broad spectrum of methods and investigated a variety of psychological constructs like mental models and trust. In addition to the subjective perception of the participants and the objective measurement of their performance in the tasks, we also investigated the emotional impact of XAI on users. The end of this chapter includes a description of the analyses of the collected data. In our studies, we used a combination of quantitative and qualitative evaluations, which helped us to gain valuable insights into users' perceptions of XAI design.

VIII. EMPIRICAL INVESTIGATION OF HUMAN-CENTERED EXPLAINABLE AI

27.1	Overview	160
27.2	Explainable AI in Industrial Human- Robot Cooperation - VR-Robot Study . . .	161
27.2.1	Highlights	161
27.2.2	Introduction	161
27.2.3	Pilot Study	162
27.2.4	Methodology	164
27.2.5	Main Experiment	167
27.2.6	Discussion	171
27.2.7	Conclusion	173
27.3	Explainable AI in Human-Agent Collaboration - Con- versational AI Study	174
27.3.1	Highlights	174
27.3.2	Introduction	174
27.3.3	Hypotheses	175
27.3.4	Pilot Studies	175
27.3.5	Main Study	176
27.3.6	Dataset	180
27.3.7	Results	181
27.3.8	Discussion	186
27.3.9	Future Work	188
27.3.10	Conclusion	188
27.4	Summary Coopera- tion & Collaboration Experiments	189

In this chapter, two experiments using AI technology in cooperation & collaboration settings are presented. The experiments described are based on the work published in:

- *VR-Robot Study*:
Hald, K.*, **Weitz, K.***, Rehm, M., & André, E. (2021). “An error occurred!” - trust repair with virtual robot using levels of mistake explanation. In K. Ogawa, T. Yonezawa, G. M. Lucas, H. Osawa, W. Johal, & M. Shiomi (Eds.), *Proceedings of the 9th international conference on human-agent interaction* (pp. 218–226). ACM. doi: [10.1145/3472307.3484170](https://doi.org/10.1145/3472307.3484170)[†]
- *Conversational AI Study*:
Weitz, K.*, Vanderlyn, L.*, Ngoc, T. V., & André, E. (2021). “It’s our fault!”: Insights into users’ understanding and interaction with an explanatory collaborative dialog system. In A. Bisazza & O. Abend (Eds.), *Proceedings of the 25th conference on computational natural language learning, conll 2021, online, november 10-11, 2021* (pp. 1-16). Association for Computational Linguistics. doi: [10.18653/v1/2021.conll-1.1](https://doi.org/10.18653/v1/2021.conll-1.1)[†]

27.1 Overview

Human-agent cooperation and collaboration in industrial settings is an expanding research field. However, when working together, agent failures are essential in decreasing trust and interfering with interaction and cooperation. In the following two experiments, we investigated the impact of textual and verbal explanations in industrial settings with an agent.

- **VR-Robot Study**: First, a scenario in human-robot cooperation with a rule-based industrial robot in VR is presented. It is unclear whether explanations help to restore human-robot trust after a robot failure. Therefore, we investigated in an online pilot and a laboratory main study the impact of textual and verbal explanations as a trust-repairing action after robot failures.
- **Conversational AI Study**: Second, an online study with a rule-based text agent in a collaborative game was conducted. Here, we investigated the impact of explanations and failures of a dialog system on end-users perception and collaborative performance.

In our VR-Robot Study, the robot and the end-user must perform a sorting task (*cooperative task*). The robot had to move blue bottles while the user had to sort red bottles (see Chapter 20 on page 125 for a detailed description of the task). Both actors shared the workspace but completed their subtasks independently of their counterparts.

*Both authors contributed equally to this work

[†]Licensed under the [Attribution 4.0 International License \(CC BY4.0\)](https://creativecommons.org/licenses/by/4.0/)

In our Conversational AI Study, the user and an AI dialogue partner had to successfully activate a series of modules equipped with switches, buttons and light displays (*collaboration task*). Here, the user and the AI had different knowledge necessary to solve the tasks. The user could see the four puzzle modules but did not know the rules about which switches and buttons had to be activated to activate the modules. On the other hand, the AI knew the rules but did not see the modules and therefore did not know what state they were in (see Chapter 20 on page 125 for a detailed description).

The two studies, their results, and their interpretation are explained in more detail in the following.

27.2 Explainable AI in Industrial Human-Robot Cooperation - VR-Robot Study

27.2.1 Highlights

- ▶ End-users demand suggestions for solutions/prevention of robot failures in explanations.
- ▶ Explanations are perceived as helpful in deciding whether to trust or distrust a robot in VR.
- ▶ Explanations did not help to repair trust after a robot failure.
- ▶ Explanations have no impact on end-users self-efficacy or emotion in a VR-robot setup.

27.2.2 Introduction

The cooperation between humans and machines in the industrial setting is becoming more and more realised due to the enormous progress in the fields of robotics and machine learning. Introducing cooperative robots in manual production can help relieve the workers of strenuous and repetitive tasks. In addition, intuitive usage and interaction by humans have become increasingly common. However, the more natural the handling of robots in the industry evolves, the more demands humans place on them. Human-robot interaction (HRI) can be disrupted if these demands are not met. In addition to reduced trust and frustration, this can have serious consequences such as accidents and production losses (de Visser et al., 2020; Hancock et al., 2011). To enable successful HRI, we must maintain human-robot trust, especially when working with a robot in close proximity. To this end, we investigate the capabilities of system communication with the human cooperator to perform trust repair through explanation in cases where the robot makes a failure while executing a shared objective. We base the research on the context of a shared task where the human and robot have to move a collection of objects within a shared tabletop workspace. To integrate system communication with non-obstructive output modalities, we base the design of the communication system on projection-based augmented reality (AR) so that messages can be displayed directly on the work surface. To sum up, we investigate how we can use failure explanations after a robot failure as a trust-repairing

1: the presented studies in this paper, as well as the gathered data have been collected with respect of the GDPR regulations

action to maintain trust during close-proximity cooperation. Rather than implementing the robots' communication system using real hardware, we test our prototype iterations using computer-generated demonstrations and VR testing environments¹. With our work, we make the following contributions:

- ▶ With our pilot study, we give novel insights about the requirements and expectations of end-users towards robot explanations after failures.
- ▶ We present a VR setup to research robot failures in close proximity cooperation tasks.
- ▶ Our results about the impact of different levels of explanations after robot failures on trust, explanation satisfaction, self-efficacy, and emotional state of end-users gain new insights regarding XAI in HRI.
- ▶ We discuss the challenges using explanations in HRI and how our findings are useful for researchers to design XAI in HRI.

Our pilot study revealed that users are more interested in solutions to failures than in why the failure happened. Therefore, in our main study, we evaluated three levels of failure explanations (no explanation, explanation, and explanation with solution) after a robot in VR made a failure in executing a shared objective. After testing with 30 participants, we found that the robot making a failure significantly affects trust toward the robot, compared to successfully completing the task. While participants found the explanations helpful in trusting or distrusting the robot, the levels of the explanation did not lead to an increase in trust towards the robot after a failure. In addition, we found no significant impact of explanations on self-efficacy and the participants' emotional state. Our results show that more than providing explanations is needed to increase human-computer trust after robot failures.

27.2.3 Pilot Study

Our work aims to investigate human-robot trust in an interaction scenario in which the robot makes a failure. In the pilot study we conducted, we first wanted to examine whether different *explanation types* (i.e., textual or auditory) are preferred by participants. In addition, we varied the *type of failure* (i.e., colour vision failure or calibration failure) (see details in Chapter 20 on page 125) as illustrated in Figure 27.1 on the facing page. For our pilot study, we presented the participants with videos of a virtual robot sorting bottles of different shapes on either side of a table.

Hypotheses

The pilot study was conducted to guarantee that the different explanation modalities and failure types did not significantly differ in their impact on trust. Furthermore, we wanted to understand whether the explanations were sufficient and whether/which additional information participants found helpful. In more detail, we formulated the following hypotheses:

- ▶ **H1: Robot Performance & Likeability:** The rating of robot performance and likeability will differ between the no-failure and

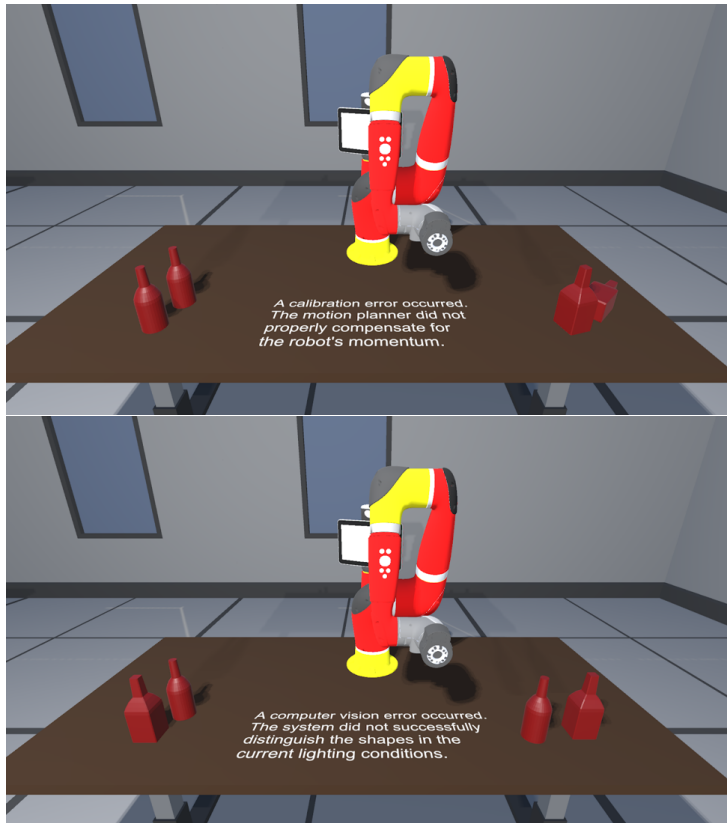


Figure 27.1: Textual explanation modality. Two robot failures were explained during the pilot study: a *calibration failure* (top) and a *computer vision failure* (bottom). Figure from Hald, Weitz, et al. (2021)

the two failure conditions, where the ratings for the no-failure robot will be higher.

- ▶ **H2: Explanation Quality:** After being presented with a robot failure in videos of a virtual robot and a given modality of explaining the failure, the user can describe the failure accurately.
- ▶ **H3: Modality of Explanation:** There will be no difference between the modality of explanation (i.e., textual and auditory) regarding likeability, performance, trustworthiness, and understanding of the robot.
- ▶ **H4: Type of Robot failure:** There will be no difference between the types of failure (i.e., calibration failure and colour vision failure) regarding likeability, performance, trustworthiness, and understanding of the robot.

To answer these hypotheses, we used a between-subjects design for the modality of explanation (i.e., textual or auditory), meaning that every participant saw one of the explanation modalities. For the two different robot failures (i.e., colour vision failure and calibration failure), we used a within-subjects design. Here, every participant saw both failures during the study².

Study Design

The pilot study took place online. Within the study, the participants were shown a series of videos of a virtual robot modelled after the Rethink Robotics Sawyer³ model. This robot had the task of sorting bottles at either end of a table based on their shape.

2: we randomized the order of the presented failures to control for sequence effects

3: More information about this robot can be found on <https://www.rethinkrobotics.com/de/sawyer> (last accessed on 29.09.2022)

- ▶ **First video:** The first video showed the robot successfully completing the sorting task, switching the positions of two bottles so that two round-based bottles are on the left side of the table and two square-based bottles are on the right. Then, the participants rated the performance of the robot and their impression of the robot. They were then asked to briefly describe the robot, its behaviour, and its task.
- ▶ **Second video:** The second video showed the robot performing the same task again but making an failure (i.e., computer vision or calibration failure). The participants then answered the same questions about the robot's performance and their impression. After that, they were asked to briefly describe the difference from the previous video.
- ▶ **First Explanation:** Subsequently, they were shown an explanation of the previously seen failure (i.e., textual or auditory explanation). The textual explanation modality is shown in Figure 27.1 on the previous page. Next, the participants had to answer several questions about the explanation shown.
- ▶ **Third video:** After answering these questions, they were shown a third video of the robot making the other type of failure.
- ▶ **Second explanation:** Here, again, an explanation was shown to them afterwards, and the participants had to evaluate it.

At the end of the online study, participants had to provide personal information about themselves.

27.2.4 Methodology

We used different scales to gain insights into the user's impressions regarding the robot failures and the explanation modalities.

Performance To evaluate the perceived robot performance, we asked the participants after every video to rate the robot's performance using a 7-point Likert scale (1 = not good, 7 = very good).

Likeability Similar to the measurement of the perceived robot performance, we asked the participants, after the no-failure video as well as after each explanation, how much they liked the robot and if they wanted to work with the robot⁴ (7-point Likert scale; 1 = not at all, 7 = totally).

4: we calculated a mean value from both ratings

Explanation Quality To measure the quality of the presented explanations, we used two items of the Explanation Satisfaction Scale (ESS)(Hoffman, Mueller, et al., 2018). Here we asked the participants (1) whether the explanations helped to trust the robot and (2) whether they helped to understand how the robot worked (5-point Likert scale; 1= I disagree strongly, 5= I agree strongly). In addition, we asked two general yes/no questions regarding the explanations, i.e., "Have you learned anything because of the explanation?" and "Was the explanation easy to understand?". Finally, we also asked for free-form feedback. Here we wanted to know from the participants which parts of the explanation were easy/not easy to understand, whether they would have needed more/additional

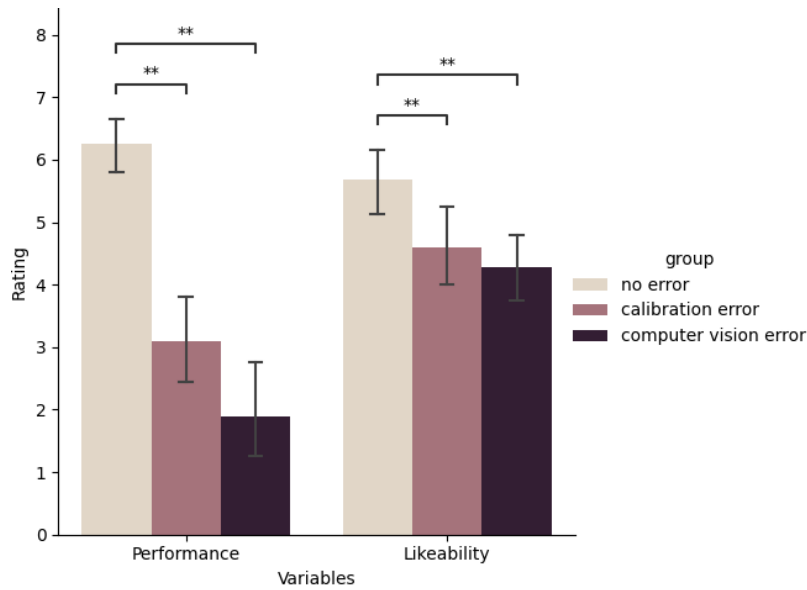


Figure 27.2: Rating of the robot in the no-failure and the two failure conditions of the pilot study. The ratings for the no-failure condition were significantly higher than for the two failure conditions. failure bars represent the 95% CI. $**p < .001$

information and which one and why the explanation was not helpful (i.e., when participants answered the “Have you learned anything because of the explanation?” question with “no”).

Demographic Information In addition, at the end of the pilot study, we collected participants’ personal information (e.g., age, gender) and their knowledge and attitudes toward AI and XAI.

Participants

In our pilot study, 20 people between 21 and 54 years ($M = 29.3$, $SD = 7.47$) participated. 11 of them were male, and nine were female. All participants had heard about the term AI, but only nine of them had heard about XAI.

Results

Rating of Robot Performance & Likeability To answer H1, we compared the variables likeability and performance between the no-failure robot and the two failure conditions. For this, we conducted paired t-tests. Here, the performance of the no-failure robot was perceived as significantly higher compared to the calibration failure robot, $t(19) = 9.20$, $p = < .001$, $d = 2.06$ (large effect) as well as the colour vision failure robot, $t(19) = 9.11$, $p = < .001$, $d = 2.04$ (large effect)⁵. Similar results were found for the likeability. The no-failure robot was liked significantly more compared to the calibration failure robot, $t(19) = 4.27$, $p = < .001$, $d = 0.95$ (large effect) as well as the colour vision failure robot, $t(19) = 6.06$, $p = < .001$, $d = 1.35$ (large effect). These results are shown in Figure 27.2. Therefore, the results support our H1 that ratings for the no-failure robot were higher.

5: interpretation of the effect size d according to J. Cohen (1988) is:
 $d < .05$: small effect;
 $d = 0.5-0.8$: medium effect;
 $d > 0.8$: large effect

Table 27.1: Independent sample t-test. No significant differences in explanation quality (trustworthiness & understandability), performance and likeability between the two different explanation modalities (textual vs auditory) for both types of robot failure. Trustworthy refers to “help to trust or distrust the robot”, Understandable refers to “helps to understand how the robot works”

Explanation Modalities	Measurement	$t(18)$	p
calibration failure text vs. audio	trustworthiness	-0.94	.36
	understandability	-1.40	.18
	performance	.30	.77
	likeability	-0.60	.55
computer vision failure text vs. audio	trustworthiness	-0.87	.39
	understandability	-0.50	.62
	performance	.74	.47
	likeability	-0.70	.49

Rating of Explanation Quality To answer H2, we asked the participants whether they had learned something because of the explanation and whether it was helpful or not to get a general impression of the explanation quality. Overall, we found evidence to support H2. Fourteen participants stated that they had learned something from the calibration failure explanation. Seventeen participants said they had learned something from the computer vision failure explanation. Besides the quantitative feedback of the participants, we also analysed the qualitative free-form feedback. Here, participants mentioned for the computer vision failure, that it would be helpful to add information on how to solve the failure (e.g., information on whether the failure occurred because the environment was too dark or too bright). Participants mentioned that the explanation was too technical for the calibration failure and that they would have needed more information on how to fix the failure or calibrate the robot correctly to avoid similar failures in future.

To evaluate the explanation quality in more detail, we used two items (“help to trust or distrust the robot” and “help to understand how the robot works”) proposed by Hoffman, Mueller, et al. (2018). Analyses on these scales will be reported in the following sections.

Comparison of Explanation Modalities For answering H3, we used independent samples t-tests to analyse the impression of the two different explanation modalities (textual vs auditory) regarding explanation quality, likeability, and performance of the robot. We found no significant differences between the conditions (see Table 27.1), supporting our H3 that explanation modalities do not differ.

Comparison of Robot-failure Types For H4, we conducted paired t-tests regarding explanation quality (i.e., trustworthiness & understandability), performance, and likeability. The analyses revealed (see Table 27.2) that the computer vision failure explanation helped more to trust or distrust the robot compared to the calibration failure explanation ($d = 0.62$ - medium effect). In addition, we found that participants rated the robot’s performance higher in the calibration failure condition ($d = 0.49$ - medium effect). However, these results do not support our H4, as we found differences between the failure types regarding performance and trustworthiness ratings.

Type of Robot failure	Measurement	<i>t</i> (19)	<i>p</i>	<i>d</i>
calibration failure vs computer vision failure	trustworthiness	-2.77	.012*	0.62
	understandability	-0.89	.38	-
	performance	2.18	.042*	0.49
	likeability	1.78	.09	-

* $p < .05$

Table 27.2: Paired t-tests. Significant differences in performance and trustworthiness between the two different failure types (calibration vs computer vision failure)

Discussion

The pilot study showed that people rated the robot significantly worse in terms of its performance and likeability when it made a failure. The general study design in terms of trust repair (comparing the trust of a correct working robot and a robot that causes an failure) was therefore maintained for the final study.

Based on the pilot study, the explanation for the calibration failure was too technical for end-users without experience in robotics. These resulted in a significantly lower trust rating and were mentioned by participants in the free-form feedback. This reflects the argument of Gerlings et al. (2021) saying that there is no generalised user to address with explanations. Instead, explanations must fit different stakeholders' abilities and preferences. To fit end-users needs, we decided to use only the computer vision failure in the final study and generate explanations for it. The free-form revealed that users are not satisfied with getting an explanation of the failure but also want a solution to prevent the failure in the future. This finding extends the work of Das et al. (2021), who stated that explanations should include environmental context and a history of successful actions of the robot in the past to support non-expert users in robot-recovery assistance. Inspired by the free-form feedback, we decided to refine the problem statement for the study and compare three different levels of failure explanation: (1) no explanation, (2) explanation of failure source and (3) explanation of failure source and a possible solution.

Since we did not find any significant differences regarding the modality of explanation (i.e., textual and auditory), we decided not to compare these factors in the final study. Instead, we decided to use only textual explanations due to better comparability.

27.2.5 Main Experiment

Study Design

To ensure high fidelity of system communication to the participants, we opted to test HRI and failure explanation using VR rather than implementing and testing with a real robot and projection-based AR overlays. This also increased the test rate, as we could test with multiple participants simultaneously, the only limit being the number of VR hardware setups. Based on the results from the pilot study, where the participants asked for more solution-oriented explanations rather than technical ones, we decided to define and test different explanation levels. The *first level* is an explanation as to why the robot made the failure, while the *second level*, in addition, explains how to solve the problem causing the failure. We compare these two levels as trust-repairing actions after a robot failure

along with a control condition, where no explanation is provided, and the user is only told that the robot failed the task. The trust repair is evaluated in terms of both trust in the robot as well as perceived quality of the explanations. Our hypotheses are as follows:

- ▶ **H1:** Providing an explanation after a robot makes a failure will yield higher levels of trust toward the robot than providing no explanation.
- ▶ **H2:** Providing different levels of explanation after a robot makes a failure will yield different levels of trust toward the robot.
- ▶ **H3:** Adding solution-oriented details to robot failure explanations will yield higher operator trust than explanations without them.

Procedure

After reading the experiment information and signing a consent form, participants were given instructions on how to complete the experiment by the test conductor. Participants were informed that they would perform a cooperative task with a virtual robot and would be given instructions via the text displayed on the table. It was emphasized that they should read the instructions carefully before they were told to put on the VR headset. Participants were introduced to the task by the text display. They were told that the robot was their teammate and that they were only supposed to move the red bottles while the robot moved the blue ones as they sorted the bottles according to the small white bottles shown next to their white platforms. The participants proceeded through the text instruction using the Menu button at the top of the Vive wands. Before starting the task, the participants were told how to move the bottles, and they were told to try them.

When the participant was told to press the Menu button to start the task, and they proceeded to push it, the robot would begin moving the blue bottles. If the participant moved the bottles before they started the task, the bottles were moved to their starting position when the task began. The task was completed when the participant sorted their bottles, and the robot moved its bottles. In the first task, the robot moved the bottles successfully, and the participant was presented with this message on the table: *“Your team succeeded at sorting the bottles. Please take off the virtual reality headset and call the test conductor”*. When they took off the headset, they were presented with the 14-item version of the Schaefer human-robot trust questionnaire (TIR) (Schaefer, 2013). Once the participants completed the questionnaire, they were instructed to put the headset back on and follow the instructions.

Once they had put the headset on, the display told them to start the task again by pressing the Menu button. In the second test, the robot would make a failure. Rather than switching round-base and square-base bottles between their platforms and sorting them correctly, it would switch two round-base bottles, leaving two blue bottles in their wrong positions. The task ended once the participant had completed their part correctly and the robot had stopped moving. The participant was then presented with this message on the table: *“Your team failed at sorting the bottles”*. If the participant was testing the condition with no explanation of the failure, they were immediately presented with the text, *“Please take off the virtual*

reality headset and call the test conductor". If the participant was testing the condition where they were given an explanation, they were presented with the message, "A computer vision failure occurred. The system did not successfully distinguish the bottles" before being told to take the headset off. Lastly, if the participant was in the condition with solution-oriented details, in addition to the previously mentioned explanation, they were presented with the message, "Better lighting conditions will help with successful sorting", before being told to take the headset off. The two tasks' lighting conditions were the same in the virtual environment. Once they had taken the headset off, the participants were presented with another TIR questionnaire as well as an additional post-test questionnaire, which they were told to fill out outside the laboratory. The approach of only doing two tasks was chosen due to the time required to answer the post-test questionnaires as well as to have the participants put the VR headsets on and off only a few times.

Methodology

To evaluate the participants' impression during and after the VR task, we used the following scales.

Trust During and after the VR task, we presented the 14-item version of the TIR questionnaire (Schaefer, 2013) at the end of each task. In the post-test questionnaire, we used the item "This explanation lets me judge when I should trust and not trust the robot" from the EES (Hoffman, Mueller, et al., 2018) to calculate an additional trust score reflecting the explanation quality.

Explanation Satisfaction We used the ESS (Hoffman, Mueller, et al., 2018) to measure the participants' subjective satisfaction with the kind of information (no explanation, explanation, or explanation with solution) that we presented after the robot failure.

Emotions We used items for the sub-scales *anger*, *happiness*, *anxiety*, and *relaxation* of the Discrete Emotions Questionnaire (DEQ) (Harmon-Jones et al., 2016) to evaluate the participants feelings after the VR task.

Self-efficacy We used two items to measure the self-efficacy towards the robot. For this, we used a variation of the item proposed by Bernacki et al. (2015) (i.e., "How confident are you that you would successfully interact with a robot like this one in the study in the future" and "How confident are you that you could solve a robot failure like this one in the study in the future?").

Participants

Thirty participants between 21 and 31 years ($M = 24.0$, $SD = 2.30$) took part in our experiment. Of these, 11 were female, and 19 were male. Twenty-nine participants had heard of the term AI, but only four had heard of the term XAI.

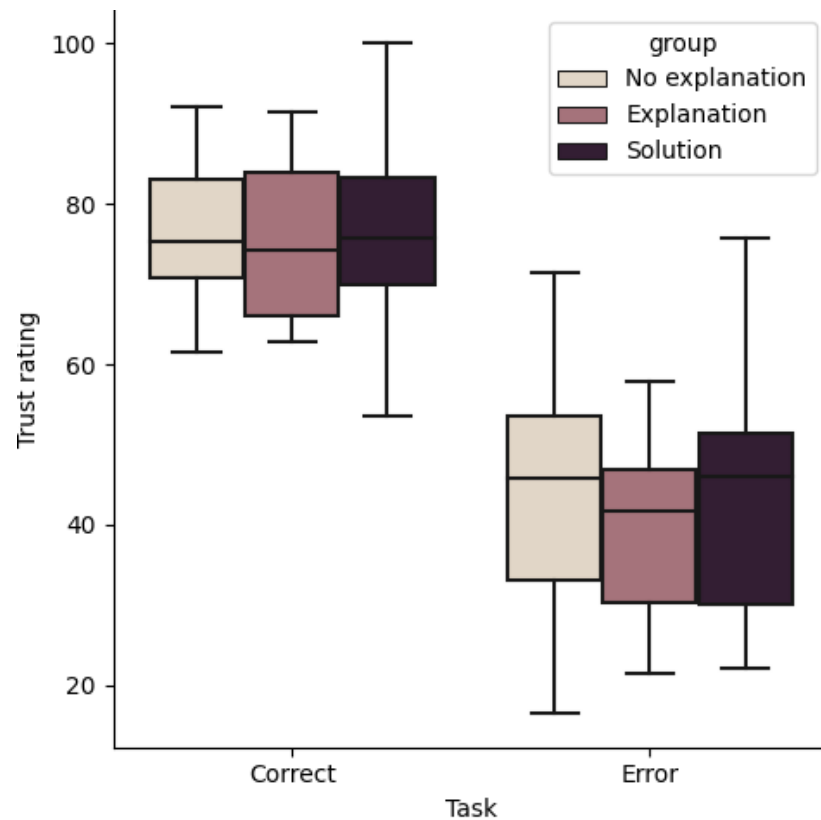


Figure 27.3: The average trust scores for the first and second bottle-sorting task between the three explanation conditions. While the difference in the trust rating between the correct and the incorrect task were significant, the three explanation groups did not differ

Results

Trust Scores The participants answered the trust questionnaire after completing each sorting task with the robot, the first being successful, while in the second task, the robot would make a failure. With all data groups being parametric, performing a paired t-test showed a significant difference in trust scores between the first and second task, whether no explanation ($t(19) = -12.5, p < .001, d = 2.79$), the base explanation ($t(19) = -11.7, p < .001, d = 2.61$) or solution-oriented explanations ($t(19) = -11.6, p < .001, d = 2.60$) were provided. However, when comparing the levels of explanation provided to the participants, performing a one-way ANOVA showed no significant effects of the type of explanation on the trust scores after the failure ($F(2, 27) = .24, p = .79$). The average trust scores are shown in Figure 27.3.

Explanation Satisfaction, Trust, and Self-efficacy After the VR experiment, all participants answered the post questionnaire, including questions about their explanation satisfaction and their trust in the explanation⁶, their general impression of the robot and their self-efficacy towards it.

To evaluate these variables between the three conditions, we conducted a one-way MANOVA. Here we found a significant statistical difference, Wilks' Lambda = 0.59, $F(10, 42) = 2.86, p = .008$. The following ANOVA revealed that only the variable trust showed significant differences between the conditions, $F(2, 25) = 5.92, p = .008$.

6: we calculated an overall explanation satisfaction value and used in addition the item for the helpfulness of explanation to trust or distrust the robot as a single variable. For details, see Chapter 23 on page 143

To determine the direction of this difference between the three conditions, we used post-hoc comparisons⁷. We found the following significant differences:

- ▶ The participants' impression of the helpfulness of the explanation to trust/distrust the system were significantly higher in the explanation & solution condition compared to the no explanation condition $t(27) = -3.73$, $p = .003$, $d = 1.67$ (large effect).
- ▶ The participants' impression of the helpfulness of the explanation to trust/distrust the system was significantly higher in the explanation condition compared to the no explanation condition $t(27) = 2.49$, $p = .04$, $d = 1.13$ (large effect).

7: we used the Holm correction for multiple testing to adjust the p-values for all post-hoc tests we calculated

Emotional state To evaluate possible differences in participants' emotional states between the three conditions, we conducted a one-way MANOVA for the emotion categories happiness, anger, anxiety, and relaxation. Here we found no significant statistical difference, Wilks' Lambda = 0.84, $F(8, 46) = 0.50$, $p = .84$.

27.2.6 Discussion

Main Findings

Based on the analyses of the trust scores, we had to reject all three hypotheses. While all three conditions yielded significant decreases in reported trust based on the scales, providing explanations for the failure, with or without suggested solutions, showed no significant difference in trust, suggesting no trust-repairing effect.

Explanations alone are insufficient to recover trust after robot failures While the ESS trust score showed that participants found the given explanations helpful in deciding whether to trust or distrust the robot, this subjective impression of the participants was not reflected in their trust ratings during the VR task. Nevertheless, the ESS trust score can be seen as a first indicator that explanations might support trust recovery in HRI but that an explanation alone is not enough to recover trust after a robot failure, even when participants retrospectively rate the explanation as helpful. Despite the effect of the helpfulness of the explanations to trust or distrust the robot, this trust can not be assumed to be transferable to trust in the robot itself, especially as scales for trust in automation and human-robot trust are not interchangeable (Kessler et al., 2017).

The effectiveness of explanations depends on various aspects. Researchers like Gerlings et al. (2021) state that explanations have to fit different stakeholders and not to "the user" in general. We extend this view by saying that it is important to differentiate between the perception of an explanation given in an actual HRI situation and the rating of an explanation afterwards. Our work contributes to the operationalisation of the taxonomy of interpretability proposed by Doshi-Velez and Kim (2017). Here the authors state that the evaluation of explanations should not be done by using only proxy tasks (i.e., studies without humans) but also include users by conducting human-grounded evaluations (for simple

tasks) as we did in our research. The next step in their taxonomy is to use the insights from these simple-task experiments to conduct application-grounded evaluations using real-world tasks with domain experts. Our results, therefore, build a baseline for real-world applications. Another essential variable is the *scenario of the task*. Contrary to our results, Nikolaidis et al. (2018) found out that in their study (i.e., a physical human-robot collaboration task), explanations greatly increased human trust to take robot's suggestions. Another important variable is the *emotional presentation* of the explanation. The affect of how an explanation is presented to the user plays a role in the effectiveness of the explanation (Klein et al., 2002; Picard & Klein, 2002). Affective feedback given by a robot leads to a more positive user impression (Hastie et al., 2016; Leite et al., 2012). The work of Robinette et al. (2015) proposes that the apology of a robot after an failure increases trust in the user.

To make the explanations for HRI more effective and improve robot trustworthiness, the recommendations of Kunkel et al. (2019) and Weld and Bansal (2019), among others, should be considered for further studies. Kunkel et al. (2019) point out that users prefer richer explanations. In addition, Weld and Bansal (2019) recommends interactive explanations. Here, the robot could provide answers to follow-up questions and actions (e.g., giving more details, changing the vocabulary, attempting to correct the failure), leading to a more social process of explanation.

Include variables such as emotions and self-efficacy to get a complete view of explanations' impact The explanations in our study did not increase participants' self-efficacy, meaning they did not feel more confident interacting with the robot in the future. In addition, the participants' emotional states in the three conditions did not differ. As Mertes et al. (2022) (see Chapter 29 on page 210 - Pneumonia Study) stated, it is important to measure the emotional state and the self-efficacy of users during human-computer interactions as they are relevant to get a complete view of the impact of XAI. They found that successful explanations (i.e. helping the user to perform better in a task and to understand the AI better) led to more positive and less negative emotions and increased self-efficacy and trust. Our study showed that participants were not emotionally affected by the explanations, and neither did the explanations change users' self-efficacy. This is in line with the fact that the explanations did not increase trust in the robot after a failure and indicates that there could be a connection between users' emotional state and their trust in robots.

For future studies, it would be valuable to explicitly ask participants about how their perception of the system communication affects their perception of the robot. In addition, investigating whether there is a separation between the robot and its operating system and communications in the participant's mental model could gain deeper insights into how users perceive the given explanations of a robot. For example, since participants showed higher trust toward the explanations relative to the robot, they may consider the robot and the communication system as separate entities.

Limitations

We conducted a VR-based instead of a real-world HRI task. This, in fact, likewise represents a limitation of the current work, but as Petrak et al. (2019) stated, VR can be a helpful tool for prototyping scenarios where humans and robots interact. We are convinced that our setup used and the associated results may prove useful in designing real-life interaction HRI studies and might be developed further and in more detail in future work.

The results of our study may have been affected by the participants' understanding of the cooperative task. Some participants seemed to need help with the task, as they would often move a bottle matching the shape of the bottle moved by the robot rather than following the instructions and sorting bottles according to the indicators on the table. The difficulty understanding the task may affect the participant's perception of the robot's failure and the explanation. If the participants need help understanding the task, when told that the team failed, they may not think to inspect the robot's work and recognize its failure, which can affect their perception of the explanations. Lastly, having the participants perform tasks simultaneously with the robot may affect how attentive they can be toward the robot and whether they can critically inspect the robot's work during the task. In future experiments, the instructions should be more precise, or the bottles should be distinguishable by more factors than their shapes while still indicating which should be moved by the robot or the participant. In addition, future studies could include physiological measures (1) as emotional indicators (see Balters and Steinert, 2017 for an overview) and (2) for a more reliable measurement of trust (e.g., eye tracking Lu and Sarter, 2019).

27.2.7 Conclusion

We set out to investigate whether system explanations as a trust-repairing action after a robot makes a failure in a cooperative task are helpful. In our conducted pilot study, we found that end-users preferred less technical explanations with a greater emphasis on how to solve the failure. Using a VR testing environment for our main study, we evaluated three levels of explanations after the robot made a failure in executing a shared objective (i.e., sorting a set of bottles by shape) in cooperation with our participants. After comparing the conditions (no explanation, explanation of robot failure, and explanation of failure with solution-oriented details) with 30 participants, we found no significant effects regarding their trust toward the robot. While participants found the explanations helpful to trust or distrust the system, we can not assume this trust to be transferable to the robot. Future studies should consider the participants' understanding of the shared task with the robot, ensuring that they recognize the nature of the robot's failure and gain the most from the explanations. In addition, special consideration should be put into investigating the participants' mental model, emotional state, and self-efficacy when interacting with a robot supported by an explanation system to understand which construct the trust is placed in.

27.3 Explainable AI in Human-Agent Collaboration - Conversational AI Study

27.3.1 Highlights

- ▶ The type of explanation (personal vs impersonal communication style) has no impact on users' trust in a collaboration task with an AI dialog partner.
- ▶ Users can have competing mental models (i.e., cognitive-reasoning and social-emotional) that lead to overestimating the AI system's capabilities. At the same time, the system's limitations can be described correctly by users.
- ▶ User attributes (i.e., age, gender, previous experience) influence the interaction with the AI-dialog partner.
- ▶ Users are responsible for solving miscommunications. For this, they are using different strategies.

27.3.2 Introduction

Human-AI collaboration, a long-standing goal in AI, refers to a partnership where a human and an AI work together towards a shared goal. A collaborative dialog allows human-AI teams to communicate and leverage strengths from both partners. To design collaborative dialog systems, it is essential to understand what mental models users form about their AI dialog partners. However, how users perceive these systems still needs to be fully understood. In this study, we designed a novel, collaborative, communication-based puzzle game and explanatory dialog system. We created a public corpus from 117 conversations and post-surveys and used this to analyse what mental models users formed. Key takeaways include: Even when users were not engaged in the game, they perceived the AI-dialog partner as intelligent and likeable, implying they saw it as a partner separate from the game. This was further supported by users often overestimating the system's abilities and projecting human-like attributes, which led to miscommunication. Creating shared mental models between users and AI systems is vital to achieving successful dialogs. Our insights on mental models and miscommunication, the game, and our corpus provide valuable tools for designing collaborative dialog systems.

The main contributions of this work are

- ▶ We provide a novel combination of two data sources: (A) Self-reports (quantitative & qualitative data) and (B) behavioural data from the dialogs to gain new insights into users' mental models about an interactive AI dialog partner.
- ▶ We demonstrate that even when users were not engaged in the task, they perceived the AI dialog partner as intelligent and likeable, implying they saw it as a partner separate from the game.
- ▶ We show users were correctly able to identify the system's abilities. Despite this, they projected human attributes and motivations onto the system, leading to miscommunications
- ▶ We find user attributes, such as age and previous experience, influenced how they interact with the system.

- We show that despite users having a sound mental model of the dialog system and employing a variety of strategies to resolve miscommunications, they were largely unable to meaningfully resolve them, implying a need for shared mental models between users and AI systems.

27.3.3 Hypotheses

It is still being determined how users perceive collaborative dialog systems and what mental models they form about them. Concretely, we then ask the following research questions:

- **RQ1: Explanation Type**
 - **RQ1a:** Does the type of explanation (i.e., personal or impersonal communication) impact participants' trust in the AI dialog partner?
 - **RQ1b:** Does the type of explanation (i.e., personal or impersonal communication) impact participants' explanation satisfaction?
- **RQ2: Mental Models**
 - **RQ2a:** How do participants perceive an AI dialog partner in a collaborative setting (e.g., trust, game engagement)?
 - **RQ2b:** What mental models do participants form about an interactive AI dialog partner in a collaborative game?
- **RQ3: Miscommunication**
 - **RQ3a:** Do participants' attributes (e.g., age, gender, previous game experience) impact their ability to collaborate successfully with an AI dialog partner?
 - **RQ3b:** Which strategies do participants use to resolve miscommunication situations?

First, we created a novel collaborative puzzle game and explanatory dialog system to answer these questions. We then conducted an experiment with 129 participants, collected logs of their conversations and survey responses of their impressions of their dialog partner. We used these to analyze what mental models they formed, how they navigated miscommunications, and the implications these models have for dialog system design.

27.3.4 Pilot Studies

First Pilot Study

The first (Wizard-of-Oz) study was intended to verify the game concept and study design were comprehensible to users. Participants completed an introduction and a tutorial and then started the main game. The dialog system was played by an experimenter using pre-defined dialog fragments from the system. After participants played the game, they completed an online post-questionnaire⁸.

Nine participants (three female, six male) between the ages of 25 and 34 took part in the study. In general, the game was well-accepted. Evaluation

8: participants were in audio contact with the experimenter throughout, in case they had any problems or questions

of the single game modules showed that the button array module relied on users to remember their previous actions, and the switches module was the most popular. Based on the free-form feedback, we exaggerated differences between similar-looking elements for the main study. In general, explanations given by the dialog system were rated above average. However, some participants also worried about their length (e.g., “If time were short, the explanations would have wasted some time.”). Based on this feedback, explanations were made more compact for the final dialog system.

Second Pilot Study

The second study’s goal was to verify that the dialog system and user interface worked correctly and that the changes from the first pilot study had been successfully implemented. This pilot was conducted online with 20 participants and was performed iteratively in batches of 5 to 10 using the Amazon Mechanical Turk platform. Each participant was required to complete the game tutorial, play the game with the dialog system, and fill in a survey. On average, the experience took 30 minutes and participants were compensated \$4. Based on the log files collected, we were able to identify and improve weaknesses in our natural language understanding and the user interface.

27.3.5 Main Study

For the main study, 129 participants were recruited from Amazon Mechanical Turk, with the same requirements and payment structure as for the second pilot study.

Study Design

All participants started with an introduction to the game, including a cover story:

It looks like you’ve gotten yourself into a bit of trouble!

All you wanted to do was take your shiny new space cruiser out for a test run, but now your being attacked by space pirates! And the worst thing is, in your rush to try out your new ship, you didn’t configure the onboard AI, so you have no idea how to activate your warp drive engine to get out of there. Luckily if you can describe your control panel well enough, the AI should be able to help you figure out what you need to do before your shield gives out.

Next, they would be introduced to the idea and goal of the game:

What is Space Jam?

Space Jam is a two player collaborative puzzle game. You will be playing the role of the hapless space pilot who forgot to configure his ship’s AI and your partner (a computer) will play the role of the AI system. As the pilot, you will be shown a control panel made of four puzzle modules, but you will have no instructions

to solve them. Your partner will be given an instruction manual, but will not be able to see the exact configuration of puzzle elements. You will need to work together to solve all four puzzle modules before time runs out.

Then the core part of the tutorial began: The introduction of the four individual puzzle modules and the corresponding “Try-it-out” task for each module (see Figure 27.4 on the next page for an illustration of the four modules):

► **Dials Module**

- **Description:** Dials are made up of a pointer and multiple panels. Each panel is labeled with a number. If there are multiple dials on the control panel these are all part of the same module and all dials need to be correctly positioned to complete the module. In the case of failures, the dial pointer(s) will be reset back to their original position.
- **Tutorial task:** Try it out: In this example there is only one dial and your partner has just told you to set it to the largest even number

► **Button Sequence Module**

- **Description:** Enabled buttons come in the following colors (shown below): Green, Blue, and Amber. You can tell if a button is activated if the fill color matches the border; inactive buttons have a gray fill color. In the case that your spaceship has any disabled buttons, those will be shown in all grey. In the case of failures, the button sequence will be reset to its original configuration.
- **Tutorial task:** Try it out: In this example your partner has just told you to activate the first amber button in the sequence.

► **Switches Module**

- **Description:** The switches module is made of four switches, each switch has a left and right side. You can tell which side of the switch (if any) is activated based on the fill color of the switch; gray filled switches are inactive. Switches come in the following colors (shown below): Green, Blue, and Amber. When describing a switch’s location, switches are numbered from the top down, i.e. the top switch is the first switch. In the case of failures, you will be immediately notified and the switch sequence will be reset to its original configuration.
- **Tutorial task:** Try it out: In this example your partner has told you to: (1) Activate the left side of the first switch, (2) Activate the right side of the second, (3) Activate the left side of the third and, (4) Activate the right side of the last switch.

► **Button Array Module**

- **Description:** Like the previous modules, the buttons in the button array module come in the following colors (shown below): Green, Blue, and Amber. When describing a button’s location, they are numbered from the top left corner, that is: the first button in the first column is the top left button. “First button” refers to the top button in a column, “last button” refers to the bottom button in a column. Solving button array

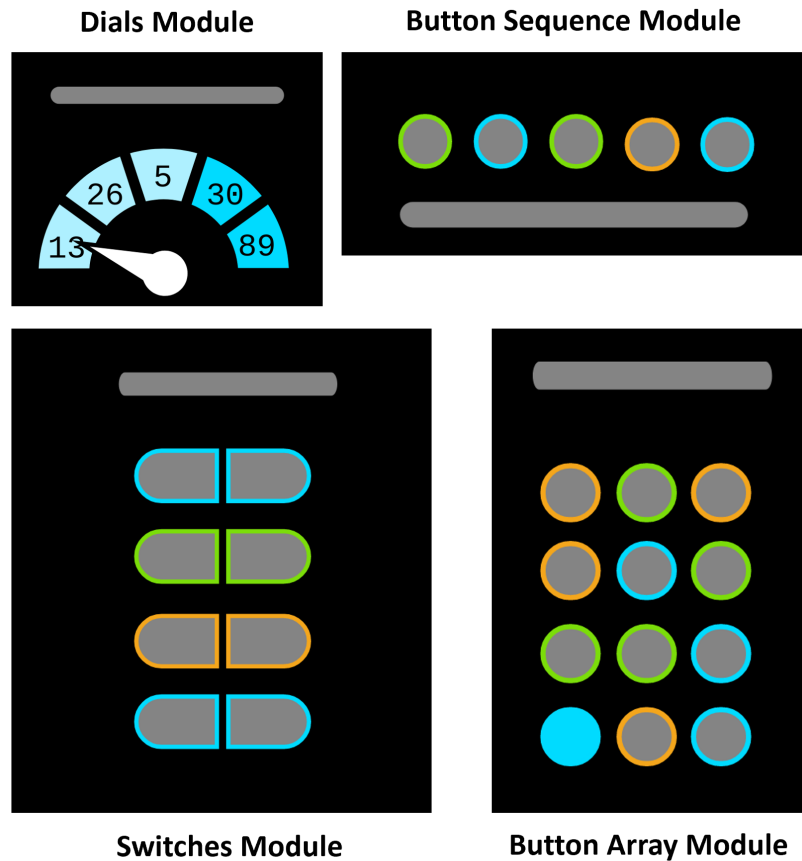


Figure 27.4: The four puzzle modules that end-users had to solve collaboratively in the Conversational AI Study

module is done in stages. Each stage starts with a button activated. If you correctly activate a button, you will move to the next stage. Some stages may require you to remember which button you pushed in an earlier stage. In the case of failures, you will be immediately notified and the button array will be reset to the start of stage.

- **Tutorial task:** Try it out: In this example your partner has told you: (1) In the first stage: activate the last button which is in the third column and has the same color as the active button, (2) In the second stage: activate the first button in the first column, (3) In the third stage: activate the last green button in the second column, (4) In the fourth stage: activate the same button that you did in the first stage.

After the successful completion of the tutorial, the main game started. Here, the participant and the AI dialog partner had to solve each of the four modules in a given time. After the game was finished, participants were redirected to the post-questionnaire.

Methodology

To understand their backgrounds, we asked participants to provide information on their age, gender, and collaborative game and video game experience.

Trust To evaluate trust in the presented dialog system, we used the Trust in Automation (TiA) questionnaire proposed by Körber (2018).

Explanation Satisfaction We used the subscale *understanding/predictability* from the TiA (Körber, 2018) to measure the participants' understanding in the two explanation style conditions.

Game Engagement To measure the participants' engagement in the collaborative game, we used the subscales *challenge*, *negative affect*, *positive affect*, *tension*, *competence* and *flow* of the Game Engagement Questionnaire (GEQ) from IJsselstein et al. (2013) (ranged from 0=not at all to 4=extremely).

System- and Self-perception To get an impression of participants' perception of the dialog partner and their emotional state, we used the Godspeed questionnaire (Bartneck et al., 2009). The Godspeed questionnaire contains 19 semantic differentials (e.g., fake-natural; 5-point Likert scale) on five subscales (i.e., anthromorphism, animacy, likeability, perceived intelligence, and perceived safety). We used four of the subscales, excluding animacy.

To qualitatively evaluate participants' impressions of their dialog partner, we used the retrospection technique proposed by Hoffman, Mueller, et al. (2018). Participants were asked four questions about their mental model of their partner after completing the game:

- ▶ “Describe your dialog partner (e.g., behavior, skills, impression)”
- ▶ “What information was important for your dialog partner?”
- ▶ “What was easy for your dialog partner?”
- ▶ “What was difficult for your dialog partner?”

Participants

One Hundred Twenty-Nine English-speaking participants were recruited. After removing those who did not meet study criteria (e.g., bad-faith participation, duplicate participation), 117 participants (36 female, 80 male, one non-binary) between the ages of 21 and 64 remained. No participants had colour vision impairments.

One Hundred Fourteen of the participants had heard about the term “AI”, but only eight had heard about the term “XAI”. After giving the participants a definition of AI and XAI to ensure they understood the terms, they were asked to rate their attitude toward each on a scale ranging from 1=extremely negative to 7=extremely positive. The attitude towards AI was rated with $M = 4.99$ ($SD = 1.19$), higher than average. The attitude towards XAI was $M = 5.88$ ($SD = 0.64$), even higher.

Participants had greater prior experience with video games ($M = 5.94$, $SD = 1.21$) compared to collaborative games ($M = 3.97$, $SD = 1.82$), measured on a 7-point Likert scale.

27.3.6 Dataset

We collected a dialog corpus investigating mental models and miscommunication in collaborative settings. The dataset, survey responses, analysis, and code are publicly available⁹.

9: GitHub repository: https://github.com/DigitalPhonetics/Our_Fault (last accessed on 29.09.2022)

Characteristics The collected corpus includes 117 dialogs with 8295 turns (system + user turns), averaging 70.9 turns per dialog, as well as user survey responses. All participant IDs were randomly assigned to ensure participant anonymity.

Annotations Each dialog consists of system and user utterances and automatically generated User Act and System Act annotations. We also provide labels for the number of puzzle modules successfully completed, the number attempted, the number of failures the user made, and the number of failures the system made per dialog. Additionally, we label the miscommunication events, per dialog, with the user ID, the start and end line, which party began the miscommunication, whether the event was an example of misunderstanding or a non-understanding, and the resolution strategy which ended the miscommunication. As a note, only one strategy is recorded per miscommunication, although other unsuccessful strategies may have been employed during the event. More information on the annotations is equipped with the dataset. In addition to the dialogs, we provide survey responses and annotations of content analysis labels from each user's free-text responses.

Example Dialog An example of a prosperous dialog snippet from our corpus is shown below. The user and system are working on the third puzzle¹⁰.

10: see Chapter 36 on page 282 for examples of miscommunication dialogs during the game

SYSTEM: Okay, the switch sequence so far should be: blue, green, amber, and the slider should be set to green. Therefore, for the second one, activate the right side of this switch.

USER: didn't work

SYSTEM: Are you sure that the color sequence is blue, green, amber, and the slider color is green?

USER: what is the slider

SYSTEM: The Slider is a vertical bar with five sections of different size and color. The color sequence from top to bottom is green, yellow, amber, orange, and red. A triangular pointer should be sitting in one of the sections. This indicates your current slider level (which is one of the five colors). Do you see it?

USER: the slider is orange

27.3.7 Results

RQ1: Explanation Type

To evaluate the impact of the explanation type (personal vs impersonal), we conducted a one-way MANOVA including the variables trust and explanation satisfaction. Here we found no significant differences between the conditions, Wilks' Lambda = 0.98, $F(2, 114) = 0.97$, $p = 0.381$. Therefore, our hypothesis regarding Trust and Explanation Satisfaction was not confirmed.

RQ2: Mental Model

Our second research questions were (RQ2a) how do participants perceive an interactive AI dialog partner? And (RQ2b) what types of mental models do participants form about it in a collaborative game? We analysed the quantitative data and the free-form feedback from our mental model questions to evaluate these research questions. For the quantitative data, we used one-sample t-tests¹¹. In addition, we prepared the qualitative data using the conventional content analysis approach proposed by Hsieh and Shannon (2005). This was performed by two independent raters, not involved in the study design or informed of the research questions.

11: we used one-sample t-tests to compare the empirical values to the mean values of the used questionnaires

Quantitative Analysis

When looking at the quantitative data, we found participants rated their dialog partner neither trustworthy nor untrustworthy ($M = 2.88$, $SD = 0.77$) and did not feel engaged to the game ($M = 1.58$, $SD = 0.52$) (see Table 27.3)¹².

For the Godspeed questionnaire (see Figure 27.5 on the following page), we found that participants perceived their dialog partner as intelligent, likeable, and safe but did not anthropomorphize it (see Table 27.3).

Interestingly, the participants rated their emotional state as calm (Godspeed item, $M = 3.37$, $SD = 1.30$) and not frustrated (GEQ item, $M = 1.86$, $SD = 1.38$) despite losing the game (see Table 27.3).

12: interpretation of the effect size d according to J. Cohen (1988) is:
 $d < .05$: small effect;
 $d = 0.5-0.8$: medium effect;
 $d > 0.8$: large effect

Measurement	$t(116)$	p	d
Trust			
GEQ	-1.68	.096	-
overall	-8.66	<.001**	0.80
frustration (item)	-1.07	.287	-
Godspeed			
intelligence	3.08	.003*	0.29
likeability	5.39	<.001**	0.50
safety	1.99	.049*	0.18
calm (item)	3.06	.003*	0.28
anthropomorphism	-4.14	<.001**	0.38

d = effect size, * $p < .05$, ** $p < .001$

Table 27.3: In the one-sample t-tests, */**significant differences were found regarding game engagement (GEQ) and all used scales of the Godspeed questionnaire

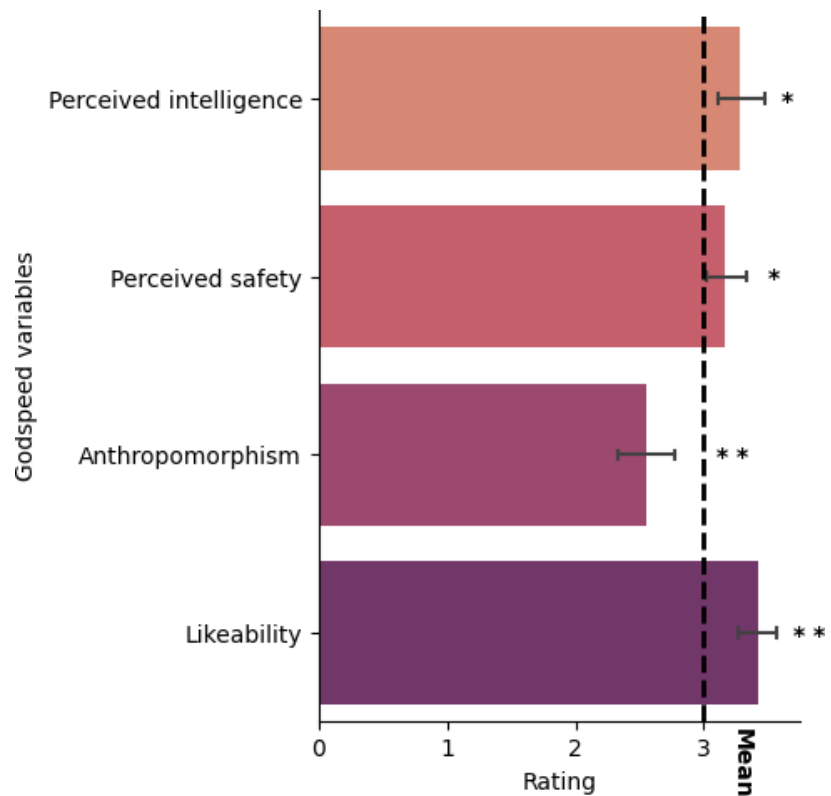


Figure 27.5: Results of the Godspeed Questionnaire (Bartneck et al., 2009) revealed that participants perceived their AI dialog partner as intelligent, likeable, and safe but not anthropomorphic. *Mean* refers to the mean value of 3, failure bars represent the 95% CI, ** $p < .001$

Qualitative Analysis

Impression of the Dialog Partner Most often, participants mentioned aspects of the dialog partner’s cognitive abilities and attributes. They also often reported the dialog partner’s impact on them and the interaction quality. For each category, we report the most common subcategories (see Table 27.4 on the next page for examples).

Additionally, several statements referred to a human-like description of the dialog partner, including positive (e.g., “polite” or “calm”) and negative attributes (e.g., “grumpy” or “frustrated”).

Important for the Dialog Partner When looking at the question “What information was important for your dialog partner?”, we found two categories: Elements and Communication.

- ▶ **Elements** Here, we found four subcategories: 1) Properties of Game Elements: e.g., colour, position, or on/off state. 2) Arrangement of Game Elements: e.g., sequence or number of elements in a puzzle 3) Other: e.g., what the player is working on. 4) Incorrect Answers: wrong or unspecific aspects (e.g., “everything” or “don’t know”).
- ▶ **Communication** Here, we also found four subcategories (three accurate mental models and one incorrect): 1) Information: important to give the dialog partner information about the modules. 2) Feedback: giving the dialog partner feedback if things worked/did not work. 3) How to give information: e.g., that things have to be said in a specific way 4) Incorrect Answers: e.g., it was important for the dialog partner to write down yes/no.

	Free-form feedback	Sub-category
A1	They seemed to use natural language and seemed to mostly know what they were doing.	competent
A2	They were very nice but, they had no idea of what they were talking about.	incompetent
A3	Whenever I said something it didn't really fit what it was looking for I suppose.	limited/ simple
B1	My dialog partner was knowledgeable and cooperative.	helpful/ cooperative
B2	I felt that the behaviour in terms of response and actions was very human like.	human-like
B3	Was an AI that was programmed to do one thing correctly. It was good at that.	machine-like
B4	They acted cold and did not try to help much.	unhelpful
B5	It was confused when I asked things that did not seem in the realm of its knowledge.	confused
C1	I was impressed with how it spoke to me.	positive
C2	Frustrating, the information they gave wasn't descriptive enough.	negative
D1	The AI kept on repeating the same questions when we were stuck.	poor dialog skills
D2	It was clear in telling me what to do to solve the puzzle.	good dialog skills

A = Dialog partners' cognitive abilities

B = Dialog partners' attributes

C = Dialog partners' impact on user

D = Dialog partners' interaction with user

Table 27.4: Main and sub-categories resulted from content analysis. For every sub-category (highlighted in bold), an example of participants' free-form feedback is given. Every example response is from a different participant

Overall, 94.02% of participants formed a correct model about what was necessary for the dialog partner. We also note, many participants learned that not only *what* but also *how* information was communicated was important.

Easy/Difficult for the Dialog Partner The last two questions we asked participants were about their mental models of what was difficult or easy for the dialog partner. Overall, participants described 130 aspects that were easy for the dialog partner and 161 that were not. Most often, simple aspects of the interaction were mentioned as easy (e.g., "giving instructions"). In contrast, most participants addressed higher cognitive abilities and complex interaction aspects as difficult for the dialog partner (e.g., "empathetic understanding of the player" or "understanding the situation").

RQ3: Miscommunication

Our third research questions were (RQ3a) Do participants' attributes impact their ability to successfully collaborate with an AI dialog partner, and (RQ3b) Which strategies do participants use to resolve miscommunication situations?

Miscommunication Events All 117 participants lost the game. When asked afterwards, “Who made failures during the game?”, 58.1% of the participants stated that they and the dialog partner made failures. 28.2% reported that only they were responsible for the failures during the game, 12.8% said it was alone the fault of the AI, and only one person stated that no one made a failure during the game. Overall, we found an average dialog length of $M = 70.9$ turns ($SD = 37.5$), from which $M = 13.6$ ($SD = 20.3$) of the turns included a miscommunication (19%).

To better understand how miscommunications were spread among participants, we divided them into three groups based on the number of turns spent on miscommunications. We found 29 users had no turns of miscommunication, 48 had few to average (1-14) turns, and 40 had greater than average (15+) turns of miscommunication. This indicates miscommunications were limited to more than just a small subset of users. Additionally, we found a significant positive relationship¹³ ($r_{sp} = .38$, $p < .001$) between dialog length and the number of turns of miscommunication, showing miscommunications led to more extended, less efficient dialogs. We also found a positive correlation between the number of times users misunderstood the system and vice versa ($r_{sp} = .69$, $p < .001$), indicating miscommunications led to further miscommunications.

13: to investigate relations between variables, Spearman’s rank correlation was used

User Attributes To investigate what caused these miscommunications, we looked at the relationships between participants’ attributes (e.g., experience with video games) and miscommunication events. We found that participants with greater experience in collaborative games were significantly more often misunderstood by the dialog system ($r_{sp} = .19$, $p = .04$). We also found a negative relationship between user age and dialog length ($r_{sp} = -.27$, $p = .003$) as well as a negative relationship between user age and the frequency the user misunderstood the system ($r_{sp} = -.21$, $p = .02$). However, the reported values only reflect correlations, not causal relationships.

Miscommunication Resolution Strategies Overall, we found 167 successful strategies belonging to nine categories (see Figure 27.6 on the facing page). The most successful user strategy was to rephrase their answer, and the most successful system strategy was to provide the user with either new or repeated information. Around 20% of dialogs were not resolved. However, it is also worth noting that several “successful” strategies (restart, agree with the system, solve without help) demonstrate that the user has given up on resolving the miscommunication and is instead focused on moving on with the task/dialog.

Our results show that the user initiated the majority of resolution strategies. However, this is not surprising, as the AI dialog partner was only equipped with simple strategies (i.e., repeating questions or information).

An example dialog for a successfully resolved miscommunication situation is shown in Figure 27.7 on the next page. The dialog system and the user tried different strategies to resolve the miscommunication. While the miscommunication was fixed at the end of the dialog snippet, there was no correct solution since the user just gave up and agreed with the system.

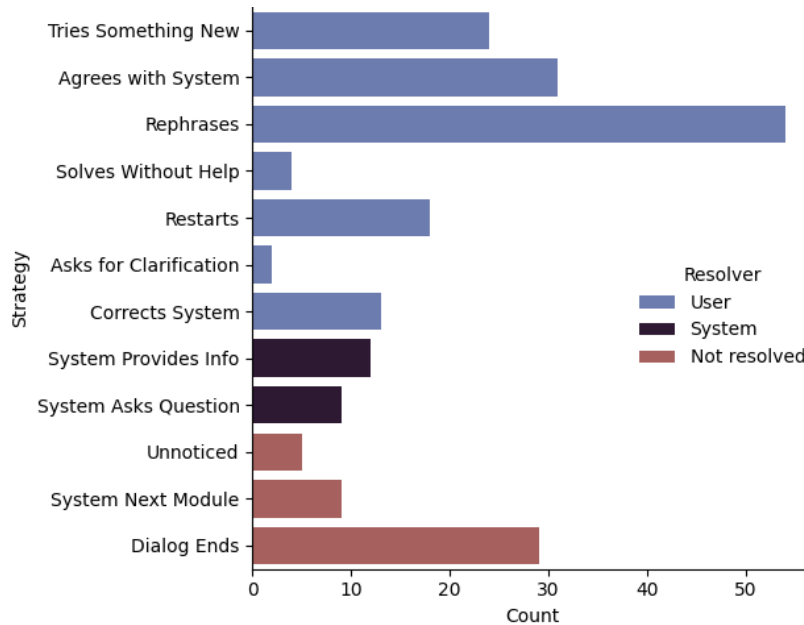


Figure 27.6: Strategies to resolve miscommunications and their frequency. *Blue* represents user strategies, *purple* represents system strategies, and *rosé* represents events that were not resolved

	DIALOG SYSTEM:	Okay, based on what you've said, the complete sequence of clickable (colorful border) buttons should be: amber, blue, amber. So you should activate the first button.
	USER:	The sequence is amber, blue, grey, amber, blue.
System provides information about itself	DIALOG SYSTEM:	Could you please slow down a bit? I can't keep track with so much info at once! Please tell me only one thing at a time.
Asks for clarification	USER:	Can you ask me the button order again?
	SYSTEM:	If I understand correctly, this is the sequence of clickable buttons: amber, blue, amber. This means you should push the first button.
Corrects system	USER:	No that is not correct
System provides information about itself	DIALOG SYSTEM:	I'm sorry, I thought that the color sequence is amber, blue, amber. Where did I go wrong?
	USER:	The third color is grey
	DIALOG SYSTEM:	Okay, based on what you've said, the complete sequence of clickable (colorful border) buttons should be: amber, blue, amber. Therefore, activate the first button.
System asked question	USER:	Okay
	DIALOG SYSTEM:	Did that work?
Rephrases	USER:	No it was the blue button
	DIALOG SYSTEM:	Alright, so the whole clickable button sequence should be amber, blue, amber. So you should activate the first button.
System asked question	USER:	Yes
	DIALOG SYSTEM:	Did that work?
Agrees with system	USER:	Yes

Figure 27.7: Example of a miscommunication situation during the study. The user misunderstood that grey buttons were disabled and, therefore, not important (highlighted in red). This misconception starts the miscommunication. The system had two strategies to handle such situations (purple): To repeat the question/task or to provide information about itself with the request to the user to reformulate the problem/question. The user tried several strategies to resolve the miscommunication (blue). In the end, the user agrees with the system without solving the task correctly

27.3.8 Discussion

We gained insights into the participants' impressions of the AI system by analysing self-reports (quantitative & qualitative) and behaviour (dialogs).

Explanation Type

With *RQ1a* and *RQ1b* we investigated the *impact of the explanation type (i.e., personal vs impersonal) on end-users trust and explanation satisfaction*. Contrary to the previous work of Kunkel et al. (2019), we did not find any significant differences in trust or explanation satisfaction between dialogs with a personal or impersonal style. There are, however, two key differences between our study design and the previous one. 1) Our scenario is more “realistic” in that there are direct consequences for taking the system’s recommendation, a time limit for accomplishing the task, and a success/failure condition for the task at the end. 2) In the work of Kunkel et al. (2019), participants received both personal and impersonal explanations and knew the personal style came from a human. In our study, participants only received one condition and knew regardless that they came from an AI. Therefore it could either be that a personal style matters less to users when they know their partner is an AI or that a personal style matters less in this type of task - either due to the time-sensitive nature or win-lose condition. Nevertheless, for future studies, we recommend a rating of personal/impersonal sentences in a pilot study to ensure that they are perceived as such by the users.

Mental Models

Quantitative Analysis Here, we wanted to know (*RQ2a*) *How do participants perceive an AI dialog partner in a collaborative setting?* Our quantitative analyses show that participants were not highly engaged by the game but perceived their AI dialog partner as intelligent and likeable. In addition, our qualitative analysis showed that the dialog partner was perceived as human-like, although this was not reflected in their quantitative rating. These results suggest that the AI is seen as a separate partner playing the game with its human counterpart rather than a part of the game itself. The difference between qualitative and quantitative results shows that qualitative analysis is an important complement to quantitative.

Qualitative Analysis Here, we asked: (*RQ2b*) *What types of mental models do participants form about an interactive AI dialog partner in a collaborative game?* Through interacting with the dialog system, participants correctly learned where it struggled; despite this, they expected more than it was capable of. This was evident when users attributed human characteristics to their dialog partner, which requires higher cognitive abilities (e.g., intentionally misleading). In particular, participants often projected human attributes/intentions into failures. For example, one user stated: “My impression is that I was directed in the wrong direction intentionally before being directed correctly at a time too late to correct the issues due to time constraints”. Similar findings were reported

by Gero et al. (2020). Here the authors found that people overestimate the AI system's abilities, particularly those who lose the game. This suggests that it is crucial when designing dialog systems to transparently convey their capabilities and limitations to users. Luger and Sellen (2016) came to a similar conclusion for speech assistants like Cortana and Siri, which suggests that our findings are transferable to other conversational agents.

Miscommunication

Miscommunication Events Although all participants lost the game, they rarely assigned fault solely to the AI in the follow-up survey, showing that users are willing to share some of the responsibility for failures in understanding. While many users (23) expressed that the dialog system could not completely understand them, they were often willing to accept partial blame. For example, one user described the system as, "Simple. As good as the person they're working with." Future work could look into what role explanation plays here as explanations are social and a part of an interactive conversation (Hilton, 1990; Miller, 2019).

User Attributes Regarding research question (RQ3a) *Do participants' attributes (e.g., age, gender, previous game experience) impact their ability to collaborate successfully with an AI dialog partner?* we have drawn attention to the fact that users' age and prior experiences can influence their interaction with the system. This was indicated by the significant negative correlation between participants' collaborative game experience and the system's misunderstandings of the user. One explanation could be that people who had prior experience with collaborative games had a mental model of how their partner should act and found it problematic that the AI partner did not match this model. Heimerl et al. (2022) (see Chapter 29 on page 210 - NOVA Study) reported that users tend to transfer their mental models into AI systems for emotion recognition and expect the system to behave as they would. However, further research would be needed to confirm this in collaborative games.

Miscommunication Resolution Strategies Our research question (2b) *addressed the strategies participants use to resolve miscommunication situations.* We found that miscommunications frequently arose in the collaborative dialogs. Although users employed various strategies to repair them, 20% of all events ended without being resolved. In a further 25% of cases, users gave up attempting to resolve the initial miscommunication, instead focusing on moving the dialog/task along. This suggests that relying solely on the user to repair miscommunication events was insufficient. Instead, having a mechanism for resolving misunderstandings or resetting back to common ground is important. With each miscommunication event lasting an average of 7.5 turns, once a misunderstanding occurred, it often precipitated a spiral, with neither party able to resolve the initial ambiguity. This limited scope of the resolution was mentioned negatively by several (12) participants, e.g., "the AI kept on repeating the same questions when we were stuck".

Our findings suggest that users can correctly identify information needed by an AI dialog partner, but this does not automatically lead to fewer miscommunications. This indicates that the user and the AI system must build an accurate mental model. For human-human interactions, Doyle and Paton (2017) stated that cooperating teams must have a “shared mental model”. A similar demand is also made in human-AI interactions (Gervits et al., 2020).

27.3.9 Future Work

To reduce miscommunication in human-AI dialogs, it could be interesting to see the impact of more sophisticated methods (e.g., pre-trained language models) and whether they would increase the AI system’s ability to understand users and lead to different user perceptions. In addition, we recommend reducing the responsibility for users. As our experiment shows, resolving miscommunication in the system we created was mainly the responsibility of the user, as the system could not recognize when the user had a poor understanding of the current problem/task. Therefore, future research should focus not only on promoting the user to develop accurate mental models of the system but also on how the AI system can develop an accurate mental model of the human counterpart to promote successful human-AI collaboration.

Regarding our correlation results, exploring a more nuanced approach to the role of user attributes (i.e., user age, experience in collaborative games) on dialogs with an AI in a collaborative setting would be interesting. In particular, it would be interesting to explore the role of user age and how pre-existing mental models (e.g., gained through prior experience with collaborative games) promote or hinder collaboration with an AI.

In addition, a comparison with other collaborative settings could gain new insights into how task-dependent or independent the user perception of an AI system is.

27.3.10 Conclusion

In this work, we designed a novel collaborative game and dialog system to collect a new corpus and investigate how users perceive an AI dialog partner and what mental models they formed.

We found that users perceived their AI dialog partner as intelligent and likeable. Almost all users could identify what information was important to share with their partner, and many also realized this information needed to be provided in a specific way. However, although users perceived the limitations of the dialog system, they tended to overestimate its abilities and attribute human characteristics to it, resulting in miscommunication.

Users were also willing to share the blame when a misunderstanding occurred, indicating they viewed their partner as sophisticated enough to be responsible for failures. Based on this, it is important for successful collaborative dialog systems that a shared (correct) mental model between the user and AI system is developed. Our game, corpus, and results

provide insight into users' mental models and miscommunications during a dialog and serve as a tool for other researchers interested in collaborative dialog.

27.4 Summary Cooperation & Collaboration Experiments

Using AI in companies and industries profitably ensures successful cooperation and collaboration between humans and machines. The impact of XAI on cooperation and collaboration success was evaluated in two conducted experiments. Here, XAI has the task of supporting humans to fulfil working tasks together with an AI entity (i.e., a robot and an AI dialog partner), especially regarding AI failures (i.e., sorting task failures and miscommunication failures). In the two conducted experiments, we found that...

- ▶ ... users tend not only to get an explanation of the AI partner but in addition recommendations for solving the failure. (*VR-Robot Study*)
- ▶ ... explanations alone are not sufficient to recover trust after robot failures. (*VR-Robot Study*)
- ▶ ... explanations are not automatically increasing users self-efficacy. (*VR-Robot Study*)
- ▶ ... the explanation style has no impact on users' trust and explanation satisfaction in a time-sensitive task. (*Conversational AI Study*)
- ▶ ... users perceive their AI partner as a separated entity from the actual task, and even when users are aware of the artificial of their partner, they give it human-like attributes. (*Conversational AI Study*)
- ▶ ... resolving failures in the collaboration is mainly part of the user, and they tend to use different strategies to solve the issues. (*Conversational AI Study*)

These experiments not only use XAI in cooperative and collaborative tasks with human counterparts but also show that explanations have shortcomings in these tasks. While the primary goal is not to deeply understand the AI system but to solve the job successfully, more than an explanation is needed to reach that goal. In addition, users are more focused on solving the task and how the AI system can help them to achieve this goal. The explanation type (i.e., personal or impersonal explanations) has no impact on end-users trust and explanation satisfaction. This task-focus of end-users indicates that more than the type of explanation, the content of the explanation (i.e., equipping explanations with concrete solutions) could help to solve failures and prevent those in the future. Also, the possibility of misinterpretation and miscommunication have to be considered in XAI design to have options for dealing with these situations if they arise.

28.1 Overview	190
28.2 Explainable AI in Human-Virtual Agent Interaction - Gloria Study	191
28.2.1 Highlights	191
28.2.2 Introduction	191
28.2.3 Research Questions . . .	191
28.2.4 Study Design	192
28.2.5 Methodology	193
28.2.6 Results	194
28.2.7 Discussion	196
28.2.8 Conclusion	200
28.3 Explainable AI with Virtual Agents in the Wild - Museum Study	200
28.3.1 Highlights	200
28.3.2 Introduction	200
28.3.3 Research Questions . . .	201
28.3.4 Study Design	202
28.3.5 Methodology	203
28.3.6 Participants	203
28.3.7 Results	204
28.3.8 Discussion	206
28.3.9 Conclusion	208
28.4 Summary Education Experiments	209

In this chapter, experiments using XAI in educational settings are presented. The chapter is based on the work published in:

- *Gloria Study*
Weitz, K., Schiller, D., Schlagowski, R., Huber, T., & André, E. (2021). “Let me explain!”: Exploring the potential of virtual agents in explainable AI interaction design. *Journal on Multimodal User Interfaces*, 15(2), 87-98. doi:[10.1007/s12193-020-00332-0](https://doi.org/10.1007/s12193-020-00332-0)*
- *Museum Study*
Weitz, K., Schlagowski, R., & André, E. (2021). Demystifying artificial intelligence for end-users: Findings from a participatory machine learning show. In S. Edelkamp, R. Möller, & E. Rueckert (Eds.), *Proceedings of KI 2021: Advances in artificial intelligence - 44th german conference on ai, virtual event, september 27 - october 1, 2021* (pp. 257-270). Springer. doi: [10.1007/978-3-030-87626-5_19](https://doi.org/10.1007/978-3-030-87626-5_19)[†]

28.1 Overview

Interactive approaches have been successfully used in education scenarios. Therefore, the inspiration for our study in the context of XAI combines the concepts of *gamification* and *virtual agents*.

Two different scenarios are presented in this block:

- **Gloria Study** First, a scenario in a classical laboratory setting was conducted. Here, participants were supposed to better understand the classification decisions of a speech recognizer with the help of the virtual agent Gloria.
- **Museum Study** Second, an experimental field study in a museum is presented. Based on the laboratory study’s promising results, we integrated Gloria into an XAI education setting at Deutsches Museum in Munich. Here, museum visitors had the opportunity to take part in a participative ML-show.

The museum study also addresses a paramount concern as Rehm (2021) stated: after evaluating XAI in a controlled lab environment to measure the impact of an XAI application in the field. This transfer helps investigate XAI’s effect in a more realistic, uncontrolled setting.

*Licensed under the [Attribution 4.0 International License \(CC BY4.0\)](https://creativecommons.org/licenses/by/4.0/)

[†]Adapted by permission from Springer Nature Customer Service Centre GmbH: Springer Nature Lecture Notes in Computer Science book series (Demystifying Artificial Intelligence for End-Users: Findings from a Participatory Machine Learning Show, Weitz, K., Schlagowski, R., & André E.), ©2021 Springer Nature Switzerland AG (2021)

28.2 Explainable AI in Human-Virtual Agent Interaction - Gloria Study

28.2.1 Highlights

- ▶ The more human-like explanations are communicated through a virtual agent, the more trustworthy the AI system is perceived.
- ▶ The perception of visual explanations (LIME) is not affected by the perception of the trustworthiness of an AI system.
- ▶ Users ask for additional information (e.g., for incorrect classified words of a keyword recognition system), interactive explanations (e.g., clicking on spectrogram areas to get more information), and comparisons (e.g., comparisons of similar sounding words).

28.2.2 Introduction

Our study evaluated which aspects of a virtual agent are relevant to support XAI visualisations. To this end, we focus on assessing the effect of different levels of anthropomorphism/human-likeness of an agent (voice, visualisation, and the content of what is said). For this evaluation, we conducted a user study in which a virtual agent presented XAI visualisations to users of a simple CNN-based speech recognition model, which classifies audio keywords based on visual representations of the audio signal (i.e., spectrograms). For the study, we split the participants into three groups, which interacted with different versions of the same virtual agent (i.e., text, voice, and visual presence) and a baseline group without a virtual agent.

Overall, our study contains three contributions:

- ▶ We present a novel XAI interaction design where we employ a virtual agent to present XAI visualisations for a simple CNN-based speech recognition model which classifies audio keywords.
- ▶ We conducted a user study to empirically verify the impact of the human-likeness of a virtual agent on the helpfulness of XAI visualisations and perceived trust in the system.
- ▶ Based on the results of this study, we are presenting suggestions to improve the integration of virtual agents in XAI interaction designs.

28.2.3 Research Questions

We examined the following research questions:

1. Does the usage of a virtual agent positively impact the perceived *trustworthiness* of a CNN?
2. Which of the *three modalities* of a virtual agent that we tested (pure information in the form of text, voice, and visual presence) are important for impacting the perceived trustworthiness of a CNN?
3. How are the presented *XAI visualisations perceived* by users?
4. How does the use of a *virtual agent affect the perception* of the presented XAI visualisations?

To answer the first and second research questions, we formulated a directional hypothesis that is evaluated within the scope of a contrast analysis. To calculate the effect size, we used the recommendations for contrast analyses from Perugini et al. (2018). For our hypothesis, we assume a linear trend, which means that the trust increases depending on the virtual agent group, where the baseline group without an agent has the lowest trust score, followed by the text agent group, the voice agent group, and the virtual agent group with the highest scores in trust.

The third and fourth research questions will be evaluated qualitatively by analysing the free-form feedback as well as quantitatively by performing an ANOVA to determine the impact of the different virtual agent modalities on the rating of XAI visualisations of the participants.

28.2.4 Study Design

To investigate the effect of agents in combination with XAI visualisations, we conducted a user study with 60 participants. None of them had a visual impairment. Each participant was given the same ten prescribed English keywords (i.e., dog, four, happy, core, on, right, eleven, two, seven, cat) to speak to our speech recognition system. Only eight of those keywords were part of the training data, whereas the remaining two words (i.e., core and eleven) were unknown to the classification system and would therefore be wrongly classified for sure. The intention was to verify that the generated explanations help the user understand correct and incorrect predictions. To reduce statistical deviations in the prediction model and the explanation framework, we chose keywords that we found reliably produce comprehensible explanations in advance. Before the test, the supervisor introduced the simple graphical user interface (GUI) to the participants. A textual cover story provided detailed instructions on how to read the systems' explanations and spectrograms. Then, every participant interacted with the GUI and spoke a predefined and fixed sequence of the ten keywords into a microphone. After each recording, the audio data was classified by the model, and an XAI visualisation for this classification was displayed together with the predicted label and the prediction accuracy of the speech recognition system. For wrong classifications, the XAI visualisations for the three predictions with the highest probability were presented. Before continuing to the next keyword, the participants rated the helpfulness ('not helpful', 'helpful', and 'don't know') of the XAI visualisation in a questionnaire. To examine the influence of the human likeness of a virtual agent on the XAI interaction design, some participants received information from a virtual agent in addition to the XAI visualisations. To this end, we split the 60 participants evenly into four test groups of 15: text agent group (only textual information), voice agent group (only information via voice), virtual embodied agent group (visual presence and voice), and a no agent group (see Table 28.1 on the next page).

The no-agent group received only the XAI visualisations without further commentary. The other three groups received additional information in varying modalities from a virtual agent named Gloria (see Figure 21.4 on page 135). The information given by the agent was selected dynamically from a set of phrases that were designed by our team in advance. These phrases were designed to communicate the following information:

Characteristic	Agent			No Agent
	Text	Voice	Virtual	
<i>n</i>	15	15	15	15
Age				
<i>M</i>	25.7	25.0	28.2	27.27
<i>SD</i>	3.99	5.6	8.6	5.19
Gender				
male	12	11	10	12
female	3	4	5	3
Experience				
Voice assistants	11	13	10	8
Audio processing	5	4	7	5
Virtual agents	5	4	6	6

Table 28.1: Demographic information of the participants, distributed to the four experimental groups

- **Acknowledgement of user inputs**, e.g., "Ok the system got that!"
- **Comments on the prediction accuracy** of the neural network, e.g., "The system was pretty sure you said seven!"
- **Comments on important phonemes** within the output of the XAI framework, e.g., "Phoneme number two was found to have a particularly positive effect towards the prediction."

The text agent group received only the textual output of Gloria's comments in a separate GUI. The voice agent group, in contrast, received the same information via text-to-speech provided by Amazon Polly ¹. The third group saw, in addition to the speech output, the virtual presence of a 3D-character designed by the Charamel GmbH ², which lip-synced the phrases and performed body gestures while communicating.

1: Amazon Polly: <https://aws.amazon.com/de/polly/> (last accessed on 29.09.2022)

2: link to the Charamel GmbH website: <https://vuppetmaster.de/> (last accessed on 29.09.2022)

28.2.5 Methodology

At the beginning of the experiment, participants were asked to provide demographic information (e.g., age, gender) and prior experience with voice assistants, audio processing, and virtual agent. After the experiment, the following items were asked:

Trust All participants rated their impression of and their trust in the system and answered the Trust in Automation (TiA) questionnaire (Jian et al., 2000).

AI System Rating All participants rated their impression of the AI system using three items (i.e., "I found the system trustworthy", "I found the system comprehensible", and "I would use the system") on a 7-point Likert scale (1=disagree, 7=agree).

Agent Rating Furthermore, the user's impressions of Gloria had queried when the participant was part of one of the virtual agent groups. The participants rated how they perceived Gloria in terms of her helpfulness (i.e., "The information Gloria gave me helped me to understand the decisions of the system"), comprehensibility (i.e., "Gloria's answers are understandable"), trustworthiness (i.e., "Gloria is trustworthy"), interaction (i.e., "I would interact with Gloria again"), and likeability (i.e., "I liked Gloria").

Participants of the text agent group also were asked to assess how often they had read the text information of Gloria on a 7-point Likert scale (1=never, 7=always). In free-form feedback, participants could state what they liked and disliked about Gloria.

XAI Rating At the end of the questionnaire, we asked participants “Where the given explanations sufficient?” on a 7-point Likert scale (1=disagree, 7=agree). In addition, they could give insights into which additional or other explanations would be helpful for them in free-form feedback.

28.2.6 Results

In this section, we describe the results of our study, starting with a comparison of trust values between the different test groups. To calculate the required sample size for the test-group comparison, we performed an a-priori power analysis. With the desired power of 0.80, an alpha value of 0.05 and an effect size of 0.45 (based on the large effect size resulted in Weitz et al. (2019)), we calculated a required sample size of 60, which would result in an expected power of 0.82. After evaluating the results, the actual effect size of 0.42 showed that an actual power of 0.75 was achieved. In addition to the group comparison, we report the evaluation of our virtual agent Gloria, followed by the ratings and the feedback for the XAI visualisations.

Test-Group Comparison on Trust

To answer our first and second research questions, we evaluated the general trust value by examining the data from the TiA questionnaire using a contrast analysis, depending on the hypothesis stated in section 28.2.3.

The results of our contrast analysis showed a linear trend $R^2 = .16$, $F(3,56) = 3.44$, $p = .023$, indicating that as the human-likeness of the agent increases, general trust increased proportionately. The planned contrast revealed that the human-likeness significantly increased the trust in the text agent ($M = 4.89$, $SD = 0.95$), voice agent ($M = 5.12$, $SD = 0.79$), and virtual embodied agent group ($M = 5.42$, $SD = 0.69$), compared to the no agent group ($M = 4.48$, $SD = 0.86$), $t(56) = 3.19$, $p = .001$, $f = 0.42$ (medium effect)³.

These findings support our hypothesis about a linear trend of the observed user trust regarding the chosen modalities, rising from no agent group over text and speech groups to the virtual embodied agent group.

Agent Evaluation

Second, we analysed how the agent Gloria was perceived by the participants in the three groups with the agent (text agent, voice agent, and virtual embodied agent). The evaluation of the agent Gloria covered the following areas: sympathy, repeated interaction, trustworthiness, comprehensibility of her statements, and helpfulness in understanding the

3: for calculating the effect size, we used the recommendations for contrast analyses from Perugini et al. (2018)

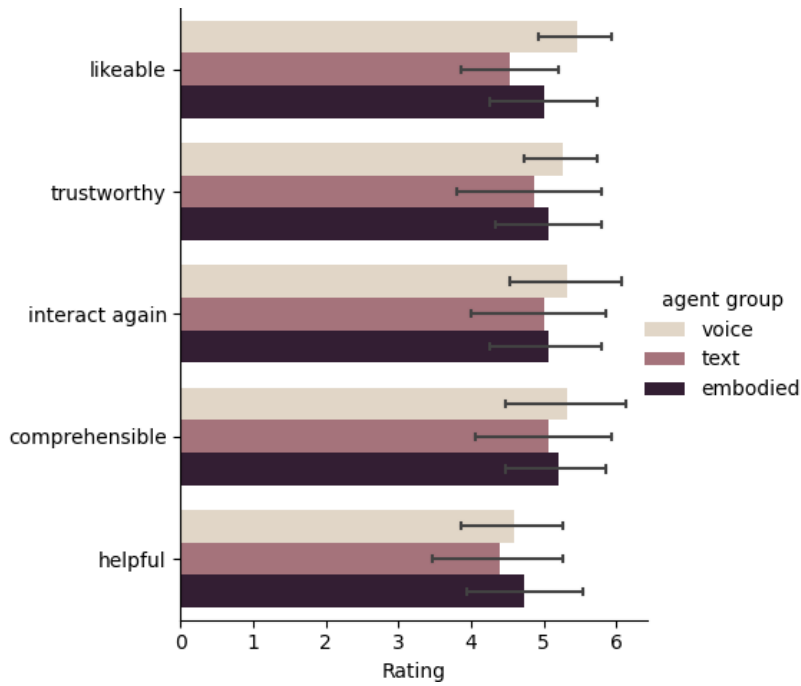


Figure 28.1: Results of the evaluation of five different aspects of the virtual agent Gloria. The rating was scaled between 1=disagree to 7=fully agree. Error bars represent a 95% CI

system's decision (see Figure 28.1). Participants evaluated each area on a 7-point Likert scale (1=disagree, 7=fully agree). For each item, Gloria received the lowest average rating among the participants of the text agent group. For being comprehensible, trustworthy, and likeable, Gloria received the highest average ratings from the voice agent group. Participants in the voice agent group also often wanted to interact with Gloria again. The virtual embodied agent group gave the highest rating for Gloria being helpful.

As a result of the evaluation of the open questions, two areas were found to be assessed positively by the participants:

- **Appearance of the virtual agent:** Facial expressions, voice, and gestures were emphasised as appealing.
- **Interactions with the virtual agent:** The participants indicated that they found verbalisation of the visualisation (e.g., the reference to relevant phonemes) supportive.

Participants within the embodied agent group mentioned that body gestures of Gloria (e.g., pointing on the spectrogram) were perceived as helpful in drawing attention to the XAI visualisation.

Evaluation of Explanations

To answer the third and fourth of our research questions, the participants gave feedback at the end of the study as to whether the given XAI visualisations were sufficient and which aspects or further explanations they would find helpful. The ANOVA reveals that the difference between the four groups was not significant, $F(1, 58) = 0.47, p = .495$, which means the ratings of the LIME visualisations do not differ between the four groups. Figure 28.2 on page 197 displays the participants' ratings on whether the given XAI visualisations were sufficient. Additionally, the average ratings

Table 28.2: Evaluation of the LIME Explanations. Answers from participants to the question which further explanations they would have found helpful. Table adapted from Weitz, Schiller, et al. (2021)

Type of information	Example feedback of participants
Additional information	<p>“Detailed answers for wrong words”</p> <p>“A verbal explanation of why some sounds were not understood”</p> <p>“Explanations for the individual case, if something is not recognized and what exactly the problem was.”</p> <p>“To tell me which phoneme had a very beneficial effect on the prediction, this could be used more.”</p> <p>“How does the system work in the background?”</p>
Comparisons	<p>“Comparisons of similar sounding words”</p> <p>“In case of incorrect predictions, additional windows with analysis of the correct label.”</p> <p>“More detailed explanation of what should be heard and what was actually heard (in the diagram).”</p> <p>“In case of wrong classification also visualisation of the actual class would be helpful.”</p> <p>“It is not clear what the word would look like if it were spoken perfectly.”</p>

of each group did not reach values above 5 (7-point Likert scale). This shows that there is still room for improvement within the XAI methods used in our study.

Many participants stated that they would have found additional and comparative information helpful (see some examples in Table 28.2). Also, the feedback analysis suggests that participants would have liked to see more interaction with the virtual agent and with the XAI visualisations (e.g., clicking on superpixels or a label to get more detailed information).

28.2.7 Discussion

The primary goal of our user study was to examine whether a user interface featuring a virtual agent positively affects the perceived trustworthiness of a CNN-based classification model for end-user. Here, we investigated whether the modalities (pure information in the form of text, voice, or visual presence) that were chosen for the communication of the classifier’s prediction results and their XAI visualisations significantly impacted the perceived trustworthiness. Furthermore, we examined the overall perceived quality of the generated XAI visualisations.

First, we discuss our findings regarding perceived user trust. Second, we discuss our findings regarding the effects of our virtual agents on user perception of the XAI visualisations.

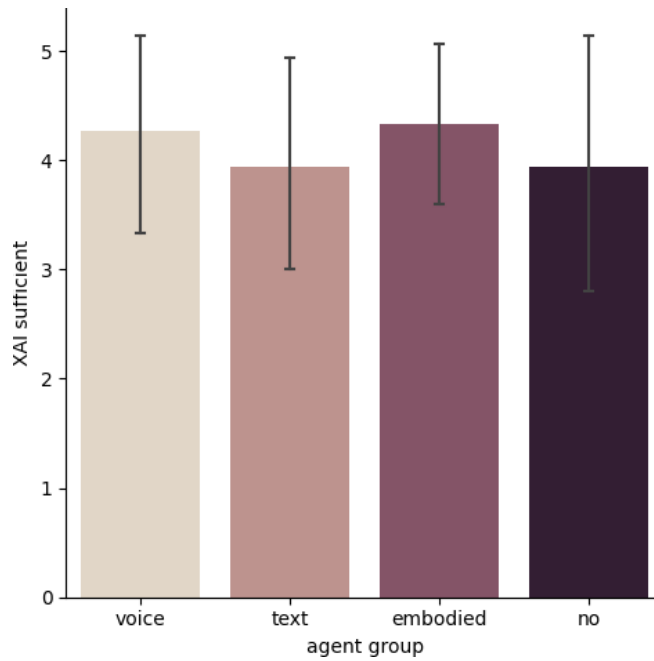


Figure 28.2: No significant differences were found in the participants' ratings on whether the displayed XAI visualisations were sufficient. The rating was scaled between 1=disagree to 7=fully agree. Error bars represent the 95% CI

Agent-User Interface Design and Perceived Trust

Examining the results of our study, we were able to support our hypotheses that end-users trust in a CNN-based classification model benefited from the

- ▶ additional text output given by a virtual agent.
- ▶ speech output provided by the virtual agent compared to text output.
- ▶ visual presence of a virtual agent performing additional lip synchronisation and body gestures compared to raw speech output.

Our results contrast the study by van Mulken et al. (1999), in which no significant increase in trustworthiness through the personification of user interfaces could be determined. They argued that the insufficient quality of virtual agents might have caused this then. This suggestion provides a possible explanation for our deviating result since technological advancements enabled us to employ a more lifelike and realistic virtual agent in our study. This is reflected in the ratings of our agent, which are all well above average in the voice agent and virtual embodied agent group (see Figure 28.1).

Our study examined the relationship between the *human-likeness* of a virtual agent and how this influences perceived user trust. The overall impression from our results is that the more human-like XAI interactions appear, the more the users tend to trust the classification model whose predictions are explained. As virtual embodied agents offer simulated human-like behaviour, such as lip synchronisation and body language, along with speech output, their potential for trust-oriented XAI interaction design seems intuitive. Still, it was not yet verified before this study. Our study indicates that a virtual agent's design choices influence humans' trust in an AI system. This information is a crucial step in establishing appropriate trust (Lee & See, 2004) in AI systems in the

future. Knowing how a user's trust can be influenced might help increase awareness towards such methods.

However, when analysing trust, one has to be careful since trust is a complex concept that various aspects can influence. Hoff and Bashir (2015) presented a three-layered framework consisting of dispositional trust, situational trust, and learned trust. Our study focused primarily on situational trust, which is strongly dependent on the situational context. This context is further divided into external and internal factors. External factors include task difficulty (i.e., spectrograms), the type of system (i.e., text-, voice-, virtual embodied agent vs. no agent), and system complexity (i.e., CNN). Among others, internal factors include subject matter (e.g., background in signal processing) and participants' self-confidence. While influences attributable to dispositional and learned trust were not explicitly addressed in our study, these could be used in further work to make more precise statements about perceived trust.

XAI Visualisation Feedback

Besides the impact of virtual agents on the perceived trustworthiness of the CNN-based classification model, we wanted to investigate (1) how the presented XAI visualisations are perceived and rated by participants and (2) how virtual agents affect this perception of XAI visualisations. We found that

- ▶ Participants wanted *additional information*.
- ▶ Participants asked for *comparative information* in visual and linguistic form.
- ▶ Participants would have preferred further *interaction* with the system (e.g., to ask questions).

As the ratings of the visual explanations were low (average around four on a 7-point Likert scale), there is still a high potential for improvement regarding the visual explanations we used in our experiment. A cause for this may be the complexity of the visual explanations, as they require some basic understanding of spectrograms and how to read them.

From the results of our study, a tendency can be observed regarding the participants' rating of the quality of the XAI visualisations (see Figure 28.2), where the no agent and text agent group rated the XAI visualisations as less sufficient than participants in the voice agent and virtual embodied agent group. This result reflects the findings on users' trust towards the system discussed in the previous subsection. A possible cause for this might be a cognitive bias such as the halo effect (Thorndike, 1920). The halo effect states that a positive impression of a person about an object in one area positively influences their opinion in other areas. In our study, the perceived trustworthiness of the CNN-based classification model could have positively influenced the participants' ratings towards the XAI visualisations. The aforementioned observation indicates that cognitive biases may occur during interaction with XAI systems. Whether and to what extent cognitive biases influence the perception of XAI should therefore be the focus of further studies.

Results from our free-form feedback showed that participants wished for additional information in an AI explanation. This aligns with the social

characteristic of explanations found by Miller (2019) since it underlines the participants' need for selective information and causality within the explanation. Our simple implementation of linguistic explanations in the text, voice, and virtual embodied agent groups, which highlight the most relevant phoneme to the user, already illustrates the usefulness of this concept. This corresponds to the findings of Siebers and Schmid (2019), who suggested that adding textual explanations can redirect the user's focus towards important areas, and of Schmid (2018), who pointed out that additional textual explanations enable the inclusion of causal relations among other information. D. H. Park et al. (2018) introduced a concept to generate such explanations for a visual question-answering system by using recurrent neural networks to generate textual explanations based on an input image, a question, and visual explanations of the predicted answer. In the same way, one could use the visual explanations we implemented in this paper to generate additional linguistic explanations for the agent which correspond to the specific input.

In addition to linguistic explanations, the supplementary use of advanced body gestures could help the agent to point at certain regions of the visualisation more precisely and thus simulate a more natural behaviour. To achieve this, one could build up on the already existing body of work that addresses the topic of automatic gesture generation (Chiu & Marsella, 2011; Gatt & Paggio, 2014; Ravenet et al., 2018).

Another aspect that emerged from the evaluation of the free-form questionnaire was that the participants wanted information that was prepared so that a particularly intuitive comparison could be made. A possible cause for this might be the specific way of integrating XAI visualisations in our system, which does not show the visualisation of the correct keyword in the case of misclassification. Instead, we displayed only three visualisations corresponding to the top predictions of our classifier. In some cases, those visualisations did not contain the word that the participant spoke. Here, participants missed additional information, which would have enabled them to interpret the explanation in the correct context. This insight supports the thesis of Miller (2019), according to which people prefer to ask why one prediction was made instead of another. To enable such a comparison, an explanation design could benefit from additionally displaying example explanations of inputs that have been classified correctly.

The participants' feedback suggests that they would have preferred to interact more with the system, for example, to ask questions when they do not understand something. This insight corresponds to Miller's findings stating that explanations have social characteristics since they represent a transfer of knowledge in the context of conversations (Miller, 2019). Conversations are one of the most important ways for people to exchange and share knowledge (Garrod & Pickering, 2004) and therefore is one of the main characteristics of human-to-human explanations. This characteristic has also been investigated in human-computer interaction by Robinson et al. (2008). They found that users most often reacted to utterances of a conversational agent with queries. It would be interesting to experiment with the application of more mature conversational agent architectures in this area since those should be able to respond adequately to questions and deal with user queries. Modern neural network-based architectures like the ones proposed by Wu et al. (2018) and Vinyals and

Le (2015) are already enabling natural user adaptive conversations with a virtual agent. Combining such conversational capabilities with the textual explanation approaches, like the one by D. H. Park et al. (2018) we mentioned before, could lead to more natural interaction and improved knowledge transfer.

28.2.8 Conclusion

Within this paper, we explored the potential of virtual agents to explain the decisions of a CNN-based classification model to end-users. To this end, we conducted a user study in which we presented XAI visualisations of the decisions from a speech recognition system to the user. While the baseline group only received the XAI visualisations, three experimental groups were presented with different modalities of a virtual agent (text, voice, or virtual presence). The results of our study show a linear trend of the user's perceived trust in the used CNN-based classification model regarding the chosen modalities, rising from the no-agent group over text and speech groups up to the virtual embodied agent group. By analysing the participants' free-form feedback, we additionally found that:

- ▶ End-users want *additional information* in AI explanations.
- ▶ End-users want explanations suitable for *intuitive comparisons*.
- ▶ End-users want to *interact with the agent*, e.g., by asking questions.

Our study's results align with our initial assumption that the end-users experience could benefit from a more human-like XAI interaction design. Based on our findings, there lies vast potential in using virtual agents to achieve this design goal.

28.3 Explainable AI with Virtual Agents in the Wild - Museum Study

28.3.1 Highlights

- ▶ A participatory machine learning show is a promising concept to enable a large group of end-users to understand the abilities and limitations of AI systems.
- ▶ Comparing machine learning show participants with museum visitors, results show that participants felt significantly more competent and positive towards technology than museum visitors.
- ▶ The explanation type of a virtual agent (i.e., first person vs third person commentary) in the machine learning show has no impact on participants' perception regarding trust of the AI system, agent rating, or rating of XAI visualisation.

28.3.2 Introduction

Interactive and collaborative approaches have been successfully used in educational scenarios. However, such techniques typically require a fair amount of technical expertise for machine learning and AI. To reach everyday users of AI technologies, we propose and evaluate a new interactive

approach to help end-users better understand AI: A *participatory machine learning show*. To enable a large group of end-users to understand the abilities and limitations of AI systems, we presented a public interactive machine learning show (ML-show) in the German museum in Munich to over 2200 visitors. Within this show, participants were able to collectively train an artificial neural network for audio keyword recognition after collecting a corpus of audio samples. After about 20 minutes of training, the audience could test how well the keyword classifier performed to understand the system's limited accuracy better. After testing the network's performance, participants were shown information detailing why the system made right or wrong predictions. This information was presented with the help of the visual XAI framework LIME (Ribeiro et al., 2016) and a virtual agent, which was previously found to have a potentially positive impact on user trust (Weitz, Schiller, et al., 2021) (see previous experiment - Gloria Study). For this purpose, we created two virtual agents with different explanation types: one agent talking in a first-person perspective (e.g., "I am sure about what you said"), another one with a third-person perspective (e.g., "The neural network is sure about what you said"). After the show, participants were asked to complete a questionnaire about their impressions of AI and XAI as well as the virtual agent.

The virtual agent and the inclusion of XAI visualisations in our edutainment show were generally rated positively by participants, even though the frameworks we used were originally designed for experts. When comparing both groups, we found that participants felt significantly more competent and positive towards technology than non-participating visitors. The impact of the explanation type of a virtual agent during the ML-show does not influence the trust in the AI system, the likeability of the virtual agent, or the rating of XAI visualisations. Our findings suggest that the consideration of specific user needs, personal background, and mental models about (X)AI systems should be included in the XAI design for end-users.

In addition to these experiment-related results, our paper presents a novel, participatory approach combining virtual agents with XAI methods to introduce machine learning topics to large user groups.

28.3.3 Research Questions

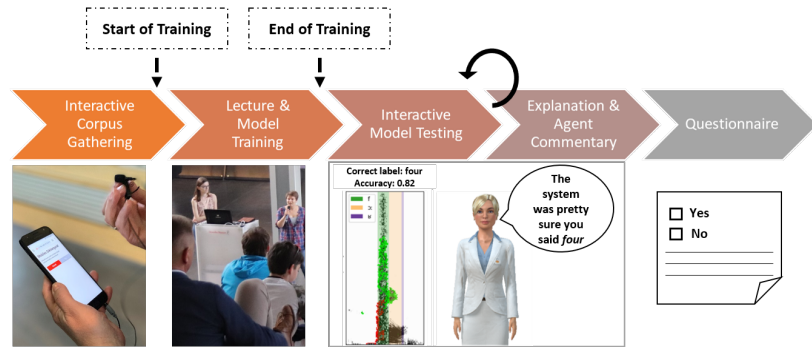
Investigating the impact of explanation type (i.e., classifier personification vs third-person commentary), we asked:

- Does the variation of the linguistic statements of an explainable virtual agent (classifier personification vs third-person commentary) influence the subjective trust in the AI system, the rating of the XAI visualizations, and the rating of the virtual agent?

By comparing the questionnaire results of ML-show attendees with baseline data, which was gathered in a separate questionnaire with non-participating museum visitors, we investigated the following research questions:

- How do end-users perceive a participatory ML-show?

Figure 28.3: Procedure during the ML-show: An interactive audio-corpus was collected and used to train a Neural Network during a lecture about Machine Learning. Afterwards, the model was tested by the participants. The virtual agent, Gloria, presented the results and XAI visualisations. At the end of the ML-show, participants had to answer a questionnaire. Figure from Weitz, Schlagowski, and André (2021)



- How do ML-show attendees differ from non-participants in terms of their attitude towards AI and self-estimated competence regarding AI?
- How do ML-show attendees differ from non-participants in terms of attitude towards technical systems in general?

28.3.4 Study Design

For six months, visitors to the museum could participate in a participatory ML-show. Here, a neural network for audio keyword speech recognition (Sainath & Parada, 2015) was trained to learn a new keyword live in the show. Simultaneously, a virtual agent named Gloria was displayed on a screen and communicated with the audience via speech output (see Chapter 21 on page 132 for details of the implementation).

The audience freely selected the new keyword trained to the neural network during a discussion at the ML-show's start (see Figure 28.3 for the study procedure). Afterwards, the visitors recorded a training dataset for the selected word by passing around the smartphone with a high-quality microphone. As soon as about 80 audio samples were recorded and transmitted to the demonstration PC, the moderator used pre-programmed software functionalities to label the samples and merge them with a subset of the speech command dataset to create the training corpus. Then, the moderator started the training process of the prediction model. While the model was trained, visitors were given a 20-minute lecture on how neural networks for speech recognition work and how the LIME framework can be used to understand the classifier decisions in this context. As soon as the lecture was finished, the moderator stopped the training. Afterwards, the network could be tested by volunteering participants multiple times by speaking known and unknown keywords into the microphone. The resulting audio samples were transmitted to the demonstration PC and passed on to the classifier. Together with prediction results, the XAI visualisations generated by the LIME framework were displayed for the audience.

In parallel to the show, the virtual agent Gloria commented on the training, communicated the classifier's prediction results, and commented on the XAI visualisations (e.g., "The most relevant phoneme for the prediction of < keyword > was...") while either personifying the classifier or commenting on the classifier's processes in the third person. For both of these situations, different versions of the virtual agent's phrases were

hard-coded in advance so that the moderator could choose which version should be presented to the visitors at the beginning of each show. We instructed the moderators to alternate between conditions A and B between the shows to equalise the number of participants for both versions.

Additionally, during three days, we gathered baseline data using a paper-based questionnaire oriented on the questions used in the Eurobarometer report (European Commission, 2017). We also recorded the affinity of museum visitors for technology using the TA-EG questionnaire (Karrer et al., 2009). We only questioned people who did not visit our participatory ML-show (for details about (X)AI attitudes of the museum visitors, see Chapter 16 on page 92).

28.3.5 Methodology

After the show, participants were asked to complete a questionnaire online or on paper. In addition to the collection of demographic information, the following questions were included:

Agent & (X)AI System Evaluation To evaluate the virtual agent Gloria, we used five items on a 7-point Likert scale (e.g., “I liked Gloria”) and free-form feedback. We collected participants’ feedback about the AI system using three items on a 7-point Likert scale (e.g., “I would use the AI system”). To gain insights into the perceived helpfulness of the XAI visualisations, we asked 1 item on a 7-point Likert scale (i.e., “Were the explanations sufficient?”) and a free-form question about which additional information would be helpful for them to understand the AI system.

Technical Affinity To measure the technical affinity of participants using the TA-EG questionnaire (Karrer et al., 2009) was queried.

Trust Subjective trust was assessed with the Trust in Automation (TiA) questionnaire (Jian et al., 2000).

Attitude towards AI At the end of the questionnaire, additional questions about the participant’s general knowledge attitude towards AI and XAI were posed (e.g., “How would you rate your knowledge of AI?” and “In general, what is your attitude towards Artificial Intelligence?”).

28.3.6 Participants

A total of 65 public participatory machine learning shows with an average of 35 participants each were held. A total of 2275 museum visitors participated in the study, of which 51 completed the subsequent questionnaire. Due to missing data in some questionnaires, 47 participants (24 male, 22 female, one non-binary) between 13 and 80 years ($M = 42.07$, $SD = 22.6$) were included in the final analyses presented in this paper (see detailed demographic information in Table 28.3 on the following page). The participants’ educational background was mixed and ranged from

Table 28.3: Descriptive information about the participants of group A (classifier personification) and group B (third person commentary)

Characteristic	Group A (first person)	Group B (third person)
<i>n</i>	31	16
Age		
<i>M</i>	43.29	41.50
<i>SD</i>	23.52	21.26
Gender		
male	13	11
female	18	4
divers	0	1
Experience		
Voice assistants	10	9
Audio processing	3	6
Virtual agents	9	4

“no degree” to “university degree”. Most participants had no previous knowledge or experience in using virtual agents, voice assistants, or audio processing. 88% of the participants stated that they had already heard of the term AI, but only 11% of them rated their AI knowledge as extensive. Most participants either had a balanced or a positive view of AI’s future impact. Most participants saw XAI as an important topic, especially for researchers, companies, and end-users. For politicians, participants rated the importance of XAI less compared to the other stakeholders.

Fifty-nine museum visitors took part in our field survey, which we used as the baseline for the comparison with the ML-show participants. Unfortunately, we had to remove the answers of one visitor due to too many unanswered questions. Therefore, for the following analyses, answers from 58 museum visitors (29 female, 29 male) between 8 and 66 years ($M = 30.3$, $SD = 16.5$) are considered. The participants’ educational background was mixed and distributed similarly as in the ML-shows.

28.3.7 Results

Results of the ML-show

Impact of Explanation Style We compared the participants’ ratings regarding (1) trust for the AI system, (2) impression of the virtual agent and (3) helpfulness of the XAI visualisations between group A (first-person commentary) and B (third-person commentary). To examine the potential effects of the virtual agent’s linguistic perspective, we conducted a one-way MANOVA. For all three dependent variables, the results were not statistically significant, Wilks’ Lambda = 0.84, $F(3, 40) = 0.27$, $p = .844$. These findings suggest that the perspective in which the virtual agent explains the decisions of the neural network had no meaningful influence on either perceived user trust or their impressions of the quality of the virtual agent or the XAI visualisations.

Agent & (X)AI Rating Participants gave the virtual agent Gloria a rating of $M = 3.9$ (7-point Likert scale). The LIME visualisations were rated with $M = 4.15$, slightly higher than the virtual agent. Investigating whether the participants would use such an AI system, the rating was

$M = 3.06$ (7-point Likert scale). In response to the free-form question about what additional information they would have liked to see in Gloria's explanations, end-users indicated that they would have liked more details (e.g., "What does Gloria calculate in the training phase?").

Correlations for ML-show Participants To examine potential connections of the educational background, gender, technical affinity, and age of the participants on questionnaire items like trust in the AI system, virtual agent impression, and the helpfulness of the XAI visualisations, we calculated spearman's product-moment correlations. We found a significant weak positive linear relationship between perceived trust in the presented AI system and educational background ($r_{sp} = .46, p < .05$), where participants with a higher educational background tend to trust the AI system more. Neither age nor gender significantly impacted subjective trust in the AI system, as we did not find any significant correlations for these variables. For the impression of the agent and the helpfulness of the XAI visualisations, we did not find correlations for the participants' age, gender, or educational background.

Comparison between Participating and Non-participating Museum Visitors

We used two one-way MANOVAs to examine if there were any significant differences compared to the non-participating museum visitors (baseline). Holm correction for multiple testing was applied.

Attitudes towards AI We conducted a MANOVA to evaluate whether there was a difference between baseline museum visitors and ML-show participants in (1) the perceived knowledge about AI as well as (2) their attitude about the impact of AI on our lives in the future and (3) in their attitude towards AI. We found no significant differences for these three variables, $F(3, 100) = 1.76, p = .16$, Pillai's Trace = 0.51.

Technical Affinity To evaluate the TA-EG questionnaire, we looked at the four subscales (excitement, competence, negativity, and positivity) using a one-way MANOVA. The MANOVA showed significant differences between the groups for the TA-EG variables, $F(4, 100) = 28.58, p < .001$, Pillai's Trace = 0.53. To find out on which subscales of the TA-EG significant differences exist, we then performed an ANOVA that revealed significant differences for the subscales competence $F(1, 103) = 23.15, p < .001$, excitement $F(1, 103) = 5.03, p < .03$, and positivity $F(1, 103) = 96.15, p < .001$.

We then used post-hoc tests to investigate the direction of these differences. For this purpose, we use t-tests or Mann-Whitney U-tests if the requirements for the t-test were not met⁴ to evaluate whether there was a difference between baseline museum visitors and ML-show participants in (1) the perceived technical competence as well as (2) their excitement towards technology and (3) their positivity towards technology. Our results show:

4: The Mann-Whitney U-test is the non-parametric equivalent of the t-test for independent samples and is used when the conditions for a parametric procedure are not met (in our case: homogeneity of variances and non-normal distribution of the data)

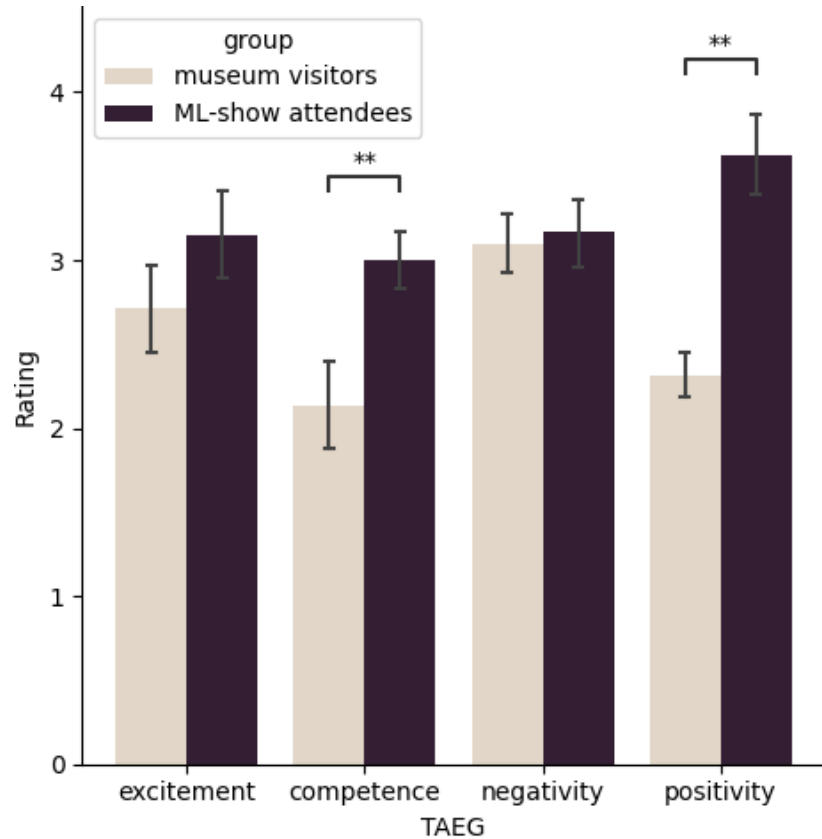


Figure 28.4: Mean TA-EG ratings by category for the ML-show participants and non-participating museum visitors. Sub-categories *competence* and *positivity* indicate significant differences between the two groups ($*p < .001$). Error bars represent the 95% CI

5: this result was no longer significant due to the alpha error correction

- **Competence:** Participants of the ML-show ($M = 3.00$, $SD = 0.63$) feel more competent about technology compared to the baseline museum visitors ($M = 2.14$, $SD = 1.08$), $U = 692$, $p < .001$.
- **Excitement:** Participants of the ML-show ($M = 3.15$, $SD = 0.94$) do not feel more excited compared to the baseline museum visitors ($M = 2.71$, $SD = 1.03$), $t(103) = -2.24$, $p = .05^5$.
- **Positivity:** Participants of the ML-show ($M = 3.62$, $SD = 0.83$) feel more positive towards technology compared to the baseline museum visitors ($M = 2.32$, $SD = 0.52$), $U = 216$, $p < .001$.

28.3.8 Discussion

Overall, end-users were receptive to XAI visualisations in our ML-show, even though visualisation methods were not specifically designed for end-users without any background knowledge of AI and XAI. Furthermore, our field study helped us to gain initial insights concerning end-users views about (X)AI and virtual agents in a participatory ML-show, which we discuss in the following.

Customisation of Virtual Agents for XAI Education Requires more Research

Our results show that the use of different representations of an AI system (i.e., classifier personification through a virtual agent vs third-person commentary about an AI system) in an XAI design does not influence the

subjective trust of end-users, their overall impression of the agent and the perceived helpfulness of XAI visualisations. These findings contradict our expectations, as in the Gloria Study, we have shown significant effects of the choice of modalities in XAI design featuring virtual agents. This discrepancy illustrates that even though evidence suggests that virtual agents can positively affect user trust in XAI applications (Weitz, Schiller, et al., 2021; Weitz et al., 2019) (see Gloria Study), it is not quite clear which factors play a role and need to be considered when designing user interfaces. For instance, Haake (2006) showed that gender stereotypes slightly influence the perception of virtual pedagogical agents. However, whether such aspects regarding the appearance of a virtual agent (e.g., female or male virtual agent) influence the subjective trust for an AI system or increase the perceived helpfulness of an XAI setting still needs to be examined.

Take Users' Attitudes and Experiences into Account

The correlation analysis of our data revealed a connection between educational background and perceived trust in our AI system. This result encourages XAI design that fits the user's educational background. As part of our study was a presentation on the basic functioning of neural networks, speech recognition and XAI, better-educated participants might have been more receptive to knowledge transfer. Thus, they might have understood the XAI visualisations better, which might have resulted in increased trust. Miller (2019) argued that explanations for AI systems have to be based on the expectations and needs of humans. In the NOVA Study (presented in Chapter 29 on page 210), we found out that more XAI information about an emotion recognition system leads not automatically to higher trust in the AI. They concluded that users tend to transfer their mental models about emotions to the AI. Therefore, having the mental model of users in mind (Rutjes et al., 2019) when personalising XAI for different stakeholders and different AI scenarios is an important step to adjust XAI to the "right amount" for individual users (Schneider & Handali, 2019). Here, trust models such as those of Sanders et al. (2011) and Hancock et al. (2011), which indicate that different components (e.g., agent characteristics, user attributes as well as situation characteristics) have an impact on trust, can be used to examine possible variables that might influence user trust in XAI scenarios.

Think About Who You Want to Reach With XAI Edutainment

The results of our study show that users who participate in an ML-show differ in aspects of technical affinity from non-participating museum visitors. Due to our study design, which did not contain a pre-study questionnaire, we cannot tell whether the differences occurred due to more technically affine museum visitors being more likely to participate in the ML-show, or whether the observed differences were a result of the ML-show itself. However, there are indications that the interaction with the AI system and the virtual agent in the ML-show could have influenced participants' technical affinity. Reich-Stiebert et al. (2019) reported in their study comparable findings. They stated that positive attitudes towards robots increased among people who had the opportunity

to be part of the prototyping process. Even though the results of our Gloria Study suggest that virtual agents can positively affect user trust in XAI applications, it needs to be clarified which factors play a role and need to be considered when designing user interfaces. According to Haake (2006), gender stereotypes are one factor that slightly influences the perception of virtual pedagogical agents. Whether the external appearance of a virtual agent (e.g., female or male virtual agent) plays a role in subjective trust for an AI system or whether they can increase the perceived helpfulness of an XAI setting still needs to be determined.

Trust and Distrust are Important Components in XAI Interaction Design

Trusting AI systems incorporating XAI and virtual agents has been previously reported in the Gloria Study. However, it demands an ethical perspective on systems that have the potential to increase user trust. In this manner, Gilpin et al. (2019) stated that XAI could not be equated with the reliability and responsibility of an AI system. Hoffman, Klein, and Mueller (2018) makes similar statements, demanding that distrust and mistrust must also be included in evaluating XAI systems to support appropriate trust. We argue that ethical XAI systems should therefore be able to (1) encourage user trust if a system performs well, (2) prevent distrust if a system performs badly, and (3) prevent overtrust if a system cannot live up to expectations.

An average prediction accuracy of about 80% after 20 minutes of training was far from perfect, so a variety of wrong classifications occurred during the show. It resulted in a demystification of AI systems. It might also have encouraged more distrust of XAI systems for users initially trusting AI systems, as they most likely used too much better prediction models in their everyday lives.

28.3.9 Conclusion

We presented a novel public participatory machine learning show where we let visitors of a museum train a neural network together in order to clarify and demystify the opportunities and limits of AI systems. During the show, we used a virtual agent and an XAI framework to provide participants with additional information about the decision-making processes of the neural network during a speech recognition task. By examining the results of a post-study questionnaire, we could deduce that the virtual agent and the inclusion of XAI visualisations in our edutainment show were generally rated positively by participants, even though the frameworks we used were originally designed for experts. We also found a correlation between trust in our AI system and the participants' educational backgrounds. Compared to non-participating museum visitors, ML-show participants felt more competent and optimistic about technology. During the discussion of our results, we pointed out possible causes and limitations of our findings. We concluded that consideration of specific user needs, personal background (e.g., education), and mental models is a promising approach for an educational XAI design for end-users.

28.4 Summary Education Experiments

AI systems that support users to learn efficiently and effectively have been based on the idea of ITS since the 1960s. Approaches like the work of Conati et al. (2021) present the combination of ITS with XAI. This idea is extended in the two shown studies in this dissertation by integrating a virtual agent into HC-XAI design. The results show that...

- ▶ ...virtual agents positively impact the end-users perception of CNN, especially in terms of trust. (*Gloria Study*)
- ▶ ... the positive impact of virtual agents is not reflected in end-users perception of the XAI visualisation, meaning that the presentation of the explanation is more important than the content of the explanations itself. (*Gloria Study*)
- ▶ ... end-users demand explanations including additional information, comparisons, and to be more interactive. (*Gloria Study*)
- ▶ ... end-users are interested in taking part in edutainment settings in public places like museums. (*Museum Study*)
- ▶ ... the explanation type (i.e., first person vs third person commentary) of a virtual agent has no impact on users' perception of users' trust in the AI system, the likeability of the virtual agent, nor the rating of XAI visualisations (LIME). (*Museum Study*)
- ▶ ... the personal background (e.g., educational background, technical affinity) of end-users should be taken into account when designing XAI in a human-centered way. (*Museum Study*)

Besides these results, the experiments present the following novelties⁶:

- ▶ Using and evaluating the impact of a virtual agent and a real CNN in an educational XAI setting in the laboratory (*Gloria Study*) and under realistic conditions in the field (*Museums Study*).
- ▶ Development of an innovative didactic education concept (i.e., a participatory ML-show) for XAI.

⁶: based on the publication dates of the papers

29.1	Overview	210
29.2	Explainable AI in Facial Emotion Classification - NOVA Study	211
29.2.1	Highlights	211
29.2.2	Introduction	211
29.2.3	Research Questions	212
29.2.4	Study Design	213
29.2.5	Methodology	213
29.2.6	Participants	214
29.2.7	Results	214
29.2.8	Discussion	221
29.2.9	Conclusion	223
29.3	Explainable AI in Pulmonary Disease Classification - Pneumonia Study	225
29.3.1	Highlights	225
29.3.2	Introduction	225
29.3.3	Hypotheses	226
29.3.4	Study Design	226
29.3.5	Methodology	227
29.3.6	Evaluation Methods	228
29.3.7	Participants	229
29.3.8	Results	230
29.3.9	Discussion	234
29.3.10	Conclusion	237
29.4	Summary Medical Decision Support Experiments	237

This chapter presents experiments using AI technology in decision support settings in healthcare. The experiments presented are based on the work published in:

- **NOVA Study**
Heimerl, A., **Weitz, K.**, Baur, T., & André, E. (2020). Unraveling ML models of emotion with NOVA: multi-level explainable AI for non-experts. *IEEE Transactions on Affective Computing* 13(3), 1155-1167. doi: [10.1109/TAFFC.2020.3043603](https://doi.org/10.1109/TAFFC.2020.3043603)*
- **Pneumonia Study**
Mertes, S., Huber, T., **Weitz, K.**, Heimerl, A., & André, E. (2022). GANterfactual-Counterfactual Explanations for Medical Non-experts Using Generative Adversarial Learning. *Frontiers in Artificial Intelligence*, 5, 825565. doi: [10.3389/frai.2022.825565](https://doi.org/10.3389/frai.2022.825565)*

29.1 Overview

Since the early days of XAI, AI-based diagnosis tools like MYCIN (Shortliffe & Buchanan, 1975; Shortliffe et al., 1975) in the 1970s gained to support medical personnel in the detection, classification, and therapy of diseases. Nowadays, DNNs are a successfully used technology for various medical-related tasks. Two experiments in the healthcare domain are presented to investigate the impact of different types and contents of XAI for medical decision making with the help of DNN. They focus on (1) facial emotion recognition, which is a relevant topic for the adequate treatment of patients and (2) pneumonia classification using X-ray images:

- **NOVA Study** The first study investigates the helpfulness of different XAI methods (LIME, confidence values, combination of both) when labelling images for facial emotion recognition. Using the software NOVA, participants had to work on a dataset and an ML-model for facial emotion recognition.
- **Pneumonia Study** The second study focuses on the impact of different XAI methods (counterfactuals, LRP, LIME) on user perception. The use case here was the detection of pneumonia in X-ray images.

29.2 Explainable AI in Facial Emotion Classification - NOVA Study

29.2.1 Highlights

- ▶ The software NOVA is an appropriate tool for end-users to interact with DNN models.
- ▶ Users create mental models about AI systems: Without XAI, they transfer their mental model of self to the AI system leading to incorrect mental models about the AI. XAI visualisations help create accurate mental models, while confidence values seem unsupportive.
- ▶ XAI does not influence the trust, self-efficacy, and cognitive workload of end-users in a facial emotion recognition setup.
- ▶ XAI perception differs from human perception: While humans focus on specific facial features when classifying emotions, DNN include larger parts of the face in its decision.

29.2.2 Introduction

In this work, we used the annotation tool NOVA for emotional behaviour analysis, which implements a workflow that interactively incorporates the ‘human in the loop’. A central aspect of NOVA is the possibility of applying semi-supervised active learning where ML techniques are used already during the annotation process by allowing pre-label data automatically (Heimerl et al., 2019). Furthermore, NOVA implements recent XAI techniques to provide users with both a confidence value of the automatically predicted annotations as well as visual explanations.

NOVA has already been evaluated in medical contexts. The work of Baur, Clausen, et al. (2020) investigates the usefulness of NOVA to annotate recordings of therapy sessions. Here, videos and audio recordings of the patient and therapist during a session for the treatment of test anxiety were recorded and annotated by eight students (i.e., psychology students) who were previously trained in using NOVA. The students indicated that NOVA was intuitive to use but that annotating longer video recordings was tiring. In another work (Terhürne et al., 2022), NOVA was used to classify the emotional expressions of patients. For this purpose, expressions of emotion on the face and posture were analysed using NOVA. The software was found to be suitable for therapists to annotate this data and then train models to recognise emotions from facial expressions.

In our study, NOVA was used in facial emotion recognition. Facial emotion recognition plays a role not only in therapy but also in medical decision support. Especially in patients who are not or no longer able to express themselves verbally (e.g., dementia patients), automated emotion recognition can help to monitor patients and determine the appropriate administration of medication (e.g., painkillers) (Hassan et al., 2019). For our study, we trained a CNN on emotion recognition on images and investigated, how XAI methods implemented in NOVA can assist end-users in terms of *trust*, *perceived self-efficacy*, *cognitive workload*, as well as creating accurate *mental models* about a CNN by conducting a user study with 53 participants. The participants had no medical background. The

results show that NOVA can easily be used by end-users and lead to high computer self-efficacy. Furthermore, the results indicate that XAI visualisations help users to create more accurate mental models about the ML system compared to the baseline condition. Nevertheless, we suggest that explanations in the field of AI have to be more focused on end-user needs, the classification task, and the model they want to explain. This work investigates the impression of ML on end-users during a Cooperative ML (CML) task. It also provides insights into whether end-users benefit from XAI information.

29.2.3 Research Questions

In our work, we propose a framework that allows end-users to employ AI techniques in their problem domain. More precisely, we introduce the NOVA tool that supports interdisciplinary researchers and end-users during the annotation process of continuous multi-modal data by incorporating ML techniques already applied during the annotation process. This way, users are enabled to interactively enhance their ML model by incrementally adding new data to the training set. At the same time, they get a better understanding of the capabilities of their model. This happens on multiple levels. First, they get a pure intuition of how well their model performs by investigating false predicted labels. They might even learn specific cases in the data when their model ‘always fails’ or when they can be sure they can ‘trust’ their model. Secondly, besides intuition, we provide XAI algorithms within the workflow that allow users to generate local post-hoc explanations on instances their model predicted. This way, we combine interactive ML techniques and XAI algorithms to involve humans in the ML process while simultaneously giving back control and transparency to users. Following the previous work of Heimerl et al. (2019), we performed a study with 53 participants to investigate how end-users can benefit from such a workflow. With this study, we want to examine the following research questions:

1. How do people with little or no ML experience rate the interaction with the NOVA software?
2. What is the impact of the XAI information presented (confidence values, LIME visualisations, both, or none) to end-users in order to develop an accurate mental model about a neural network model for facial emotion expression recognition?
3. How do end-users rate the presented information (confidence values, LIME visualisations) in terms of simplicity of understanding and support for explaining the ML model?
4. How does the relevant image information of the XAI method LIME for facial emotion expression classification differ from humans?

We investigate the first research question by descriptively evaluating the end-users feedback. For the second and third questions, we calculated comparisons between different groups. Finally, we contrast LIME visualisations with non-expert drawings of relevant areas in face images to answer the fourth question.

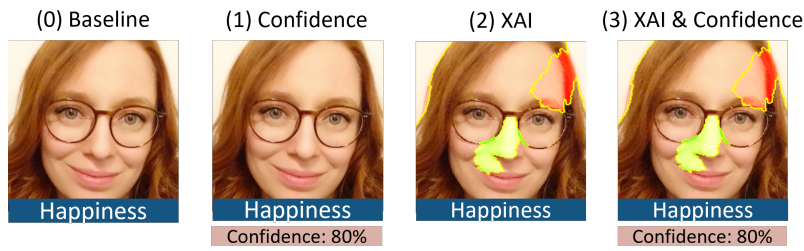


Figure 29.1: Overview of the experimental conditions of the NOVA Study. (0) Participants of the *baseline condition* received only images with facial emotion expressions with a classification result (e.g., happiness). In addition to this information, (1) participants of the *confidence value condition* received confidence value for each image, (2) participants of the *XAI visualisation condition* received images with highlighted superpixels relevant for the classification, (3) participants of the *XAI & confidence condition* got all available information

29.2.4 Study Design

We conducted a study to investigate the influence of different types of XAI information (confidence values and LIME visualisations) on task performance, computer self-efficacy, cognitive workload, and subjective trust of NOVA users with no or little ML background. The participants should help improve the model's performance by identifying as many wrongly classified images as possible given a five-minute time frame. Additionally, in five minutes, they had to find as many images as likely that the model had already been classified well. For this purpose, the participants were presented with 254 images and the corresponding classifications of a CNN in NOVA. The 254 images were equally distributed between the five classes and shown in an unsorted way. They were supposed to navigate freely through these images to get an overview of the model.

After filling out a questionnaire, the participants were shown pictures with emotional facial expressions. Then, they had to classify each facial expression and report how sure they were with their decision. After this, they had to draw which areas were necessary to classify each image.

29.2.5 Methodology

After interacting with NOVA, the participants completed a questionnaire, including the following items:

Personal Information At the beginning of the questionnaire, we asked the participant about personal information. These questions included age, gender, educational background, experience with ML in general and NOVA, and their knowledge and attitudes about AI and XAI.

Impression of NOVA After finishing the task using NOVA, we asked the participants to indicate their overall impression of NOVA. For this purpose, we used two questions, i.e., "The information NOVA provides are easy to understand", and "The information provided by NOVA helps to understand the model". These questions were rated on a 7-point Likert scale (1 = don't agree, 7 = totally agree).

Impression of XAI Methods In addition to the general impression of NOVA, the participants were asked to rate the helpfulness and explainability of the presented XAI methods. The first question was “The < *XAI visualisations* > NOVA provides are easy to understand”, and the second question was “The < *XAI visualisations* > provided by NOVA help to explain the model”. The phrases in *italic* were changed depending on the experimental condition. Again, these questions were rated on a 7-point Likert scale (1= don’t agree, 7=totally agree).

Mental Model To gain insight into the users’ mental model, we used the task reflection method, an approach recommended by Hoffman, Mueller, et al. (2018). The free-form feedback of the task reflection was combined with a 7-point Likert scale (1=unsure, 7=sure) that allows users to evaluate their confidence in their statement. In addition, we measured the mental model of self of the participants. At the end of the study, we showed all participants five images with people expressing emotions (i.e., anger, neutral, disgust, sadness, happiness) and asked them to (1) classify the emotion themselves and rate their confidence in the decision (8-point Likert scale) and (2) draw the areas on the face that were relevant for them to identify the emotion.

Trust For the assessment of the trustworthiness of the AI system, we used the TiA questionnaire (Jian et al., 2000).

Computer Self-Efficacy To measure the computer self-efficacy of the participants, we used the CSE scale (Compeau & Higgins, 1995).

Cognitive Workload We also collected data about their subjective workload using the NASA-TLX questionnaire (Hart & Staveland, 1988).

29.2.6 Participants

In total, 53 participants took part in the study (see Table 29.1 on the facing page for detailed demographic information). We conducted a power analysis to encounter the required study size. All participants stated that they had heard the term Artificial Intelligence before. On average, they rated their impression of AI with 4.77 ($SD = 0.91$) positive (range from 1 = extremely negative, 7 = extremely positive). In contrast, only two participants stated that they had heard about XAI. After giving the participants the information about what the goal of XAI is, on average, participants saw XAI as most important for end-users ($M = 5.60$, $SD = 1.45$), followed by researchers ($M = 5.47$, $SD = 1.35$), companies ($M = 5.40$, $SD = 1.39$), and politicians ($M = 5.17$, $SD = 1.61$). Most participants had no experience with ML, and none had used the software NOVA before.

29.2.7 Results

In the following, the results of the study will be presented. Starting with the software NOVA evaluation, followed by comparing the experimental

Characteristic	Conditions				Total
	0	1	2	3	
<i>n</i>	13	13	14	13	53
Age					
<i>M</i>	22.46	22.85	22.36	22.23	22.47
<i>SD</i>	2.47	3.02	2.59	2.89	2.68
Gender					
male	2	7	3	5	17
female	11	6	11	8	36
Experience					
NOVA	0	0	0	0	0
ML	1	4	1	2	8

Table 29.1: Demographic information of the participants. 0=Baseline condition; 1=Confidence values condition; 2=LIME condition; 3=LIME and confidence values condition

groups' results. Afterwards, the LIME visualisations are compared with the human areas of interest.

End-Users Impression of NOVA

All 53 participants interacted with the NOVA software for the first time. The one-sample t-tests revealed that the overall impression of NOVA was quite high. Participants rated NOVA as easy to understand ($M = 6.02$, $SD = 0.84$), $t(52) = 17.43$, $p < .001$, $d = 2.39$ (large effect)¹. They also rated NOVA to help understand the ML model ($M = 5.34$, $SD = 1.06$), $t(52) = 9.24$, $p < .001$, $d = 1.27$ (large effect). The evaluation of the self-efficacy scale showed that the participants were confident that they would be able to cope successfully with the given tasks when interacting with NOVA again ($M = 7.62$, $SD = 1.18$), $t(52) = 13.0$, $p < .001$, $d = 1.79$ (large effect).

1: interpretation of the effect size d according to J. Cohen (1988) is:

$d < .05$: small effect;

$d = 0.5-0.8$: medium effect;

$d > 0.8$: large effect

Subjective Trust, Self-Efficacy, and Cognitive Workload of End-Users

A one-way MANOVA was calculated to investigate the differences between the four conditions regarding subjective trust, computer-self efficacy, and cognitive workload. The result of the MANOVA was not statistically significant, Wilks' Lambda = 0.80, $F(9, 115) = 1.21$, $p = .293$, which means there were no statistical differences between the conditions regarding the TiA, CSE and NASA-TLX ratings of the participants.

End-Users' Impression of XAI Methods

After the participants interacted with NOVA and described their impression of NOVA, participants in the three XAI information conditions were asked about the simplicity and helpfulness of this information using two items. Overall, the results show that confidence values as well as LIME visualisations both reached values beyond 5 (1 = disagree, 7 = fully agree), which means they tend to be helpful and easy to understand (see Table 29.2 on the next page).

To evaluate the two items, we conducted two one-way MANOVAs. The first MANOVA compared the impressions (simplicity and helpfulness) of the two conditions, which saw the LIME visualisations. Here we found

Table 29.2: Rating of participants, if the confidence values and LIME visualisations are helpful and easy to understand (Conditions: 1=confidence values; 2=LIME visualisations; 3=LIME visualisations and confidence values)

Characteristic	Conditions			Total
	1	2	3	
<i>n</i>	13	14	13	40
Confidence values (easy)				
<i>M</i>	6.77	-	6.23	6.50
<i>SD</i>	0.44	-	0.83	0.71
Confidence values (helpful)				
<i>M</i>	6.00	-	6.31	6.15
<i>SD</i>	0.91	-	0.75	0.83
LIME visualisations (easy)				
<i>M</i>	-	5.43	5.85	5.63
<i>SD</i>	-	1.22	1.41	1.31
LIME visualisations (helpful)				
<i>M</i>	-	5.71	5.62	5.67
<i>SD</i>	-	0.99	1.12	1.04

no significant differences between the conditions, Wilk's Lambda = 0.82, $F(2, 24) = 2.57$, $p = .097$. The second MANOVA compared the impressions (simplicity and helpfulness) of the two conditions, which saw the confidence values. Here we found a significant difference between the conditions, Wilk's Lambda = 0.71, $F(2, 23) = 4.80$, $p = .018$. The following ANOVA revealed that there was a significant difference in the variable "easy to understand", $F(1,24) = 4.26$, $p = 0.05$, where participants of condition 1 (confidence values) rated the confidence values as easier to understand compared to participants of condition 3 (XAI & confidence values).

End-Users' Mental Model about the CNN

In order to determine the end-users mental model of the CNN for facial emotion expression recognition, the participants were given the task of finding correctly and incorrectly classified images. Subsequently, they had to explain what aspects were relevant for the classification by the model. In addition, participants should state how confident they were in their explanation. Overall, the participants were as confident in their explanations about the relevant aspects of the neural network for correctly classified images ($M = 5.09$) as for incorrectly classified images ($M = 5.03$). When considering the confidence of the statements in the four conditions, a reasonably equal rating between the conditions can be seen (see Figure 29.2 on the facing page). To evaluate the ratings between the four conditions statistically, we conducted a one-way MANOVA. Here we found no statistical difference for all four groups, Wilks' Lambda = 0.94, $F(6, 96) = 0.51$, $p = .806$.

The similar, quite good ratings between the conditions, even in the baseline condition without objective information in the form of XAI, evaluate the open questions on the participants' reasons even more interesting. The lack of XAI information did not disturb the participants of the baseline condition to generate interpretations about the models' behaviour. Instead, they justified the behaviour of the neural network with the arguments they use for emotion classification (see Table 29.3 on page 218 for examples of participants' feedback). For instance, in the baseline condition, most participants described their assumptions about the models'

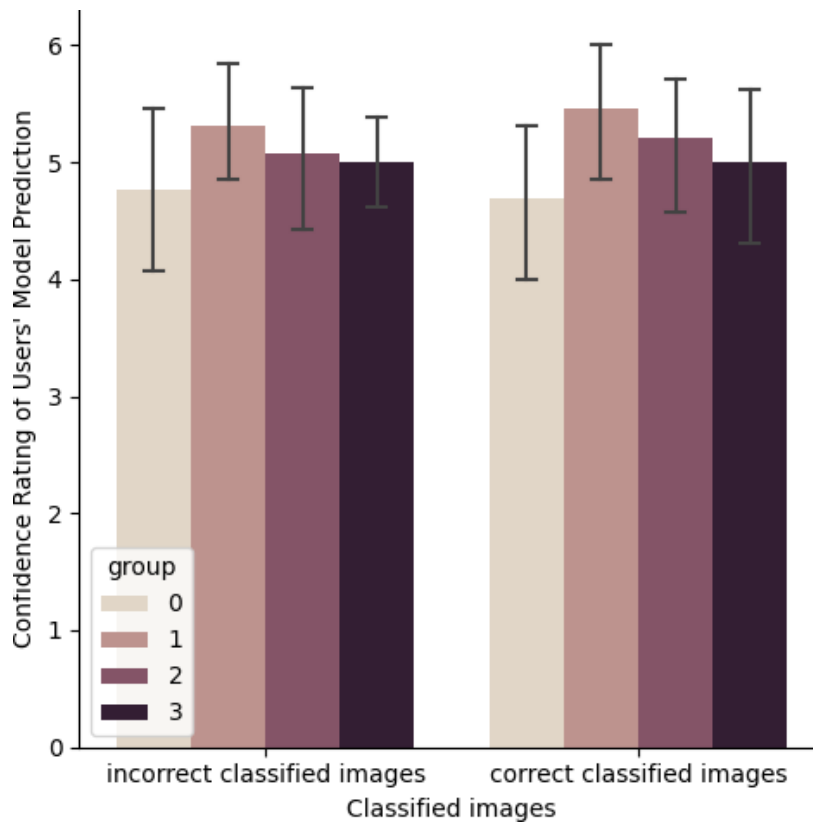


Figure 29.2: Rating of the participants to what extent they are confident in their description of the behaviour of the neural network model (1 = disagree to 7 = fully agree). 0=Baseline condition; 1 = Confidence values condition; 2 = LIME condition; 3 = LIME and confidence values condition. No significant differences between the conditions were found. Error bars represent the 95% CI

behaviour for the emotion happiness, followed by descriptions for the emotion sadness. Here, prototypical facial expressions (e.g., for happiness: pull up of the corners of the mouth, show teeth) were used as explanations. Furthermore, participants often used their own assumptions to reference the model's behaviour (e.g., "corresponded to my own opinion" or "for me, she looked disgusted"). In contrast to this, in the two conditions with the LIME visualisations, it can be seen that the participants described their own strategies for emotion recognition less and used the XAI information instead. For example, they refer to superpixel areas presented to them by LIME. In Figure 29.3 on the following page, two images presented in the study using NOVA are shown. The superpixels generated by LIME for the classification of happiness are displayed. In the left image, the CNN model focuses on the mouth to classify happiness. In the right image, the model focuses on the background to classify happiness. This faulty learning of the neural network with simultaneous correct prediction was only recognised and mentioned as a problem by participants in the two LIME visualisation conditions.

Interestingly, in the two conditions where confidence values were displayed, the participants did not use the information about the uncertainty of the model to explain its behaviour. The decisive factor was whether people were additionally shown LIME visualisations or whether they only saw confidence values. If they saw LIME visualisations combined with confidence values, the answers were similar to the condition that only saw LIME visualisations. If they only saw confidence values, the responses were very similar to the baseline condition, who assumed their own assumptions were those of the model.

Figure 29.3: XAI visualisation generated by LIME for two images classified as happy by a neural network model. While in the left image the network focused on the mouth region, in the right picture the background seems to have had an impact on the model's decision. Figure from Heimerl et al. (2022)



Table 29.3: Explanations were given by the participants about the behaviour of the neural network. Sentences in *italics* refer to the network's behaviour when classifying images correctly, non-italic statements to incorrect classifications. Table from Heimerl et al. (2022)

Condition	Example feedback of the participants
Baseline (no XAI information)	<ul style="list-style-type: none"> • <i>The emotions were clearly recognizable. the facial features were clear</i> • <i>Happy was especially recognized by a laughing mouth; also corresponded most often to my own opinion</i> • Images and emotions have been well matched. Only neutral faces did not always fit perfectly • The Indian woman's eyes were so full of make-up that the system predicted she would be happy, but for me she looked neutral to disgusted
Confidence values	<ul style="list-style-type: none"> • <i>For the pictures classified as Happy: on the smile, teeth often shown; Neutral: Few facial expressions"</i> • <i>Often happy, because of teeth & smiling</i> • Large eyes are classified as aggressive in pictures • Sad often did not correspond to a sad expression. Apart from that nothing special noticed
LIME visualisations	<ul style="list-style-type: none"> • <i>I think the model recognized the pictures correctly, especially because it looked at the mouth and/or eyes</i> • <i>The model had focused the XAI visualisation on relevant areas of the face. eye area, mouth area</i> • I believe that the model has misclassified the images because it has often focused the mouth - and just because the mouth is open does not mean that the image shows someone "Happy" or "Angry" • In some cases, the XAI visualisation did not refer to the face at all, but marked the background or clothing
Confidence values & LIME visualisations	<ul style="list-style-type: none"> • <i>On certain parts of the face, the model was able to easily identify the appropriate emotions</i> • <i>The model focuses on the eye and mouth area</i> • Sometimes the eyes are not taken into account, e.g. if the teeth are seen, the person can still be sad • Unnecessary areas such as the background are taken into account, mouth and eyes are hardly or not at all considered. Why the program does not concentrate on these areas is not understandable

Areas of Interests for LIME and Humans

As the final task of the study, the participants were asked to highlight areas of relevance for classifying emotions in images. They were explicitly told to mark areas that they think have been important for their recognition of a specific emotion. In the following, we will compare heatmaps generated from participants' reported areas with XAI visualisations generated by LIME. Figure 29.4 on the next page displays heatmaps and LIME visualisations for five faces (A to E) that correspond to a specific emotion. The following emotions are present: A: anger, B: neutral, C: disgust, D: sadness, E: happiness. The top row covers the heatmaps. The brightness of the colouring describes the importance of the facial areas, as marked by the participants. The bottom row covers the XAI visualisation from LIME. The spaces defined by the yellow bounds describe the areas of the face that have been important for classifying a specific emotional expression. When analysing the heatmaps, it is conspicuous that the participants identified the eye and mouth area for all faces to be most important. Little to no attention has been paid to other facial regions. For the angry face (A), emphasis was placed on the area between the eyes and the eyes themselves. This is most likely due to the presence of wrinkles. The mouth region played a subordinate role. In the neutral face (B), especially the area around the mouth has been considered important. In addition to that, the eyes have been given attention. The disgusted face (C), similar to the angry face, displays wrinkles between the eyes, which the participants identified as a relevant area. Moreover, the specific shape of the mouth, with the corners of the mouth facing downwards, was marked as very important. For the sad facial expression (D), the mouth and the wrinkled chin have been recognised as valuable information. It is noteworthy that for this facial expression, the eyes themselves have been considered exceptionally important. That is most likely because tears are present in the corners of the woman's eyes in this image. In the happy-looking face (E), the region around the mouth, displaying a big smile and the corresponding wrinkles around the cheeks, have been identified as the most important. Little attention was given to the area around the eyes. In contrast, the automatically generated LIME visualisations cover larger areas of importance. This difference is especially evident in the angry, neutral and disgusted face. Humans generally focus more on specific facial features, whereas the trained model takes a rather holistic approach by emphasising larger areas of the face.

Following the question about essential areas for emotion recognition, we asked the participants after each shown image how certain they were with their decision (8-point Likert scale: 1 = unsure to 8 = entirely sure). The corresponding results are displayed in Figure 29.5 on the following page. Overall, the participants have been very confident in their own decisions. None of the different emotions resulted in a score below 6. However, there have been differences between the emotions. The participants have been most certain with their judgement of the happy and sad faces. They have been most uncertain of the disgusted facial expression, followed by neutral—anger placed in the middle regarding their certainty. Also, no one of the participants did classify any of the presented emotions wrong.

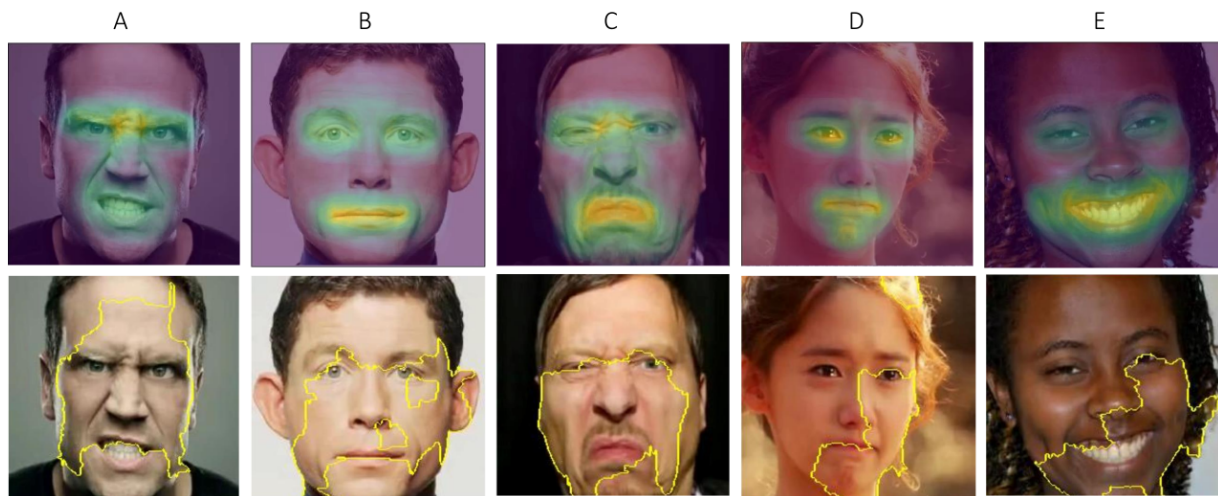


Figure 29.4: Comparison between the average areas of interest according to the study participants and model agnostic explanations generated with LIME. The different faces show varying emotions. A: anger, B: neutral, C: disgust, D: sadness, E: happiness. Figure from Heimerl et al. (2022)

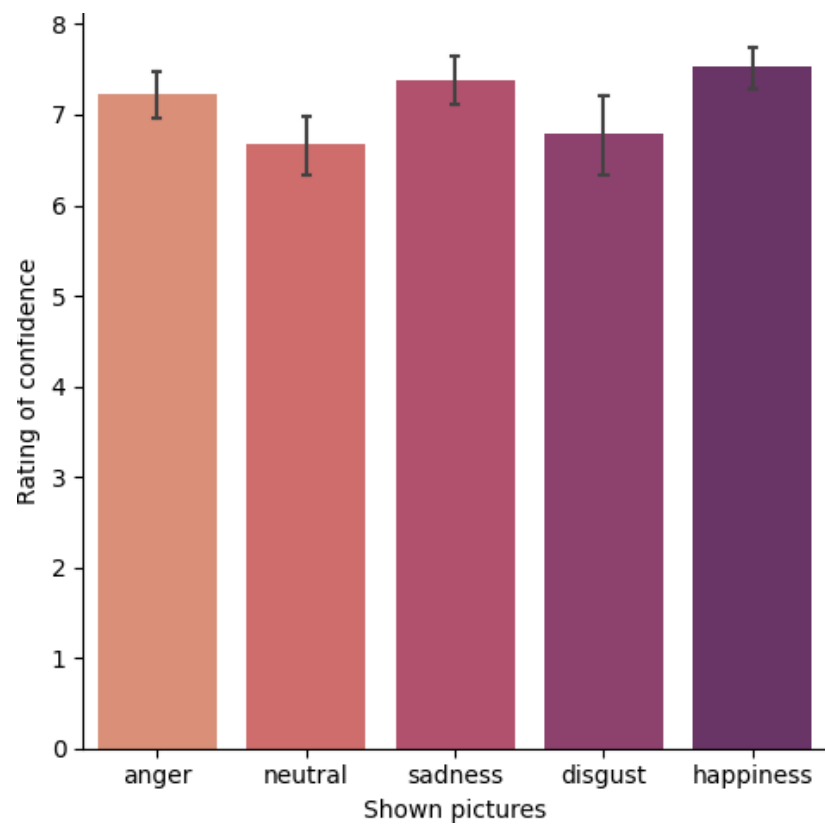


Figure 29.5: Rating of the participants to what extent they are confident in their classification of the emotional pictures (1=unsure to 8=sure). Error bars represent the 95% CI

29.2.8 Discussion

Our study aimed to gain insights into the interaction between end-users and ML models using NOVA. In the following, the results obtained will be discussed.

NOVA Is Helpful for End-Users

Our study results show that users with little or no experience with ML can use NOVA for labelling data in the revision step of the CML workflow. Furthermore, even though all participants had never worked with NOVA before, they found it easy to use and had the impression that NOVA helps them understand the ML model. Similar results were found for the given XAI information. Confidence values, as well as XAI visualisations generated by LIME, were rated as easy to understand and helpful by the participants. Also, the self-efficacy values of the CSE show that the participants have a high computer self-efficacy towards NOVA. They believed that they could do similar tasks with NOVA in the future.

XAI Does Not Automatically Influence Users' Perceptions

We found no difference in the CSE values between the four experimental conditions. Instead, the participants in all conditions achieved high CSE values. Similar results were found for subjective trust and cognitive workload. A cause for this could be the easy handling and usage of NOVA as well as the domain of the classification task of the neural network model. Emotion expression recognition is a task where (most) humans perform pretty well. This could have led to more self-confidence and increased trust in the system, compared to our work presented in Weitz, Schiller, et al. (2021) and Weitz et al. (2019) (see Chapter 28 on page 190 - Gloria Study & Museum Study), where the domain explained with XAI visualisations (spoken words in the form of spectrograms) was not familiar to humans. But the use of different XAI methods influences the perceived simplicity of the specific method. For example, participants found the use of confidence values harder to understand when they also saw LIME visualisations. This result indicates that XAI visualisations give users the impression of being easier to be interpreted. de Graaf and Malle (2017) assumed that people apply human traits to AI systems. This leads to the expectation that the AI system should explain its behaviour in a human-like manner. Miller (2019) points out that explanations including probabilities are not necessarily the best for a user.

End-Users Create Assumptions About AI

A compelling result we observed is that even without XAI information, participants in the baseline condition formulated extensive explanations about the behaviour of the neural network model and were also very confident in their reasoning. This indicates that with high computer self-efficacy and a well-known application domain (e.g., emotion expression recognition), end-users tend to equate their own assumptions with those of the ML model. However, this assumption can have devastating consequences if it does not hold because people do not question whether

the model has learned what it should have learned (see Figure 29.3 on page 218). We found a difference regarding the users' mental models of the AI system and their assessment of how helpful and easy to understand the XAI methods were. Although the users had the impression that the XAI methods were helpful and easy to understand, only the two conditions with LIME visualisations helped the users to create more accurate mental models. Even if the participants' explanations about the behaviour of the CNN model in the conditions of the LIME visualisations were more accurate and correct than in the other conditions, it must be pointed out that visualisations alone are insufficient to generate exhaustive explanations. For example, many participants in the two LIME visualisation conditions still assumed additional information that is not part of the visualisations (e.g., image sharpness, image exposure). XAI visualisations alone do not explain anything. They only provide information that has to be interpreted by the user. But the interpretation itself may again be flawed. Therefore, it is necessary to go beyond visualisations and provide additional information, for example, in the form of combining LIME with linguistic explanations about relational concepts (Rabold et al., 2019) (e.g., "The classification was happiness because the raised corners of the mouth were relevant") in order not to leave the interpretation entirely to the imagination of the user.

XAI Perception Differs from Human Perception

We presented the results for manually highlighting facial regions that are supposedly relevant for a specific emotion and compared those with the marked regions generated by LIME, in which the output of LIME describes areas that have been crucial for the classification. We found that the participants identified the eye and mouth areas to be the most important. However, depending on the presented emotion, they weighted those areas differently, e.g., for the angry facial expression, the eye region was considered more important. In contrast, for the happy face, the focus was on the mouth. Moreover, they tended to value specific facial features more than a holistic approach to recognising emotions in facial expressions. Those findings are interesting when contextual with the research of Bombari et al. (2013). They investigated the role of featural (e.g., the shape of the mouth) and configural face information (relational information, e.g., the distance between the nose and the mouth) when recognising emotions. For their experiments, they used faces representing four different emotions (happiness, sadness, anger, and fear). They reported that happiness had been recognised more easily and rapidly when compared to other emotions. Also, they stated that the mouth region has been crucial for recognising happiness. This is in line with our finding that the participants have been most confident in their classification for the happy facial expression (see Figure 29.5 on page 220), and they highlighted the mouth as most relevant for their classification. It is important to note that in our study, we explicitly asked the participants what they think the important regions for recognising a specific emotion are, whereas Bombari et al. (2013) gathered that information by using eye tracker systems. When we compare the results of Bombari et al. (2013) with the generated heatmaps of the facial expressions in Figure 29.4 on page 220, it seems that when asked what the relevant information for recognising a specific emotion is, humans tend to focus more on the featural aspects of

faces rather than the configural information. We mentioned earlier that our trained neural network model follows a rather holistic approach to recognising emotional expressions. When we inspect the visualisation for the relevant areas generated by LIME, it is visible that a large area of the face is marked as especially important for classification. The participants identified specific features as important, whereas the CNN model focuses on larger facial regions. It is essential to understand that depending on the emotion, either configural or featural information is more relevant for humans to classify facial expressions visually (Bombari et al., 2013). Still, when asked, people tend to state that mainly featural information is considered important. This should be kept in mind when providing additional information to humans about the inner workings of ML models. It could be similar to our case that the model imitates a human-like holistic perception behaviour, but the users may not appreciate the explanation as they feel like irrelevant information is considered important. Further, generating explanations should align with human expectations while mapping the actual behaviour of ML models.

Implications for Other Emotion Recognition Domains

In our proposed study, we investigated how XAI techniques can assist end-users in terms of trust, perceived self-efficacy, cognitive workload, and creating an accurate mental model of a system. However, we solely considered a non-verbal aspect of affective computing, namely the recognition of emotional facial expressions. Recent studies in the field of affective computing also focus on sentiment analysis and natural language processing (Cambria, 2016). As a result, innovative approaches emerged like using stacked ensemble to predict the intensity of sentiments and emotions (Akhtar et al., 2020) or applying novel semi-supervised learning techniques to extract knowledge from unstructured social data (Hussain & Cambria, 2018). For future work, it would be interesting to examine how XAI methods perform on black-box models that predict emotion from the text.

29.2.9 Conclusion

In our study, we showed that interactive ML applications like NOVA are helpful for tasks involving end-users. Even end-users found NOVA easy to use and understand. Moreover, the participants were confident about their ability to employ NOVA for similar affective computing annotation tasks. We have further shown that XAI information is comprehensible and helpful to our participants with little or no expertise in data annotation and ML. We conclude that incorporating such techniques in healthcare applications offers value to users in the interactive ML loop to guarantee a correct working AI system for emotion and pain recognition. One of our work's key revelations is that humans create assumptions about AI. In our study, we found that when end-users get presented with little to no additional information about the system's inner workings, they apply their own mental model to the ML model. This became evident when investigating the reported feedback of the participants about the predictions the system made. We argue that this is connected to the high levels of self-efficacy and a domain (emotion expression recognition) the

participants are familiar with. Further, such behaviour is to be seen as critical, especially when the computational model does not align with the end-users mental model. In those cases, the users might stop questioning what the computational model has learned.

Moreover, it became evident that explanations in the form of visualisations help create an accurate mental model but more is needed to provide more transparency and insight about a given system. This claim is grounded on the fact that the participants in the two LIME conditions - even though they referred in their feedback to the given visualisations - still made up additional reasons that were not accessible from the information they were provided. Further, we argue that such visualisations themselves are not sufficient explanations but offer additional information that has to be interpreted by the user. Therefore we recommend combining this kind of visual feedback with additional information or interpretation to provide the user with more holistic explanations. For example, a possible implementation for our use case could be to add some textual or verbal explanation in the form of “The person seems happy because the raised corners of the mouth were of high relevance and indicate a smile”. In such a case, the user would have access to the actual image with the marked areas considered important by the ML model and an interpretation of what the model actually focused on.

At last, the context and domain of a classification task might influence how XAI visualisations are perceived and interpreted. For example, analysing the task results where participants were asked to identify important information in given images of facial expressions, we found a discrepancy between what people consider important and how they actually process certain emotions. This could lead to less acceptance of an ML model even though the behaviour might align with the human approach to processing information. Therefore we suggest generating explanations that align with human perception of a given problem domain. This is highly connected to our earlier recommendation to provide textual and visual explanations that are easier for the user to comprehend and assist them in interpreting the presented behaviour.

In this work, we applied the CML workflow incorporating explanations in an affective computing problem domain. We strongly believe that disciplines such as healthcare, psychotherapy, and others may benefit from such technologies. Especially in high-risk environments that apply AI, it is crucial to fully understand the underlying processes that led to a classification result when relying on high prediction accuracies. Tools such as NOVA prove valuable as they can potentially help domain experts (e.g., physicians, psychotherapists) with little to no expertise in ML to better assess the behaviour of their ML models. In addition, our results show that end-users benefit from XAI visualisations and such methods could be used to explain to them a classification of a medical decision support system.

29.3 Explainable AI in Pulmonary Disease Classification - Pneumonia Study

29.3.1 Highlights

- ▶ Counterfactual explanations are more helpful for end-users to trust an AI system than saliency maps.
- ▶ Counterfactual explanations increase end-users self-efficacy and confidence in the CNN compared to saliency maps.
- ▶ XAI visualisations have an emotional impact on end-user: Counterfactual explanations support positive emotions (i.e., relaxation) toward an AI system and reduce negative emotions (i.e., anger) compared to LRP.

29.3.2 Introduction

In image classification tasks like in the NOVA Study that was presented before, many state-of-the-art methods to explain such classifiers (e.g., LIME, LRP) rely on visual highlighting of important areas of the input data. On the contrary, counterfactual explanation systems try to enable counterfactual reasoning by modifying the input image so that the classifier would make a different prediction. By doing so, the users of counterfactual explanations are equipped with different explanatory information. Especially in medical contexts, where relevant information often consists of textural and structural information, high-quality counterfactual images have the potential to give meaningful insights into decision processes. We used the system proposed in Mertes et al. (2022) to create counterfactual explanations for a classifier that was trained on a classification task to predict if X-ray images of the human upper body are showing lungs that are suffering from pneumonia or not. In addition to being a highly relevant application for explanations, this scenario is suitable for evaluating explanations for end-users since they are not expected to have in-depth knowledge of that domain, i.e., they are entirely reliant on the explanation that the XAI system gives to follow the AI's decisions. Furthermore, pneumonia in X-ray images is predominantly reflected by opacity in the shown lungs. Opacity is textural information that can not be explained sufficiently enough by the spatial information provided by common saliency map approaches.

To validate our assumptions, we compare the counterfactuals generated with the GAN-based approach proposed in Mertes et al. (2022) against two established saliency map methods, namely LIME and LRP. For this, we conduct a user study in an exemplary medical use case. We evaluate the three visual explanations through a user study inspired by a healthcare scenario. Our results show that in the chosen medical use case for decision support, counterfactual explanations lead to significantly better results regarding mental models, explanation satisfaction, trust, emotions, and self-efficacy compared to LIME and LRP. With our work, we make the following contributions:

- ▶ We evaluate our approach in a user study and gain insights into the applicability of counterfactual explanations for non-ML experts in an exemplary medical context.

- We compare counterfactual explanations against two state-of-the-art explanation systems that use saliency maps.

29.3.3 Hypotheses

All our hypotheses target end-users in healthcare and AI. Since we aim to evaluate our proposed counterfactual approach, we do not investigate differences between the saliency map conditions (LRP and LIME). For our user study, we formulated the following hypotheses:

1. **Explanation Satisfaction** Participants are more satisfied with counterfactuals' explanatory quality than LIME and LRP.
2. **Mental Models** Participants use counterfactuals to create more accurate mental models about the AI than with LIME and LRP.
3. **Trust** Participants have more trust in the AI system if it is explained with counterfactuals than if it is explained with LRP or LIME.
4. **Emotions** The intuitive and simple interpretation of counterfactuals makes participants feel happier, more relaxed, and less angry compared to LRP and LIME.
5. **Self-efficacy** If counterfactuals are a more satisfying XAI method than LRP or LIME, participants also feel strengthened in their self-efficacy towards the AI system, compared to participants in the LRP and LIME conditions.

29.3.4 Study Design

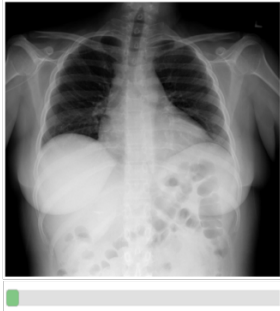
The study was conducted online. After providing some demographic information, the participants received a short tutorial explaining the X-ray images and the XAI visualisations they would interact with in the experiment. After the tutorial, each participant had to answer a quiz. Here, questions were asked to ensure that the participants carefully read the tutorial and understood how to interpret the X-ray images (e.g., "Which part of the body is marked in this picture?") and the XAI visualisations (e.g., "What do green areas in images tell you?" for the LIME and LRP conditions). Only participants who successfully solved the quiz were allowed to participate in the experiment.

After the quiz, the main experiment with the prediction task started. We randomly assigned each participant to one of three conditions. The participants in each condition only interacted with a single visual explanation method (between-subjects design). The participants were asked to predict the AI's diagnosis for 12 images. In addition to the original image, the participants were provided with a slider to interact with the XAI visualisations. Moving the slider to the right linearly interpolated the original image with either the counterfactual image or a version of the image that is augmented with an LRP or LIME heatmap, depending on the condition of the user².

2: by tracking if the participants used the slider, we additionally know whether they looked at the XAI visualisations

In our pilot study ($N = 10$), we found that participants often project their own reasoning on the AI. The participants in the main study were asked whether they *themselves* would classify the given image as pneumonia/not pneumonia and how confident they were in this diagnosis to differentiate between their own and the AI's diagnoses. Then they were

Task Description and AI Diagnosis for the last image.



Do you think the original x-ray (on the left side of the slider) shows a person suffering from pneumonia or not?

How confident are you that your diagnosis is right?

Not at all confident
☐
☐
☐
☐
☐
☐
☐
Very confident

What do you think will the AI decide? (Base your prediction on the Explanation)

How confident are you that you predicted the AI correctly?

Not at all confident
☐
☐
☐
☐
☐
☐
☐
Very confident

Please briefly explain your selection:

Figure 29.6: Simplified illustration of our prediction task used in the Pneumonia Study. Figure from Mertes et al. (2022)

asked to predict whether *the AI* will classify the image as pneumonia/not pneumonia, based on the given XAI visualisation. Here too, they had to give a confidence rating in their prediction and could give an optional justification for their prediction. After each prediction, they were told the actual decision of the AI for the last image. A schematic of the complete task is shown in Figure 29.6.

After predicting the AI's decision for all 12 X-ray images, the task reflection, where they had to describe their understanding of the AI's reasoning, followed. Then the questionnaires about explanation satisfaction, trust, self-efficacy and emotion were provided.

29.3.5 Methodology

To evaluate our hypotheses, we used the following measurements:

Mental Model AI We used two metrics to evaluate the mental models that the participants formed through our XAI methods. Quantitatively, we conducted a *prediction task*, as proposed by Hoffman, Mueller, et al. (2018). First, participants had to predict what the AI model would decide for a given X-ray image (i.e., pneumonia/no pneumonia). Second, they were asked how confident they were in their decision on a 7-point

Likert scale (1 = not at all confident to 7 = very confident). In addition, they could give a justification for their prediction if they wanted to. For a more qualitative evaluation, we used a form of *task reflection*, also proposed by Hoffman, Mueller, et al. (2018). The participants were asked to describe their understanding of the AI's reasoning after completing the prediction task. For this, the participants were asked two questions about their mental model of the AI: "What do you think the AI pays attention to when it predicts pneumonia?" and "What do you think the AI pays attention to when it predicts healthy lungs?". In addition, we asked participants about their confidence in detecting pneumonia using the presented explanation in the future (10-point Likert scale, 1 = not at all confident to 10 = completely confident).

Mental Model Self To gain insights into participants' mental models about themselves, they had to classify the given image as pneumonia/no pneumonia and how confident they were in their own diagnosis on a 7-point Likert scale (1 = not at all confident to 7 = very confident).

Explanation Satisfaction We used the Explanation Satisfaction Scale, proposed by Hoffman, Mueller, et al. (2018) to measure the participants' subjective satisfaction with the visual explanations (LRP, LIME, or counterfactuals) that we presented.

Trust To evaluate the trust in the presented AI system, we used two items (i.e., "I trust the system" and "I can rely on the AI system") from the Trust in Automation (TiA) questionnaire proposed by Körber (2018).

Emotions We used items for the subscales *anger*, *happiness*, and *relaxation* of the Discrete Emotions Questionnaire (DEQ) (Harmon-Jones et al., 2016) to evaluate the participants' feelings during solving the tasks.

Self-efficacy We used one item to measure self-efficacy towards the AI system. For this, we used a variation of one item proposed by Bernacki et al. (2015) (i.e., "How confident are you that you could detect pneumonia using the presented explanations in the future?").

29.3.6 Evaluation Methods

Quantitative Evaluation of the Data We calculated the mean of the correct predictions of the AI and the participant's confidence in their predictions of the AI. To ensure that we only used responses where the participants at least saw the visual explanations, we excluded answers where the participant did not move the slider. If, for example, a participant did not use the slider 4 times, we only calculated the mean for the remaining 8 answers. For the DEQ, we calculated the mean for the emotion subscales happy, anger, and relaxation. For the TiA, we calculated an overall trust score from the two questions presented.

Qualitative Evaluation of the Participants' Mental Model of the AI

Similar to Anderson et al. (2019) and Huber et al. (2021), we used a form of summative content analysis (Hsieh & Shannon, 2005) to qualitatively evaluate the participants' free text answers to the questions "What do you think the AI pays attention to when it predicts pneumonia?" and "What do you think the AI pays attention to when it predicts healthy lungs?". Our classifier was trained on a dataset consisting of X-ray images of normal lungs and X-ray images that contain lung opacity, which is a crucial indicator of lungs suffering from pneumonia. Since we only told the participants that our model classifies pneumonia, we can score their responses based on whether they correctly identified lung opacity as a key feature for our model. To this end, two annotators independently went through the answers and assigned concepts to each answer (e.g., opacity, clarity, contrast, and other organs than the lung).

- ▶ **1 point** were received for answers to the pneumonia question that contained at least one concept related to opacity, like *opacity*, *white colour in the X-ray* and *lung shadows*. Similarly, answers to the healthy lungs question that contained at least one concept related to clarity, like *clarity*, *black colour in the X-ray* or *no lung shadows*, received 1 point.
- ▶ **0.5 points** was received for answers for both questions that contained a concept related to contrast, like *contrast* or *clear edges*.
- ▶ **0 points** were received for all other answers.

For 21 out of all 236 responses, the two annotators differed in the given score. Here, a third annotator was asked to assign 0, 0.5 or 1 points to the answer and the final points were calculated by a majority vote between the three annotators. By adding the points for those two questions, each participant was given a score between 0 and 2, approximating the correctness of their description of the AI

29.3.7 Participants

Aiming for an 80 % power in a one-way between-subject MANOVA (three conditions, $\alpha=.05$), the conducted a-priori power analysis suggested that we would need 37 participants in each condition ($N = 111$) to detect an effect of $\eta_p^2 = 0.04$. In order to compensate for possible drop-outs, we collected data from 122 participants using the Clickworker online platform³. Participation was limited to US, UK, Australian, or Canadian users whose native language is English to ensure a sufficient English level. Since LRP and LIME are not designed with colour-blind people in mind, the participants were also asked if they were colour-blind and stopped from participating if they were. To ensure that the participants understood the provided information about the task correctly, we used a quiz they had to complete to participate in the study. As an incentive to diligently do the task, the participants received a bonus payment in addition to the base payment if they correctly predicted at least 2/3 of the AI model's prediction. In addition to these precautions, we subsequently excluded 4 participants because they never looked at the XAI visualisations or their responses did not reflect a severe engagement with the study (e.g., free text answers which are not related to the question at all). For our final analyses, we used data from 118 participants between 18 and 67 years ($M = 38.5$, $SD = 10.9$). Sixty-three of them were male, 53 were female, and

3: Clickworker: <https://www.clickworker.com/clickworker/>

two were non-binary. All in all, only 8 participants reported experience in healthcare. Forty-three participants stated that they had experience in AI. While the participants were randomly separated in the three XAI visualisation conditions, the level of AI and healthcare experience was evenly distributed between the three conditions.

29.3.8 Results

Impact of XAI Methods on Explanation Satisfaction, Trust, and Prediction Accuracy

To gain an impression of their mental models of the AI, the participants had to predict the decision of the CNN (pneumonia/no pneumonia). At the end of the study, they rated their trust in the AI and their explanation satisfaction. To evaluate these variables between the three conditions, we conducted a one-way MANOVA. Here we found a significant statistical difference, Wilks' Lambda = 0.59, $F(6, 226) = 11.2$, $p < .001$. The following ANOVA revealed that all three variables showed significant differences between the conditions:

- **Prediction accuracy:** $F(2, 115) = 30.18$, $p < .001$,
- **Explanation satisfaction:** $F(2, 115) = 5.87$, $p = .004$,
- **Trust:** $F(2, 115) = 3.89$, $p = .02$,

To determine the direction of the differences between the three XAI method conditions, we used post-hoc comparisons for each variable⁴. We found the following differences:

- **Prediction accuracy:** The participants' predictions of the AI's decisions were significantly more correct in the counterfactual condition compared to the LRP condition $t(115) = -6.48$, $p = .001$, $d = 1.47$ (large effect)⁵ as well as compared to the LIME conditions $t(115) = -6.92$, $p = .001$, $d = 1.55$ (large effect) (see Figure 29.7 on the facing page).
- **Explanation satisfaction:** Participants were significantly more satisfied with the explanation quality of the counterfactual explanations compared to the LRP saliency maps, $t(115) = -3.05$, $p = .008$, $d = 0.70$ (medium effect) and the LIME visualisations, $t(115) = -2.85$, $p = 0.01$, $d = 0.64$ (medium effect) (see Figure 29.7 on the next page).
- **Trust:** The AI was rated as significantly more trustworthy in the counterfactual condition compared to the LRP condition, $t(115) = -2.56$, $p = .03$, $d = 0.58$ (medium effect) but not to the LIME condition, $t(115) = -0.29$, $p = .07$ (see Figure 29.7 on the facing page).

4: we used the Holm correction for multiple testing to adjust the p-values for all post-hoc tests we calculated

5: interpretation of the effect size d according to J. Cohen (1988) is:
 $d < .05$: small effect;
 $d = 0.5-0.8$: medium effect;
 $d > 0.8$: large effect

Impact of XAI Methods on End-Users' Self-Efficacy

The first analysis reported above showed that (1) the quality of counterfactual explanations was rated significantly higher, and (2) participants predicted the decisions of the AI were significantly more accurate compared to LIME and LRP. Therefore, based on our last hypothesis, we examined whether these positive assessments were also reflected in the self-efficacy and prediction confidence of the participants. For this purpose, we conducted a one-way MANOVA. Here, we found a significant statistical

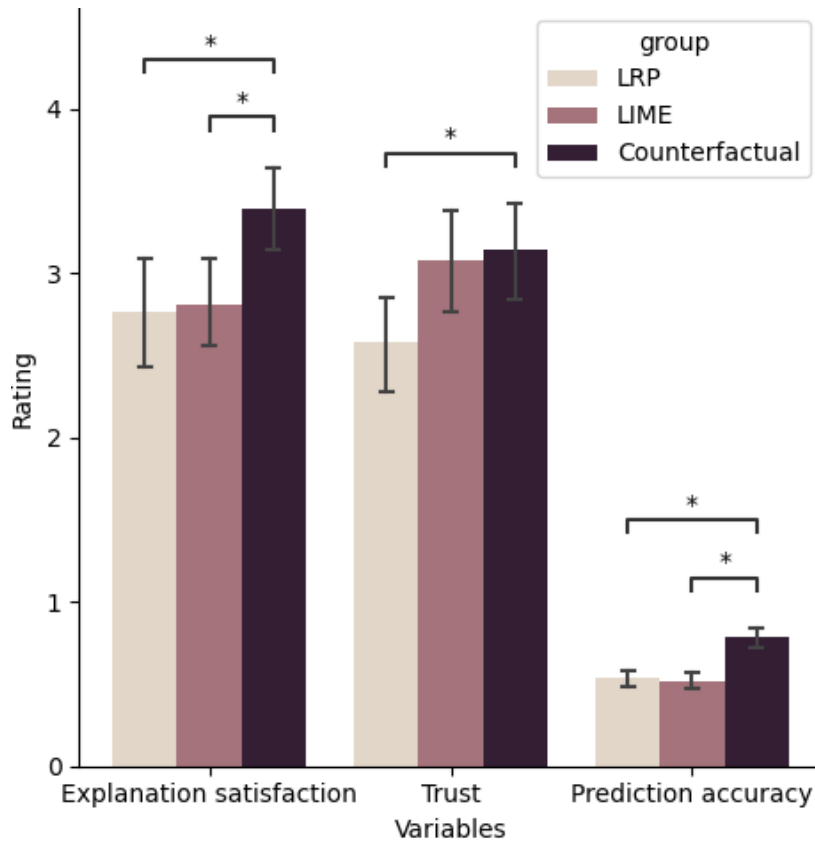


Figure 29.7: There are significant differences in explanation satisfaction, trust, and prediction accuracy between the counterfactual and saliency map conditions (LRP and LIME). Error bars represent the 95% CI., * $p < .05$

difference, Pillai's Trace = 0.15, $F(4, 230) = 4.69$, $p = .001$. The following ANOVA revealed a statistical difference for self-efficacy $F(2, 115) = 6.93$, $p = .001$ and prediction confidence $F(2, 115) = 7.68$, $p < .001$ between the conditions.

The post-hoc comparisons showed that counterfactuals lead to significantly higher self-efficacy compared to LRP $t(115) = -3.44$, $p = .002$, $d = 0.78$ (medium effect) as well as LIME, $t(115) = -2.94$, $p = .01$, $d = 0.66$ (medium effect) (see Figure 29.8 on the next page).

The same pattern was found for the prediction confidence, where counterfactuals lead to significantly higher prediction confidence compared to LRP $t(115) = -3.45$, $p = .002$, $d = 0.78$ (medium effect) as well as LIME, $t(115) = -3.32$, $p = .003$, $d = 0.74$ (medium effect) (see Figure 29.8 on the following page).

A closer look reveals that these significant differences stem from the confidence in the correct predictions and not the confidence in the incorrect ones (see Figure 29.9 on the next page).

Result of the Qualitative Evaluation of End-Users' Mental Models

Subsequently, to the significant differences in the prediction accuracy as the first impression of participants' mental model, we analysed the content analysis results of the task reflection responses. For this, we conducted a one-way ANOVA. Here we found a significant statistical difference, $F(2, 115) = 7.91$, $p < .001$. To determine the direction of the

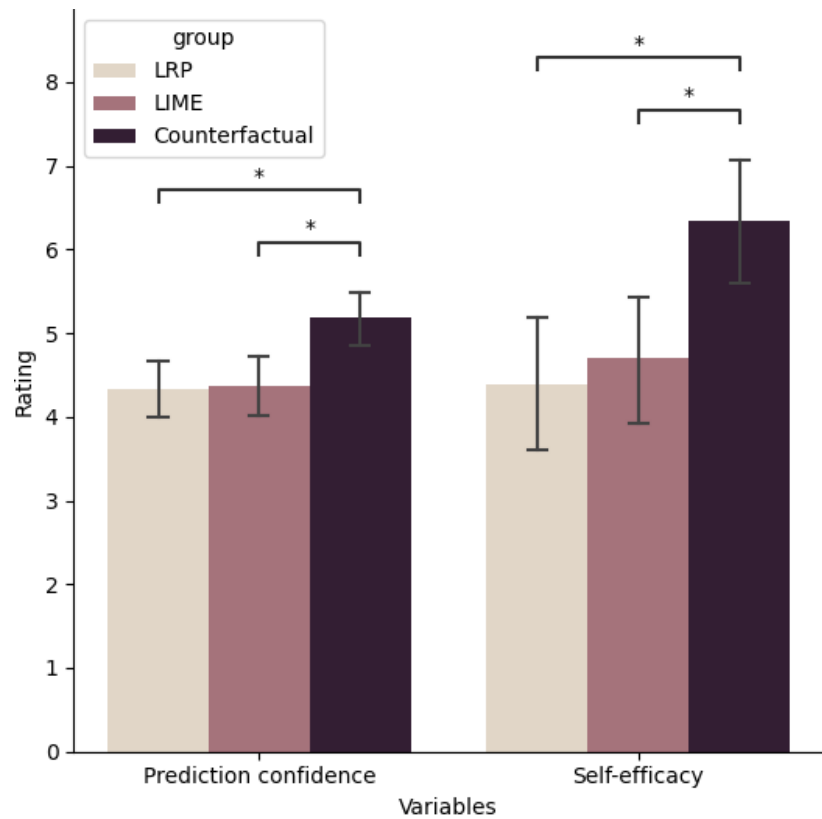


Figure 29.8: We found significant differences regarding self-efficacy and general confidence of the participants in their predictions of the AI between the counterfactual condition and the saliency map conditions (LRP and LIME). Error bars represent the 95% CI. * $p < .05$

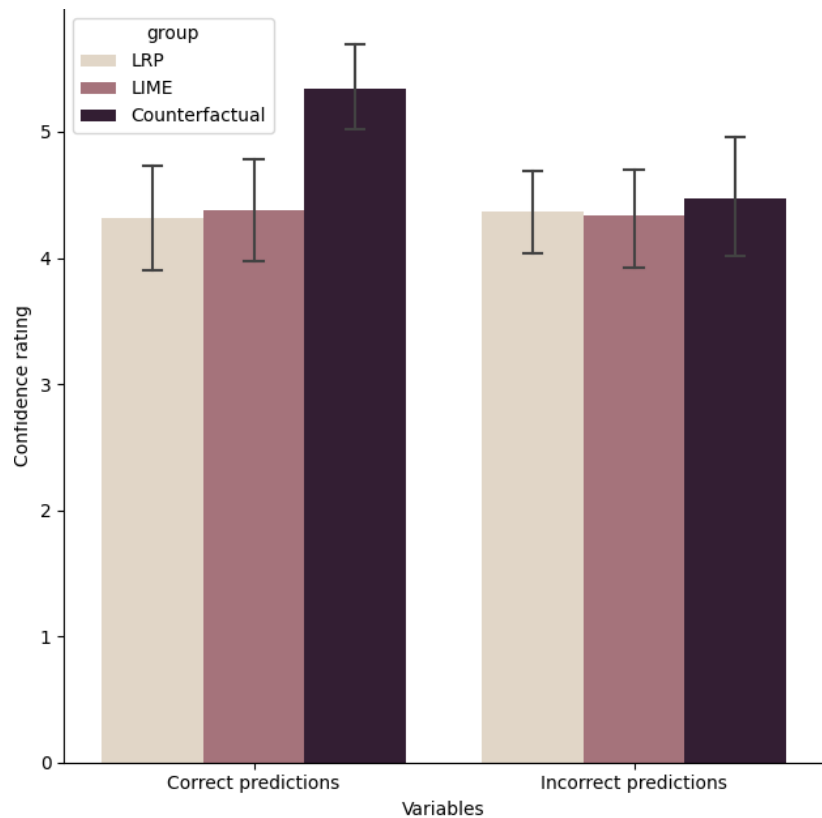


Figure 29.9: Confidence rating of the participants for correct and false predictions. The significant difference between the counterfactual and saliency map conditions is based on the confidence in correct predictions, not incorrect ones. Error bars represent the 95% CI

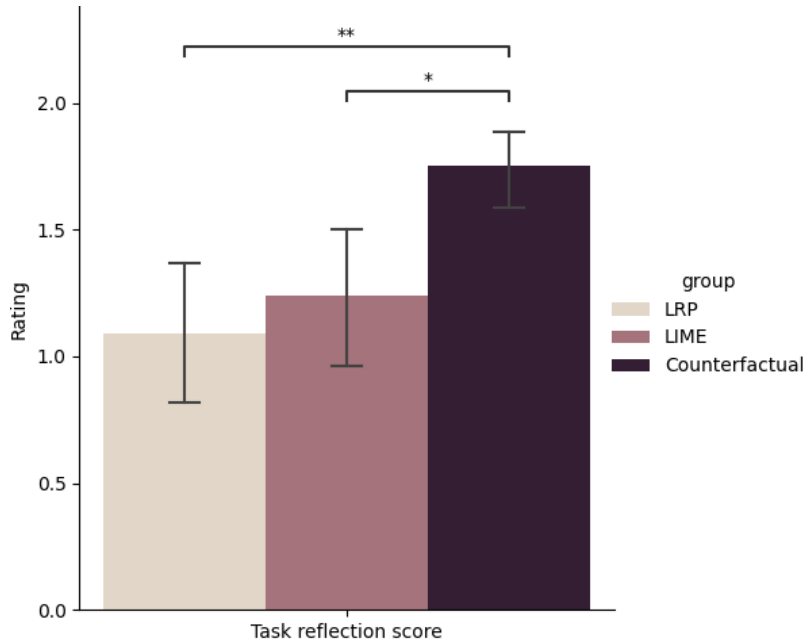


Figure 29.10: Results of the task reflection questions. Error bars represent the 95% CI. * $p < .05$, ** $p < .001$

difference between the three conditions, we used post-hoc comparisons (see Figure 29.10): Participants were asked to describe the AI's reasoning in three different conditions: counterfactual, LRP, and LIME. Out of these, participants created correct descriptions significantly more often in the counterfactual condition compared to the LRP condition, $t(115) = -3.76$, $p < .001$, $d = 0.85$ (large effect) and the LIME condition, $t(115) = -2.97$, $p = .01$, $d = 0.66$ (medium effect).

Impact of XAI Methods on End-Users' Emotional State

We also wanted to investigate whether working with the XAI methods influenced the participants' emotional states. To analyse possible effects, we conducted a one-way MANOVA. Here we found a significant statistical difference, Pillai's Trace = 0.20, $F(6, 228) = 4.26$, $p < .001$. The following ANOVA revealed that the emotion anger, $F(2, 115) = 6.75$, $p = .002$ and relaxation, $F(2, 115) = 9.07$, $p < .001$ showed significant differences between the conditions. Happy showed no significant differences between the conditions, $F(2, 115) = 2.06$, $p = .13$. The post-hoc comparisons⁶ showed the following differences (see Figure 29.11 on the next page):

- **Anger:** Participants in the counterfactual condition felt significantly less angry than in the LRP condition, $t(115) = 3.68$, $p = .001$, $d = 0.83$ (large effect). No differences were found for the LIME condition, $t(115) = 1.83$, $p = .12$.
- **Relaxation:** Participants in the counterfactual condition were significantly more relaxed than in the LRP condition, $t(115) = -4.26$, $p < .001$, $d = 0.96$ (large effect). No differences were found for the LIME condition, $t(115) = -2.12$, $p = .06$ ⁷

6: We used the Holm correction for multiple testing to adjust the p-values

7: This p-value was no longer significant due to the Holm correction.

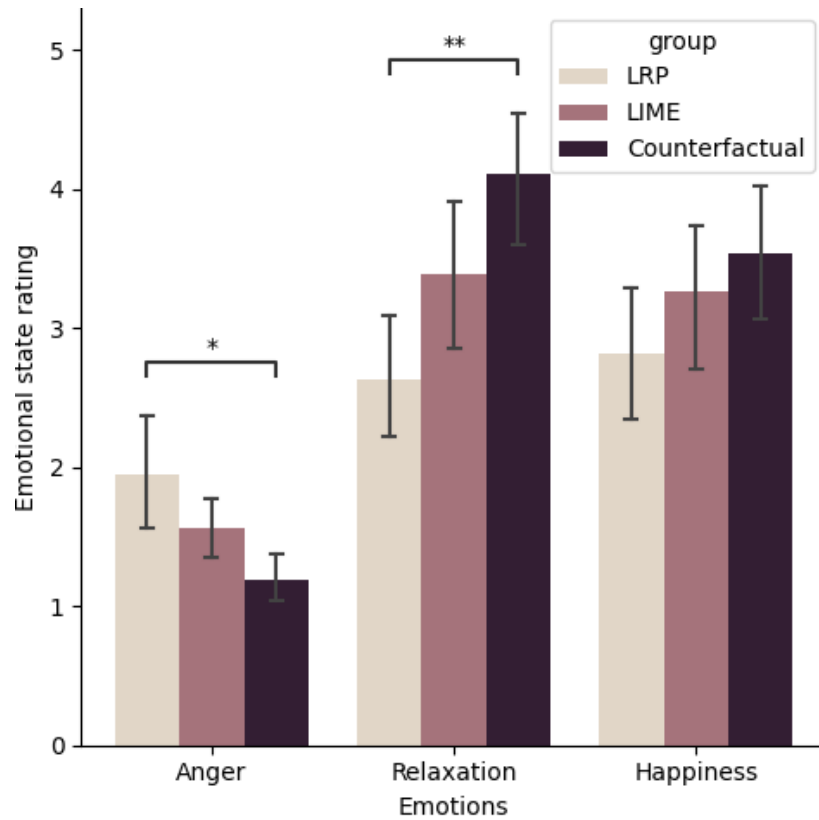


Figure 29.11: Results of the emotion questionnaires. Participants in the counterfactual condition felt significantly less angry and more relaxed compared to the LRP saliency map condition. For LIME, no significant differences were found. Error bars represent the 95% CI. * $p < .05$, ** $p < .001$

29.3.9 Discussion

The study described in the previous sections was conducted to verify our hypotheses. With this in mind, we discuss our results in this section.

Counterfactual Explanations Were More Satisfying

The counterfactual explanation images that were generated by the use of our novel approach provided the participants with significantly more satisfying explanations than both of the saliency map approaches. Saliency map methods like LIME and LRP only show which pixels were important for the AI's decision. The users are left alone with the task of building a bridge between the information of *where* the AI looked at and *why* it looked there. On the contrary, our system's counterfactual explanations directly show, *how* the input image would have to be modified to alter the AI's decision. Thus, the participants did not have to come up with an interpretation of the semantics of important areas by themselves. As the results of our study show, this difference plays a significant role in how satisfying the explanations are to end-users, validating our first hypothesis that participants are more satisfied with counterfactual explanations.

Counterfactual Explanations Support More Accurate Mental Models

Different methods were used to evaluate if the explanation systems allowed the participants to build up an appropriate mental model of the

classifier to test our second hypothesis of creating more accurate mental models with the help of counterfactual explanations. First, the participants had to do a prediction task of 12 images, where they had to decide if the AI would classify each of those images either as *Pneumonia* or *No Pneumonia*. Our results show that the participants were significantly better at performing those prediction tasks when they were shown counterfactual images created by our system than they were when provided with LIME or LRP saliency maps. Again, this advantage is likely caused by the fact that the counterfactual images give more than just spatial information about the regions of importance. The actual decision of the AI was highly dependent on the blurriness of some lung regions. A crucial thing to mention is that the absence of blurriness (i.e., the clarity of X-ray images that do not show lungs that are infected by pneumonia) occurs at similar places where cloudy areas would appear in the case of pneumonia. Thus, the visual highlighting created by LIME or LRP predominantly shows where this distinction between opaque and not opaque lungs is made. However, the information is missing, to which degree the AI thinks there is an opacity in the lung. In contrast, the counterfactual images give this information by increasing or decreasing that opacity respectively. In general, our counterfactual explanations have the most advantage in these kinds of tasks, where the important regions are not distinct for different decisions. Specifically, our approach excels in tasks where the AI's decision is directed by different textural characteristics rather than the position of particular objects in the image. The content analysis of the task reflection strengthens this assumption. Here, participants from the LRP and LIME conditions often referred to certain organs or regions in the image instead of focusing on the key feature of opacity. Examples of this are: "The AI pays attention not to just the lungs but the surrounding areas as well. The Abdomen seems to be an area of focus.", "From the heatmap, I noticed the AI paying attention to the surrounding areas of the lungs, the spine, heart, abdomen, and the armpits often when it predicted pneumonia." and "I think the AI needs to see the green near the bottom of the chest to think healthy lungs."

Counterfactuals Increased Trust

Our results show that counterfactual explanations encouraged the participants to trust the AI system more and partially support our third hypothesis that end-users trust counterfactual explanations more. However, this only became apparent in comparison to LRP, not LIME. This result indicates, on the one hand, that the type of explanation (counterfactual explanation vs feature importance/saliency maps) influences the perceived trust of users. On the other hand, it also shows that even explanations of one XAI type (here: saliency map approaches) are perceived differently by users. This finding is important because it indicates that the type of visualisation (pixel-wise or superpixel-based) also influences the users' trust rating. Our study examined the general influences of three XAI methods on trust. Based on the results, further analyses are necessary. For example, whether there is a correlation between the participant's predictions and the trust rating arises. One interesting observation in our results is that participants in the LIME condition trusted the system similarly to those in the counterfactual condition, even though they did significantly worse in the mental model evaluation. This indicates that

their trust might not be justified. While this is interesting, the question of whether the trust of the participants in the AI system was actually justified needs to be examined more closely in the future.

XAI Has an Impact on End-Users' Emotions

In our user study, we investigated the impact of XAI visualisations on trust and mental models and, for the first time in this domain, the participants' emotional states. With hypothesis five, we assumed that because of the more intuitive and more straightforward interpretation of counterfactual explanations brings end-users in a more positive mood than LRP or LIME. On the same side, it reduces anger regarding non-understandable XAI visualisations. The result shows that XAI influences users' understanding and trust but also impacts users' affective states. Counterfactual explanations promote positive emotions (i.e., relaxation) and reduce negative emotions (i.e., anger). Kaptein et al. (2017) argue in their paper that emotions should be included as an important component of AI explanations (e.g., self-explanatory cognitive agents). Based on our results, we extend this argument by stating that users' emotions should also be taken into account in XAI designs

Counterfactual Explanation Support End-Users' Confidence and Self-Efficacy

Finally, in our sixth and last hypothesis, we assumed that counterfactual explanations strengthen end-users self-efficacy. Our results show that participants were able to correctly assess the predictions of the AI with the help of the counterfactual explanations and were very confident in their judgements. Upon closer inspection, we found that this boost in confidence only stems from the correct participants' predictions. This indicates that they were not overconfident but justified in their confidence. While this is an interesting observation, it needs further investigation. The increase in confidence is also reflected in a significant increase in participants' self-efficacy in the counterfactual condition, compared to LIME and LRP. Already in our work presented in Heimerl et al. (2022) (see Chapter 29 on page 210 - NOVA Study) assumed that the use of XAI could be a valuable support to improve self-efficacy towards AI. This assumption was empirically proven for the first time in our study and contributed to a more human-centred AI.

Limitations

It has to be investigated further how our proposed counterfactual generation method performs in other use cases. We believe that the advantage of our system in this pneumonia detection scenario, to some degree, results from the fact that the relevant information in the images is of a rather textural structure. A further noteworthy observation is that, although the study showed that the produced counterfactuals lead to good results in our chosen end-users task, our system modifies relevant features in a very strong way, i.e., features that are relevant for the classifier are modified to such a degree that the classifier is *sure* that the produced image belongs to the respective other class. As these strong image modifications point out

the relevant features in a very emphasized way, they lead to satisfactory explanations for end-users unfamiliar with the fine details of the problem domain.

29.3.10 Conclusion

In this work, we conducted a user study to compare counterfactual explanations with two state-of-the-art XAI approaches, namely LIME and LRP. As an evaluation use case, we chose the explanation of a classifier that distinguishes between X-ray images of lungs suffering from pneumonia and lungs not infected. The counterfactual approach outperformed the standard XAI techniques in this particular use case. Firstly, the counterfactual explanations that were generated by our system led to significantly more satisfying results than the two other systems that are based on saliency maps. Secondly, the participants formed significantly better mental models of the AI based on our counterfactual approach than on the two saliency map approaches. Also, participants had more trust in the AI after being confronted with the counterfactual explanations than with the LRP visualisations. Furthermore, end-users who were shown counterfactual images felt less angry and more relaxed than those who were shown LRP images. All in all, we showed that our approach is promising and offers excellent potential for being applied in similar domains. However, it has to be investigated further how the system performs in other use cases. The advantage of our system in this specific scenario results from the relevant information of the images being of a rather textural structure (e.g., opacity). Thus, as provided by LIME and LRP, raw spatial information about important areas does not carry enough information to understand the AI's decisions. Therefore, we recommend applying our approach in similar use cases where relevant class-defining features are expected to have a textural structure.

29.4 Summary Medical Decision Support Experiments

The use of XAI in the domain of medical decision support represents an important and challenging application scenario for XAI, as the target audience is end-users who have little or no knowledge of AI but, at the same time, have to make decisions in a vulnerable area (i.e., diagnosis, monitoring of patients). The two studies carried out indicate that...

- ▶ ... end-users form mental models about a CNN which do not necessarily correspond to the actual system performance and thus overestimate the performance of the system. (*NOVA Study*)
- ▶ ... a "the more, the better" approach to presenting explanations does not lead to more accurate mental models (*NOVA Study*)
- ▶ ... XAI visualisations (LIME) can help detect CNN learning errors. (*NOVA Study*)
- ▶ ... the type of XAI visualisation significantly impacts how well people can predict the decision of CNN. (*Pneumonia Study*)

- ... compared to LIME and LRP, counterfactual explanations not only have a positive impact on the end-users perception of an AI system (i.e., trust, self-efficacy, and confidence) and users' performance but also affect users emotionally in a positive way (i.e., decreasing anger and increasing relaxation) compared to LRP. (*Pneumonia Study*)

8: based on the publication dates of the papers

In addition to these research findings, the novelty of the presented studies should also be emphasised⁸: The NOVA Study was one of the first studies that investigated different types of explanations during a CML task with a software system, including a real CNN. In the Pneumonia Study, different standard visual explanation algorithms (LIME, LRP) were compared with a counterfactual explanation approach. For the first time, users' emotions regarding and XAI system were investigated in this context.

IX. CONCLUSION

30

Contribution

30.1	Conceptual Contribution	240
30.2	Empirical Contribution	241
30.2.1	XAI Design Related Research Questions . . .	241
30.2.2	End-User Related Research Questions	244
30.3	Technical Contribution	248

The following is a summary of the contribution of this dissertation. Thereby, the *conceptual*, *empirical*, and *technical* contribution will be further clarified. The conceptual contribution summarises the interdisciplinary HC-XAI concept presented in this dissertation. The empirical contribution answers the research questions formulated at the beginning regarding end-user and XAI design. The results and findings from the user surveys and experiments serve as the basis for this. Finally, the technical contribution provides an overview of the AI systems developed for these experiments and the implementation of XAI in these systems.

30.1 Conceptual Contribution

This dissertation follows an interdisciplinary approach and uses concepts from HCI and psychology, AI architectures (white-box and black-box approaches), and XAI (i.e., explainable model, explanation interface). From these interdisciplinary research areas, the following aspects were considered and incorporated into the HC-XAI concept:

- ▶ **Artificial Intelligence** In Chapter 4 on page 16, essential knowledge about the functioning of knowledge-based white-box and data-driven black-box approaches was given. Understanding how these systems work is necessary to understand the opportunities and limitations of the respective architecture (e.g., the black-box character of CNN). In addition, a basic understanding is required to understand the XAI methods used in these systems. I want to point out that rule-based systems and CNN are only two examples of AI systems. Many other white-box and black-box approaches (e.g., semantic nets and reinforcement learning) have not been discussed in detail in this dissertation. I used rule-based systems as an example of white-box approaches and CNN for black-box systems for this dissertation's technical and empirical part.
- ▶ **Explainable Artificial Intelligence** In Chapter 5 on page 22, the basics of the relevance of explanations in human life were first explained before the thesis turned to XAI. Here again, after a definition of the term and adjacent and overlapping terms, beneficial and harmful functions of explanations were explained. Subsequently, concrete XAI methods for rule-based systems and CNN were presented as they serve as examples for white-box and black-box AI systems. Then, different approaches were presented in the Related Work Chapter 9 on page 47 to help design XAI. Finally, in Chapter 10 on page 52, related work was presented to investigate XAI for the three AI purpose scenarios covered in this dissertation.
- ▶ **Psychology** In Chapter 6 on page 36 an introduction to the psychological constructs of mental models, trust, self-efficacy & cognitive workload, and emotions were given. In addition, the influence of

these constructs in the interaction with different AI systems was illustrated by presenting relevant studies.

- **Human-Computer Interaction** The persona approach for XAI, presented in detail in Chapter 15 on page 85, has its origins in HCI. In this dissertation, the insights gained from using personas in established HCI research serve as a basis to further develop the persona approach for design in HC-XAI. The result is a persona template that can be used for HC-XAI studies. Using such a template was then exemplarily applied in the context of three end-user surveys, which differed in their objectives and complexity. The three surveys showed that the developed persona concept for HC-XAI can be used flexibly and for different use cases.

These domains provide the basis for a conceptual model of an HC-XAI system. It consists of three components: (a) an AI system (i.e., white-box, black-box, hybrid approach), (b) explanation design (explainable model and explanation interface), and (c) user evaluation. The interdisciplinary concept serves as a basis for a five-step approach, which should support researchers and practitioners in the design of HC-XAI. In this dissertation, this step-by-step approach was presented conceptually on the one hand and empirically investigated for three concrete AI purposes from the areas of cooperation & collaboration, education, and medical decision support on the other. Different levels of interactive XAI were presented. These ranged from low interactivity by presenting XAI visualisations in the Pneumonia Study and the NOVA Study to using a medium interactive XAI, using a virtual agent, communicating to users and presenting XAI visualisation in the Gloria and Museum Study. The most interactive XAI presented in this dissertation was investigated in the cooperation and collaboration experiments in the VR-Robot Study and the Conversational AI Study, where end-users were presented with explanations while solving a task with the AI. Thus, the concept did not only remain theoretical, but possible empirical implementations were also illustrated. The findings from these investigations will be discussed in the next section.

30.2 Empirical Contribution

The empirical contribution based on the interdisciplinary HC-XAI concept is the core of this dissertation. Experiments with XAI systems were conducted in six user studies for three AI application scenarios. In addition, three surveys were conducted to investigate end-user attributes and build personas based on the results. For this purpose, six research questions were pursued, divided into questions for the empirical investigation regarding the XAI design and the end-users.

30.2.1 XAI Design Related Research Questions

The XAI-related research questions were:

- **RQ-XAI-1:** What are the *requirements and demands* for explanations in AI scenarios depending on the context of use?

- **RQ-XAI-2:** Which *aspects of an explanation* (i.e., type, content, interface) are helpful to end-users to *appropriate trust* and build *accurate mental models* about AI systems?
- **RQ-XAI-3:** How does XAI impact the *cognitive load*, *self-efficacy*, and *emotions* of end-users?

RQ-XAI-1: Requirements and Demands In the technical realisation chapter, I described the designs of the used AI systems and the related XAI methods that we used for our studies. The designs were related to the respective application purpose: We investigated cooperative/collaborative interaction between humans and machines in the VR-Robot Study and the Conversational AI Study, presented in Chapter 27 on page 160. In these two studies, the focus was on the interaction between humans and machines. Therefore, verbal explanations were used for both scenarios. For both studies, we based tasks and challenges on industrial contexts. In the VR-Robot Study, we investigated the influence of explanations in the context of robot errors. Here, an industrial robot caused an error in a sorting task it performed with a human user. In the Conversational AI Study, we investigated the influence of explanations in a collaborative game that required humans and AI to work together to win. Collaborative tasks inspired the game as they occur in teams in control centres (Schulze Kissing & Bruder, 2016). In both studies, we used rule-based AI systems that allowed us to develop initial prototypes that could be used in the studies.

In two experiments tailored to the education application domain (Gloria Study and Museum Study) that are described in Chapter 28 on page 190, we investigated the effect of integrating a virtual agent in an HC-XAI design. Similar to the studies presented for cooperation & collaboration, an interactive HC-XAI approach was chosen to communicate explanations to end-users. The virtual agent had the role of a teacher, explaining the decisions of a CNN with the help of XAI visualisations generated by LIME. For this purpose, the agent explained the classification decisions of a CNN for keyword recognition, a task that is challenging for users who are not familiar with interpreting spectrograms. This setup was suitable for explaining to people the decisions of CNN. In the context of a large-scale ML-show at the Deutsches Museum in Munich, we communicated this to more than 2000 interested museum visitors for one year (of which 47 successfully participated in our study survey). Our interactive ML show took up the museum's idea that knowledge must be made tangible and experienceable. In our ML show, the participants were allowed to participate actively and generate voice recordings for the data set for a CNN, test the trained CNN with their voice recordings at the end and have the explanations of the virtual agent presented to them. Besides the challenge of developing an interactive concept for museum visitors, which we successfully met, we also managed to adapt the ML show to the needs of the museum's staff, who had little knowledge of how to use CNN and were also end-users. As a result, they could present the ML show almost maintenance-free for a year in the museum.

Regarding medical decision support (see Chapter 29 on page 210), we investigated the influence of explanations in a medical context in two scenarios. In the NOVA Study, users interacted with the NOVA software

to gain insight into CNN's emotion recognition classifications and explanations (i.e. LIME, confidence values). In the Pneumonia Study, end-users were asked to judge the classifications of a CNN for pneumonia detection based on X-ray images. They were shown different visual explanations (i.e., LRP, LIME, counterfactuals). Both studies investigated the influence of the type and content of explanations. In particular, we focused on which type of explanation helps end-user make correct judgments about the AI systems they use. In contrast to the other studies presented in this dissertation, in the NOVA Study and the Pneumonia Study, the design of the XAI system was less interactive. Instead, we investigated whether type and content help end-users understand an AI's decisions better and thus assess the system better. This characteristic is very relevant for users to make correct assessments with the help of the decision support system. Our studies are one step in creating more human-centered medical decision support systems explanations, thus more beneficial for end-users and, in a future step, for domain experts (e.g., physicians).

RQ-XAI-2: Trust & Mental Models The studies outlined here examined various aspects regarding the content and type of explanation and their influence on trust and mental models. It was found that explanations helped to form accurate mental models. For example, users detected and named CNN errors using LIME-generated visual explanations in the NOVA Study. However, numerical explanations in the form of confidence values did not help promote accurate mental models from CNN. The usefulness of visual explanations, especially counterfactuals, was demonstrated in the Pneumonia Study. With them, users performed better in task prediction and reflection tasks than with LIME or LRP-based visualizations. In the Conversational AI Study, we showed that trust and explanation satisfaction were not influenced by the explanation type (i.e., personal vs impersonal style of explanations). However, users attributed human characteristics to the AI that were contrary to their described capabilities of the system (e.g., dialog partner capability: only understands simple sentences vs dialog partner characteristic: intentionally tries to deceive me). Regarding trust in AI systems, the VR-Robot Study showed that explanations alone are insufficient to restore trust after a robot error. However, users indicated that explanations helped assess whether to trust or distrust the robot. We found a significant influence in trusting CNN for keyword recognition in the Gloria Study. The more natural (i.e., virtual embodied) a virtual agent presented the CNN's explanations, the more users trusted the CNN.

RQ-XAI-3: Cognitive Load, Self-Efficacy, & Emotions In some studies, we also investigated the influence of explanations on users' cognitive load (NOVA Study), self-efficacy (VR-Robot Study, NOVA Study, Pneumonia Study), and emotions (VR-Robot Study, Pneumonia Study). While we found no effect of explanations for cognitive load, self-efficacy, and emotions in the NOVA Study and VR-Robot Study, a significant effect was found for self-efficacy and emotions in the Pneumonia Study. End-users presented with counterfactual explanations had a significantly higher score in self-efficacy than in the LIME and LRP condition. Further analysis showed that users in the counterfactual condition also had higher prediction confidence (i.e., user prediction of the AI's decision) than in the

LIME and LRP condition. A closer look showed that this significant effect was due to users' confidence in their correct predictions. This means that users were significantly more confident in their correct predictions about the model than users in the LIME and LRP condition. Moreover, in addition to the influence of counterfactual explanations on participants' prediction performance, we found an effect on their emotions. Users in the counterfactual condition indicated they felt more relaxed and less angry than in the LRP condition. No significant differences were found for the LIME condition.

Summary In summary, the empirical results of this dissertation support the theoretical assumptions made by researchers such as Miller et al. (2017) for DNN and Clancey (1983) for rule-based systems: not all presented explanations turn out to be helpful explanations for users. This dissertation shows that visual explanations can be more beneficial for end-users than numerical explanations (i.e., confidence values). Still, there are differences in the usefulness of the visual explanation concerning the algorithm used (i.e., LRP, LIME, counterfactuals) as well as the application domain (e.g., benefit from XAI in the Pneumonia Study, but not in the NOVA Study). Linguistic explanations that provide simplification in terms of less algorithmic *understanding* communicated to end-users using explainable agents (Anjomshoae et al., 2019) represent a promising option for the future design of XAI systems. For example, the Gloria Study showed that the more human-like/natural the explanations are designed using a virtual agent, the more trustworthy they appear. Our Museum Study shows that such a concept can be used to teach large groups of end-users about how a CNN works, thus promoting AI and data literacy. Our pilot experiments in the Conversational AI Study and the VR-Robot Study also showed that when users are not directly confronted with explanations in an interaction situation, they want the most complete and comprehensive explanations possible. Short explanations are preferred in interactions where other primary goals (e.g., completing a task without error or under time pressure) are in the foreground. Extensive explanations such as the ones in the VR-Robot Study, which in addition to a verbal explanation, also included a solution to prevent the robot error in the future, did not help to restore trust in the robot. Accordingly, the research field of HC-XAI needs to focus on personalized explanations adapted to user groups. Here, finding the appropriate balance between complete explanations of AI and simplified explanations in the sense of *understanding* is a challenge. The investigation of personal attitudes and opinions about (X)AI of users was the focus of the second part of the empirical investigation, the results of which are summarized below.

30.2.2 End-User Related Research Questions

In addition to the empirical investigation of XAI in studies with end-users, the opinions and attitudes of 200 users towards (X)AI in three application scenarios regarding *companies, education, and mobile health* were investigated in three surveys (see Chapter 15 on page 85). These surveys differ in their objectives and complexity:

- The less complex survey investigates **end-users in education** (see Chapter 16 on page 92). The literature research shows few studies

on using XAI in the context of education, and even fewer works deal with the end-user in such application areas. Therefore, the survey aimed to get the first insight into typical users' attitudes towards (X)AI in education. The personas developed from the survey can now serve as a basis for further research to establish HC-XAI for this user group.

- The survey of **end-users in companies** represented the next level of complexity (see Chapter 17 on page 99). Here, employees from German companies were asked about the use of (X)AI in their companies. Compared to the survey in the educational context, the goal here was more tangible: The survey asked about concrete applications of AI in the company, further training opportunities in the area of AI for employees, and plans for the use of AI. However, only general questions were asked about XAI in this survey.
- The last survey asked **end-users about a specific mobile health application** (see Chapter 18 on page 108). This survey is the most complex in this dissertation. For one particular (but fictional) application (i.e., an app that detects a person's stress level), users were asked how they would like the application to explain the classification result (i.e., stressed/not stressed). For this, preferences for different aspects of XAI were asked: different types and content of explanations and distinct explanation interfaces were rated by end-users. This type of survey and the resulting personas serve as concrete recommendations for the design of HC-XAI systems.

All three surveys served to address the requirement to create more user-centric explanations, as called for by researchers like Conati et al. (2021), Miller (2019), and Schneider and Handali (2019).

To this end, three research questions were addressed in the dissertation:

- **RQ-User-1:** What are end-users *knowledge, experiences, and attributes* towards AI and XAI? What do they expect from such systems?
- **RQ-User-2:** How are end-users demographic characteristics (e.g., age, educational background) related to the *knowledge, experiences, and attributes* toward AI and XAI?
- **RQ-User-3:** Which *personas* for human-centered XAI can be derived from empirical data about end-users? How do they *differ* regarding the application scenario?

RQ-User-1: Knowledge & Attributes In all three application scenarios, end-users have similar knowledge about AI and XAI. Most users have already heard of AI, which, in contrast to the European Commission's study in 2017 (European Commission, 2017), shows that the topic of AI is a socially discussed topic that also affects end-users who have no specific expertise in computer science or AI. On the other hand, XAI is mainly unknown to users, very few of whom have heard of the topic. This is not surprising due to the young discipline of XAI for DNN, which gained momentum around the 2015s. In companies, XAI already seems to be a more critical topic. 62% of the survey participants ($N = 50$) stated that they had already heard of XAI (in comparison: Education: 24%, Healthcare: 4.3%). Participants in the company survey reported the chances and

risks of using AI in the company. Particularly an increase in productivity and flexibility was mentioned positively. Risks like financial aspects and employees' qualifications were mainly stated. The employees were able to name and describe concrete AI applications that are used in their company. Employees generally perceive the AI technology used in their companies as useful, reliable, operable, comprehensible, and transparent. In the education survey, users stated that they would reject the usage of AI in education. At the same time, they were more in favour of AI usage in the household, transportation, safety, and care sector. In the healthcare survey, it became apparent that while users have a positive attitude towards (X)AI, they prefer a data-based app for stress recognition via their smartphone. The main reason for this was privacy issues.

RQ-User-2: Relationships In the companies survey, we found no correlation between the age, gender, or educational background of company employees with the perception of AI technology used in their companies. Regarding employees' roles in the company, we found a significant correlation with the perception of AI technology used in the company. Furthermore, we found that a higher position in the company leads to a less positive attitude towards AI. At the same time, the educational background seems to positively impact the knowledge and attitude towards XAI. In addition, we found that participants who had a positive general attitude towards AI perceived the AI used in the company as positive. The answers to the education survey revealed that while the technical affinity of people was related to their age and gender, no correlation was found regarding their knowledge of AI, indicating that the knowledge about AI has not to be associated with a general technical affinity. In the healthcare survey, we found that the main drivers for investing time to interact with an explanatory system are users' positive attitude towards AI and XAI, their personality (i.e., high values in Openness to Experiences and Extraversion), and their technical affinity, where users with higher values in technical affinity tend to invest more time in understanding an explanation. In addition, users stated that they would spend more time with a stress recognition app when it provides an interactive interface for explanations (e.g., by asking questions to the app).

RQ-User-3: Personas To create prototypical users, I used the persona approach, a well-known method in HCI (Castro et al., 2008). This approach could also benefit AI design (Holzinger et al., 2022). I adapted the persona approach for the context of HC-XAI (see Chapter 15 on page 85). The resulting template addresses the aspects of *context*, *problem*, *needs*, *existing solutions*, and *goal* of (X)AI. This template was in the next step used to collect survey data to design personas about end-users of XAI. The personas derived from the three surveys differ in the context of AI used, the problems, needs, and the goal of XAI.

Based on the results of the companies survey, I created the persona of the **company leader Wolfgang**. He faces the problem that his employees have to be trained to handle the AI software used in the company (*problem*). Therefore he wants to support his employees to work efficiently with an AI system (*needs*). The *goal in the context of XAI* is to help Wolfgang to offer appropriate (X)AI in-house training courses for his employees based on the already existing training program *existing solutions*. Whether XAI

can be beneficial as part of a training program would be an interesting research question for the future and is not answered in this dissertation.

Based on the education survey results, two personas were described: the **critical stakeholder Regina** and the **uninformed stakeholder Dirk**. Dirk represents a young user with little technical affinity and no knowledge about XAI. Nevertheless, he is confronted with AI in his private life (*problem*). To understand the AI he is using is relevant for him to assess its risks and benefits (*needs*). To support him in gaining knowledge about (X)AI, educational methods that include his lack of technical affinity and age have to be considered (*goal of XAI*). Here, existing educational concepts using non-digital teaching material could be a low-threshold entry point (*existing solution*). Regina represents a user who uses AI in her household but is also unaware of the topic of XAI. While she actively uses an AI-based vacuum cleaner in her home, she does not perceive AI education as an essential topic (*problem*). While she rates her AI knowledge as experienced, she needed to learn about the benefits of XAI (*needs*). Giving her an overview of the benefits and limitations of AI and how XAI addresses this issue could be one *goal of XAI* in this scenario. For this, already existing (online) courses could be used (*existing solutions*). For both personas, courses and teaching materials are necessary to impart XAI knowledge. In my description of the personas, the responsibility for perception remains with them. However, this should differ from the aim of (X)AI education. Educational institutions and schools should include education about (X)AI in their courses. This aspect is not within the scope of this dissertation but should be mentioned here as a note for the sake of completeness.

We also derived three personas based on our mobile health survey data: the **power user Anni**, the **sceptical user Michael**, and the **casual user Karl**. Anni represents an extraverted person who uses AI-based health apps. In doing so, she is interested in gaining more insights into how the AI comes to its decision (*problem*). While she is already using wearables to track her fitness level (*existing solutions*), she is interested in a mobile health app that can answer questions about the classification that was made (*needs*). To address this, XAI needs to offer an interactive interface (*goal of XAI*). Michael is similar to Dirk, not technically affine but using AI technology daily. He has privacy concerns because of not knowing what happens with the collected data about him (*problem*). He is interested in gaining knowledge about his data usage (*needs*). The apps he is using already give, in regards to the GDPR, information about that (*existing solutions*). XAI could help him not only understand which data the app is using but which impact this data has on the classification results of the app (*goal of XAI*). Karl is prototypical for a casual user. For him, the intuitive usage of technology is essential. Therefore, he likes to see information at a glance and does not want to waste time with an app that explains unnecessary details (*problem*). Based on his preferences, he needs a mobile health app that only shows explanations he wants in a good overview (*needs*). Inspiration for the design could be popular interfaces of other apps he is already using (*existing solutions*). To satisfy Karl, the HC-XAI mobile health app should have a visually appealing and easy-to-understand interface (*goal of XAI*).

Summary The personas in this dissertation represent different genders, ages, technical affinity levels, and attitudes towards AI and XAI. The goal of XAI can be subsumed in the usage of XAI as a tool to gain knowledge about AI in general or a specific AI application. The driver of this goal can be *intrinsic*: e.g., personas of Anni and Michael, who want to understand their AI-based apps, or *extrinsic*: Regina and Dirk, who should gain AI competence; Wolfgang, who wants his employees to gain AI competence to work more efficiently and effectively with AI technology in his company. How intrinsic and extrinsic motivation plays a role in the impact and perception of XAI still needs to be answered.

30.3 Technical Contribution

This dissertation provides different levels of interactive XAI. The technical contribution is characterised by using current black-box AI methods (see Chapter 21 on page 132) and classical white-box AI approaches (see Chapter 20 on page 125). For these AI systems, XAI was designed to be oriented to the requirements of the application context and based on user needs. Furthermore, the systems used for the experiments in this dissertation are, characterised by using executable systems in all user studies. The implemented explanations in these HC-XAI systems vary in content, type, and interface. Intrinsic explanation procedures, as well as post-hoc explanations, are used.

- **Post-hoc Explainability:** Post-hoc methods were used in this dissertation to generate visual explanations that highlight the relevant areas in an image (LRP, LIME) or generate images that represent contrastive examples (i.e., counterfactuals) to the classification. In addition, numerical approaches (i.e., confidence values) were used. These different types of explanations were compared or combined with an interactive interface (i.e., virtual agent) to investigate the impact of such systems on end-users.
- **Intrinsic Explainability:** Rule-based systems are characterised by an intrinsic explainability (Molnar, 2019). Here, the explanations arise directly from the design of the rule-based system (e.g., the if-else structure). They were communicated verbally, investigating their impact in cooperative & collaborative scenarios.

The explanations used in this dissertation vary in *explanation content* by providing answers to the questions: “Why?”, “Why not?”, “How?”, “What?”, and “What if?”. In addition, different *types of explanations* are used in the experiments presented (i.e., text, speech, visual, numerical). Regarding the *interface of the explanation* (i.e., how is XAI presented to the user), we implemented for the Gloria Study and the Museum Study an HC-XAI system that combines visual explanations generated by LIME with a virtual agent. We used this implementation to investigate in the two studies whether the personification of an AI system with the help of a virtual agent influenced end-users perception. Combining a virtual agent with XAI visualisation methods for CNN was a new way to communicate with end-users about black-box decisions. Besides our in-the-lab implementation, we successfully test this approach in a vast museum set up to educate big groups of end-users about CNN and XAI. Furthermore, for the two cooperation & collaboration studies, we used

two *explanation interfaces* for communicating with the participants of our studies via text. This was done during the machine and end-user were conducting a task together. The challenge was to investigate the impact of explanations during a cooperative or collaborative task.

31

Limitations

In this dissertation, the impact of explanations on end-users trust, mental models, self-efficacy, cognitive load, and emotions have been studied. But there are other relevant variables in XAI design. For example, aspects such as privacy and fairness were not the focus of the studies but are also goals of XAI that need to be considered when designing explanations (Barredo Arrieta et al., 2020).

I also presented three user surveys to create prototypical XAI users (i.e., personas). These personas were developed based on the results of the gathered survey data. Nevertheless, the found correlations do not imply a causal relationship. In addition, the group of respondents needs to be more representative, and the questionnaire mainly asked Western users. Cultural differences were not considered. Future research should further expand the persona approach by including diversity. However, the developed personas highlight different needs of users depending on user attributes (e.g., age, gender, attitude towards AI) and the application scenario (e.g., security concerns in the mobile health survey). These findings can serve as a valuable basis for user studies in application-grounded scenarios. Holzinger et al. (2022) point out that human-centered research should investigate all possible stakeholders (e.g., decision-makers, domain experts). While the dissertation focuses on end-users, further research should also focus on the needs and attitudes towards (X)AI of other stakeholders (e.g., for healthcare: domain experts like physicians) to develop HC-XAI that serves different user groups.

The scenarios studied in this dissertation are very application-specific. Every application scenario presents other challenges that XAI has to deal with. For example, we investigated collaborative & cooperative scenarios where users and AI work together and where AI errors occur, scenarios where end-users had to handle demanding tasks like spectrograms, and more easy ones like facial emotion recognition. The dissertation addresses these challenges and their effect on users. The different scenarios and explanation designs reduce the generalisation of the findings. However, creating human-centred AI is precisely about developing XAI solutions adapted to the users and the application scenarios. Creating one XAI system for all AI possible application purposes and domains seems less likely and reasonable at this point. Instead, this dissertation presents a general, interdisciplinary HC-XAI concept. Based on this, three end-user surveys of various complexity and six experiments for different AI purposes were successfully conducted. Therefore, the HC-XAI concept presented in this dissertation can be used meaningfully for various XAI application contexts and purposes.

Except for the VR-Robot Study, we used explanations coupled with the respective AI system. For example, the explanations for DNN were generated using the XAI methods of LIME, LRP, and counterfactuals. In contrast, for the Conversational AI Study, the explanations were integrated into our rule-based system. Wick and Thompson (1989) emphasise that decoupling an AI system from the explanation module could result

in an even better fit for end-users, as the explanation module would be more independent from the AI system and thus more flexible.

32.1 Interactive HC-XAI . .	252
32.2 HC-XAI for Application- Grounded Scenarios .	253
32.3 Ethical Considerations	254

What should future research on HC-XAI look like? The following sections provide possible pointers based on the results of this dissertation.

32.1 Interactive HC-XAI

This dissertation presented the first empirical studies investigating interactive HC-XAI by mediating XAI in natural language communication (i.e., VR-Robot Study, Conversational AI Study), allowing users to engage with the AI system independently (i.e., NOVA Study) or by presenting speech and XAI visualisations to end-users in combination (i.e., Gloria Study, Museum Study). The feedback from the participants strongly indicates a desire for interactive HC-XAI:

- ▶ Clickable areas in visual explanations
- ▶ Possibility to ask (back) questions
- ▶ Tentative communication: the more natural, the better
- ▶ Interested not only in explanations but also in solutions, e.g., to avoid mistakes of the AI in the future
- ▶ Personalised explanations are preferred

Combining different modalities to communicate explanations to end-users is a promising way forward. Integrating a virtual agent in the explanation design, as we did in work presented in this dissertation, is one step into a multimodal XAI design, like Anjomshoe et al. (2019) demands. A possible next step in this design is described by D. H. Park et al. (2018). In their approach, they combine visual question answering (e.g. “Is there a firefighter in the picture?” -> “Yes”) with a textual justification (e.g., “...because the person is wearing a firefighter uniform”) and visual pointing (e.g., highlighting the uniform on the image). In addition, such an interactive XAI approach could be combined with the persona approach for XAI presented in this dissertation. This would help to select the appropriate modalities for the respective user group.

Researchers, as well as society and politics, need to think about what role AI and, thus, XAI should play in our lives. Chromik and Butz (2021) show what roles interactive XAI can play. For example, XAI interaction as information transmission focuses on transporting complete and correct information from sender to receiver. Another role is XAI interaction as a dialogue like I presented in the studies regarding cooperation & collaboration. As the results of this dissertation indicate, there is no generally suitable explanation interface; it depends on the user group, the application scenario, and the goal XAI should fulfil. Nevertheless, the insights gained in this dissertation support the idea that interactive XAI interfaces could be beneficial to end-users. Future research can use my work as a starting point to explore new approaches with the fast development of current AI systems or to investigate different user groups (e.g., domain experts). The question of what role interactive XAI should play is connected

to what role AI should play in our lives. This is about the general question of how we want to use AI. Shneiderman (2020c) points out that some see AI as a partner. He advises against it, pointing out that “computers are not people and people are not computers.” (Shneiderman, 2020c, p. 113). This view leads to the argument that AI should be seen as a tool that should not act human-like (Shneiderman, 2020c). Instead, AI should have well-designed user interfaces that allow users to maintain control and make sense of the tool (Shneiderman, 2020c). Future research should clearly define the role of AI in its application to create well-designed XAI user interfaces.

32.2 HC-XAI for Application-Grounded Scenarios

As outlined in the Introduction chapter, this dissertation focuses on a human-grounded evaluation oriented towards the taxonomy of Doshi-Velez and Kim (2017). For this purpose, user studies were developed to investigate the impact of XAI on end-users. These studies were designed as *simple tasks*, which were characterised by a controlled experimental setup and therefore had a higher internal validity than *real tasks* in application-grounded scenarios. The findings serve as a basis for application-grounded scenarios, as proposed by (Doshi-Velez & Kim, 2017).

Cheng et al. (2021) show a possible design approach that involves experts already in the design of the XAI system. Cheng et al. (2021) developed an XAI system called VBridge, which predicted complications after cardiac surgery. They conducted a structured interview with six clinicians to assess their doubts and requirements for an XAI system. For this, they demonstrated a low-fidelity prototype of the XAI system. This first demonstrator was developed into a high-fidelity prototype over the next three months. They showed the current status of the prototype to the participants every week to incorporate feedback iteratively. The final prototype was presented to two of the clinicians. Both clinicians understood the explanations generated by the system and were also able to successfully use the system to initiate personalised treatments for the two use case patients. Also, in the interviews, the clinicians indicated that VBridge would be an excellent way to help medical students make more accurate diagnoses. Nevertheless, besides quality issues (e.g., quality of used data, visual scalability), cognitive biases could lead physicians to misinterpret their diagnosis by supporting evidence through the XAI system (Cheng et al., 2021).

The work of Evans et al. (2022) investigates the effect of XAI methods in the context of digital pathology. They used a parallel evaluation approach: (1) They created an online questionnaire using five different XAI approaches to generate explanations for images (i.e., saliency maps, concept activation examples, prototype examples, counterfactuals, and trust scores that represent the confidence in the annotations). The images used were nuclei of cells stained with the Ki-67 method. They had these explanations evaluated by 25 pathologists via social network platforms such as Twitter. (2) At the same time, they also conducted 60-minute semi-structured interviews with six board-certified pathologists. Their

results show that participants' social and cognitive biases influence their interactions with XAI systems. For example, they found that clinicians attributed comprehensible causal reasoning to AI. In this way, the clinicians tried to understand the explanations given. For example, inference of causal factors responsible for high and low trust scores was popularly used as a basis for the perceived trustworthiness of results - along the lines of "the AI seemed to have the same difficulties as I do." (Evans et al., 2022, p. 291).

These two examples exemplify application-based scenarios, which are gradually being studied more and more. They indicate that the results found in the experiments conducted in this dissertation (e.g., NOVA Study: end-users transfer their mental model to those of the AI) are transferable to real-world scenarios, meaning that they can be found in application-grounded scenarios with domain experts. However, the studies also show that the design of HC-XAI systems presents researchers with demanding challenges. This dissertation contributes to these research efforts by

- ▶ presenting an interdisciplinary HC-XAI concept that serves as a step-by-step approach for researchers to develop HC-XAI by including stakeholder from the very beginning
- ▶ providing a structure to evaluate XAI regarding their content, type, and interface of explanation while at the same time taking into account technical requirements of white-box and black-box AI systems
- ▶ presenting the findings of three user surveys with 200 people and experiments with 483 participants.

32.3 Ethical Considerations

This dissertation explored the topic of HC-XAI in more detail by presenting an interdisciplinary concept for the design of HC-XAI. This was empirically filled with life through user surveys and experiments. In particular, the experiments focused on end-users mental models, trust, self-efficacy, cognitive workload, and emotions. However, these are only some aspects that play a role in the study of HC-XAI. In addition to explainability, fairness, data protection, privacy, and responsibility also play a major role. Particularly under the great influence of AI on many areas of our lives, as explained at the beginning, we have to be careful with the influence of such systems on users. End-users are critical in using such systems, mainly when vulnerable data is used, as the results of our Mobile Health Survey show. In future research, these aspects should be addressed and investigated to consider privacy and ethical aspects in the design for HC-XAI. It should also be emphasised that researchers in this field have a high responsibility: Their research influences the design of future AI systems. This fact should be kept in mind when designing the studies and interpreting the results. For example, effects such as those obtained in our Gloria Study (i.e., XAI is considered more trustworthy in combination with virtual agents) can be used for harmful applications, such as manipulating end-users. Therefore, researchers should draw attention to this problem and communicate these issues, particularly informing users of the systems about it.

It may still be long until an HC-XAI system can promote values like transparency, trust, fairness, and reliability. But I think it is worth going this way so that all users of AI benefit from it.

X. APPENDIX

Publications & Contribution

33.1 Publications Relevant for This Dissertation

In the following, my contribution for the publications used in this dissertation are described. All papers except of Weitz (2018), André et al. (2021), Weitz, Dang, and André (2022), Weitz, Zellner, and André (2022), and Zellner (2021) were peer-reviewed and published on scientific conferences or in scientific journals/books.

- ▶ André, E., Bauer, W., Braun, M., Dang, C. T., Peissner, M., **Weitz, K.** (2021). Projektbericht: KI-Kompetenzentwicklung bei Sach- und Produktionsarbeit. *Lernende Systeme - Die Plattform für Künstliche Intelligenz*
Own contribution: I defined the research questions and designed the online survey. I did the statistical analyses to answer the research questions. I contributed to manuscript revision and reading and approved the submitted version.
- ▶ Hald, K.*, **Weitz, K.***, Rehm, M., & André, E. (2021). “An error occurred!” - trust repair with virtual robot using levels of mistake explanation. In K. Ogawa, T. Yonezawa, G. M. Lucas, H. Osawa, W. Johal, & M. Shiomi (Eds.), *Proceedings of the 9th international conference on human-agent interaction* (pp. 218-226). ACM. doi: [10.1145/3472307.3484170](https://doi.org/10.1145/3472307.3484170)
Own contribution: I had a central role in planning the user studies (i.e., measurement, statistic evaluation, study design, research questions). I was in charge of conducting the pilot study as an online survey. I did parts of the statistical analyses and wrote large parts of the paper (i.e., pilot study, results, discussion). I was in charge of manuscript revision and reading and approved the submitted version.
- ▶ Heimerl, A., **Weitz, K.**, Baur, T., & André, E. (2020). Unraveling ML models of emotion with NOVA: multi-level explainable AI for non-experts. *IEEE Transactions on Affective Computing* 13(3), 1155-1167. doi: [10.1109/TAFFC.2020.3043603](https://doi.org/10.1109/TAFFC.2020.3043603)
Own contribution: I had a central role in defining the research questions, planning and designing the user study, and planning as well as to conduct the statistical analyses. I did large parts of the study conduction. I wrote large parts of the paper. I contributed to manuscript revision and reading and approved the submitted version.
- ▶ Mertes, S., Huber, T., **Weitz, K.**, Heimerl, A., & André, E. (2022). GANterfactual-Counterfactual Explanations for Medical Non-experts Using Generative Adversarial Learning. *Frontiers in Artificial Intelligence*, 5, 825565. doi: [10.3389/frai.2022.825565](https://doi.org/10.3389/frai.2022.825565)
Own contribution: I helped with choosing the constructs to measure, the selection of appropriate questionnaires, and the evaluation of the study. I wrote certain sections of the paper. I contributed to manuscript revision and reading and approved the submitted version.
- ▶ Schmid, U., & **Weitz, K.** (2022). Künstliche Intelligenz und Psychologie- Von Kognitiver Modellierung bis Erklärbarkeit. In A. Schütz, M. Brand, & S. Steins-Löber (Eds.), *Psychologie: eine Einführung in ihre Grundlagen und Anwendungsfelder*, (pp. 219-231). Kohlhammer.
Own contribution: I wrote parts of the chapter, especially the sections about DNN, XAI, and Human-Centered AI. I was in charge of manuscript revision and reading and approved the submitted version.
- ▶ **Weitz, K.** (in publishing process). Erklärbare Künstliche Intelligenz im Kontext Bildung und Lernen. In C. deWitt, C. Gloerfeld, & S. Wrede (Eds.), *Künstliche Intelligenz in der Bildung*. Springer. doi: [10.1007/978-3-658-40079-8](https://doi.org/10.1007/978-3-658-40079-8)

*Both authors contributed equally to this work

Own contribution: I structured and wrote the whole paper. I was in charge of manuscript revision and reading and approved the submitted version.

- **Weitz, K.** (2021). Vertrauen und Vertrauenswürdigkeit bei sozialen Robotern - Stärkung von Mensch-Roboter-Vertrauensbeziehungen mithilfe Erklärbarer Künstlicher Intelligenz. In O. Bendel (Ed.), *Soziale Roboter* (pp. 309-323). Springer Gabler, Wiesbaden.

Own contribution: I structured and wrote the whole paper. I was in charge of manuscript revision and reading and approved the submitted version.

- **Weitz, K.** (2021). Towards Human-Centered AI: Psychological concepts as foundation for empirical XAI research. *it-Information Technology*, 64(1-2), 71-75. doi: [10.1515/itit-2021-0047](https://doi.org/10.1515/itit-2021-0047)

Own contribution: I structured and wrote the whole paper. I was in charge of manuscript revision and reading and approved the submitted version.

- **Weitz, K.** (2018). *Applying Explainable Artificial Intelligence for Deep Learning Networks to Decode Facial Expressions of Pain and Emotions* (Master's thesis), University of Bamberg.

Own contribution: Based on the given topic, I trained the CNN to investigate different XAI approaches. I implemented the XAI approaches. I wrote the thesis.

- **Weitz, K., Dang, C. T., André, E.** (2022). Do We Need Explainable AI in Companies? Investigation of Challenges, Expectations, and Chances from Employees' Perspective. *arXiv preprint arXiv: 2210.03527* <https://doi.org/10.48550/arXiv.2210.03527>

Own contribution: I defined the research questions and designed the online survey. I did the statistical analyses and wrote the paper. I was in charge of manuscript revision and reading and approved the submitted version.

- **Weitz, K., Schiller, D., Schlagowski, R., Huber, T., & André, E.** (2019). "Do you trust me?": Increasing user-trust by integrating virtual agents in explainable ai interaction design. In C. Pelachaud, J. Martin, H. Buschmeier, G. M. Lucas, & S. Kopp (Eds.), *Proceedings of the 19th ACM international conference on intelligent virtual agents, IVA 2019, paris, france, july 2-5, 2019* (pp. 7-9). ACM. doi: [10.1145/3308532.3329441](https://doi.org/10.1145/3308532.3329441)

Own contribution: I had a central role in planning the user study (i.e., define research questions, study design, statistic evaluation). I did large parts of testing participants. I did the statistical analyses and wrote parts of the paper. I was in charge of manuscript revision and reading and approved the submitted version.

- **Weitz, K., Schlagowski, R., & André, E.** (2021). Demystifying artificial intelligence for end-users: Findings from a participatory machine learning show. In S. Edelkamp, R. Möller, & E. Rueckert (Eds.), *Proceedings of KI 2021: Advances in artificial intelligence - 44th german conference on ai, virtual event, september 27 - october 1, 2021* (pp. 257-270). Springer. doi: [10.1007/978-3-030-87626-5_19](https://doi.org/10.1007/978-3-030-87626-5_19)

Own contribution: I had a central role in planning the user study (i.e., measurement, statistic evaluation, study design). I prepared and visited the museum to introduce the ML-show. I did the statistical analyses and wrote large parts of the paper. I was in charge of manuscript revision and reading and approved the submitted version.

- **Weitz, K.*, Vanderlyn, L.*, Ngoc, T. V., & André, E.** (2021). "It's our fault!": Insights into users' understanding and interaction with an explanatory collaborative dialog system. In A. Bisazza & O. Abend (Eds.), *Proceedings of the 25th conference on computational natural language learning, conll 2021, online, november 10-11, 2021* (pp. 1-16). Association for Computational Linguistics. doi: [10.18653/v1/2021.conll-1.1](https://doi.org/10.18653/v1/2021.conll-1.1)

Own contribution: I had a central role in planning and conducting the user studies (i.e., define research questions, study design, statistic evaluation). I was in charge of running the first pilot study. I did the statistical analyses and wrote large parts of the paper. I contributed to manuscript revision and reading and approved the submitted version.

- **Weitz, K., Zellner, A., & André, E.** (2022). What Do End-Users Really Want? Investigation of Human-Centered XAI for Mobile Health Apps. *arXiv preprint arXiv: 2210.03506* <https://doi.org/10.48550/arXiv.2210.03506>

Own contribution: I structured and wrote main parts of the paper and did all the statistical analyses in the paper. I was in charge of manuscript revision and reading and approved the submitted version.

- Zellner, A. (2021). *Towards personalized explanations in digital health - user-centered explanations for*

*Both authors contributed equally to this work

mobile health applications (Master's thesis). University of Augsburg.

Own contribution: I co-supervised the thesis with Prof. Dr. Elisabeth André. I defined the topic and the direction of research. I supported the student by defining the research questions and providing relevant literature. I monitored the preparation of the online study and conducted the online study via MTurk.

33.2 Other Publications

In the following, other publications that are not part of the dissertation but were conducted or written during the doctorate are listed:

- ▶ Aslan, I., **Weitz, K.**, Schlagowski, R., Flutura, S., Valesco, S. G., Pfeil, M., & André, E. (2019). Creativity support and multimodal pen-based interaction. *2019 International Conference on Multimodal Interaction*, (pp. 135-144).
- ▶ Flutura, S., Seiderer, A., Huber, T., **Weitz, K.**, Aslan, I., Schlagowski, R., André, E., & Rathmann, J. (2020). Interactive machine learning and explainability in mobile classification of forest-aesthetics. *Proceedings of the 6th EAI International Conference on Smart Objects and Technologies for Social Good*, (pp. 90-95).
- ▶ Huber, T., **Weitz, K.**, André, E., & Amir, O. (2021). Local and global explanations of agent behavior: Integrating strategy summaries with saliency maps. *Artificial Intelligence*, 301, 103571.
- ▶ Schiller, D., **Weitz, K.**, Janowski, K., & André, E. (2019). Human-inspired socially-aware interfaces. *International Conference on Theory and Practice of Natural Computing*, (pp. 41-53). Springer, Cham.
- ▶ Petrak, B., Sopper, G., **Weitz, K.**, & André, E. (2021). Do you mind if I pass through? Studying the appropriate robot behavior when traversing two conversing people in a hallway setting. *30th IEEE International Conference on Robot & Human Interactive Communication (RO-MAN)*, (pp. 369-375).
- ▶ Petrak, B., Stapels, J. G., **Weitz, K.**, Eyssel, F., & André, E. (2021). To move or not to move? Social acceptability of robot proxemics behavior depending on user emotion. *30th IEEE International Conference on Robot & Human Interactive Communication (RO-MAN)*, (pp. 975-982).
- ▶ Petrak, B., **Weitz, K.**, Aslan, I., & Andre, E. (2019). Let me show you your new home: studying the effect of proxemic-awareness of robots on users' first impressions. In *28th IEEE international conference on robot and human interactive communication (RO-MAN)*, (pp. 1-7).
- ▶ Weber, K., Janowski, K., Rach, N., **Weitz, K.**, Minker, W., Ultes, S., & André, E. (2020). Predicting persuasive effectiveness for multimodal behavior adaptation using bipolar weighted argument graphs. *Proceedings of the 19th International Conference on Autonomous Agents and Multi Agent Systems*, (pp. 1476-1484).

Activities, Teaching, Awards & Roles

While conducting research for my dissertation at the University of Augsburg, I took on other typical tasks in the academic life of an university. These are listed in the following overview.

34.1 Peer-Review Activities & Editorials

I was a reviewer for several papers submitted to the following journals and conferences:

- ▶ **Journals**

Applied Artificial Intelligence Journal (AAI)
 Journal on Multimodal User Interfaces (JMUI)
 Transactions on Affective Computing (TAFFC)

- ▶ **Conferences**

International Conference of Intelligent Virtual Agents (IVA)
 International Symposium on Robot and Human Interactive Communication (RO-MAN)
 Conference of the Cognitive Society (CogSci)
 International Conference on Human-Robot Interaction (HRI)

- ▶ **Editorials**

Program committee member of the 10th International Conference on Affective Computing & Intelligent Interaction (ACII)
 Program committee member of the Workshop “Functions of Emotions for Socially Interactive Agents” at the 9th International Conference on Affective Computing & Intelligent Interaction (ACII)

34.2 Teaching

34.2.1 Masters Theses

- ▶ **Isabella Kohlbus (2022)**. Ich sehe was, was du nicht siehst: Entwicklung einer interaktiven, erklärbaren Anwendung von Visual Question Answering
- ▶ **Alexander Zellner (2021)**. Towards Personalized Explanations in Digital Health: User-Centered Explanations for Mobile Health Applications
- ▶ **Julia Maria Brenner (2021)**. Voice-centric interaction with wealth technology - Design and integration of a voice user interface into an online wealth-management system

34.2.2 Bachelors Theses

- ▶ **Engelbert Arrosquipa (2021)**. Gamifizierung von Anwendungen des maschinellen Lernens für Endnutzer am Beispiel des MNIST-Datensatzes
- ▶ **Simon Maihöfner (2021)**. Wie fair ist Künstliche Intelligenz? Entwicklung eines Prototyps zur Vermittlung von KI-Wissen an Endnutzer:innen

- ▶ **Tobias Liepert (2021).** Auf dem Weg zu menschenfreundlichen Erklärungen? Verwendung von Konzeptaktivierungsfunktionen zur Erklärung neuronaler Netze zur Gesichts- und Emotionserkennung

34.2.3 Student Projects

- ▶ **Simon Krüger (2021).** Die Informationslandschaft in der COVID-19 Pandemie: Erlauben digitale Medien eine auf Fakten basierende Entscheidung zur Schutzimpfung?

34.2.4 Lectures

- ▶ **Winter term 2022/2023**
Seminar Menschzentrierte Künstliche Intelligenz
Seminar Menschzentrierte Künstliche Intelligenz (Elite-Masterstudiengang)
- ▶ **Summer term 2022**
Vorlesung & Übung Human-Computer-Interaction (Elite-Masterstudiengang)
Seminar Menschzentrierte Künstliche Intelligenz
- ▶ **Winter term 2021/2022**
Seminar Menschzentrierte Künstliche Intelligenz
Seminar Menschzentrierte Künstliche Intelligenz (Elite-Masterstudiengang)
- ▶ **Summer term 2021**
Vorlesung & Übung Human-Computer-Interaction (Elite-Masterstudiengang)
- ▶ **Winter term 2020/2021**
Seminar Menschzentrierte Künstliche Intelligenz (Elite-Masterstudiengang)
- ▶ **Summer term 2020**
Vorlesung & Übung Human-Computer-Interaction (Elite-Masterstudiengang)

34.3 Awards & Roles

- ▶ **Since January 2022**
Elected member of the board of the Gesellschaft für Informatik - the largest professional society for computer science in the German-speaking area
- ▶ **Since December 2020**
Second deputy women's representative of the Faculty of Applied Computer Science, University of Augsburg
- ▶ **Since October 2020**
Awarded as GI Junior-Fellow 2020 (Award for outstanding, young talents) - for my work and knowledge transfer in the field of Human-Centered AI
- ▶ **Since June 2020**
Member of the winner team of the #wirfuerschule Hackathon with the project "In Data we Trust?" - combining Big Data with AI and Gamification to support data and ethics competence and political education

34.4 Invited Talks

When researching how XAI affects end-users, it was important for me that the findings of our work reach a broad audience. I have been very fortunate to present our research in various formats. The following is a list of all the invited talks I gave during my dissertation time.

► 2022

- **Weitz, K. (2022).** Pixel & Co. Informatik in der Grundschule. Webinar followed by a Q&A session at the Online Digital Symposium. 04.11.2022, online.
- **Weitz, K. (2022).** Human-Centered XAI: Investigating the Impact of Explainable AI on End-Users. Talk at the Workshop “Interdisciplinary Conversations on the Fairness, Explainability and Uncertainty of AI” of Die Junge Akademie, from 19.09.-20.09.22. 19.09.2022, Berlin.
- **Weitz, K. (2022).** Intelligente Maschinen – gestern, heute, morgen. Künstliche Intelligenz zum Ausprobieren. Workshop at the Seminar “Narrativen Künstlicher Intelligenz” of the Konrad Adenauer Stiftung, from 08.09.-11.09.22. 08.09.2022, Munich.
- **Weitz, K. (2022).** Herausforderungen der Menschzentrierten Künstlichen Intelligenz für Unternehmen: Die Rolle von Erklärbarkeit und Vertrauen in der Mensch-Technik Interaktion. Talk at the Workshop „Digitalisierung und Zukunft der Arbeit“ of the Technical University of Munich (TUM), 10.08.2022, Munich.
- **Weitz, K. (2022).** Mensch und Maschine – ein Traum-Paar? Aktuelle Forschung zur Künstlichen Intelligenz. Talk followed by a Q&A session at the Evangelische Akademie Anhalt and the Evangelische Akademie Sachsen-Anhalt e.V.. 30.06.2022, Dessau-Roßlau.
- **Weitz, K. (2022).** Menschzentrierte KI Science oder Fiction? Talk at the Computer Science information day, University of Augsburg. 28.04.2022, Augsburg.
- **Weitz, K. (2022).** “Sollten wir Maschinen vertrauen?” Chancen und Herausforderungen auf dem Weg zur Menschzentrierten KI. Online presentation followed by a Q&A session at the AI Production network, University of Augsburg. 21.04.2022, Augsburg.
- **Weitz, K. (2022).** Mehr Science, weniger Fiction: Keine Angst vor der Künstlichen Intelligenz! Einblicke in die Funktionsweise von KI. Talk followed by a Q&A session at the Evangelisches Forum Annahof. 28.03.2022, Augsburg.
- **Weitz, K. (2022).** Neugierde für Informatik wecken - Ideen für die anschauliche Vermittlung von Informatikkonzepten an Kinder und Jugendliche. Talk followed by a Q&A session at the Leipzig book fair, forum teaching practice of the Verband Bildungsmedien e.V.. 19.03.2022, Leipzig. –cancelled due to the Covid-19 pandemic–
- **Weitz, K. (2022).** “Algorithmus ist keine Krankheit” - Informatikkonzepte anschaulich und lebensweltbezogen an Kinder und Jugendliche vermitteln. Online presentation with discussion for student teachers at the University of Augsburg. 26.01.2022, online.

► 2021

- **Weitz, K. (2021).** Von “Versteh ich nicht!” bis “Das ist unfair!” - Herausforderungen auf dem Weg zur Menschzentrierten Künstlichen Intelligenz. Online talk with discussion in the lecture series “Artificial Intelligence - Past, Present and Future” in cooperation with Deutsches Museum Munich as part of the Hi!A Festival for Art and Science in Bavaria. 15.12.2021, online.
- **Weitz, K. (2021).** Erklärbare Künstliche Intelligenz (XAI) als Schlüssel für mehr Transparenz? – Einblicke in die Informatik. Talk followed by a Q&A session in the seminar “Transparency, Ethics & Publicity in (Digital) Democracy” of the University of Duisburg Essen. 13.12.2021, online.
- **Weitz, K. (2021).** Von A wie Algorithmus bis K wie Künstliche Intelligenz: Anregungen für die praxistaugliche Vermittlung von Informatikkonzepten. Talk followed by a Q&A session at the #WestermannLogin@Kassel. 26.11.-27.11.2021, Kassel. –cancelled due to the Covid-19 pandemic–
- **Weitz, K. (2021).** Pixel & Co. Informatik in der Grundschule. Talk followed by a Q&A session at the digital-day of the Westermann group. 27.10.2021, München. –cancelled due to the Covid-19 pandemic–
- **Weitz, K. (2021).** “Algorithmus ist keine Krankheit” - Informatikkonzepte anschaulich und lebensweltbezogen an Kinder und Jugendliche vermitteln. Online talk with discussion for student

teachers at the University of Augsburg. 09.07.2021, online.

- **Weitz, K. (2021).** Pixel & Co. Informatik in der Grundschule. Webinar followed by a Q&A session at the Online Digital Symposium. 12.06.2021, online.
- **Weitz, K. (2021).** Pixel & Co. Informatik in der Grundschule. Webinar of the Westermann publisher followed by a Q&A session. 20.05.2021, online.
- **Weitz, K. (2021).** Was hast du gelernt? Wie Erklärbare KI uns hilft, Neuronale Netze besser zu verstehen. Keynote followed by a Q&A session at the M3- Minds Mastering Machines conference, 27.04.2021, online.
- **Weitz, K. (2021).** Erklärbare Künstliche Intelligenz - Ziele, Methoden und Herausforderungen auf dem Weg zur Menschzentrierten KI. Webinar followed by a Q&A session as part of the AI-Camp organized by the National Society for Informatics and the BMBF, 16.02.2021.
- **Weitz, K. (2021).** Menschenzentrierte KI - Wie erklärable KI unsere Vorstellungen und Erwartungen verändert. Online presentation with discussion at the National Society for Informatics Regional Group Würzburg, 20.01.2021.

► 2020

- **Weitz, K. (2020).** PRIVACY ON STAND BY? Talk and public discussion with the artists Yu Zhang and Matthew Mosher at the "DIGITAL SPRING Festival", 21.03.2020, Salzburg. –*cancelled due to the Covid-19 pandemic*–

► 2019

- André, E., **Weitz, K. (2019).** Künstliche Intelligenz...und ich? Public discussion with the Minister of Science Bernd Sibler as part of the event "Siblers Denkräume", 03.12.2019, Augsburg.
- **Weitz, K. (2019).** Was "denkt" Künstliche Intelligenz? Wie wir sichtbar machen, wie intelligent KI wirklich ist. Science Slam Talk at the turmdersinne Symposium, 12.10.2019, Fürth.
- André, E., **Weitz, K.**, Weber, K., & Janowski, K. (2019) Sind Computer schlauer als wir?. Talk with Live Demo as part of the children's university, 05.07.2019, Augsburg.
- **Weitz, K. (2019).** Genial naiv. Wie schlau ist künstliche Intelligenz?. Talk within the scope of the Wissenschaftsjahr 2019 at the German Museum, 02.06.2019, Munich.
- **Weitz, K. (2019).** Gotta catch 'em all! Wenn die KI zum Pokemotrainer wird. Talk as part of the Girls' Day at the German Museum, 28.03.2019, Munich.

► 2018

- **Weitz, K. (2018).** Lachst du noch oder weinst du schon? Worauf neuronale Netze achten, wenn sie uns anschauen. Talk at the Science Slam of the Center Digitalisierung.Bayern (ZD.B), 21.11.2018, Munich.

Surveys

In the following, the three (online) surveys reported in this dissertation can be found. They served as a basis for the persona approach, described in Chapter 15 on page 85.

35.1 Survey XAI in Companies - Translated Version

The survey presented was conducted for the work in André et al. (2021).

35.1.1 Personal Information

What is your age? _____

To which gender identity do you most identify?

☐ male ☐ female ☐ other: _____

What is the highest level of education you have achieved?

- ☐ No degree
- ☐ Secondary school degree (Hauptschule)
- ☐ Secondary school degree (Realschule)
- ☐ High school diploma
- ☐ Bachelor's/Master's degree
- ☐ Completed doctorate / completed post-doctoral qualification

35.1.2 Information About the Company and Your Work Area

To which industry sector does your company or institution belong? [[single choice]]

- ☐ Agriculture and forestry
- ☐ Manufacturing industry
- ☐ Energy and water supply, sewage and waste disposal
- ☐ Trade
- ☐ Transport and logistics
- ☐ Financial and insurance services
- ☐ Technical services
- ☐ Public administration, social security
- ☐ Health and social work
- ☐ Other: _____

In which field of business do you work yourself? [[single choice]]

Production work, like:

- ☐ Automotive industry

- ☐ Energy supply companies
- ☐ Precision mechanics and optical industry
- ☐ Aerospace
- ☐ Mechanical and plant engineering
- ☐ Medical technology
- ☐ Metal construction companies
- ☐ Tool manufacturing
- ☐ Other: _____

Office work, like:

- ☐ Engineering office, technical services
- ☐ Associations and organizations
- ☐ Public administration, tax and accounting firms
- ☐ Banking and insurance
- ☐ Law company
- ☐ Real estate industry
- ☐ Human resources services
- ☐ Industrial companies
- ☐ Trade
- ☐ Other: _____

What is your position in the company? [[single choice]]

- ☐ Managing Director
- ☐ Executive
- ☐ Subject matter expert
- ☐ Administrator
- ☐ Temporary employee
- ☐ Trainee
- ☐ Other: _____

How many employees does your company/organization have? [[single choice]]

- ☐ 1 to 10 employees
- ☐ 11 to 50 employees
- ☐ 51 to 250 employees
- ☐ More than 250 employees

35.1.3 AI Applications in Your Company - General Information

We would first like to look at the strategic planning in your company before we go into specific AI applications in the next set of questions.

In which areas does your company plan to make changes with the help of Artificial Intelligence in the next few years? (Multiple answers possible)

- ☐ Organization
- ☐ Processes
- ☐ Develop/invest in new technologies
- ☐ Respond to market changes
- ☐ Expand regular customer business
- ☐ Acquire new customers
- ☐ Develop other innovations

- Comply with laws/regulations
- Improve cost management/implement controlling
- Other: _____
- No changes planned

What is driving AI development in your company? (Multiple answers possible)

- Requirements of customers
- Adaptation of business models
- Investments
- Increase of productivity
- Increase of flexibility
- Competition/market situation
- Other: _____
- AI development is not being driven forward

What do you see as challenges, obstacles or problems for your company in the implementation of AI-related changes? (Multiple answers possible)

- Financial issues
- Acceptance by employees
- Qualification of employees
- Shortage of skilled workers
- Technical compatibility problems
- Lack of technical equipment
- Insufficient speed rate of the company's Internet access
- Other: _____
- No challenges/barriers/problems

35.1.4 AI Applications in Your Company - Concrete Applications

Are AI technologies already being used in your company (as a prototype or in application)?

- Yes
- No
- Don't know

If 'Yes': In the following, we would like to know more about the AI technologies used. Therefore, answer the following questions for the most relevant AI technology used in your organization.

Please describe in bullet points what is the mission/goal of the AI technology? [[free text]]

Where will the AI technology be used? (e.g., department, manufacturing area, etc.) [[free text]]

non-autonomous fully autonomous

How autonomous is the application? ○ ○ ○ ○ ○

Is the application a prototype or is it already integrated into everyday business?

- Prototype
- Everyday business

- ☐ Yes
- ☐ No

Explainable AI will enable people to understand, appropriately trust, and effectively manage AI technologies.

What is your general attitude towards Explainable Artificial Intelligence (XAI)?

Extremely negative

Extremely positive

☐ ☐ ☐ ☐ ☐ ☐

With the help of Explainable Artificial Intelligence (XAI), it should be possible to better understand Artificial Intelligence (AI). **What is your opinion on this?**

I don't agree

I agree

XAI is important for end-users
(who have no experience with AI)

○ ○ ○ ○ ○ ○ ○

XAI is important for researcher

○ ○ ○ ○ ○ ○ ○

XAI is important for companies

○ ○ ○ ○ ○ ○ ○

XAI is important for politicians

○ ○ ○ ○ ○ ○ ○

35.1.6 End

Would you like to receive further information about the results of the survey? If 'yes', please provide your mail address where we can send you the results of our study. [[Free text]]

If you are interested in participating in an expert panel as part of our study, enter your name as well as your mail address. We will then get in touch with you. [[Free text]]

Thank you for your participation!

35.2 Survey XAI in Education - Translated Version

The presented survey served as a basis for the work presented in Weitz, Schlagowski, and André (2021).

35.2.1 Personal Information

Age: _____

Gender

- ☐ male ☐ female ☐ divers

Highest level of education

- ☐ No degree
☐ Secondary school degree (Hauptschule)
☐ Secondary school degree (Realschule)
☐ Highschool diploma (Gymnasium)
☐ Finished training
☐ Bachelor's/Master's degree (FH)
☐ Bachelor's degree (University)
☐ Master's degree (University)
☐ Completed doctorate

35.2.2 AI and XAI Knowledge

Have you ever heard about the term Artificial Intelligence (AI)?

- ☐ Yes
☐ No

The term 'Artificial Intelligence' is often used to describe machines (or computers) that mimic 'cognitive' functions that humans associate with the human mind, such as 'learning' and 'problem solving'

How would you rate your knowledge of Artificial Intelligence?

- ☐ I can explain the term accurately
☐ I can roughly explain the term ☐ I have heard the term before, but I can't think of anything specific about it.

What is your personal attitude towards Artificial Intelligence in general?

Extremely negative ☐ ☐ ☐ ☐ ☐ ☐ ☐ Extremely positive

In which areas should Artificial Intelligence be used? (Multiple answers possible)

- ☐ Household
☐ Care work
☐ Education
☐ Transport
☐ Leisure
☐ Art

- Safety

What future do you think we will have with Artificial Intelligence?

(Please select only one answer)

- Our life is getting worse
- Our life is getting better
- Negative and positive effects balance each other
- Don't know

Have you ever heard about the term Explainable Artificial Intelligence (XAI)?

- Yes
- No

With the help of explainable Artificial Intelligence (XAI), it should be possible to better understand Artificial Intelligence (AI). **What is your opinion on this?**

	I don't agree					I agree	
XAI is important for end-users (who have no experience with AI)	○	○	○	○	○	○	○
XAI is important for researcher	○	○	○	○	○	○	○
XAI is important for companies	○	○	○	○	○	○	○
XAI is important for politicians	○	○	○	○	○	○	○

The survey was developed in the work of Zellner (2021).

What is your gender? ☐ female ☐ male ☐ diverse

In which country are you currently living in? _____

- ☐ Did not complete school
- ☐ Elementary school
- ☐ Secondary school degree
- ☐ High school degree
- ☐ Completed apprenticeship
- ☐ College/University degree
- ☐ Doctoral degree
- ☐ other: _____

- Smartphone
- Smart-watch/fitness-tracker

- Visual (studies with picture, mindmaps)
- Aural (studies via listening, audio books)
- Communicative (studies through exchange with others, discussions)
- Motor (studies through haptics, memory cards, models)

(Affinity for Technology Interaction Short scale from Wessel et al. (2019))

[illegible]

Here are a number of personality traits that may or may not apply to you. You should rate the extent to which the pair of traits applies to you, even if one characteristic applies more strongly than the other.

(*Ten-Item Personality Inventory (TIPI) from Gosling et al. (2003)*)

	completely disagree	largely disagree	slightly disagree	slightly agree	largely agree	completely agree
extraverted, enthusiastic	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
critical, quarrelsome	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
dependable, self-disciplined	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
anxious, easily upset	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
open to new experiences, complex	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
reserved, quiet	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
sympathetic, warm	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
disorganized, careless	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
calm, emotionally stable	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
conventional, uncreative	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Were you working in a medical related job? [[single choice]]

- ☐ Yes
- ☐ No

The following questions refer to "self-monitoring" apps. These are usually found on a smartphone and/or smart-wearable, they allow the user to track or register body features, activities or behaviour. For example Sleep-, Food-tracker, brain training, fitness, and meditation apps. Possible example use-cases could be: liver spots, steps, nutrition, mood, sleep, menstrual cycle

In the last 30 days, how often did you use mobile self-monitoring apps, which are related to health and/or fitness? [[single choice]]

- ☐ never
- ☐ less than once a week
- ☐ once a week
- ☐ multiple times per week
- ☐ daily

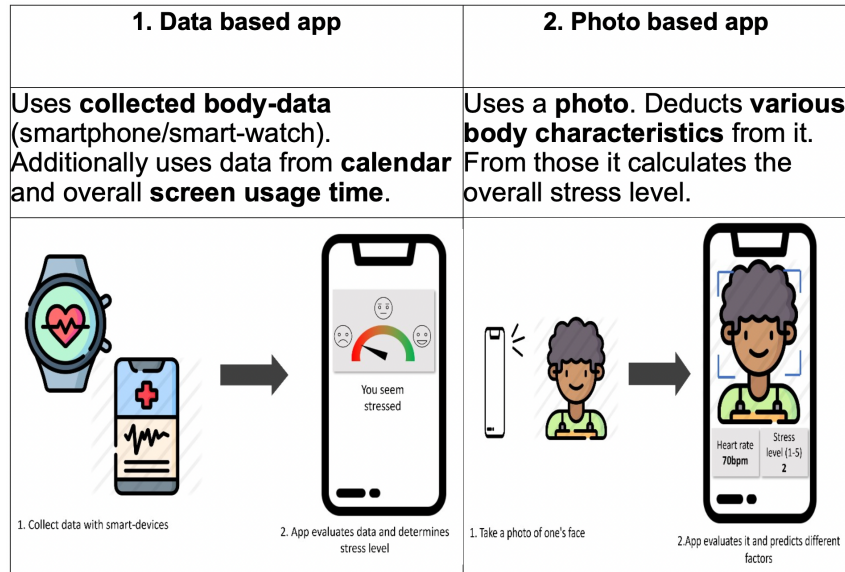
In which category would you place the health app(s) that you use? (Multiple answers possible)

- ☐ Fitness (workouts, activity-tracking)
- ☐ Wellbeing (sleep, meditation)
- ☐ Nutrition (nutritional values, fasting)
- ☐ Mental training (brain training, memory)
- ☐ other: _____

Although you are not using self-monitoring apps, which of the following categories would interest you the most to use such an app from? (Multiple answers possible)

	1 - totally unimportant				5 - very important			
Fitness (workouts, activity-tracking)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Wellbeing (sleep, meditation)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Nutrition (nutritional values, fasting)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Mental Training (brain training, memory)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

We will present you with two app-concepts from the area of self-monitoring of well-being. Both apps present the user with a prediction of the stress level.



Which of these two example applications would you rather use? [[single choice]]

- ☐ Data-based application
- ☐ Photo-based application

Why would you prefer to use this application over the other one? -voluntary information- [[free text]]

From the presented apps, please evaluate how important an explanation for the result would be for you.

	1 - totally unimportant				5 - very important
Data-based app	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Photo-based app	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

For the [[data-based app/photo-based app]]: How likely would one of the following questions come to your mind while using the app?

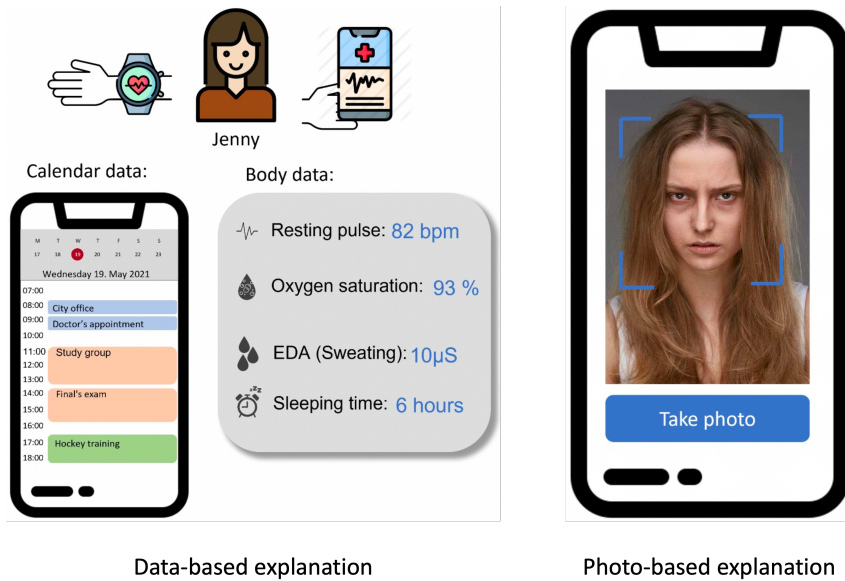
	1 - totally unimportant				5 - very important
Why do I get this prediction?	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
How does the system come to this prediction?	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Why do I not get predicted in the other class?	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
What do I have to change to get predicted differently?	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

You will be shown a scenario from the [[data-based/photo-based]] stress app. For this you will get the input, which the app gets.

Assume for the following questions, that you are explaining this decision to somebody else. For this you can you use all the data the app provided you with

Does the user of this app seem stressed for you?

- ☐ Yes
- ☐ No



Please describe briefly why the person [[seems stressed/does not seem stressed]]. -voluntary information- [[free text]]

[[Data-based question]] Please evaluate for the individual aspects, how likely you would use them in your explanation.

	1 - totally unimportant				5 - very important			
Calendar entries	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Resting pulse	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Oxygen saturation	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
EDA (sweating)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Sleeping time	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

[[Photo-based question]] Please evaluate the importance of the following picture sections for your explanation.

	1 - totally unimportant				5 - very important			
Eye region	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Hair	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Skin	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Gaze, facial expression	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Mouth region	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Forehead	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

[[optional - based on the answer of the previous question]]

Why are you not using some of the aspects from the question before?

	1 - I strongly disagree				5 - I strongly agree			
I know their meaning but it is not informative enough	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I don't know their meaning	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

[[optional - based on the answer of the previous question]]

Previously you stated you know the meaning of some [[aspects/picture sections]] but they are not informative

enough for your explanation.

How do you decide which [[aspect/picture section]] is more informative? -voluntary information-
[[free text]]

Would you adapt your explanation depending on who receives it?

- ☐ Yes
- ☐ No

[[optional - based on the answer of the previous question]]

Previously you answered that you would change your explanation depending on who is the recipient. **What aspects are important to you when adapting your explanation?** (Multiple answers possible)

- ☐ Age
- ☐ Cognitive skills of the recipient
- ☐ Technology knowledge of recipient
- ☐ other: _____

[[Data-based question]] **Assume you will receive the following explanations for the scenario, which one do you like more?**

When comparing your results to the average, non-stressed user, your heart rate is higher, and your sleep duration is below average.

You have a high pulse and too little sleep.

Which explanation do you prefer?

Left both the same Right
☐ ☐ ☐

You have a lot of appointments in your calendar, this indicates a stressed everyday life.
In terms of your physical characteristics, the increased heart rate, which can also be seen in quiet situations, is a sign of stress.
You also have little sleep, which means that your body may not be able to recover sufficiently.

You are stressed because you have a busy schedule, a high pulse and little sleep.

Which explanation do you prefer?

Left both the same Right
☐ ☐ ☐

[[Photo-based question]] **Assume you will receive the following explanations for the scenario, which one do you like more?**

Compared to the non-stressed average user, you have more pronounced dark circles.
The average user heart rate is lower than yours.

You have pronounced dark circles and your recognized pulse is increased.

Which explanation do you prefer?

Left both the same Right
☐ ☐ ☐

You have very pronounced dark circles, this can be a sign of lack of sleep and a stressed environment. Then there is the tousled hair, which can also be a sign of a stressed environment. In addition, an increased pulse was measured repeatedly. So one can conclude that you are stressed.

You are stressed because your hair is tousled, your heart rate is increased, and you have pronounced dark circles.

Which explanation do you prefer? Left both the same Right

☐ ☐ ☐

35.3.2 Presentation of Three Explanation Types

In the following you will be shown 3 representations of explanations for the data-based stress applications. Please answer the questions always referring to the current representation. The hand shows a user action is performed

Live Explanation

The user can change parameters to see how it would influence the app's prediction result.

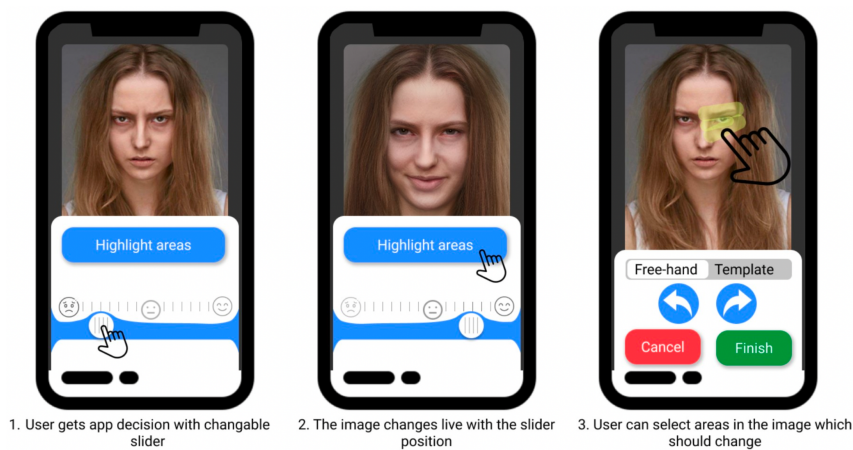
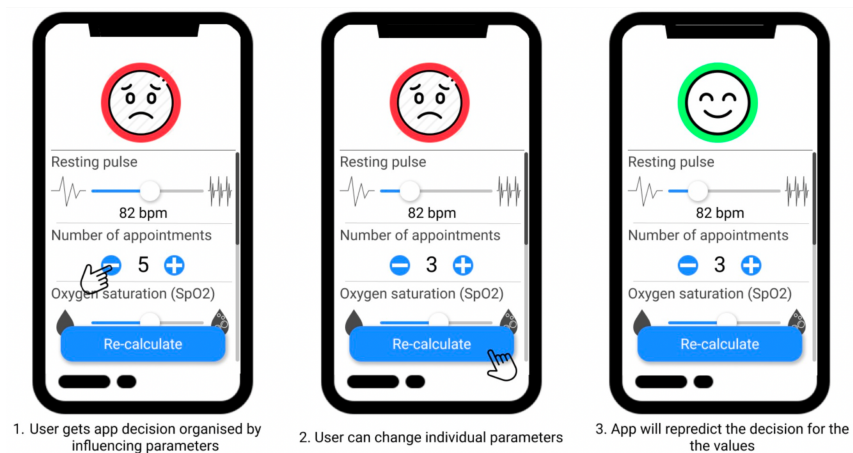


Photo-based explanation



Data-based explanation

The text below the smartphone is only for understanding and should not be taken into account when answering.

(Five items of the Explanation Satisfaction Scale (ESS) from Hoffman et al. (2018))

	1 - I strongly disagree	2 - I disagree somewhat	3 - I am neutral about it	4 - I agree somewhat	5 - I strongly agree
From the live explanation, I understand how the app works	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
This live explanation of how the app works is satisfying	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
This live explanation of how the app works has sufficient detail	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
This live explanation of how the app works is useful to my goals	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
This live explanation of the app shows me how accurate the app is	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I would like to try out the live explanation	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I would like to determine for myself which factors are taken into account for the live explanation	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

What do you like about the live-explanation, what do you not like? [[free text]]

Feature Explanation

The user gets a feature (cloud), which shows different features influencing the app's prediction.

The text below the smartphone is only for understanding and should not be taken into account when answering.
(Five items of the Explanation Satisfaction Scale (ESS) from Hoffman et al. (2018))

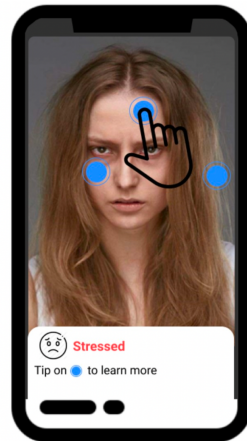
	1 - I strongly disagree	2 - I disagree somewhat	3 - I am neutral about it	4 - I agree somewhat	5 - I strongly agree
From the feature-cloud explanation, I understand how the app works	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
This feature-cloud explanation of how the app works is satisfying	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
This feature-cloud explanation of how the app works has sufficient detail	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
This feature-cloud explanation of how the app works is useful to my goals	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
This feature-cloud explanation of the app shows me how accurate the app is	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I would like to try out the feature-cloud explanation	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I would like to determine for myself which factors are taken into account for the feature-cloud explanation	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

What do you like about the feature-cloud explanation, what do you not like? [[free text]]

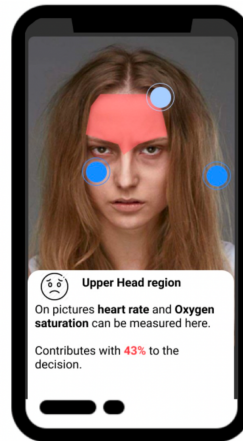
Ask-the-App Explanation

The user sees the app's decision and an interface in which he can ask questions about the result.

The text below the smartphone is only for understanding and should not be taken into account when answering.
(Five items of the Explanation Satisfaction Scale (ESS) from Hoffman et al. (2018))

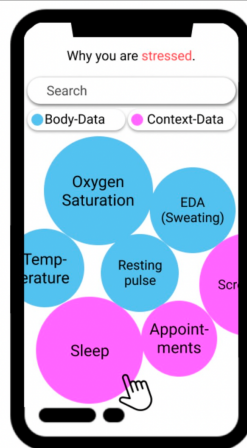


1. Markers in regions that have influenced the app's decision



2. Tapping a marker opens a more detailed explanation

Photo-based explanation



1. Features, which have influenced the app's decision, increased size means more influence



2. The user can read through details to each feature

Data-based explanation

	1 - I strongly disagree	2 - I disagree somewhat	3 - I am neutral about it	4 - I agree somewhat	5 - I strongly agree
From the Ask-the-App explanation, I understand how the app works	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
This Ask-the-App explanation of how the app works is satisfying	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
This Ask-the-App explanation of how the app works has sufficient detail	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
This Ask-the-App explanation of how the app works is useful to my goals	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
This Ask-the-App explanation of the app shows me how accurate the app is	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I would like to try out the Ask-the-App explanation	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I would like to determine for myself which factors are taken into account for the Ask-the-App explanation	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

What do you like about the feature-cloud explanation, what do you not like? [[free text]]

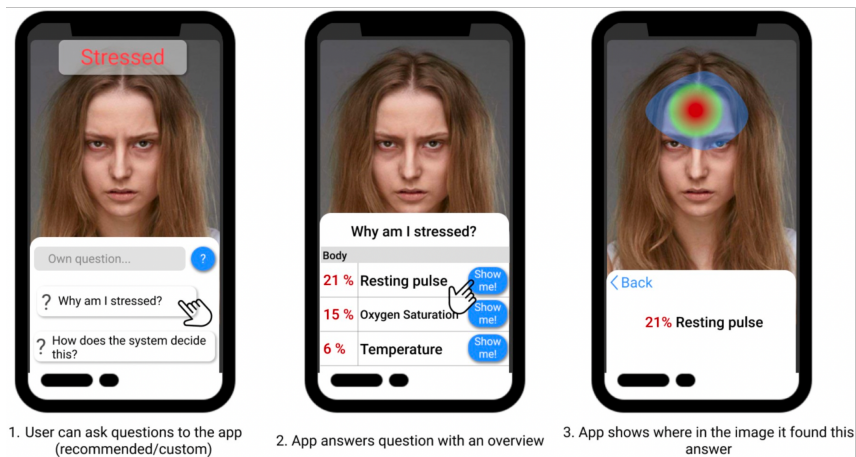
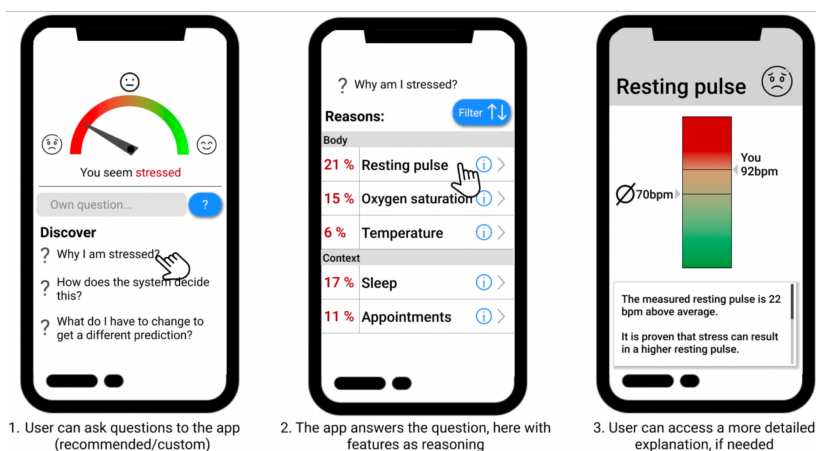


Photo-based explanation



Data-based explanation

Which of the explanations would you most likely use in your daily life?

Rate 1 - Live explanation 2 - feature-cloud explanation Ask-the-App explanation

○ ○ ○

Apart from the [[live explanation/feature-cloud explanation/Ask-the-app explanation]] could you imagine using one of the other explanations?

- ☐ Yes
- ☐ No

[[optional - based on the answer of the previous question]]

Please describe briefly why you would switch between the explanation representations in daily life? -voluntary information- [[free text]]

Did the explanations leave open questions at your end?

- ☐ Yes
- ☐ No

[[optional - based on the answer of the previous question]]

Which ones would that be? -voluntary information- [free text]]

How much time would you invest to understand the App's explanation for its decision?

- ☐ Less than 1 minute [[single choice]]
- ☐ 1-2 minutes
- ☐ 2-5 minutes
- ☐ more than 5 minutes

Would you be willing to spend more time on the explanation when you were able to play around with the explanation interactively or ask follow-up questions towards the app?

1 - completely disagree 5 - I completely agree
 Rate ☐ ☐ ☐ ☐ ☐

35.3.3 AI & XAI Knowledge

Have you ever heard about the term Artificial Intelligence (AI) before?

- ☐ Yes
- ☐ No

Does the definition of Artificial Intelligence given above match your idea of AI?

- ☐ Yes
- ☐ No

[[optional - based on the answer of the previous question]]

Your idea of Artificial Intelligence differs from the definition above. Briefly describe what you mean by Artificial Intelligence. -voluntary information- [[free text]]

What is your personal attitude towards Artificial Intelligence (AI)?

Very negative Very positive
☐ ☐ ☐ ☐ ☐ ☐

Have you ever heard of the term Explainable Artificial Intelligence (XAI)?

- ☐ Yes
- ☐ No

Explainable AI will enable people to understand AI technologies, trust them appropriately, and manage them effectively.

Does the definition of Explainable Artificial Intelligence given above match your idea of Explainable AI?

- ☐ Yes
- ☐ No

What is your personal attitude towards Explainable Artificial Intelligence (XAI)?

Extremely negative Extremely positive

○ ○ ○ ○ ○ ○ ○

With the help of Explainable Artificial Intelligence (XAI), it should be possible to improve the understanding of Artificial Intelligence. What is your personal assessment?

	Disagree strongly						Agree strongly
XAI is important for laypersons (which do not have any experience with AI)	○	○	○	○	○	○	○
XAI is important for scientists	○	○	○	○	○	○	○
XAI is important for businesses	○	○	○	○	○	○	○
XAI is important for politicians and decision maker	○	○	○	○	○	○	○

Please evaluate to what extent you agree with the following statements referring to XAI.

	1 - completely disagree			4 - completely agree
I like AI explanations with a lot of details	○	○	○	○
I would like to be able to ask questions back to the AI concerning the given explanation	○	○	○	○
I prefer AI explanations which are fitted towards my preferences and priorities	○	○	○	○

[[optional - based on the answer of the question regarding personalized explanations]]

Why would you not want a personalized explanation from the AI for its decision? -voluntary information- [[free text]]

[[optional - based on the answer of the question regarding explanation adaption]]

Previously you have stated you would adapt your own explanations for the recipient. But you do not want to receive a personalized explanation from the AI. If you want you can state why is this so. -voluntary information- [[free text]]

You completed the survey. Thank you!

User Strategies to Resolve Miscommunication

The following examples of user strategies to resolve miscommunication in the Conversational AI Study (see Chapter 27 on page 160) are illustrated. The observed strategies were published in Weitz, Vanderlyn, et al. (2021).

36.1 Agrees With System

The user gives up trying to solve the puzzle and just agrees with the system.

SYSTEM: Okay, based on what you've said there should be two half-circle dials and the slider should point to green. Therefore, you should move both pointers to the middle position.

USER: i only can get one to work

SYSTEM: If I have this correct, there are two half-circle dials and the slider is set to green. So you should set both pointers to the middle position.

USER: I can only set one of the dials

SYSTEM: Alright, so there should be one dial and the slider level should be green. So you should set the pointer to the far left position.

USER: okay

36.2 Corrects System

The user informs the system there is a fourth button it didn't know about.

SYSTEM: I'm sorry, I didn't understand. Please rephrase?

USER: the button order

SYSTEM: The complete sequence of clickable (colorful border) buttons is: amber, blue, amber. Therefore, the first button should be activated.

USER: there are 4 buttons

SYSTEM: What color is the fourth clickable button? Clickable buttons have a blue, amber, or green border.

36.3 Asks for Clarification

The user realizes they don't know what the slider is and asks for clarification.

USER: What should we do next?

SYSTEM: I'm sorry, I didn't understand. Please rephrase?

USER: Hello?

SYSTEM: There are 2 half dials and the slider is set to green. Therefore, both pointers should be set to the middle position.

USER: What slider is set to green?

SYSTEM: The Slider is a vertical bar with five sections ... (truncated to save space)

36.4 Solves Without Help

The user solves the puzzle without help from the system.

SYSTEM: What color is the first clickable button? Clickable buttons have a blue, amber, or green border.

USER: yse

SYSTEM: Sorry, I didn't understand that. Could you please say it again in other words?

USER: i did that module

36.5 Restarts

The user doesn't know how to correct the system, so just restarts the module.

SYSTEM: I'm sorry, I didn't understand. Please rephrase?

USER: The sequence I gave you was wrong

SYSTEM: The complete sequence of clickable (colorful border) buttons is: amber, green, amber. Therefore, the first button should be activated.

USER: restart

36.6 Tries Something New

After ignoring system question, the user tries different phrases until their input is recognized again.

USER: lets play

SYSTEM: Sorry, I didn't understand that. Could you please say it again in other words?

USER: hello?

SYSTEM: How many dials are there on the control panel? There might be one or two of them.

36.7 Rephrases

The user gives too much information at once, rephrases to only give one piece per turn.

USER: There are two half dials, but the slider color is yellow

SYSTEM: I can't process the information this fast, please only tell me one thing at a time.

USER: the slider is yellow

SYSTEM: What number is the pointer on the left dial pointing to?

Bibliography

- Acuña, S. T., Castro, J. W., & Juristo, N. (2012). A HCI technique for improving requirements elicitation. *Information and Software Technology*, 54(12), 1357–1375. <https://doi.org/10.1016/j.infsof.2012.07.011>
- Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access*, 6, 52138–52160. <https://doi.org/10.1109/ACCESS.2018.2870052>
- Akhtar, M. S., Ekbal, A., & Cambria, E. (2020). How intense are you? predicting intensities of emotions and sentiments using stacked ensemble [application notes]. *IEEE Computational Intelligence Magazine*, 15(1), 64–75. <https://doi.org/10.1109/MCI.2019.2954667>
- Alizadeh, F., Esau, M., Stevens, G., & Cassens, L. (2020). Explainable AI: Take one step back, move two steps forward. In C. Hansen, A. Nürnberger, & B. Preim (Eds.), *Mensch und computer 2020 - workshopband, magdeburg, germany, september 6-9, 2020*. Gesellschaft für Informatik e.V. <https://doi.org/10.18420/muc2020-ws111-369>
- Alqaraawi, A., Schuessler, M., Weiß, P., Costanza, E., & Berthouze, N. (2020). Evaluating saliency map explanations for convolutional neural networks: A user study (F. Paternò, N. Oliver, C. Conati, L. D. Spano, & N. Tintarev, Eds.), 275–285. <https://doi.org/10.1145/3377325.3377519>
- Alvarado, O., Heuer, H., Vanden Abeele, V., Breiter, A., & Verbert, K. (2020). Middle-aged video consumers' beliefs about algorithmic recommendations on youtube. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW2), 1–24. <https://doi.org/10.1145/3415192>
- Amsterlaw, J., & Wellman, H. M. (2006). Theories of mind in transition: A microgenetic study of the development of false belief understanding. *Journal of Cognition and Development*, 7(2), 139–172. https://doi.org/10.1207/s15327647jcd0702_1
- Anderson, A., Dodge, J., Sadarangani, A., Juozapaitis, Z., Newman, E., Irvine, J., Chattopadhyay, S., Fern, A., & Burnett, M. (2019). Explaining reinforcement learning to mere mortals: An empirical study. In S. Kraus (Ed.), *Proceedings of the twenty-eighth international joint conference on artificial intelligence, IJCAI 2019, macao, china, august 10-16, 2019* (pp. 1328–1334). International Joint Conferences on Artificial Intelligence Organization. <https://doi.org/10.24963/ijcai.2019/184>
- André, E., Bauer, W., Braun, M., Dang, C. T., Peissner, M., & Weitz, K. (2021). *KI-Kompetenzentwicklung bei Sach und Produktionsarbeit*. Lernende Systeme Die Plattform für Künstliche Intelligenz. https://www.plattform-lernende-systeme.de/files/Downloads/Publikationen/AG2_WP_Projektbericht_Kompetenzentwicklung_KI.pdf
- Anjomshoe, S., Najjar, A., Calvaresi, D., & Främling, K. (2019). Explainable agents and robots: Results from a systematic literature review. In E. Elkind, M. Veloso, N. Agmon, & M. E. Taylor (Eds.), *Proceedings of the 18th international conference on autonomous agents and multiagent systems, AAMAS '19, montreal, qc, canada, may 13-17, 2019* (pp. 1078–1088). International Foundation for Autonomous Agents; Multiagent Systems.
- Arras, L., Osman, A., & Samek, W. (2022). CLEVR-XAI: A benchmark dataset for the ground truth evaluation of neural network explanations. *Information Fusion*, 81, 14–40. <https://doi.org/10.1016/j.inffus.2021.11.008>
- Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., & Samek, W. (2015). On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7), e0130140. <https://doi.org/10.1371/journal.pone.0130140>
- Bach, S., Binder, A., Müller, K.-R., & Samek, W. (2016). Controlling explanatory heatmap resolution and semantics via decomposition depth. *Proceedings of the 2016 IEEE International Conference on Image Processing, ICIP 2016, Phoenix, AZ, USA, September 25-28, 2016*, 2271–2275. <https://doi.org/10.1109/ICIP.2016.7532763>
- Balaji, K., & Lavanya, K. (2019). Chapter 5 - medical image analysis with deep neural networks. In A. K. Sangaiah (Ed.), *Deep learning and parallel computing environment for bioengineering systems* (pp. 75–97). Academic Press. <https://doi.org/10.1016/B978-0-12-816718-2.00012-9>
- Balters, S., & Steinert, M. (2017). Capturing emotion reactivity through physiology measurement as a foundation for affective engineering in engineering design science and engineering practices. *Journal of Intelligent Manufacturing*, 28(7), 1585–1607. <https://doi.org/10.1007/s10845-015-1145-2>

- Bandura, A. (1977). Self-efficacy: Toward a unifying theory of behavioral change. *Psychological review*, 84(2), 191–215. <https://doi.org/10.1037/0033-295X.84.2.191>
- Bandura, A. (2010). Self-efficacy. *The Corsini Encyclopedia of Psychology*, 1–3. <https://doi.org/10.1002/9780470479216.corpsy0836>
- Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Benetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58, 82–115. <https://doi.org/10.1016/j.inffus.2019.12.012>
- Bartneck, C., Kulić, D., Croft, E., & Zoghbi, S. (2009). Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *International Journal of Social Robotics*, 1(1), 71–81. <https://doi.org/10.1007/s12369-008-0001-3>
- Baur, T., Clausen, S., Heimerl, A., Lingenfeller, F., Lutz, W., & André, E. (2020). Nova: A tool for explanatory multimodal behavior analysis and its application to psychotherapy. *Proceedings of the International Conference on Multimedia Modeling*, 577–588. https://doi.org/10.1007/978-3-030-37734-2_47
- Baur, T., Damian, I., Lingenfeller, F., Wagner, J., & André, E. (2013). Nova: Automated analysis of nonverbal signals in social interactions. In A. A. Salah, H. Hung, O. Aran, & H. Gunes (Eds.), *Human behavior understanding - 4th international workshop, HBU 2013, barcelona, spain, october 22, 2013. proceedings* (pp. 160–171). https://doi.org/10.1007/978-3-319-02714-2_14
- Baur, T., Heimerl, A., Lingenfeller, F., Wagner, J., Valstar, M. F., Schuller, B., & André, E. (2020). Explainable cooperative machine learning with NOVA. *KI - Künstliche Intelligenz*, 34(2), 143–164. <https://doi.org/10.1007/s13218-020-00632-3>
- BBC News. (2000). Pc rage hits uk. Retrieved October 31, 2022, from <http://news.bbc.co.uk/2/hi/science/nature/623341.stm>
- Beck, J., Stern, M., & Haugsjaa, E. (1996). Applications of AI in education. *XRDS: Crossroads, The ACM Magazine for Students*, 3(1), 11–15. <https://doi.org/10.1145/332148.332153>
- Bell, T., Alexander, J., Freeman, I., & Grimley, M. (2009). Computer science unplugged: School students doing real computing without computers. *The New Zealand Journal of Applied Computing and Information Technology*, 13(1), 20–29.
- Bernacki, M. L., Nokes-Malach, T. J., & Aleven, V. (2015). Examining self-efficacy during learning: Variability and relations to behavior, performance, and learning. *Metacognition and Learning*, 10(1), 99–117. <https://doi.org/10.1007/s11409-014-9127-x>
- Berner, E. S. (2007). *Clinical decision support systems* (Vol. 233). Springer.
- Biran, O., & Cotton, C. (2017). Explanation and justification in machine learning: A survey. *IJCAI-17 workshop on explainable AI (XAI)*, 8(1), 8–13.
- Bitcom Research. (2019). Welche Hemmnisse sehen Sie beim Einsatz von Industrie-4.0-Anwendungen in Ihrem Unternehmen? Retrieved February 1, 2022, from <https://de.statista.com/statistik/daten/studie/990517/umfrage/umfrage-zu-hemmnissen-bei-industrie-40-anwendungen-in-industrieunternehmen/>
- Bitcom Research. (2020). Was sind aus Sicht Ihres Unternehmens die wichtigsten Vorteile von Künstlicher Intelligenz im Kontext von Industrie 4.0? Retrieved February 1, 2022, from <https://de.statista.com/statistik/daten/studie/990505/umfrage/umfrage-zu-vorteilen-kuenstlicher-intelligenz-in-deutschen-industrieunternehmen/>
- Bland, J. M., & Altman, D. G. (1997). Statistics notes: Cronbach's alpha. *BMJ*, 314(7080), 572. <https://doi.org/10.1136/bmj.314.7080.572>
- Bluethmann, W., Ambrose, R., Diftler, M., Askew, S., Huber, E., Goza, M., Rehnmark, F., Lovchik, C., & Magruder, D. (2003). Robonaut: A robot designed to work with humans in space. *Autonomous Robots*, 14(2-3), 179–197. <https://doi.org/10.1023/A:1022231703061>
- Bombari, D., Schmid, P., Mast, M., Birri, S., Mast, F., & Lobmaier, J. (2013). Emotion recognition: The role of featural and configural face information. *The Quarterly Journal of Experimental Psychology*, 66(12), 2426–2442. <https://doi.org/10.1080/17470218.2013.789065>
- Brock, D. C. (2018). Learning from artificial intelligence's previous awakenings: The history of expert systems. *AI Magazine*, 39(3), 3–15. <https://doi.org/10.1609/aimag.v39i3.2809>
- Brooks, D. J. (2017). *A human-centric approach to autonomous robot failures* [Doctoral dissertation]. University of Massachusetts Lowell.

- Bruckert, S., Finzel, B., & Schmid, U. (2020). The next generation of medical decision support: A roadmap toward transparent expert companions. *Frontiers in Artificial Intelligence*, 3, 507973. <https://doi.org/10.3389/frai.2020.507973>
- Budiu, R. (2019). Mental models for intelligent assistants. Retrieved October 5, 2022, from <https://www.nngroup.com/articles/mental-model-ai-assistants>
- Bueff, A., Papantonis, I., Simkute, A., & Belle, V. (2022). Explainability in machine learning: A pedagogical perspective. *arXiv preprint arXiv:2202.10335*. <https://arxiv.org/abs/2202.10335>
- Buxmann, P., Hess, T., & Lehmann, S. (2008). Software as a service. *Wirtschaftsinformatik*, 50(6), 500–503. <https://doi.org/10.1007/s11576-008-0095-0>
- Cacioppo, J. T., & Petty, R. E. (1982). The need for cognition. *Journal of Personality and Social Psychology*, 42(1), 116–131. <https://doi.org/https://doi.org/10.1037/0022-3514.42.1.116>
- Cai, C. J., Winter, S., Steiner, D., Wilcox, L., & Terry, M. (2019). “Hello ai”: Uncovering the onboarding needs of medical practitioners for human-ai collaborative decision-making. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW), 1–24. <https://doi.org/10.1145/3359206>
- Cambria, E. (2016). Affective computing and sentiment analysis. *IEEE Intelligent Systems*, 31(2), 102–107. <https://doi.org/10.1109/MIS.2016.31>
- Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research*. Houghton Mifflin Company.
- Castelfranchi, C., & Falcone, R. (2010). *Trust theory: A socio-cognitive and computational model*. John Wiley & Sons.
- Castelli, N., Ogonowski, C., Jakobi, T., Stein, M., Stevens, G., & Wulf, V. (2017). What happened in my home? an end-user development approach for smart home data visualization. In G. Mark, S. R. Fussell, C. Lampe, M. C. Schraefel, J. P. Hourcade, C. Appert, & D. Wigdor (Eds.), *Proceedings of the 2017 CHI conference on human factors in computing systems, denver, co, usa, may 06-11, 2017* (pp. 853–866). ACM. <https://doi.org/10.1145/3025453.3025485>
- Castro, J. W., Acuña, S. T., & Juristo, N. (2008). Integrating the personas technique into the requirements analysis activity. In A. F. Gelbukh & M. E. Adiba (Eds.), *Proceedings of the 9th mexican international conference on computer science, ENC 2008, 6-10 october 2008, mexicali, baja california, mexico* (pp. 104–112). IEEE Computer Society. <https://doi.org/10.1109/ENC.2008.40>
- Cenefetelli, R. T. (2004). Getting in touch with our feelings towards technology. 2004(1), F1–F6. <https://doi.org/10.5465/ambpp.2004.13862511>
- Cheng, F., Liu, D., Du, F., Lin, Y., Zytek, A., Li, H., Qu, H., & Veeramachaneni, K. (2021). Vbridge: Connecting the dots between features and data to explain healthcare models. *IEEE Transactions on Visualization and Computer Graphics*, 28(1), 378–388. <https://doi.org/10.1109/TVCG.2021.3114836>
- Chiu, C.-C., & Marsella, S. (2011). How to train your avatar: A data driven approach to gesture generation. In H. H. Vilhjálmsson, S. Kopp, S. Marsella, & K. R. Thórisson (Eds.), *Proceedings of the intelligent virtual agents - 11th international conference, IVA 2011, reykjavik, iceland, september 15-17, 2011* (pp. 127–140, Vol. 6895). Springer. https://doi.org/10.1007/978-3-642-23974-8_14
- Choo, J., & Liu, S. (2018). Visual analytics for explainable deep learning. *IEEE Computer Graphics and Applications*, 38(4), 84–92. <https://doi.org/10.1109/MCG.2018.042731661>
- Chromik, M., & Butz, A. (2021). Human-xai interaction: A review and design principles for explanation user interfaces. *Human-Computer Interaction – INTERACT 2021*, 619–640. https://doi.org/10.1007/978-3-030-85616-8_36
- Chromik, M., Eiband, M., Völkel, S. T., & Buschek, D. (2019). Dark patterns of explainability, transparency, and user control for intelligent systems. In C. Trattner, D. Parra, & N. Riche (Eds.), *Joint proceedings of the ACM IUI 2019 workshops co-located with the 24th ACM conference on intelligent user interfaces (ACM IUI 2019), los angeles, usa, march 20, 2019* (Vol. 2327).
- Clancey, W. J. (1983). The epistemology of a rule-based expert system—a framework for explanation. *Artificial Intelligence*, 20(3), 215–251. [https://doi.org/10.1016/0004-3702\(83\)90008-5](https://doi.org/10.1016/0004-3702(83)90008-5)
- Clocksin, W. F., & Mellish, C. S. (2003). *Programming in prolog*. Springer Berlin Heidelberg.
- Clore, G. L., & Huntsinger, J. R. (2007). How emotions inform judgment and regulate thought. *Trends in Cognitive Sciences*, 11(9), 393–399. <https://doi.org/10.1016/j.tics.2007.08.005>

- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Taylor & Francis Ltd. <https://doi.org/10.4324/9780203771587>
- Cohen, P. R. (2020). Back to the future for dialogue research. *Proceedings of the 34th AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, 34(09), 13514–13519.
- Compeau, D. R., & Higgins, C. A. (1995). Computer self-efficacy: Development of a measure and initial test. *MIS Quarterly*, 19(2), 189–211. <https://doi.org/10.2307/249688>
- Conati, C., Barral, O., Putnam, V., & Rieger, L. (2021). Toward personalized XAI: A case study in intelligent tutoring systems. *Artificial Intelligence*, 298, 103503. <https://doi.org/10.1016/j.artint.2021.103503>
- Cotter, K., Cho, J., & Rader, E. (2017). Explaining the news feed algorithm: An analysis of the "news feed fyi" blog. *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems, Denver, CO, USA, May 06-11, 2017, Extended Abstracts*, 1553–1560. <https://doi.org/10.1145/3027063.3053114>
- Craik, K. J. W. (1967). *The nature of explanation*. Cambridge University Press.
- Cummings, M. L. (2004). Automation bias in intelligent time critical decision support systems. In *Aiaa 1st intelligent systems technical conference* (pp. 289–294). <https://doi.org/10.2514/6.2004-6313>
- Damasio, A. R. (1994). *Descartes' Irrtum: Fühlen, Denken und das menschliche Gehirn*. List Verlag, München.
- Das, D., Banerjee, S., & Chernova, S. (2021). Explainable AI for robot failures: Generating explanations that improve user assistance in fault recovery. In C. L. Bethel, A. Paiva, E. Broadbent, D. Feil-Seifer, & D. Szafir (Eds.), *Proceedings of HRI '21: ACM/IEEE international conference on human-robot interaction, boulder, co, usa, march 8-11, 2021* (pp. 351–360). ACM. <https://doi.org/10.1145/3434073.3444657>
- de Graaf, M. M. A., & Malle, B. F. (2017). How people explain action (and autonomous intelligent systems should too). *2017 AAAI Fall Symposia, Arlington, Virginia, USA, November 9-11, 2017*, 19–26.
- Deterding, S., Dixon, D., Khaled, R., & Nacke, L. (2011). From game design elements to gamefulness: Defining "gamification". In A. Lugmayr, H. Franssila, C. Safran, & I. Hammouda (Eds.), *Proceedings of the 15th international academic mindtrek conference: Envisioning future media environments, mindtrek 2011, tampere, finland, september 28-30, 2011* (pp. 9–15). ACM. <https://doi.org/10.1145/2181037.2181040>
- de Visser, E. J., Peeters, M. M., Jung, M. F., Kohn, S., Shaw, T. H., Pak, R., & Neerincx, M. A. (2020). Towards a theory of longitudinal trust calibration in human–robot teams. *International Journal of Social Robotics*, 12(2), 459–478. <https://doi.org/10.1007/s12369-019-00596-x>
- Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, 144(1), 114–126. <https://doi.org/10.1037/xge0000033>
- Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*. <https://arxiv.org/abs/1702.08608>
- Doyle, E. E., & Paton, D. (2017). Decision-making: Preventing miscommunication and creating shared meaning between stakeholders. In *Observing the volcano world: Volcano crisis communication* (pp. 549–570). Springer. https://doi.org/10.1007/11157_2016_31
- Dreiseitl, S., & Ohno-Machado, L. (2002). Logistic regression and artificial neural network classification models: A methodology review. *Journal of Biomedical Informatics*, 35(5-6), 352–359. [https://doi.org/10.1016/S1532-0464\(03\)00034-0](https://doi.org/10.1016/S1532-0464(03)00034-0)
- Dreyfus, H. L. (1992). *What computers still can't do: A critique of artificial reason*. MIT press.
- Efron, B., Hastie, T., Johnstone, I., & Tibshirani, R. (2004). Least angle regression. *The Annals of Statistics*, 32(2), 407–499. <https://doi.org/10.1214/009053604000000067>
- Ehsan, U., Liao, Q. V., Passi, S., Riedl, M. O., & Daume III, H. (2022). Seamful xai: Operationalizing seamful design in explainable ai. <https://doi.org/10.48550/arXiv.2211.06753>
- Ekman, P., & Friesen, W. V. (1971). Constants across cultures in the face and emotion. *Journal of Personality and Social Psychology*, 17(2), 124–129. <https://doi.org/10.1037/h0030377>
- Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639), 115–118. <https://doi.org/10.1038/nature21056>

- European Commission. (2017). Special eurobarometer 460—attitudes towards the impact of digitisation and automation on daily life. eurobarometer report. Retrieved October 4, 2022, from <https://europa.eu/eurobarometer/surveys/detail/2160>
- European Commission. (2018a). Artificial intelligence for europe. Retrieved October 10, 2022, from <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=COM%3A2018%3A237%3AFIN>
- European Commission. (2018b). General data protection regulation. Retrieved March 14, 2022, from <https://gdpr.eu/tag/gdpr/>
- Evans, T., Retzlaff, C. O., Geißler, C., Kargl, M., Plass, M., Müller, H., Kiehl, T.-R., Zerbe, N., & Holzinger, A. (2022). The explainability paradox: Challenges for xAI in digital pathology. *Future Generation Computer Systems*, 133, 281–296. <https://doi.org/10.1016/j.future.2022.03.009>
- Felzenszwalb, P. F., & Huttenlocher, D. P. (2004). Efficient graph-based image segmentation. *International Journal of Computer Vision*, 59(2), 167–181. <https://doi.org/10.1023/B:VISI.0000022288.19776.77>
- Ferreira, B., Silva, W., Barbosa, S. D., & Conte, T. (2018). Technique for representing requirements using personas: A controlled experiment. *IET Software*, 12(3), 280–290. <https://doi.org/10.1049/iet-sen.2017.0313>
- Field, A., Miles, J., & Field, Z. (2012). *Discovering statistics using R*. Sage Publications.
- Finzel, B., Tafler, D. E., Scheele, S., & Schmid, U. (2021). Explanation as a process: User-centric construction of multi-level and multi-modal explanations. In S. Edelkamp, R. Möller, & E. Rueckert (Eds.), *Proceedings of KI 2021: Advances in artificial intelligence - 44th german conference on ai, virtual event, september 27 - october 1, 2021* (pp. 80–94, Vol. 12873). Springer. https://doi.org/10.1007/978-3-030-87626-5_7
- Fiok, K., Farahani, F. V., Karwowski, W., & Ahram, T. (2022). Explainable artificial intelligence for education and training. *The Journal of Defense Modeling and Simulation*, 19(2), 133–144. <https://doi.org/10.1177/15485129211028651>
- Fogel, D., Hanson, J. C., Kick, R., Malki, H. A., Sigwart, C., Stinson, M., & Turban, E. (1993). The impact of machine learning on expert systems. In S. C. Kwasny & J. F. Buck (Eds.), *Proceedings of the ACM 21th conference on computer science, CSC '93, indianapolis, in, usa, february 16-18, 1993* (pp. 522–527). ACM. <https://doi.org/10.1145/170791.171158>
- Franke, T., Attig, C., & Wessel, D. (2019). A personal resource for technology interaction: Development and validation of the affinity for technology interaction (ATI) scale. *International Journal of Human-Computer Interaction*, 35(6), 456–467. <https://doi.org/10.1080/10447318.2018.1456150>
- Franquet, T. (2018). Imaging of community-acquired pneumonia. *Journal of Thoracic Imaging*, 33(5), 282–294. <https://doi.org/10.1097/RTI.0000000000000347>
- Friedman, S., Forbus, K., & Sherin, B. (2018). Representing, running, and revising mental models: A computational model. *Cognitive Science*, 42(4), 1110–1145. <https://doi.org/10.1111/cogs.12574>
- Fukushima, K. (1988). Neocognitron: A hierarchical neural network capable of visual pattern recognition. *Neural Networks*, 1(2), 119–130. [https://doi.org/10.1016/0893-6080\(88\)90014-7](https://doi.org/10.1016/0893-6080(88)90014-7)
- Fulton, L. B., Lee, J. Y., Wang, Q., Yuan, Z., Hammer, J., & Perer, A. (2020). Getting playful with explainable ai: Games with a purpose to improve human understanding of AI. In R. Bernhaupt, F. ' . Mueller, D. Verweij, J. Andres, J. McGrenere, A. Cockburn, I. Avellino, A. Goguey, P. Bjøn, S. Zhao, B. P. Samson, & R. Kocielnik (Eds.), *Extended abstracts of the 2020 CHI conference on human factors in computing systems, CHI 2020, honolulu, hi, usa, april 25-30, 2020* (pp. 1–8). ACM. <https://doi.org/10.1145/3334480.3382831>
- Garcia, F. J. C., Robb, D. A., Liu, X., Laskov, A., Patron, P., & Hastie, H. (2018a). Explain yourself: A natural language interface for scrutable autonomous robots. *arXiv preprint arXiv:1803.02088*. <http://arxiv.org/abs/1803.02088>
- Garcia, F. J. C., Robb, D. A., Liu, X., Laskov, A., Patron, P., & Hastie, H. (2018b). Explainable autonomy: A study of explanation styles for building clear mental models. In E. Krahmer, A. Gatt, & M. Goudbeek (Eds.), *Proceedings of the 11th international conference on natural language generation, tilburg university, the netherlands, november 5-8, 2018* (pp. 99–108). Association for Computational Linguistics. <https://doi.org/10.18653/v1/w18-6511>
- Garrod, S., & Pickering, M. J. (2004). Why is conversation so easy? *Trends in Cognitive Sciences*, 8(1), 8–11. <https://doi.org/10.1016/j.tics.2003.10.016>
- Gatt, A., & Paggio, P. (2014). Learning when to point: A data-driven approach. *COLING 2014, 25th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, August 23-29, 2014, Dublin, Ireland, 2007–2017*.

- Gaudiello, I., Zibetti, E., Lefort, S., Chetouani, M., & Ivaldi, S. (2016). Trust as indicator of robot functional and social acceptance. an experimental study on user conformation to iCub answers. *Computers in Human Behavior*, 61, 633–655. <https://doi.org/10.1016/j.chb.2016.03.057>
- Gerlings, J., Jensen, M. S., & Shollo, A. (2021). Explainable AI, but explainable to whom? *arXiv preprint arXiv:2106.05568*. <https://arxiv.org/abs/2106.05568>
- Gero, K. I., Ashktorab, Z., Dugan, C., Pan, Q., Johnson, J., Geyer, W., Ruiz, M., Miller, S., Millen, D. R., Campbell, M., Kumaravel, S., & Zhang, W. (2020). Mental models of ai agents in a cooperative game setting. In R. Bernhaupt, F. ' Mueller, D. Verweij, J. Andres, J. McGrenere, A. Cockburn, I. Avellino, A. Goguey, P. Bjøn, S. Zhao, B. P. Samson, & R. Kocielnik (Eds.), *Proceedings of CHI '20: CHI conference on human factors in computing systems, honolulu, hi, usa, april 25-30, 2020* (pp. 1–12). ACM. <https://doi.org/10.1145/3313831.3376316>
- Gervits, F., Thurston, D., Thielstrom, R., Fong, T., Pham, Q., & Scheutz, M. (2020). Toward genuine robot teammates: Improving human-robot team performance using robot shared mental models. In A. E. F. Seghrouchni, G. Sukthankar, B. An, & N. Yorke-Smith (Eds.), *Proceedings of the 19th international conference on autonomous agents and multiagent systems, AAMAS '20, auckland, new zealand, may 9-13, 2020* (pp. 429–437). International Foundation for Autonomous Agents; Multiagent Systems. <https://doi.org/10.5555/3398761.3398815>
- Gettier, E. L. (1963). Is justified true belief knowledge? *Analysis*, 23(6), 121–123. <https://doi.org/10.1093/analysis/23.6.121>
- Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., & Kagal, L. (2018). Explaining explanations: An overview of interpretability of machine learning. In F. Bonchi, F. J. Provost, T. Eliassi-Rad, W. Wang, C. Cattuto, & R. Ghani (Eds.), *Proceedings of the 5th IEEE international conference on data science and advanced analytics, DSAA 2018, turin, italy, october 1-3, 2018* (pp. 80–89). IEEE. <https://doi.org/10.1109/DSAA.2018.00018>
- Gilpin, L. H., Testart, C., Fruchter, N., & Adebayo, J. (2019). Explaining explanations to society. *arXiv preprint arXiv:1901.06560*. <http://arxiv.org/abs/1901.06560>
- Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press. <http://www.deeplearningbook.org/>
- Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, & K. Q. Weinberger (Eds.), *Advances in neural information processing systems 27: Annual conference on neural information processing systems 2014, december 8-13 2014, montreal, quebec, canada* (pp. 2672–2680).
- Gosling, S. D., Rentfrow, P. J., & Swann Jr, W. B. (2003). A very brief measure of the big-five personality domains. *Journal of Research in Personality*, 37(6), 504–528. [https://doi.org/10.1016/S0092-6566\(03\)00046-1](https://doi.org/10.1016/S0092-6566(03)00046-1)
- Gray, C. M., Kou, Y., Battles, B., Hoggatt, J., & Toombs, A. L. (2018). The dark (patterns) side of UX design. In R. L. Mandryk, M. Hancock, M. Perry, & A. L. Cox (Eds.), *Proceedings of the 2018 CHI conference on human factors in computing systems, CHI 2018, montreal, qc, canada, april 21-26, 2018* (pp. 1–14). ACM. <https://doi.org/10.1145/3173574.3174108>
- Greca, I. M., & Moreira, M. A. (2000). Mental models, conceptual models, and modelling. *International Journal of Science Education*, 22(1), 1–11. <https://doi.org/10.1080/095006900289976>
- Gu, J., Wang, Z., Kuen, J., Ma, L., Shahroudy, A., Shuai, B., Liu, T., Wang, X., Wang, G., Cai, J., & Chen, T. (2018). Recent advances in convolutional neural networks. *Pattern Recognition*, 77, 354–377. <https://doi.org/10.1016/j.patcog.2017.10.013>
- Gunning, D., & Aha, D. (2019). Darpa's explainable artificial intelligence (XAI) program. *AI Magazine*, 40(2), 44–58. <https://doi.org/10.1609/aimag.v40i2.2850>
- Gunning, D., Stefik, M., Choi, J., Miller, T., Stumpf, S., & Yang, G.-Z. (2019). XAI—explainable artificial intelligence. *Science Robotics*, 4(37). <https://doi.org/10.1126/scirobotics.aay7120>
- Guo, F. Y., Shamdasani, S., & Randall, B. (2011). Creating effective personas for product design: Insights from a case study. In P. P. Rau (Ed.), *Internationalization, design and global development - 4th international conference, IDGD 2011, held as part of HCI international 2011, orlando, fl, usa, july 9-14, 2011. proceedings* (pp. 37–46, Vol. 6775). https://doi.org/10.1007/978-3-642-21660-2_5
- Haake, M. (2006). *Virtual pedagogical agents-beyond the constraints of the computational approach* [Master's thesis]. Ergonomics; Aerosol Technology, Department of Design Sciences, Lund University.

- Hald, K. (2021). *Human-robot trust assessment from physical apprehension signals* [Doctoral dissertation]. Aalborg Universitetsforlag.
- Hald, K., Rehm, M., & Moeslund, T. B. (2021). Human-robot trust assessment using top-down visual tracking after robot task execution mistakes. *Proceedings of the 30th IEEE International Conference on Robot & Human Interactive Communication, RO-MAN 2021, Vancouver, BC, Canada, August 8-12, 2021*, 892–898. <https://doi.org/10.1109/RO-MAN50785.2021.9515501>
- Hald, K., Weitz, K., Rehm, M., & André, E. (2021). “An error occurred!” - trust repair with virtual robot using levels of mistake explanation. In K. Ogawa, T. Yonezawa, G. M. Lucas, H. Osawa, W. Johal, & M. Shiomi (Eds.), *Proceedings of the 9th international conference on human-agent interaction* (pp. 218–226). ACM. <https://doi.org/10.1145/3472307.3484170>
- Halvey, M. J., & Keane, M. T. (2007). An assessment of tag presentation techniques. In C. L. Williamson, M. E. Zurko, P. F. Patel-Schneider, & P. J. Shenoy (Eds.), *Proceedings of the 16th international conference on world wide web, WWW 2007, banff, alberta, canada, may 8-12, 2007* (pp. 1313–1314). ACM. <https://doi.org/10.1145/1242572.1242826>
- Hancock, P. A., Billings, D. R., Schaefer, K. E., Chen, J. Y., De Visser, E. J., & Parasuraman, R. (2011). A meta-analysis of factors affecting trust in human-robot interaction. *Human factors*, 53(5), 517–527.
- Hanney, K., & Keane, M. T. (1997). The adaptation knowledge bottleneck: How to ease it by learning from cases. *International Conference on Case-Based Reasoning*, 1266, 359–370. https://doi.org/10.1007/3-540-63233-6_506
- Harmon-Jones, C., Bastian, B., & Harmon-Jones, E. (2016). The discrete emotions questionnaire: A new tool for measuring state self-reported emotions. *PloS one*, 11(8), 1–25. <https://doi.org/10.1371/journal.pone.0159915>
- Hart, S. G., & Staveland, L. E. (1988). Development of nasa-tlx (task load index): Results of empirical and theoretical research. In *Advances in psychology* (pp. 139–183, Vol. 52). Elsevier. [https://doi.org/10.1016/S0166-4115\(08\)62386-9](https://doi.org/10.1016/S0166-4115(08)62386-9)
- Hassan, T., Seuß, D., Wollenberg, J., Weitz, K., Kunz, M., Lautenbacher, S., Garbas, J.-U., & Schmid, U. (2019). Automatic detection of pain from facial expressions: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(6), 1815–1831. <https://doi.org/10.1109/TPAMI.2019.2958341>
- Hassenzahl, M. (2008). User experience (ux) towards an experiential perspective on product quality. *Proceedings of the 20th Conference on l'Interaction Homme-Machine*, 11–15. <https://doi.org/10.1145/1512714.1512717>
- Hastie, H. F., Dente, P., Küster, D., & Kappas, A. (2016). Sound emblems for affective multimodal output of a robotic tutor: A perception study. In Y. I. Nakano, E. André, T. Nishida, L. Morency, C. Busso, & C. Pelachaud (Eds.), *Proceedings of the 18th ACM international conference on multimodal interaction, ICMi 2016, tokyo, japan, november 12-16, 2016* (pp. 256–260). ACM. <https://doi.org/10.1145/2993148.2993169>
- Hayes-Roth, F. (1985). Rule-based systems. *Communications of the ACM*, 28(9), 921–932. <https://doi.org/10.1145/4284.4286>
- Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica*, 47(1), 153–161. <https://doi.org/10.2307/1912352>
- Heider, F., & Simmel, M. (1944). An experimental study of apparent behavior. *The American Journal of Psychology*, 57(2), 243–259. <https://doi.org/10.2307/1416950>
- Heimerl, A., Baur, T., Lingenfelser, F., Wagner, J., & André, E. (2019). NOVA - a tool for explainable cooperative machine learning. *Proceedings of the 8th International Conference on Affective Computing and Intelligent Interaction, ACII 2019, Cambridge, United Kingdom, September 3-6, 2019*, 109–115. <https://doi.org/10.1109/ACII.2019.8925519>
- Heimerl, A., Weitz, K., Baur, T., & André, E. (2022). Unraveling ML models of emotion with NOVA: multi-level explainable AI for non-experts. *IEEE Transactions on Affective Computing*, 13(3), 1155–1167. <https://doi.org/10.1109/TAFFC.2020.3043603>
- Hellström, T., & Bensch, S. (2018). Understandable robots-what, why, and how. *Paladyn, Journal of Behavioral Robotics*, 9(1), 110–123. <https://doi.org/10.1515/pjbr-2018-0009>
- Hendler, J. (2008). Avoiding another AI winter. *IEEE Intelligent Systems*, 23(02), 2–4. <https://doi.org/10.1109/MIS.2008.20>

- Hill, T., Smith, N. D., & Mann, M. F. (1987). Role of efficacy expectations in predicting the decision to use advanced technologies: The case of computers. *Journal of Applied Psychology*, 72(2), 307–313. <https://doi.org/10.1037/0021-9010.72.2.307>
- Hilton, D. J. (1990). Conversational processes and causal explanation. *Psychological Bulletin*, 107(1), 65–81. <https://doi.org/10.1037/0033-2909.107.1.65>
- Hoff, K. A., & Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust. *Human Factors*, 57(3), 407–434. <https://doi.org/10.1177/0018720814547570>
- Hoffman, R. R., & Clancey, W. (2021). Methods and standards for research on explainable artificial intelligence: Lessons from intelligent tutoring systems. *Authorea Preprints*. <https://doi.org/10.22541/au.162317004.45114437/v1>
- Hoffman, R. R., Klein, G., & Mueller, S. T. (2018). Explaining explanation for “explainable AI”. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 62(1), 197–201. <https://doi.org/10.1177/1541931218621047>
- Hoffman, R. R., Mueller, S. T., Klein, G., & Litman, J. (2018). Metrics for explainable AI: challenges and prospects. <http://arxiv.org/abs/1812.04608>
- Holliday, D., Wilson, S., & Stumpf, S. (2016). User trust in intelligent systems: A journey over time. *Proceedings of the 21st International Conference on Intelligent User Interfaces, IUI 2016, Sonoma, CA, USA, March 7-10, 2016*, 164–168. <https://doi.org/10.1145/2856767.2856811>
- Holzinger, A., Biemann, C., Pattichis, C. S., & Kell, D. B. (2017). What do we need to build explainable AI systems for the medical domain? *arXiv preprint arXiv:1712.09923*. <http://arxiv.org/abs/1712.09923>
- Holzinger, A., Kargl, M., Kipperer, B., Regitnig, P., Plass, M., & Müller, H. (2022). Personas for artificial intelligence (AI) an open source toolbox. *IEEE Access*, 10, 23732–23747. <https://doi.org/10.1109/ACCESS.2022.3154776>
- Honig, S., & Oron-Gilad, T. (2018). Understanding and resolving failures in human-robot interaction: Literature review and model development. *Frontiers in Psychology*, 9, 861. <https://doi.org/10.3389/fpsyg.2018.00861>
- Hsieh, H.-F., & Shannon, S. E. (2005). Three approaches to qualitative content analysis. *Qualitative Health Research*, 15(9), 1277–1288. <https://doi.org/10.1177/1049732305276687>
- Hubel, D. H., & Wiesel, T. N. (1959). Receptive fields of single neurones in the cat’s striate cortex. *The Journal of Physiology*, 148(3), 574–591. <https://doi.org/10.1113/jphysiol.1959.sp006308>
- Hubel, D. H., & Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cat’s visual cortex. *The Journal of Physiology*, 160(1), 106–154. <https://doi.org/10.1113/jphysiol.1962.sp006837>
- Huber, T., Weitz, K., André, E., & Amir, O. (2021). Local and global explanations of agent behavior: Integrating strategy summaries with saliency maps. *Artificial Intelligence*, 301, 103571. <https://doi.org/10.1016/j.artint.2021.103571>
- Hussain, A., & Cambria, E. (2018). Semi-supervised learning for big social data analysis. *Neurocomputing*, 275, 1662–1673. <https://doi.org/10.1016/j.neucom.2017.10.010>
- Ibáñez-Espiga, M., Serio, A. D., & Kloos, C. D. (2014). Gamification for engaging computer science students in learning activities: A case study. *IEEE Transactions on Learning Technologies*, 7(3), 291–301. <https://doi.org/10.1109/TLT.2014.2329293>
- IJsselstein, W. A., de Kort, Y. A., & Poels, K. (2013). *The game experience questionnaire* (tech. rep.). Technische Universiteit Eindhoven.
- Jarrett, K., Kavukcuoglu, K., Ranzato, M. A., & LeCun, Y. (2009). What is the best multi-stage architecture for object recognition? *Proceedings of the IEEE 12th International Conference on Computer Vision, ICCV 2009, Kyoto, Japan, September 27 - October 4, 2009*, 2146–2153. <https://doi.org/10.1109/ICCV.2009.5459469>
- Jian, J.-Y., Bisantz, A. M., & Drury, C. G. (2000). Foundations for an empirically determined scale of trust in automated systems. *International Journal of Cognitive Ergonomics*, 4(1), 53–71. https://doi.org/10.1207/S15327566IJCE0401_04
- Jin, S.-A. A. (2010). The effects of incorporating a virtual agent in a computer-aided test designed for stress management education: The mediating role of enjoyment. *Computers in Human Behavior*, 26(3), 443–451. <https://doi.org/10.1016/j.chb.2009.12.003>

- Johnson, J., Alahi, A., & Fei-Fei, L. (2016). Perceptual losses for real-time style transfer and super-resolution. In B. Leibe, J. Matas, N. Sebe, & M. Welling (Eds.), *Computer vision - ECCV 2016 - 14th european conference, amsterdam, the netherlands, october 11-14, 2016, proceedings, part II* (pp. 694–711, Vol. 9906). https://doi.org/10.1007/978-3-319-46475-6_43
- Johnson-Laird, P. N. (1983). *Mental models: Towards a cognitive science of language, inference, and consciousness*. Harvard University Press.
- Junior, P. T. A., & Filgueiras, L. V. L. (2005). User modeling with personas. In M. C. C. Baranauskas & O. Mayora-Ibarra (Eds.), *CLIHIC '05, proceedings of the 2005 latin american conference on human-computer interaction, cuernavaca, mexico, october 23-26, 2005* (pp. 277–282). ACM. <https://doi.org/10.1145/1111360.1111388>
- Kakogeorgiou, I., & Karantzalos, K. (2021). Evaluating explainable artificial intelligence methods for multi-label deep learning classification tasks in remote sensing. *International Journal of Applied Earth Observation and Geoinformation*, 103, 102520. <https://doi.org/10.1016/j.jag.2021.102520>
- Kapaniaris, A. G. (2020). Folk culture and enriched digital teaching: Designing educational scenarios with the use of ict. In M. Kalogiannakis & S. Papadakis (Eds.), *Handbook of research on tools for teaching computational thinking in p-12 education* (pp. 484–505). IGI Global. <https://doi.org/10.4018/978-1-7998-4576-8.ch019>
- Kaptein, F., Broekens, J., Hindriks, K., & Neerincx, M. (2017). The role of emotion in self-explanations by cognitive agents. *Proceedings of the 7th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos, ACII Workshops 2017, San Antonio, TX, USA, October 23-26, 2017*, 88–93. <https://doi.org/10.1109/ACIIW.2017.8272595>
- Karrer, K., Glaser, C., Clemens, C., & Bruder, C. (2009). Technikaffinität erfassen—der fragebogen ta-eg. *Der Mensch im Mittelpunkt technischer Systeme*, 8, 196–201.
- Kass, R., & Finin, T. (1988). The need for user models in generating expert system explanations. *International Journal of Expert Systems*, 1(4).
- Kassam, K. S., Markey, A. R., Cherkassky, V. L., Loewenstein, G., & Just, M. A. (2013). Identifying emotions on the basis of neural activation. *PloS One*, 8(6), e66032. <https://doi.org/10.1371/journal.pone.0066032>
- Kaynak, O. (2021). The golden age of artificial intelligence. <https://doi.org/10.1007/s44163-021-00009-x>
- Keil, F. C. (2006). Explanation and understanding. *Annual Review of Psychology*, 57, 227–254. <https://doi.org/10.1146/annurev.psych.57.102904.190100>
- Kessler, T. T., Larios, C., Walker, T., Yerdon, V., & Hancock, P. (2017). A comparison of trust measures in human–robot interaction scenarios. In P. Savage-Knepshield & J. Chen (Eds.), *Advances in human factors in robots and unmanned systems* (pp. 353–364, Vol. 499). Springer. https://doi.org/10.1007/978-3-319-41959-6_29
- Khemlani, S., & Johnson-Laird, P. N. (2009). Disjunctive illusory inferences and how to eliminate them. *Memory and Cognition*, 37(5), 615–623. <https://doi.org/10.3758/MC.37.5.615>
- Khemlani, S., & Johnson-Laird, P. N. (2017). Illusions in reasoning. *Minds and Machines*, 27(1), 11–35. <https://doi.org/10.1007/s11023-017-9421-x>
- Khosravi, H., Shum, S. B., Chen, G., Conati, C., Tsai, Y.-S., Kay, J., Knight, S., Martinez-Maldonado, R., Sadiq, S., & Gašević, D. (2022). Explainable artificial intelligence in education. *Computers and Education: Artificial Intelligence*, 3, 100074. <https://doi.org/10.1016/j.caeai.2022.100074>
- Kim, M.-Y., Atakishiyev, S., Babiker, H. K. B., Farruque, N., Goebel, R., Zaïane, O. R., Motallebi, M.-H., Rabelo, J., Syed, T., Yao, H., et al. (2021). A multi-component framework for the analysis and design of explainable artificial intelligence. *Machine Learning and Knowledge Extraction*, 3(4), 900–921. <https://doi.org/10.3390/make3040045>
- Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. In Y. B. and Yann LeCun (Ed.), *Proceedings of the 3rd international conference on learning representations, ICLR 2015, san diego, ca, usa, may 7-9, 2015, conference track proceedings* (pp. 1–15).
- Kisler, T., Reichel, U., & Schiel, F. (2017). Multilingual processing of speech via web services. *Computer Speech & Language*, 45, 326–347. <https://doi.org/10.1016/j.csl.2017.01.005>
- Klein, J., Moon, Y., & Picard, R. W. (2002). This computer responds to user frustration: Theory, design, and results. *Interacting with Computers*, 14(2), 119–140. [https://doi.org/10.1016/S0953-5438\(01\)00053-4](https://doi.org/10.1016/S0953-5438(01)00053-4)

- Knezic, D., Wubbels, T., Elbers, E., & Hajer, M. (2010). The socratic dialogue and teacher education. *Teaching and Teacher Education*, 26(4), 1104–1111. <https://doi.org/https://doi.org/10.1016/j.tate.2009.11.006>
- Kohlbrenner, M. H. (2017). *On the stability of neural network explanations* [Bachelor's Thesis]. Berlin Institute of Technology.
- Kondracki, N. L., Wellman, N. S., & Amundson, D. R. (2002). Content analysis: Review of methods and their applications in nutrition education. *Journal of Nutrition Education and Behavior*, 34(4), 224–230. [https://doi.org/10.1016/S1499-4046\(06\)60097-3](https://doi.org/10.1016/S1499-4046(06)60097-3)
- Körber, M. (2018). Theoretical considerations and development of a questionnaire to measure trust in automation. In S. Bagnara, R. Tartaglia, S. Albolino, T. Alexander, & Y. Fujita (Eds.), *Proceedings of the 20th congress of the international ergonomics association (iea 2018)* (pp. 13–30, Vol. 823). Springer International Publishing. https://doi.org/10.1007/978-3-319-96074-6_2
- Körber, M., Baseler, E., & Bengler, K. (2018). Introduction matters: Manipulating trust in automation and reliance in automated driving. *Applied Ergonomics*, 66, 18–31. <https://doi.org/10.1016/j.apergo.2017.07.006>
- Kraus, T., Ganschow, L., Eisenträger, M., & Wischmann, S. (2021). *Erklärbare Künstliche Intelligenz - Anforderungen, Anwendungen, Lösungen*. Technologieprogramm KI-Innovationswettbewerb des Bundesministeriums für Wirtschaft und Energie, Begleitforschung iit-Institut für Innovation und Technik in der VDI/VDE Innovation Technik GmbH.
- Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6), 84–90. <https://doi.org/10.1145/3065386>
- Kulesza, T., Burnett, M., Wong, W.-K., & Stumpf, S. (2015). Principles of explanatory debugging to personalize interactive machine learning. In O. Brdiczka, P. Chau, G. Carenini, S. Pan, & P. O. Kristensson (Eds.), *Proceedings of the 20th international conference on intelligent user interfaces, IUI 2015, atlanta, ga, usa, march 29 - april 01, 2015* (pp. 126–137). ACM. <https://doi.org/10.1145/2678025.2701399>
- Kulesza, T., Stumpf, S., Burnett, M., Yang, S., Kwan, I., & Wong, W.-K. (2013). Too much, too little, or just right? ways explanations impact end-user' mental models. In C. Kelleher, M. M. Burnett, & S. Sauer (Eds.), *Proceedings of the 2013 IEEE symposium on visual languages and human centric computing, san jose, ca, usa, september 15-19, 2013* (pp. 3–10). IEEE Computer Society. <https://doi.org/10.1109/VLHCC.2013.6645235>
- Kunkel, J., Donkers, T., Michael, L., Barbu, C.-M., & Ziegler, J. (2019). Let me explain: Impact of personal and impersonal explanations on trust in recommender systems. In S. A. Brewster, G. Fitzpatrick, A. L. Cox, & V. Kostakos (Eds.), *Proceedings of the 2019 CHI conference on human factors in computing systems, CHI 2019, glasgow, scotland, uk, may 04-09, 2019* (pp. 1–12). ACM. <https://doi.org/10.1145/3290605.3300717>
- Laato, S., Tiainen, M., Islam, A. N., & Mäntymäki, M. (2022). How to explain ai systems to end-user: A systematic literature review and research agenda. *Internet Research*, 32(7), 1–31.
- Langer, E. J. (1992). Matters of mind: Mindfulness/mindlessness in perspective. *Consciousness and Cognition*, 1(3), 289–305. [https://doi.org/10.1016/1053-8100\(92\)90066-J](https://doi.org/10.1016/1053-8100(92)90066-J)
- Langer, E., Blank, A., & Chanowitz, B. (1978). The mindlessness of ostensibly thoughtful action: The role of “placebic” information in interpersonal interaction. *Journal of Personality and Social Psychology*, 36(6), 635–642. <https://doi.org/10.1037/0022-3514.36.6.635>
- Langer, M., Oster, D., Speith, T., Hermanns, H., Kästner, L., Schmidt, E., Sesing, A., & Baum, K. (2021). What do we want from explainable artificial intelligence (XAI)?—a stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research. *Artificial Intelligence*, 296, 103473. <https://doi.org/10.1016/j.artint.2021.103473>
- Lapuschkin, S., Binder, A., Müller, K.-R., & Samek, W. (2017). Understanding and comparing deep neural networks for age and gender classification. *Proceedings of the 2017 IEEE International Conference on Computer Vision Workshops, ICCV Workshops 2017, Venice, Italy, October 22-29, 2017*, 1629–1638. <https://doi.org/10.1109/ICCVW.2017.191>
- Latikka, R., Turja, T., & Oksanen, A. (2019). Self-efficacy and acceptance of robots. *Comput. Hum. Behav.*, 93, 157–163. <https://doi.org/10.1016/j.chb.2018.12.017>
- Lawrence, C. (2019). Requiem for rethink robotics. *Mechanical Engineering*, 141(02), 40–45. <https://doi.org/10.1115/1.2019-FEB3>
- LeCun, Y., et al. (1989). Generalization and network design strategies. *Connectionism in Perspective*, 19, 143–155.

- LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521, 436–444. <https://doi.org/10.1038/nature14539>
- LeCun, Y., Bottou, L., Orr, G. B., & Müller, K.-R. (1998). Efficient backprop. In *Neural networks: Tricks of the trade* (pp. 9–50). Springer Berlin Heidelberg. https://doi.org/10.1007/3-540-49430-8_2
- LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11), 2278–2324. <https://doi.org/10.1109/5.726791>
- Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors*, 46(1), 50–80. <https://doi.org/10.1518/hfes.46.1.50.30392>
- Leite, I., Castellano, G., Pereira, A., Martinho, C., & Paiva, A. (2012). Modelling empathic behaviour in a robotic game companion for children: An ethnographic study in real-world settings. *Proceedings of the International Conference on Human-Robot Interaction, HRI'12, Boston, MA, USA - March 05 - 08, 2012*, 367–374. <https://doi.org/10.1145/2157689.2157811>
- Lester, J. C., Converse, S. A., Kahler, S. E., Barlow, S. T., Stone, B. A., & Bhogal, R. S. (1997). The persona effect: Affective impact of animated pedagogical agents. In S. Pemberton (Ed.), *Human factors in computing systems, CHI '97 conference proceedings, atlanta, georgia, usa, march 22-27, 1997* (pp. 359–366). ACM/Addison-Wesley. <https://doi.org/10.1145/258549.258797>
- Lewis, J. D., & Weigert, A. (1985). Trust as a social reality. *Social Forces*, 63(4), 967–985. <https://doi.org/10.1093/sf/63.4.967>
- Li, C.-Y., Ortega, D., Văth, D., Lux, F., Vanderlyn, L., Schmidt, M., Neumann, M., Völkel, M., Denisov, P., Jenne, S., Karacevic, Z., & Vu, N. T. (2020). ADVISER: A toolkit for developing multi-modal, multi-domain and socially-engaged conversational agents. In A. Celikyilmaz & T. Wen (Eds.), *Proceedings of the 58th annual meeting of the association for computational linguistics: System demonstrations, ACL 2020, online, july 5-10, 2020*. Association for Computational Linguistics. <https://doi.org/10.18653/v1/2020.acl-demos.31>
- Liao, Q. V., Pribić, M., Han, J., Miller, S., & Sow, D. (2021). Question-driven design process for explainable AI user experiences. *arXiv preprint arXiv:2104.03483*. <https://arxiv.org/abs/2104.03483>
- Lighthill, P. S. J. (1973). Artificial intelligence: A general survey. *Artificial Intelligence: a paper symposium*.
- Lim, B. Y., Dey, A. K., & Avrahami, D. (2009). Why and why not explanations improve the intelligibility of context-aware intelligent systems. In D. R. Olsen Jr., R. B. Arthur, K. Hinckley, M. Ringel Morris, S. E. Hudson, & S. Greenberg (Eds.), *Proceedings of the 27th international conference on human factors in computing systems, CHI 2009, boston, ma, usa, april 4-9, 2009* (pp. 2119–2128). ACM. <https://doi.org/10.1145/1518701.1519023>
- Limakrisna, N., Wahyuni, S., & Badriah, P. C. (2021). The influence of self efficacy on cognitive load in online learning in the pandemic covid-19 on nursing students. *European Journal of Molecular & Clinical Medicine*, 7(11), 2562–2575.
- Lin, M., Chen, Q., & Yan, S. (2014). Network in network. *arXiv preprint arXiv:1312.4400*. <http://arxiv.org/abs/1312.4400>
- Lindner, A., Seegerer, S., & Romeike, R. (2019). Unplugged activities in the context of AI. In S. N. Pozdniakov & V. Dagiene (Eds.), *Informatics in schools. new ideas in school informatics - 12th international conference on informatics in schools: Situation, evolution, and perspectives, ISSEP 2019, larnaca, cyprus, november 18-20, 2019, proceedings* (pp. 123–135, Vol. 11913). https://doi.org/10.1007/978-3-030-33759-9_10
- Lipton, Z. C. (2018). The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 16(3), 31–57. <https://doi.org/10.1145/3236386.3241340>
- Lombrozo, T. (2006). The structure and function of explanations. *Trends in Cognitive Sciences*, 10(10), 464–470. <https://doi.org/10.1016/j.tics.2006.08.004>
- Lombrozo, T. (2007). Simplicity and probability in causal explanation. *Cognitive Psychology*, 55(3), 232–257. <https://doi.org/10.1016/j.cogpsych.2006.09.006>
- Lombrozo, T. (2012). Explanation and abductive inference. In K. J. Holyoak & R. G. Morrison (Eds.), *The oxford handbook of thinking and reasoning* (pp. 260–276). Oxford University Press.
- Long, D., & Magerko, B. (2020). What is AI literacy? competencies and design considerations. In R. Bernhaupt, F. ' Mueller, D. Verweij, J. Andres, J. McGrenere, A. Cockburn, I. Avellino, A. Goguey, P. Bjøn, S. Zhao, B. P. Samson, & R. Kocielnik (Eds.), *Proceedings of CHI '20: CHI conference on human factors in computing systems, honolulu, hi, usa, april 25-30, 2020* (pp. 1–16). ACM. <https://doi.org/10.1145/3313831.3376727>

- Loyola-Gonzalez, O. (2019). Black-box vs. white-box: Understanding their advantages and weaknesses from a practical point of view. *IEEE Access*, 7, 154096–154113. <https://doi.org/10.1109/ACCESS.2019.2949286>
- Lu, Y., & Sarter, N. B. (2019). Eye tracking: A process-oriented method for inferring trust in automation as a function of priming and system reliability. *IEEE Transactions on Human-Machine Systems*, 49(6), 560–568. <https://doi.org/10.1109/THMS.2019.2930980>
- Lucieri, A., Bajwa, M. N., Dengel, A., & Ahmed, S. (2020). Achievements and challenges in explaining deep learning based computer-aided diagnosis systems. *arXiv preprint arXiv:2011.13169*. <https://arxiv.org/abs/2011.13169>
- Luckin, R., Holmes, W., Griffiths, M., & Forcier, L. B. (2016). Intelligence unleashed: An argument for AI in education.
- Luger, E., & Sellen, A. (2016). “Like having a really bad PA” the gulf between user expectation and experience of conversational agents. In J. Kaye, A. Druin, C. Lampe, D. Morris, & J. P. Hourcade (Eds.), *Proceedings of the 2016 CHI conference on human factors in computing systems, san jose, ca, usa, may 7-12, 2016* (pp. 5286–5297). ACM. <https://doi.org/10.1145/2858036.2858288>
- Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, & R. Garnett (Eds.), *Proceedings of the advances in neural information processing systems 30: Annual conference on neural information processing systems 2017, december 4-9, 2017, long beach, ca, USA* (pp. 4768–4777). Curran Associates, Inc.
- Lyons, J. B. (2013). Being transparent about transparency: A model for human-robot interaction. *Trust and Autonomous Systems, Papers from the 2013 AAAI Spring Symposium, Palo Alto, California, USA, March 25-27, 2013, SS-13-07*.
- Madsen, M., & Gregor, S. (2000). Measuring human-computer trust. *Proceedings of the 11th Australasian Conference on Information Systems*, 53, 6–8.
- Malik, A. A., & Bilberg, A. (2019). Developing a reference model for human-robot interaction. *International Journal on Interactive Design and Manufacturing*, 13(4), 1541–1547. <https://doi.org/10.1007/s12008-019-00591-6>
- Mandinach, E. B., & Gummer, E. S. (2013). A systemic view of implementing data literacy in educator preparation. *Educational Researcher*, 42(1), 30–37. <https://doi.org/10.3102/0013189X12459803>
- Marsh, S., & Dibben, M. R. (2005). Trust, untrust, distrust and mistrust – an exploration of the dark(er) side. In P. Herrmann, V. Issarny, & S. Shiu (Eds.), *Trust management, 3rd international conference, itrust 2005, paris, france, may 23-26, 2005, proceedings* (pp. 17–33, Vol. 3477). Springer Berlin Heidelberg. https://doi.org/10.1007/11429760_2
- McCarthy, J., Minsky, M. L., Rochester, N., & Shannon, C. E. (2006). A proposal for the dartmouth summer research project on artificial intelligence, august 31, 1955. *AI Magazine*, 27(4), 12–12. <https://doi.org/10.1609/aimag.v27i4.1904>
- McCorduck, P. (2004). *Machines who think: A personal inquiry into the history and prospects of artificial intelligence*. CRC Press.
- Merritt, S. M., & Ilgen, D. R. (2008). Not all trust is created equal: Dispositional and history-based trust in human-automation interactions. *Human Factors*, 50(2), 194–210. <https://doi.org/10.1518/001872008X288574>
- Mertes, S., Huber, T., Weitz, K., Heimerl, A., & André, E. (2022). GANterfactual-counterfactual explanations for medical non-experts using generative adversarial learning. *Frontiers in Artificial Intelligence*, 5, 825565. <https://doi.org/10.3389/frai.2022.825565>
- Millecamp, M., Htun, N. N., Conati, C., & Verbert, K. (2019). To explain or not to explain: The effects of personal characteristics when explaining music recommendations. In W. Fu, S. Pan, O. Brdiczka, P. Chau, & G. Calvary (Eds.), *Proceedings of the 24th international conference on intelligent user interfaces, IUI 2019, marina del ray, ca, usa, march 17-20, 2019* (pp. 397–407). ACM. <https://doi.org/10.1145/3301275.3302313>
- Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267, 1–38. <https://doi.org/10.1016/j.artint.2018.07.007>
- Miller, T., Howe, P., & Sonenberg, L. (2017). Explainable AI: Beware of Inmates Running the Asylum. *arXiv preprint arXiv:1712.00547*. <http://arxiv.org/abs/1712.00547>
- Mitchell, T. M. (1997). *Machine learning*.

- Mohorovicic, S., & Strcic, V. (2011). An overview of computer programming teaching methods. *Proceedings of the 22nd Central European Conference on Information and Intelligent Systems, Varaždin, Croatia, September 21-12, 2011*, 47–52.
- Mollahosseini, A., Hassani, B., & Mahoor, M. H. (2019). Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, 10(1), 18–31. <https://doi.org/10.1109/TAFFC.2017.2740923>
- Molnar, C. (2019). *Interpretable machine learning: A guide for making black box models explainable*. <https://christophm.github.io/interpretable-ml-book/>
- Montavon, G., Binder, A., Lapuschkin, S., Samek, W., & Müller, K. (2019). Layer-wise relevance propagation: An overview. In W. Samek, G. Montavon, A. Vedaldi, L. K. Hansen, & K. Müller (Eds.), *Explainable AI: interpreting, explaining and visualizing deep learning* (pp. 193–209, Vol. 11700). Springer. https://doi.org/10.1007/978-3-030-28954-6_10
- Montavon, G., Samek, W., & Müller, K.-R. (2018). Methods for interpreting and understanding deep neural networks. *Digital Signal Processing*, 73, 1–15. <https://doi.org/10.1016/j.dsp.2017.10.011>
- Mueller, S. T., Veinott, E. S., Hoffman, R. R., Klein, G., Alam, L., Mamun, T., & Clancey, W. J. (2021). Principles of explanation in human-ai systems. *arXiv preprint arXiv:2102.04972*. <https://arxiv.org/abs/2102.04972>
- Nam, J. G., Park, S., Hwang, E. J., Lee, J. H., Jin, K.-N., Lim, K. Y., Vu, T. H., Sohn, J. H., Hwang, S., Goo, J. M., et al. (2019). Development and validation of deep learning–based automatic detection algorithm for malignant pulmonary nodules on chest radiographs. *Radiology*, 290(1), 218–228. <https://doi.org/10.1148/radiol.2018180237>
- Nass, C., & Moon, Y. (2000). Machines and mindlessness: Social responses to computers. *Journal of Social Issues*, 56(1), 81–103. <https://doi.org/https://doi.org/10.1111/0022-4537.00153>
- Neerincx, M. A., Waa, J. v. d., Kaptein, F., & Diggelen, J. v. (2018). Using perceptual and cognitive explanations for enhanced human-agent team performance. In D. Harris (Ed.), *Engineering psychology and cognitive ergonomics - 15th international conference, EPCE 2018, held as part of HCI international 2018, las vegas, nv, usa, july 15-20, 2018, proceedings* (pp. 204–214). https://doi.org/10.1007/978-3-319-91122-9_18
- Nikolaidis, S., Kwon, M., Forlizzi, J., & Srinivasa, S. (2018). Planning with verbal communication for human-robot collaboration. *ACM Transactions on Human-Robot Interaction*, 7(3), 1–21. <https://doi.org/10.1145/3203305>
- Norman, D. A. (1983). Some observations on mental models. In D. Gentner & A. L. Stevens (Eds.), *Mental models*. Psychology Press. <https://doi.org/10.4324/9781315802725>
- Nourani, M., Roy, C., Block, J. E., Honeycutt, D. R., Rahman, T., Ragan, E., & Gogate, V. (2021). Anchoring bias affects mental model formation and user reliance in explainable ai systems. In T. Hammond, K. Verbert, D. Parra, B. P. Knijnenburg, J. O'Donovan, & P. Teale (Eds.), *Proceedings of IUI '21: 26th international conference on intelligent user interfaces, college station, tx, usa, april 13-17, 2021* (pp. 340–350). <https://doi.org/10.1145/3397481.3450639>
- Olds, B. M., Moskal, B. M., & Miller, R. L. (2005). Assessment in engineering education: Evolution, approaches and future collaborations. *Journal of Engineering Education*, 94(1), 13–25. <https://doi.org/10.1002/j.2168-9830.2005.tb00826.x>
- Oxford Learner's Dictionary. (2022). Oxford learner's dictionary definition of power user. Retrieved September 15, 2022, from <https://www.oxfordlearnersdictionaries.com/definition/english/power-user>
- Park, D. H., Hendricks, L. A., Akata, Z., Rohrbach, A., Schiele, B., Darrell, T., & Rohrbach, M. (2018). Multimodal explanations: Justifying decisions and pointing to the evidence. *Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, 8779–8788. <https://doi.org/10.1109/CVPR.2018.00915>
- Park, E. H., Werder, K., Cao, L., & Ramesh, B. (2022). Why do family members reject ai in health care? competing effects of emotions. *Journal of Management Information Systems*, 39(3), 765–792. <https://doi.org/10.1080/07421222.2022.2096550>
- Partridge, D. (1991). *A new guide to artificial intelligence*. Intellect Books.
- Pedreschi, D., Giannotti, F., Guidotti, R., Monreale, A., Ruggieri, S., & Turini, F. (2019). Meaningful explanations of black box ai decision systems. *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth*

- AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019, 33(01), 9780–9784. <https://doi.org/10.1609/aaai.v33i01.33019780>
- Perugini, M., Gallucci, M., & Costantini, G. (2018). A practical primer to power analysis for simple experimental designs. *International Review of Social Psychology*, 31(1), 1–23. <https://doi.org/10.5334/irsp.181>
- Peters, D., Vold, K., Robinson, D., & Calvo, R. A. (2020). Responsible ai—two frameworks for ethical design practice. *IEEE Transactions on Technology and Society*, 1(1), 34–47. <https://doi.org/10.1109/TTS.2020.2974991>
- Petrak, B., Weitz, K., Aslan, I., & André, E. (2019). Let me show you your new home: Studying the effect of proxemic-awareness of robots on users' first impressions. *Proceedings of the 28th IEEE International Conference on Robot and Human Interactive Communication, RO-MAN 2019, New Delhi, India, October 14-18, 2019*, 1–7. <https://doi.org/10.1109/RO-MAN46459.2019.8956463>
- Phelps, E. A. (2004). Human emotion and memory: Interactions of the amygdala and hippocampal complex. *Current Opinion in Neurobiology*, 14(2), 198–202. <https://doi.org/10.1016/j.conb.2004.03.015>
- Phelps, E. A., Ling, S., & Carrasco, M. (2006). Emotion facilitates perception and potentiates the perceptual benefits of attention. *Psychological Science*, 17(4), 292–299. <https://doi.org/10.1111/j.1467-9280.2006.01701.x>
- Picard, R. W., & Klein, J. (2002). Computers that recognise and respond to user emotion: Theoretical and practical implications. *Interacting with Computers*, 14(2), 141–169. [https://doi.org/10.1016/S0953-5438\(01\)00055-8](https://doi.org/10.1016/S0953-5438(01)00055-8)
- Plutchik, R. (1982). A psychoevolutionary theory of emotions. 21, 529–553. <https://doi.org/10.1177/053901882021004003>
- Polyak, S. T., von Davier, A., & Peterschmidt, K. (2017). Analyzing game-based collaborative problem solving with computational psychometrics. *Frontiers in Psychology*, 8, 1–16. <https://doi.org/10.3389/fpsyg.2017.02029>
- Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences*, 1(4), 515–526. <https://doi.org/10.1017/S0140525X00076512>
- Pylyshyn, Z. W. (1978). When is attribution of beliefs justified? [p&w]. *Behavioral and Brain Sciences*, 1(4), 592–593. <https://doi.org/10.1017/S0140525X00076895>
- Rabold, J., Deininger, H., Siebers, M., & Schmid, U. (2019). Enriching visual with verbal explanations for relational concepts - combining LIME with aleph. In P. Cellier & K. Driessens (Eds.), *Machine learning and knowledge discovery in databases - international workshops of ECML PKDD 2019, würzburg, germany, september 16-20, 2019, proceedings, part I* (pp. 180–192, Vol. 1167). Springer. https://doi.org/10.1007/978-3-030-43823-4_16
- Ravenet, B., Clavel, C., & Pelachaud, C. (2018). Automatic nonverbal behavior generation from image schemas. In E. André, S. Koenig, M. Dastani, & G. Sukthankar (Eds.), *Proceedings of the 17th international conference on autonomous agents and multiagent systems, AAMAS 2018, stockholm, sweden, july 10-15, 2018* (pp. 1667–1674). ACM.
- Redifer, J. L., Bae, C. L., & Zhao, Q. (2021). Self-efficacy and performance feedback: Impacts on cognitive load during creative thinking. *Learning and Instruction*, 71, 101395. <https://doi.org/10.1016/j.learninstruc.2020.101395>
- Rehm, M. (2021). Multimodal and multicultural field agents: Considerations for “outside-the-lab” studies. In J. Miehle, W. Minker, E. André, & K. Yoshino (Eds.), *Multimodal agents for ageing and multicultural societies: Communications of nii shonan meetings* (pp. 71–85). Springer Singapore. https://doi.org/10.1007/978-981-16-3476-5_4
- Rehse, J.-R., Mehdiyev, N., & Fettke, P. (2019). Towards explainable process predictions for industry 4.0 in the DFKI-smart-lego-factory. *KI-Künstliche Intelligenz*, 33(2), 181–187. <https://doi.org/10.1007/s13218-019-00586-1>
- Reich-Stiebert, N., Eyssel, F., & Hohnemann, C. (2019). Involve the user! changing attitudes toward robots by user participation in a robot prototyping process. *Comput. Hum. Behav.*, 91, 290–296. <https://doi.org/10.1016/j.chb.2018.09.041>
- Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Why should I trust you?: Explaining the predictions of any classifier. In B. Krishnapuram, M. Shah, A. J. Smola, C. C. Aggarwal, D. Shen, & R. Rastogi (Eds.), *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*,

- san francisco, ca, usa, august 13-17, 2016 (pp. 1135–1144). ACM. <https://doi.org/10.1145/2939672.2939778>
- Ribera, M., & Lapedriza, A. (2019). Can we do better explanations? a proposal of user-centered explainable AI. In C. Trattner, D. Parra, & N. Riche (Eds.), *Joint proceedings of the ACM IUI 2019 workshops co-located with the 24th ACM conference on intelligent user interfaces (ACM IUI 2019), los angeles, usa, march 20, 2019* (p. 38, Vol. 2327). CEUR-WS.org.
- Riedl, M. O. (2019). Human-centered artificial intelligence and machine learning. *Human Behavior and Emerging Technologies*, 1(1), 33–36. <https://doi.org/10.1002/hbe2.117>
- Robinette, P., Howard, A. M., & Wagner, A. R. (2015). Timing is key for robot trust repair. In A. Tapus, E. André, J. Martin, F. Ferland, & M. Ammi (Eds.), *Social robotics - 7th international conference, ICSR 2015, paris, france, october 26-30, 2015, proceedings* (pp. 574–583, Vol. 9388). Springer. https://doi.org/10.1007/978-3-319-25554-5_57
- Robinson, S., Traum, D. R., Ittycheriah, M., & Henderer, J. (2008). What would you ask a conversational agent? observations of human-agent dialogues in a museum setting. *Proceedings of the International Conference on Language Resources and Evaluation, LREC 2008, 26 May - 1 June 2008, Marrakech, Morocco*.
- Roschelle, J., & Teasley, S. D. (1995). The construction of shared knowledge in collaborative problem solving. In C. O'Malley (Ed.), *Computer supported collaborative learning* (pp. 69–97). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-85098-1_5
- Rudd, J., Stern, K., & Isensee, S. (1996). Low vs. high-fidelity prototyping debate. *interactions*, 3(1), 76–85. <https://doi.org/10.1145/223500.223514>
- Russell, S., & Norvig, P. (2016). *Artificial intelligence: A modern approach, global edition*. Pearson.
- Rutjes, H., Willemsen, M., & IJsselstein, W. (2019). Considerations on explainable ai and users' mental models. *Proceedings of the CHI 2019 Workshop: Where is the Human? Bridging the Gap Between AI and HCI*, 1–5.
- Sainath, T. N., & Parada, C. (2015). Convolutional neural networks for small-footprint keyword spotting. *Proceedings of the INTERSPEECH 2015, 16th Annual Conference of the International Speech Communication Association, Dresden, Germany, September 6-10, 2015*, 1478–1482.
- Sakai, T., Miyazawa, K., Horii, T., & Nagai, T. (2021). A framework of explanation generation toward reliable autonomous robots. *Advanced Robotics*, 35(17), 1054–1067. <https://doi.org/10.1080/01691864.2021.1946423>
- Salem, M., Lakatos, G., Amirabdollahian, F., & Dautenhahn, K. (2015). Would you trust a (faulty) robot? effects of error, task type and personality on human-robot cooperation and trust. In J. A. Adams, W. D. Smart, B. Mutlu, & L. Takayama (Eds.), *Proceedings of the tenth annual ACM/IEEE international conference on human-robot interaction, HRI 2015, portland, or, usa, march 2-5, 2015* (pp. 141–148). ACM. <https://doi.org/10.1145/2696454.2696497>
- Sanders, T., Oleson, K. E., Billings, D. R., Chen, J. Y. C., & Hancock, P. A. (2011). A model of human-robot trust: Theoretical model development. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 55(1), 1432–1436. <https://doi.org/10.1177/1071181311551298>
- Schaefer, K. (2013). *The perception and measurement of human-robot trust* [Doctoral dissertation]. University of Florida.
- Schallner, L., Rabold, J., Scholz, O., & Schmid, U. (2019). Effect of superpixel aggregation on explanations in lime—a case study with biological data. *Machine Learning and Knowledge Discovery in Databases - International Workshops of ECML PKDD 2019, Würzburg, Germany, September 16-20, 2019, Proceedings, Part I*, 1167, 147–158. https://doi.org/10.1007/978-3-030-43823-4_13
- Schmid, U. (2018). Inductive programming as approach to comprehensible machine learning. In C. Beierle, G. Kern-Isberner, M. Ragni, F. Stolzenburg, & M. Thimm (Eds.), *Proceedings of the 7th workshop on dynamics of knowledge and belief (DKB-2018) and the 6th workshop KI & kognition (KIK-2018) co-located with 41st german conference on artificial intelligence (KI 2018), berlin, germany, september 25, 2018* (Vol. 2194). CEUR-WS.org.
- Schmid, U., & Finzel, B. (2020). Mutual explanations for cooperative decision making in medicine. *KI-Künstliche Intelligenz*, 34(2), 227–233. <https://doi.org/10.1007/s13218-020-00633-2>

- Schmid, U., Tresp, V., Bethge, M., Kersting, K., & Stiefelhagen, R. (2021). Künstliche Intelligenz–die dritte Welle. In R. H. Reussner, A. Koziol, & R. Heinrich (Eds.), *50. Jahrestagung der Gesellschaft für Informatik, INFORMATIK 2020 - back to the future, Karlsruhe, Germany, 28. September - 2. Oktober 2020* (pp. 91–95, Vol. P-307). GI. https://doi.org/10.18420/inf2020_08
- Schneider, J., & Handali, J. (2019). Personalized explanation in machine learning: A conceptualization. *arXiv preprint arXiv:1901.00770*. <https://arxiv.org/abs/1901.00770>
- Schulze Kissing, D., & Bruder, C. (2016). Der Einsatz Synthetischer Aufgabenumgebungen zur Untersuchung kollaborativer Prozesse in Leitzentralen am Beispiel der “Generic Control Center Task Environment”(ConCent). *Kognitive Systeme*, 2016(1). <https://doi.org/10.17185/dupublico/42384>
- Scott, S. K., Young, A. W., Calder, A. J., Hellawell, D. J., Aggleton, J. P., & Johnsons, M. (1997). Impaired auditory recognition of fear and anger following bilateral amygdala lesions. *Nature*, 385, 254–257. <https://doi.org/10.1038/385254a0>
- Selvaraju, R. R., Das, A., Vedantam, R., Cogswell, M., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. *Proceedings of the IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, 618–626. <https://doi.org/10.1109/ICCV.2017.74>
- Sheh, R. K.-M. (2017). “why did you do that?” explainable intelligent robots. *The Workshops of the The Thirty-First AAAI Conference on Artificial Intelligence, Saturday, February 4-9, 2017, San Francisco, California, USA*, WS-17.
- Shneiderman, B. (2000). The limits of speech recognition. *Communications of the ACM*, 43(9), 63–65. <https://doi.org/10.1145/348941.348990>
- Shneiderman, B. (2020a). Bridging the gap between ethics and practice: Guidelines for reliable, safe, and trustworthy human-centered AI systems. *ACM Transactions on Interactive Intelligent Systems*, 10(4), 1–31. <https://doi.org/10.1145/3419764>
- Shneiderman, B. (2020b). Human-centered artificial intelligence: Reliable, safe & trustworthy. *International Journal of Human–Computer Interaction*, 36(6), 495–504. <https://doi.org/10.1080/10447318.2020.1741118>
- Shneiderman, B. (2020c). Human-centered artificial intelligence: Three fresh ideas. *AIS Transactions on Human-Computer Interaction*, 12(3), 109–124. <https://doi.org/10.17705/1thci.00131>
- Shortliffe, E. H., & Buchanan, B. G. (1975). A model of inexact reasoning in medicine. *Mathematical Biosciences*, 23(3-4), 351–379. [https://doi.org/10.1016/0025-5564\(75\)90047-4](https://doi.org/10.1016/0025-5564(75)90047-4)
- Shortliffe, E. H., Davis, R., Axline, S. G., Buchanan, B. G., Green, C. C., & Cohen, S. N. (1975). Computer-based consultations in clinical therapeutics: Explanation and rule acquisition capabilities of the mycin system. *Computers and Biomedical Research*, 8(4), 303–320. [https://doi.org/10.1016/0010-4809\(75\)90009-9](https://doi.org/10.1016/0010-4809(75)90009-9)
- Siebers, M., & Schmid, U. (2019). Please delete that! why should I? explaining learned irrelevance classifications of digital objects. *KI - Künstliche Intelligenz*, 33(1), 35–44. <https://doi.org/10.1007/s13218-018-0565-5>
- Siegler, R. S., et al. (2002). Microgenetic studies of self-explanation. In N. Granott & J. Parziale (Eds.), *Microdevelopment: Transition processes in development and learning* (pp. 31–58).
- Simonyan, K., Vedaldi, A., & Zisserman, A. (2014). Deep inside convolutional networks: Visualising image classification models and saliency maps. In Y. Bengio & Y. LeCun (Eds.), *Proceedings of the 2nd international conference on learning representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, workshop track proceedings* (pp. 1–8).
- Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*. <http://arxiv.org/abs/1409.1556>
- Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1), 1929–1958. <https://doi.org/10.5555/2627435.2670313>
- Stange, S., Buschmeier, H., Hassan, T., Ritter, C., & Kopp, S. (2019). Towards self-explaining social robots. verbal explanation strategies for a needs-based architecture. *Workshop on Cognitive Architectures for Human–Robot Interaction, AAMAS 2019, Montreal, Canada*.
- Stange, S., & Kopp, S. (2020). Effects of a social robot’s self-explanations on how humans understand and evaluate its behavior. In T. Belpaeme, J. E. Young, H. Gunes, & L. D. Riek (Eds.), *Proceedings of the HRI*

- '20: ACM/IEEE international conference on human-robot interaction, cambridge, united kingdom, march 23-26, 2020 (pp. 619–627). ACM. <https://doi.org/10.1145/3319502.3374802>
- Stangl, B. (2012). Emotional mental models. In N. M. Seel (Ed.), *Encyclopedia of the Sciences of Learning* (pp. 1131–1133). Springer US. https://doi.org/10.1007/978-1-4419-1428-6_728
- Stanton, B., & Jensen, T. (2021). Trust and artificial intelligence. <https://doi.org/10.6028/NIST.IR.8332-draft>
- Stieler, F., Rabe, F., & Bauer, B. (2021). Towards domain-specific explainable ai: Model interpretation of a skin image classifier using a human approach. *IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2021, virtual, June 19-25, 2021*, 1802–1809. <https://doi.org/10.1109/CVPRW53098.2021.00199>
- Sundar, S. S., Jung, E. H., Waddell, T. F., & Kim, K. J. (2017). Cheery companions or serious assistants? role and demeanor congruity as predictors of robot attraction and use intentions among senior citizens. *International Journal of Human-Computer Studies*, 97, 88–97. <https://doi.org/10.1016/j.ijhcs.2016.08.006>
- Sutton, R. T., Pincock, D., Baumgart, D. C., Sadowski, D. C., Fedorak, R. N., & Kroeker, K. I. (2020). An overview of clinical decision support systems: Benefits, risks, and strategies for success. *NPJ digital medicine*, 3(1), 1–10. <https://doi.org/10.1038/s41746-020-0221-y>
- Sweller, J., Van Merriënboer, J. J., & Paas, F. G. (1998). Cognitive architecture and instructional design. *Educational Psychology Review*, 10(3), 251–296. <https://doi.org/10.1023/A:1022193728205>
- Tenhundfeld, N. L., Barr, H. M., Emily, H., & Weger, K. (2021). Is my Siri the same as your Siri? an exploration of users' mental model of virtual personal assistants, implications for trust. *IEEE Transactions on Human-Machine Systems*, 52(3), 512–521. <https://doi.org/10.1109/THMS.2021.3107493>
- Terhürne, P., Schwartz, B., Baur, T., Schiller, D., Eberhardt, S. T., André, E., & Lutz, W. (2022). Validation and application of the non-verbal behavior analyzer: An automated tool to assess non-verbal emotional expressions in psychotherapy. *Frontiers in Psychiatry*, 2488. <https://doi.org/10.3389/fpsy.2022.1026015>
- Thorndike, E. L. (1920). A constant error in psychological ratings. *Journal of Applied Psychology*, 4(1), 25–29. <https://doi.org/https://doi.org/10.1037/h0071663>
- Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58, 267–288. <https://doi.org/10.1111/j.2517-6161.1996.tb02080.x>
- Tiddi, I., d'Aquin, M., & Motta, E. (2015). An ontology design pattern to define explanations. In K. Barker & J. M. Gómez-Pérez (Eds.), *Proceedings of the 8th international conference on knowledge capture, K-CAP 2015, palisades, ny, usa, october 7-10, 2015* (pp. 1–8). ACM. <https://doi.org/10.1145/2815833.2815844>
- Tjoa, E., & Guan, C. (2020). A survey on explainable artificial intelligence (XAI): toward medical XAI. *IEEE Transactions on Neural Networks and Learning Systems*, 32(11), 4793–4813. <https://doi.org/10.1109/tnnls.2020.3027314>
- Tomsett, R., Braines, D., Harborne, D., Preece, A., & Chakraborty, S. (2018). Interpretable to whom? a role-based model for analyzing interpretable machine learning systems. *arXiv preprint arXiv:1806.07552*. <http://arxiv.org/abs/1806.07552>
- Tomsett, R., Harborne, D., Chakraborty, S., Gurram, P., & Preece, A. (2020). Sanity checks for saliency metrics. *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, 34(04), 6021–6029.
- Tonekaboni, S., Joshi, S., McCradden, M. D., & Goldenberg, A. (2019). What clinicians want: Contextualizing explainable machine learning for clinical end use. In F. Doshi-Velez, J. Fackler, K. Jung, D. C. Kale, R. Ranganath, B. C. Wallace, & J. Wiens (Eds.), *Proceedings of the machine learning for healthcare conference, MLHC 2019, 9-10 august 2019, ann arbor, michigan, USA* (pp. 359–380, Vol. 106). PMLR.
- van Lent, M., Fisher, W., & Mancuso, M. (2004). An explainable artificial intelligence system for small-unit tactical behavior. In D. L. M. and George Ferguson (Ed.), *Proceedings of the nineteenth national conference on artificial intelligence, sixteenth conference on innovative applications of artificial intelligence, july 25-29, 2004, san jose, california, USA* (pp. 900–907). AAAI Press / The MIT Press.
- van Mulken, S., André, E., & Müller, J. (1999). An empirical study on the trustworthiness of life-like interface agents. In H. Bullinger & J. Ziegler (Eds.), *Human-computer interaction: Communication, cooperation, and application design, proceedings of HCI international '99 (the 8th international conference on human-computer interaction), munich, germany, august 22-26, 1999, volume 2* (pp. 152–156). Lawrence Erlbaum.

- van Waveren, S., Carter, E. J., & Leite, I. (2019). Take one for the team: The effects of error severity in collaborative tasks with social robots. In C. Pelachaud, J. Martin, H. Buschmeier, G. M. Lucas, & S. Kopp (Eds.), *Proceedings of the 19th ACM international conference on intelligent virtual agents, IVA 2019, paris, france, july 2-5, 2019* (pp. 151–158). ACM. <https://doi.org/10.1145/3308532.3329475>
- Vasile, C., Marhan, A.-M., Singer, F. M., & Stoicescu, D. (2011). Academic self-efficacy and cognitive load in students. *Procedia-Social and Behavioral Sciences*, 12, 478–482. <https://doi.org/10.1016/j.sbspro.2011.02.059>
- Vermeire, T., Laugel, T., Renard, X., Martens, D., & Detyniecki, M. (2021). How to choose an explainability method? towards a methodical implementation of XAI in practice. *Machine Learning and Principles and Practice of Knowledge Discovery in Databases - International Workshops of ECML PKDD 2021, Virtual Event, September 13-17, 2021, Proceedings, Part I*, 1524, 521–533. https://doi.org/10.1007/978-3-030-93736-2_39
- Vilone, G., & Longo, L. (2021). Notions of explainability and evaluation approaches for explainable artificial intelligence. *Information Fusion*, 76, 89–106. <https://doi.org/10.1016/j.inffus.2021.05.009>
- Vinyals, O., & Le, Q. V. (2015). A neural conversational model. *arXiv preprint arXiv:1506.05869*. <https://arxiv.org/pdf/1506.05869.pdf>
- Virtue, P. (2021). GANs unplugged. *Proceedings of the 35th AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, 35(17), 15664–15668.
- von Ahn, L., & Dabbish, L. (2008). Designing games with a purpose. *Communications of the ACM*, 51(8), 58–67. <https://doi.org/10.1145/1378704.1378719>
- Vosniadou, S., & Brewer, W. F. (1992). Mental models of the earth: A study of conceptual change in childhood. *Cognitive psychology*, 24(4), 535–585. [https://doi.org/10.1016/0010-0285\(92\)90018-W](https://doi.org/10.1016/0010-0285(92)90018-W)
- Vytal, K., & Hamann, S. (2010). Neuroimaging support for discrete neural correlates of basic emotions: A voxel-based meta-analysis. *Journal of Cognitive Neuroscience*, 22(12), 2864–2885. <https://doi.org/10.1162/jocn.2009.21366>
- Wachter, S., Mittelstadt, B., & Russell, C. (2018). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard Journal of Law & Technology*, 31, 841–887.
- Wang, N., Pynadath, D. V., & Hill, S. G. (2016a). The impact of POMDP-generated explanations on trust and performance in human-robot teams. In C. M. Jonker, S. Marsella, J. Thangarajah, & K. Tuyls (Eds.), *Proceedings of the 2016 international conference on autonomous agents & multiagent systems, singapore, may 9-13, 2016* (pp. 997–1005). ACM.
- Wang, N., Pynadath, D. V., & Hill, S. G. (2016b). Trust calibration within a human-robot team: Comparing automatically generated explanations. In C. Bartneck, Y. Nagai, A. Paiva, & S. Sabanovic (Eds.), *Proceedings of the 11th ACM/IEEE international conference on human robot interaction, HRI 2016, christchurch, new zealand, march 7-10, 2016* (pp. 109–116). IEEE/ACM. <https://doi.org/10.1109/HRI.2016.7451741>
- Warden, P. (2018). Speech commands: A dataset for limited-vocabulary speech recognition. *arXiv preprint arXiv:1804.03209v1*. <http://arxiv.org/abs/1804.03209>
- Weiner, B. (1972). Attribution theory, achievement motivation, and the educational process. *Review of Educational Research*, 42(2), 203–215. <https://doi.org/10.2307/1170017>
- Weitz, K. (2018). *Applying explainable artificial intelligence for deep learning networks to decode facial expressions of pain and emotions* [Master's thesis]. Otto-Friedrich-Universitaet Bamberg.
- Weitz, K., Dang, C. T., & André, E. (2022). Do we need explainable ai in companies? investigation of challenges, expectations, and chances from employees' perspective. <https://doi.org/10.48550/arXiv.2210.03527>
- Weitz, K., Gärtig-Daug, A., Knauf, D., & Schmid, U. (2017). Computer science in early childhood education: Pedagogical beliefs and perceived self-confidence in preschool teachers. In E. Barendsen & P. Hubwieser (Eds.), *Proceedings of the 12th workshop on primary and secondary computing education, wipsce 2017, nijmegen, the netherlands, november 8-10, 2017* (pp. 117–118). ACM. <https://doi.org/10.1145/3137065.3144573>
- Weitz, K., Schiller, D., Schlagowski, R., Huber, T., & André, E. (2019). “Do you trust me?”: Increasing user-trust by integrating virtual agents in explainable AI interaction design. In C. Pelachaud, J. Martin, H. Buschmeier, G. M. Lucas, & S. Kopp (Eds.), *Proceedings of the 19th ACM international conference on*

- intelligent virtual agents, IVA 2019, paris, france, july 2-5, 2019 (pp. 7–9). ACM. <https://doi.org/10.1145/3308532.3329441>
- Weitz, K., Schiller, D., Schlagowski, R., Huber, T., & André, E. (2021). “Let me explain!”: Exploring the potential of virtual agents in explainable AI interaction design. *Journal on Multimodal User Interfaces*, 15(2), 87–98. <https://doi.org/10.1007/s12193-020-00332-0>
- Weitz, K., Schlagowski, R., & André, E. (2021). Demystifying artificial intelligence for end-users: Findings from a participatory machine learning show. In S. Edelkamp, R. Möller, & E. Rueckert (Eds.), *Proceedings of KI 2021: Advances in artificial intelligence - 44th german conference on ai, virtual event, september 27 - october 1, 2021* (pp. 257–270). Springer. https://doi.org/10.1007/978-3-030-87626-5_19
- Weitz, K., Vanderlyn, L., Ngoc, T. V., & André, E. (2021). It’s our fault!”: Insights into users’ understanding and interaction with an explanatory collaborative dialog system. In A. Bisazza & O. Abend (Eds.), *Proceedings of the 25th conference on computational natural language learning, conll 2021, online, november 10-11, 2021* (pp. 1–16). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2021.conll-1.1>
- Weitz, K., Zellner, A., & André, E. (2022). What do end-users really want? investigation of human-centered xai for mobile health apps. <https://doi.org/10.48550/arXiv.2210.03506>
- Weld, D. S., & Bansal, G. (2019). The challenge of crafting intelligible intelligence. *Communications of the ACM*, 62(6), 70–79. <https://doi.org/10.1145/3282486>
- Wellman, H. M. (2011). Reinvigorating explanations for the study of early cognitive development. *Child Development Perspectives*, 5(1), 33–38. <https://doi.org/10.1111/j.1750-8606.2010.00154.x>
- Wessel, D., Attig, C., & Franke, T. (2019). ATI-S-an ultra-short scale for assessing affinity for technology interaction in user studies. In F. Alt, A. Bulling, & T. Döring (Eds.), *Proceedings of mensch und computer 2019, hamburg, germany, september 8-11, 2019* (pp. 147–154). GI / ACM. <https://doi.org/10.1145/3340764.3340766>
- Wick, M. R., & Thompson, W. B. (1989). Reconstructive explanation: Explanation as complex problem solving. In N. S. Sridharan (Ed.), *Proceedings of the 11th international joint conference on artificial intelligence, IJCAI 1989, detroit, mi, usa, august 1989* (pp. 135–140). Morgan Kaufmann.
- Wiggins, J. B., Grafsgaard, J. F., Boyer, K. E., Wiebe, E. N., & Lester, J. C. (2017). Do you think you can? the influence of student self-efficacy on the effectiveness of tutorial dialogue for computer science. *International Journal of Artificial Intelligence in Education*, 27(1), 130–153. <https://doi.org/10.1007/s40593-015-0091-7>
- Wimmer, H., & Perner, J. (1983). Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children’s understanding of deception. *Cognition*, 13(1), 103–128. [https://doi.org/10.1016/0010-0277\(83\)90004-5](https://doi.org/10.1016/0010-0277(83)90004-5)
- Wolf, C. T. (2019). Explainability scenarios: Towards scenario-based XAI design. In W. Fu, S. Pan, O. Brdiczka, P. Chau, & G. Calvary (Eds.), *Proceedings of the 24th international conference on intelligent user interfaces, IUI 2019, marina del ray, ca, usa, march 17-20, 2019* (pp. 252–257). ACM. <https://doi.org/10.1145/3301275.3302317>
- Wright, P., & McCarthy, J. (2010). Experience-centered design: Designers, users, and communities in dialogue. *Synthesis lectures on human-centered informatics*, 3(1), 1–123.
- Wu, J., Ghosh, S., Chollet, M., Ly, S., Mozgai, S., & Scherer, S. (2018). NADiA: Neural network driven virtual human conversation agents. In E. André, S. Koenig, M. Dastani, & G. Sukthankar (Eds.), *Proceedings of the 17th international conference on autonomous agents and multiagent systems, AAMAS 2018, stockholm, sweden, july 10-15, 2018* (pp. 2262–2264). ACM. <https://doi.org/10.1145/3267851.3267860>
- Xie, Y., Gao, G., & Chen, X. ’. (2019). Outlining the design space of explainable intelligent systems for medical diagnosis. *arXiv preprint arXiv:1902.06019*. <http://arxiv.org/abs/1902.06019>
- Yang, Q., Steinfeld, A., & Zimmerman, J. (2019). Unremarkable AI: Fitting intelligent decision support into critical, clinical decision-making processes. In S. A. Brewster, G. Fitzpatrick, A. L. Cox, & V. Kostakos (Eds.), *Proceedings of the 2019 CHI conference on human factors in computing systems, CHI 2019, glasgow, scotland, uk, may 04-09, 2019* (pp. 1–11). ACM. <https://doi.org/10.1145/3290605.3300468>
- Yin, M., Wortman Vaughan, J., & Wallach, H. (2019). Understanding the effect of accuracy on trust in machine learning models. In S. A. Brewster, G. Fitzpatrick, A. L. Cox, & V. Kostakos (Eds.), *Proceedings of the*

- 2019 CHI conference on human factors in computing systems, CHI 2019, glasgow, scotland, uk, may 04-09, 2019 (pp. 1–12). ACM. <https://doi.org/10.1145/3290605.3300509>
- Yu, K., Berkovsky, S., Conway, D., Taib, R., Zhou, J., & Chen, F. (2016). Trust and reliance based on system accuracy. In J. Vassileva, J. Blustein, L. Aroyo, & S. K. D'Mello (Eds.), *Proceedings of the 2016 conference on user modeling adaptation and personalization, UMAP 2016, halifax, ns, canada, july 13 - 17, 2016* (pp. 223–227). ACM. <https://doi.org/10.1145/2930238.2930290>
- Yu, K., Berkovsky, S., Taib, R., Conway, D., Zhou, J., & Chen, F. (2017). User trust dynamics: An investigation driven by differences in system performance. In G. A. Papadopoulos, T. Kuflik, F. Chen, C. Duarte, & W. Fu (Eds.), *Proceedings of the 22nd international conference on intelligent user interfaces, IUI 2017, limassol, cyprus, march 13-16, 2017* (pp. 307–317). ACM. <https://doi.org/10.1145/3025171.3025219>
- Zellner, A. (2021). *Towards personalized explanations in digital health - user-centered explanations for mobile health applications* [Master's thesis]. University of Augsburg.
- Zhou, Y., & Chellappa, R. (1988). Computation of optical flow using a neural network. *Proceedings of International Conference on Neural Networks (ICNN'88), San Diego, CA, USA, July 24-27, 1988, 1998*, 71–78. <https://doi.org/10.1109/ICNN.1988.23914>
- Zhu, J.-Y., Park, T., Isola, P., & Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. *Proceedings of the IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, 2223–2232. <https://doi.org/10.1109/ICCV.2017.244>
- Zhu, L., & Williams, T. (2020). Effects of proactive explanations by robots on human-robot trust. In A. R. Wagner, D. Feil-Seifer, K. S. Haring, S. Rossi, T. E. Williams, H. He, & S. S. Ge (Eds.), *Social robotics - 12th international conference, ICSR 2020, golden, co, usa, november 14-18, 2020, proceedings* (pp. 85–95, Vol. 12483). Springer. https://doi.org/10.1007/978-3-030-62056-1_8