

From Competition to Complementarity: Foundations and Evidence for Effective Human-AI Collaboration

Zur Erlangung des akademischen Grades eines

**Doktors der Ingenieurwissenschaften
(Dr.-Ing.)**

von der KIT-Fakultät für Wirtschaftswissenschaften
des Karlsruher Instituts für Technologie (KIT)

genehmigte

Dissertation

von

Max Richard Schemmer

Tag der mündlichen Prüfung: 7. November 2023

Referent: Prof. Dr. Gerhard Satzger

Korreferent: Prof. Dr. Andreas Oberweis

Abstract

The rapid development of artificial intelligence (AI) and its growing ability to autonomously perform a wide range of tasks raises questions about the future role of humans in various domains. In this thesis, we argue that despite the potential ability of AI to automate tasks, there is complementarity between humans and AI that can be harnessed to achieve complementary team performance (CTP), i.e., a level of performance that neither can achieve individually. Therefore, we propose that in the presence of complementarity, humans and AI should work together, a situation commonly referred to as human-AI collaboration. However, existing empirical work shows that the realization of CTP seems to be inconsistent and elusive. Some studies confirm its achievement, while others contradict this finding, essentially weakening human-AI collaboration as an alternative to automation. Therefore, this thesis investigates the challenges of achieving CTP and aims to provide guidance on how to obtain it.

Since the current overview of CTP is rather anecdotal, we begin the thesis by conducting a structured literature review and a meta-analysis to provide a comprehensive and statistical perspective on the current state of empirical studies in the field of human-AI collaboration. Our results show that human-AI collaboration can outperform human individuals but often does not exceed the AI's individual performance. These findings underscore the need for a deeper understanding of the factors that influence CTP.

To address the identified shortcomings, we derive a theoretical foundation of the impact factors of CTP, which we refer to as the human-AI complementarity concept. Our concept, anchored in existing justificatory knowledge from the information systems field, provides a formalization, highlights possible sources of complementarity potential, and introduces a classification of mechanisms to fuse human and AI decisions during collaboration. Based on this concept, the remainder of the thesis focuses on how existing complementarity potential can be harnessed.

We first explore the process of harnessing complementarity potential in AI-assisted decision-making, the most common form of human-AI collaboration. In this setting, a human receives advice from an AI agent and then decides whether to accept or adjust it. To achieve CTP in this form of collaboration, it is critical that humans rely

appropriately on the advice provided by AI. This area of research is still in its infancy and lacks a solid foundation and established measures. To address this research gap, we develop a new measure and a research model. In addition, we conduct two empirical studies to validate and test both the measure and the model. We then shift our focus beyond AI-assisted decision-making to show how complementarity can be harnessed in other forms of human-AI collaboration. Finally, we address the challenge of preserving complementarity over time by conceptualizing a new form of human-AI collaboration—intelligent decision assistance.

This thesis enriches the research on human-AI collaboration by conducting a comprehensive analysis of the current state of empirical studies, introducing new theoretical concepts, developing a measure and research models, and addressing critical challenges. From a practical standpoint, we enable organizations to maximize AI's utility and yield higher return on investment by offering clear guidance to consistently achieve CTP. Finally, we contribute to the ubiquitous discourse on the future of work, which predominantly focuses on either full automation or pure human work, by providing foundations and evidence for effective human-AI collaboration.

Acknowledgement

This thesis represents not only the culmination of my academic journey, but also the realisation of a lifelong ambition to be an inventor. I am deeply grateful to a number of people whose support and guidance have been instrumental in shaping this thesis.

First, I would like to express my sincere gratitude to my supervisor, Gerhard Satzger. Your balanced approach of allowing me freedom while providing essential guidance has been invaluable, not only in this academic pursuit, but also in shaping my personal and professional narrative. Your expertise and insights have contributed significantly to my growth in areas beyond this research. I am equally grateful to Niklas Kühl for his unwavering support throughout my journey as a PhD student and beyond.

I am also indebted to my colleagues at KSRI. Working alongside such brilliant minds has broadened my understanding of the nuances of human-AI collaboration. Their perspectives and critiques have been instrumental in refining my thoughts and approaches throughout this dissertation.

A special thank you goes to my family, friends, and most importantly, my wife Anna, and my daughter Ronja. Anna, the unwavering support, understanding and love you have given me is the foundation upon which I have built my aspirations and achievements. Ronja, your presence in my life has been a constant source of motivation. This thesis is not only a reflection of my academic pursuits, but also a testament to the values and dreams I hope to pass on to you.

Finally, I reflect on this journey with a sense of pride and accomplishment. The journey of creating something innovative, especially in the field of AI and its collaboration with humanity, has been both challenging and exhilarating. This thesis is a small step in the vast landscape of AI development, and I am honoured to contribute to this evolving field.

Contents

List of Figures	xiii
List of Tables	xvii
List of Abbreviations	xix
I Introduction and Foundations	1
1 Introduction	3
1.1 Motivation	3
1.2 Research Objective and Research Questions	5
1.3 Research Design and Structure	8
1.4 Integration of Articles	14
2 Foundations	25
2.1 Artificial Intelligence and Machine Learning	25
2.1.1 Terminology	27
2.1.2 The Role of Rational Agents in Information Systems	30
2.1.3 A Typology for Machine Learning in AI Systems	34
2.2 Human-AI Collaboration	36
2.3 Explainable Artificial Intelligence	38
II Analysis of the Current State of Empirical Work on Human-AI Collaboration	41
3 Structured Literature Review of Empirical Studies on Human-AI Collaboration	43
3.1 Introduction	43
3.2 Conceptual Foundations	45
3.2.1 Hybrid Intelligence	45
3.2.2 Explainable Artificial Intelligence	46

3.3	Methodology	47
3.4	Results	48
3.4.1	Overview	49
3.4.2	Collaboration Characteristics	50
3.4.3	Task Characteristics	51
3.4.4	Artificial Intelligence Characteristics	52
3.4.5	Human Characteristics	54
3.5	Research Implications for Human-AI Collaboration	54
3.5.1	Collaboration Characteristics	54
3.5.2	Task Characteristics	55
3.5.3	Artificial Intelligence Characteristics	56
3.5.4	Human Characteristics	57
3.6	Conclusion	59
4	A Meta-Analysis of the Impact of Explainable AI on Decision Performance	61
4.1	Introduction	61
4.2	Related Work	63
4.3	Methodology	64
4.3.1	Data Collection	64
4.3.2	Statistical Analysis	66
4.4	Results	67
4.4.1	Data Collection	67
4.4.2	AI Assistance Versus XAI Assistance	69
4.4.3	Human Versus XAI Assistance	69
4.4.4	Tabular Versus Text Data	71
4.4.5	Summary of the Articles	73
4.5	Limitations	79
4.6	Discussion and Future Work	79
4.7	Conclusion	81
III	Conceptualization of Human-AI Complementarity	83
5	Conceptualization of Human-AI Complementarity and the Influence of Information Asymmetry	85
5.1	Introduction	85
5.2	Theoretical Foundations and Related Work	88
5.2.1	Human-AI Collaboration	88
5.2.2	Human-AI Complementarity	89

5.2.3	Information Asymmetry	91
5.3	Conceptualization of Human-AI Complementarity	92
5.3.1	Purpose and Scope of Human-AI Complementarity	92
5.3.2	Formalization of Complementarity Potential	93
5.3.3	Sources of Theoretical Complementarity Potential	101
5.3.4	Integration Mechanisms: Realizing the Theoretical Complementarity Potential	103
5.4	Experimental Design	105
5.4.1	Task and AI Model	107
5.4.2	Study Design and Procedure	108
5.4.3	Evaluation Measures	110
5.5	Results	110
5.6	Discussion	114
5.6.1	Contributions	114
5.6.2	Theoretical Implications	116
5.6.3	Managerial and Political Implications	117
5.6.4	Limitations	118
5.6.5	Future Research	118
5.7	Conclusion	119

IV Harnessing Complementarity Potential in AI-Assisted Decision-Making 121

6	Harnessing Complementarity: The Role of Appropriate Reliance	123
6.1	Introduction	123
6.2	Related Work	126
6.3	Conceptualization of Appropriate Reliance	129
6.3.1	Reliance and Appropriateness	129
6.3.2	Towards a Measurement Concept—Appropriateness of Reliance	130
6.3.3	Definition of Appropriate Reliance	133
6.4	Theory Development And Hypotheses	134
6.5	Experimental Design	138
6.5.1	Task, Model, and Explanations	138
6.5.2	Study Design and Procedure	138
6.5.3	Evaluation Measures	140
6.6	Results	141
6.6.1	Descriptive Analysis	142
6.6.2	Appropriateness of Reliance & Appropriate Reliance	142
6.6.3	Structural Equation Model	143

6.7	Discussion	145
6.8	Conclusion	148
7	Harnessing Complementarity: The Influence of Human Learning on Appropriate Reliance	149
7.1	Introduction	149
7.2	Theoretical Foundations and Related Work	152
7.2.1	Human-AI Collaboration	152
7.2.2	Explainable Artificial Intelligence	153
7.2.3	Appropriate Reliance in Artificial Intelligence	153
7.2.4	Learning from Artificial Intelligence	154
7.3	Theoretical Development	155
7.4	Methodology	158
7.4.1	Task, Model and Explanations	159
7.4.2	Experimental Design	160
7.4.3	Measurements	161
7.4.4	Participants	162
7.5	Results	163
7.5.1	Descriptive Results and Appropriateness of Reliance	163
7.5.2	Structural Equation Modeling	165
7.5.3	Explorative Sub-Group Analysis	166
7.6	Discussion	167
7.6.1	RQ1: The Effect of Learning on Appropriateness of Reliance	168
7.6.2	RQ2: The Effect of Explanations on Learning	168
7.6.3	Implications for Theory and Practice	169
7.6.4	Limitations and Future Work	169
7.7	Conclusion	170
V	Harnessing Complementarity Potential beyond AI-Assisted Decision- Making	171
8	Harnessing Complementarity in Anomaly Detection	173
8.1	Introduction	173
8.2	Related Work	176
8.2.1	Anomaly Definition	176
8.2.2	Anomaly Detection	177
8.2.3	Anomaly Investigation	178
8.2.4	Explainable Anomaly Detection	179

8.2.5	Explainable Autoencoder-Based Anomaly Detection in Multi-variate Time Series	179
8.3	Conceptualization of Explainable Anomaly Detection for Anomaly Investigation	180
8.4	Experiment Design	183
8.4.1	Dataset, Task and Data Preprocessing	183
8.4.2	Explainable Anomaly Detection System	186
8.4.3	Experimental Design	187
8.5	Results	193
8.5.1	Qualitative Interpretation of Detected Anomalies	193
8.5.2	Experiment Results	194
8.6	Discussion	195
8.6.1	Implications	196
8.6.2	Limitations and Future Work	196
8.7	Conclusion	197
9	Harnessing Complementarity in Intelligent Decision Assistance	199
9.1	Introduction	199
9.2	Literature Review	201
9.2.1	Decision Support Systems	201
9.2.2	Automation	203
9.3	Conceptualization of Intelligent Decision Assistance	205
9.4	Validation Study	209
9.5	Discussion	212
9.6	Conclusion	213
VI	Finale	215
10	Conclusion	217
10.1	Summary and Theoretical Contributions	217
10.2	Managerial Implications	226
10.3	Limitations and Future Research	229
A	Appendix	231
A.1	Appendix Human-AI Complementarity Conceptualization	231
A.1.1	In-Person Pilot Study	231
A.1.2	Task and AI Tutorial	232
A.1.3	Participant Statistics	234
	Bibliography	235

List of Figures

2.1	Appearance of the terms “artificial intelligence” and “machine learning” in AIS Senior Scholars’ Basket journals.	28
2.2	Conceptual framework describing the general architecture for intelligent agents in AI-based information systems.	32
2.3	Typology of AI-based information systems.	34
3.1	Key elements of human-AI collaboration.	48
3.2	Structured literature review results.	50
4.1	Flowchart describing the data collection and article selection procedure.	67
4.2	Forest plot of the standardized mean difference between AI-assisted and XAI-assisted performance.	70
4.3	Forest plot of the standardized mean difference between human and XAI-assisted performance.	72
4.4	Forest plot of the standardized mean difference between human and XAI-assisted performance considering the subgroups.	74
5.1	Conceptual overview of the notion of complementarity potential.	94
5.2	Case 1: Deviation of human prediction, AI prediction and integrated prediction from the ground truth for a single instance.	98
5.3	Case 2: Deviation of human prediction, AI prediction and integrated prediction from the ground truth for a single instance.	98
5.4	Case 3: Deviation of human prediction, AI prediction and integrated prediction from the ground truth for a single instance.	100
5.5	Case 4: Deviation of human prediction, AI prediction and integrated prediction from the ground truth for a single instance.	101
5.6	Conceptual overview of the human-AI collaboration process.	103
5.7	Sequence of the individual steps in the experiment	108
5.8	Experiment user interface.	109
5.9	Experiment performance results.	111
6.1	Combinatorics of initial human decisions, AI advice and human reliance for a single task instance in a sequential task setting.	131

6.2	Two-dimensional depiction of appropriateness of reliance (AoR). . .	133
6.3	Illustrative example of how humans can increase their relative AI reliance (<i>RAIR</i>) based on explanations of the AI advice.	136
6.4	Research model on the effect of explanations on appropriateness of reliance (AoR).	137
6.5	Online experiment graphical user interface.	139
6.6	Illustration of appropriateness of reliance (AoR)	143
6.7	Structural equation modeling results.	144
7.1	Distinction between related work and our contribution.	151
7.2	Combinatorics of initial human decisions, AI advice and human re- liance for a single task instance in a sequential task setting (Schemmer et al., 2023d).	154
7.3	Research model on the effect of human learning on appropriateness of reliance (AoR).	159
7.4	Online experiment graphical user interface for the example-based explanation condition.	161
7.5	Appropriateness of reliance analysis.	165
7.6	Structural equation modeling results.	166
8.1	Different anomaly types (Choi et al., 2021). The three examples depict uni-variate time series, with the three types of anomalies in time series.	177
8.2	Representation of the proposed method. Explanations generated from the anomaly detection support the human anomaly investigation. . .	182
8.3	Excerpt of test data with highlighted event days.	185
8.4	Exemplary explanations for extreme weather event and public holiday.	188
8.5	Example of an extreme weather event.	190
8.6	Distribution of the effectiveness of participants.	194
8.7	Distribution of the efficiency of participants.	195
9.1	Positioning of Intelligent Decision Assistance on the two dimensions of explainability and degree of automation	207
9.2	Comparison of traditional automated decision-making and Intelligent Decision Assistance (IDA)	209
10.1	Research contribution addressing RQ1: Empirical results of structured literature review.	218
10.2	Research contribution addressing RQ2: Conceptualization of Human- AI complementarity.	220

10.3	Research contribution addressing RQ3: Measurement of appropriateness of reliance.	222
10.4	Research contribution addressing RQ3: Research Model including task-specific human learning mediator.	223
10.5	Research contribution addressing RQ4: Our method of using explainable anomaly detection to support anomaly investigation.	225
A.1	Online experiment graphical user interface for the task tutorial. . . .	232
A.2	Online experiment graphical user interface for the AI tutorial.	233

List of Tables

1.1	Overview of conducted research studies and research methodologies.	9
1.2	Overview of the integrated articles in the structure of the thesis. . . .	15
1.3	Summary of the paper included in Chapter 2	16
1.4	Summary of the paper included in Chapter 3	17
1.5	Summary of the paper included in Chapter 4	18
1.6	Summary of the paper included in Chapter 5	19
1.7	Summary of the paper included in Chapter 6	20
1.8	Summary of the paper included in Chapter 7	21
1.9	Summary of the paper included in Chapter 8	22
1.10	Summary of the paper included in Chapter 9	23
2.1	AI research streams	28
3.1	Experimental conditions of XAI and their comparisons in existing studies.	53
4.1	Overview of articles.	68
5.1	Conceptual overview of the human-AI collaboration process.	105
5.2	Conceptualization of human-AI complementarity	106
5.3	Summary of human-AI complementarity results.	115
6.1	Related work on explainable AI (XAI) and appropriate reliance (AR).	128
6.2	Summary of participants' characteristics.	141
6.3	Descriptive results.	142
6.4	Structural equation model fitting index.	144
6.5	Analysis results of the structural equation model.	145
7.1	Descriptive results.	164
7.2	Correlation analysis.	164
7.3	Structural equation model fitting index.	166
7.4	Subgroup analysis.	167
8.1	Summary of participants' characteristics.	192
8.2	Descriptive outcomes.	194

9.1	Validation study results	211
A.1	Summary of participants' characteristics.	234

List of Abbreviations

ADS	anomaly detection system
AI	artificial intelligence
AoR	appropriateness of reliance
AR	appropriate reliance
CAIR	correct AI reliance
CI	confidence interval
CS	computer science
CSR	correct self-reliance
CTP	complementary team performance
DSS	decision support system
HCI	human-computer interaction
HI	hybrid intelligence
IDA	intelligent decision assistance
IS	information system
JAS	judge-advisor system
ML	machine learning
RAIR	relative AI reliance
ROI	return on investment
RQ	research question
RSR	relative self-reliance
SEM	structural equation modelling
SLR	structured literature review
SMD	standardized mean difference
SVM	support vector machine
XAI	explainable AI

Part I

Introduction and Foundations

“*Today’s workforce should prepare to work hand in hand with AI.*”

— **Arvind Krishna**
(IBM CEO)

1.1 Motivation

In recent years, artificial intelligence (AI) has made significant strides in various domains such as medicine (Wu et al., 2020), finance (Day et al., 2018) or manufacturing (Stauder & Köhl, 2022), even including those involving high-stakes decision-making. For example, AI applications now assist doctors in making diagnoses (Leibig et al., 2022), aid recruiters in the hiring process (Peng et al., 2022), and even provide decision support within the judicial sector (Kleinberg et al., 2018). This widespread adoption of AI stems from ongoing advancements in machine learning (ML), leading to improved capabilities and enhanced performance (Janiesch et al., 2021; Ren et al., 2015). Especially novel developments in the architectures of deep neural networks (Dong et al., 2021) have led to cases where ML models outperformed many existing benchmarks based on traditional methods (Sarker, 2021).

Across a growing range of tasks, the increasing capabilities of AI have begun to surpass human performance (Takeda et al., 2023). This includes mastering complex games such as poker (Brown & Sandholm, 2019) and Go (Silver et al., 2018), accurately identifying categories in image recognition tasks (He et al., 2015), and detecting medical conditions such as breast cancer (Pisano, 2020). Especially with the recent rise of foundation models—pre-trained large ML models (Bommasani et al., 2021)—the number of tasks that both humans and AI can perform effectively and independently is growing. Examples of such tasks include material discovery (Takeda et al., 2023), social media content creation (Bommasani et al., 2021), and programming code generation (Mozannar et al., 2022). The increasing ability of AI to perform tasks autonomously distinguishes today’s AI from the past, when AI

provided selective input for broader downstream decisions that only humans could make. For example, in credit allowance decisions, AI only provided an aggregated probability of default that the lender used in the downstream task of making the final credit decision. Today, however, AI is increasingly capable of making such decisions independently, meaning that more and more tasks could be automated. As automation increases, AI is essentially competing with humans and raising questions about the future role of humans. To summarize, we¹ can say that in more and more application cases, AI is matching human performance, which essentially changes the role of AI from a decision-making input provider to an autonomous *AI agent* (Qian & Qian, 2020).

In this thesis, we argue that modern AI agents, despite being able to automate, also have the potential to enhance humans through complementary capabilities (Bansal et al., 2021; Dellermann et al., 2019a). For example, in the medical domain, both, human and AI agents are able to conduct the diagnosis of diseases on their own (Pisano, 2020; Reverberi et al., 2022). However, it has been demonstrated that they typically show different performances on individual task instances (Geirhos et al., 2021; Steyvers et al., 2022). For example, human and AI agents in cancer detection show different detection qualities on individual CT images (Jussupow et al., 2021). In this context, AI agents may detect patterns in large amounts of data that humans will find challenging to discover, while humans, in turn, excel at the causal interpretation and intuition required to interpret these patterns (Lake et al., 2017; Li et al., 2019b). By leveraging this complementarity potential between these two agents, a superior performance beyond the human or AI agent alone can be reached.² This desired superior performance is referred to as *complementary team performance (CTP)* (Bansal et al., 2021).

We argue that in the presence of complementarity potential, human involvement, as opposed to automation, should be considered to enable CTP. In this thesis, we refer to the interplay between human and AI agents as *human-AI collaboration*. This concept is defined as a setting where two or more agents, including at least one human and one AI agent, collaborate to achieve mutual goals (Terveen, 1995; Vössing et al., 2022). For the purpose of this thesis, as a starting point, we restrict

¹Linguistically, this thesis uses the first-person plural using “we”. Two rationales underpin this practice. First, it promotes the readability of this work, as “we” better engages the reader and invites to a shared journey. Second, research always takes place in a community that collectively seeks to explore a research area.

²Some people argue that when AI agents reach superintelligence, that is, when they become better than humans in every way, this complementarity will disappear. However, as Keynes said: “In the long run, we are all dead.” (Keynes, 1923, p. 80). Keynes was arguing for some focus on the short and medium term. In the short and medium term, no AI is likely to reach the cognitive level of superintelligence, which ensures the relevance of human-AI complementarity in the now.

our study of human-AI collaboration to scenarios involving a single human and one AI agent. The goal of achieving CTP through human-AI collaboration has recently gained attention in various disciplines, such as information system (IS) (Fuegener et al., 2022), human-computer interaction (HCI) (Bansal et al., 2021; Zhang et al., 2022), and computer science (CS) (Bansal et al., 2019b; Siu et al., 2021).

Various empirical studies have demonstrated that human-AI collaboration can outperform human individuals, but they often do not exceed the AI agents' individual performance (Bansal et al., 2021). The observation that CTP is often not attained in empirical studies raises questions about which factors could contribute to achieving CTP. It illustrates that the current knowledge of how the respective capabilities of humans and AI agents can be utilized to create joint synergies has yet to be sufficiently developed. This means there is a need for additional concepts that foster an in-depth understanding of complementarity and how to harness it in human-AI collaboration. In this thesis, we aim to understand what is necessary to create *effective* human-AI collaboration, i.e., to achieve CTP consistently and thereby contribute to human-AI collaboration research in IS, HCI, and CS.

1.2 Research Objective and Research Questions

Building on the outlined motivation, the objective of this thesis is to explore and establish foundations that can guide *effective* human-AI collaboration. The evolution of AI toward autonomous agents requires deeper insights into the unique and complementary capabilities of human and AI agents. This work aims to contribute to a better understanding of the mechanisms that promote effective human-AI collaboration (i.e., achieving CTP), which leads to the following research objective of this thesis:

Research Objective

The research objective of this thesis is to derive foundations and evidence for understanding and designing effective human-AI collaboration.

Our primary research objective is divided into four research questions (RQs). In the following, the individual research questions are introduced.

With the ongoing development of AI agents and the increasing interest in human-AI collaboration, researchers have begun to assess whether the improvement in decision performance from human-AI collaboration can be quantified (Bansal et al.,

2021; Buçinca et al., 2020). While some researchers find a benefit from human-AI collaboration in user studies (Buçinca et al., 2020), others find negligible evidence (Carton et al., 2020). These differences could be due to different experimental setups and tasks tailored to different research objectives. It lacks a unified view of related work and statistical analysis. In particular, there is no overview of the achievement of CTP across experimental studies. Therefore, we aim to clarify the current state of the art of AI-assisted decision-making performance in empirical user studies and formulate:

Research Question 1 (RQ1)

How effectively do human agents and AI agents collaborate?

Answering our first research question highlights a severe ambiguity in current research on human-AI collaboration (Hemmer et al., 2021; Schemmer et al., 2022b). More specifically, we show that CTP is often not attained in empirical studies. Most studies investigate whether CTP can be achieved through appropriate human reliance on the AI agents' advice. In these studies, researchers supplement the AI agents' recommendation with additional information aiming to enable human agents to assess better its reliability, e.g., explanations or confidence measures (Bansal et al., 2021; Liu et al., 2021; Schreiber et al., 2020; Zhang et al., 2020) which is commonly referred to as explainable AI (XAI) (Wanner et al., 2020). However, our research shows that this additional information does not reliably have positive effects on decision-making performance (Hemmer et al., 2021; Schemmer et al., 2022b). The fact that CTP is often not reached and the most common design feature of effective human-AI collaboration (XAI) does not reliably improve performance illustrates that the current understanding of human-AI collaboration is not yet sufficiently developed and that there is a need for a deeper understanding. Therefore, we formulate:

Research Question 2 (RQ2)

What are the key factors that influence effective human-AI collaboration?

We derive and formalize two key factors of effective human-AI collaboration, namely the existence of complementarity potential and the effective realization of this potential (Schemmer et al., 2023b). To illustrate complementarity potential, in classifications, it refers to the proportion of tasks that can be effectively accomplished by either entity (such as a human or an AI) but not by both together. It essentially signifies the unique problem-solving abilities possessed by each agent.

It is important to note that even if complementarity potential is available, it is useless if it cannot be harnessed. Current research does not show robust ways to improve decision performance through human-AI collaboration (Bansal et al., 2021; Hemmer et al., 2021; Schemmer et al., 2022b). Therefore, while ensuring sufficient complementarity potential is essential, in this thesis, we focus on realizing existing complementarity potential. More specifically, we focus on human agents receiving input from AI agents—so-called human ex-post integration. Here, the most common instantiation of this integration is AI-assisted decision-making (Lai et al., 2023). AI-assisted decision-making refers to a setting in which a human receives AI advice and then is asked to either follow or adjust the advice. Since this is the most common integration, we want to explore it further and formulate the following third research question:

Research Question 3 (RQ3)

How can complementarity potential in AI-assisted decision-making be harnessed?

To harness complementarity in AI-assisted decision-making, human decision-makers need to appropriately rely on the advice recommended by the AI agent (Schemmer et al., 2023d). However, it is unclear how to design for appropriate reliance. Therefore, we develop and validate a measure of appropriate reliance and a research model to investigate appropriate reliance in AI-assisted decision-making. However, the spectrum of human-AI collaboration is much broader than AI-assisted decision-making (Vössing et al., 2022), which encompasses classification and regression tasks. AI agents can address a magnitude of additional tasks (Carbonell et al., 1983), such as clustering, anomaly detection, etc. Especially the generation of text or images, commonly referred to as generative AI (Jo, 2023), is growing. Other forms of collaboration and different tasks than classification and regression may require different foundations. Therefore, we extend our perspective and research beyond AI-assisted decision-making. Thus, as the final research question of this thesis, we want to explore how to potentially harness complementarity potential beyond AI-assisted decision-making:

Research Question 4 (RQ4)

How can complementarity potential beyond AI-assisted decision-making be harnessed?

In the following, we introduce the research design of this thesis to answer the research questions.

1.3 Research Design and Structure

This thesis consists of six parts. In the current Part I, we introduce the research questions and outline the research design. Furthermore, the foundations of the thesis are presented. In Part II, we analyze the current state of empirical studies on human-AI collaboration to address RQ1. Based on the insights gained in Part III, we conceptualize the foundation of effective human-AI collaboration to address RQ2. Next, in Part IV, we derive insights on how to harness human-AI complementarity in AI-assisted decision-making, addressing RQ3. In Part V, we broaden our perspective to address RQ4, which focuses on leveraging complementarity beyond AI-assisted decision-making. Finally, Part VI, we summarize the findings, discuss implications, and outline limitations and potential future research. Table 1.1 on page 9 visualizes the structure of this thesis.

The type of research questions and the nature of the study determine the choice of research methodology. Following the structure of the thesis and based upon the research questions introduced in Section 1.2, in the following, we illustrate the research methodology.

In **Part I**, in **Chapter 2**, we introduce the foundations of this thesis. We analyze the role of ML in AI agents. We do so by taking an ML perspective on AI agents' capabilities and their relevant implementation. To this end, we review the relevant literature for both terms and synthesize and conceptualize the results.

In **Part II**, the first research question (**RQ1**) of this thesis is addressed in two chapters. To answer the question of how effective human agents and AI agents collaborate, we collect studies based on a structured literature review (SLR) that empirically analyze human-AI collaboration and study them with two approaches—first qualitatively and then quantitatively using a meta-analysis. The nature of SLRs and meta-analyses require a certain degree of homogeneity within the sample (Borenstein et al., 2021; vom Brocke et al., 2009). Therefore, we first determine the scope of the articles to analyze. A common trait found amongst a considerable sample of these empirical studies is the emphasis on equipping humans with insights into the decision-making processes of AI, a concept often denoted as XAI. Consequently, our search strategy is tailored to incorporate studies that revolve around the theme of XAI.

In **Chapter 3**, we conduct a structured literature review following vom Brocke et al. (2009) to collect the body of empirical studies of human-AI collaboration. We analyze the collected data (29 articles) qualitatively using a socio-technical perspective (Maedche et al., 2019) and cluster the identified factors according to

Part I	Introduction and Foundations
	Chapter 1 <i>Introduction</i>
Part II - RQ1 -	Chapter 2 <i>Foundations</i> Method: Literature review and conceptualization
	Analysis of the Current State of Empirical Work on Human-AI Collaboration
Part III - RQ2 -	Chapter 3 <i>Structured Literature Review of Empirical Studies on Human-AI Collaboration.</i> Method: Structured literature review
	Chapter 4 <i>A Meta-Analysis of the Impact of Explainable Artificial Intelligence on Decision Performance</i> Method: Meta-analysis
Part IV - RQ3 -	Conceptualization of Human-AI Complementarity
	Chapter 5 <i>Conceptualization of Human-AI Complementarity and the Influence of Information Asymmetry</i> Method: Formalization, conceptualization, and behavioral experiment
Part V - RQ4 -	Harnessing Complementarity in AI-assisted Decision-Making
	Chapter 6 <i>Harnessing Complementarity: The Role of Appropriate Reliance</i> Method: Conceptualization, theory development, behavioral experiment, and structural equation modeling
Part VI	Chapter 7 <i>Harnessing Complementarity: The Influence of Human Learning on Appropriate Reliance</i> Method: Theory development, behavioral experiment, and structural equation modeling
	Harnessing Complementarity Beyond AI-assisted Decision-Making
Part VII	Chapter 8 <i>Harnessing Complementarity in Anomaly Detection</i> Method: Conceptualization and behavioral experiment
	Chapter 9 <i>Harnessing Complementarity in Intelligent Decision Assistance</i> Method: Conceptualization and structured literature review
Part VIII	Finale
	Chapter 10 <i>Conclusion</i>

Table 1.1.: Overview of conducted research studies and research methodologies.

human agents, AI agents, and tasks. Our results highlight that only a small number of experiments show CTP. Most of the time, the human decision performance with AI assistance is inferior to the AI performance if the AI agent had performed the task alone. This leaves the question unanswered why CTP—also exceeding this AI performance—could not have been accomplished. Furthermore, our results show that XAI has ambiguous effects on team performance in general and CTP in specific.

Some studies show positive results, while others show negative results. Finally, we derive 12 hypotheses about factors that may have a potential impact on CTP.

In **Chapter 4**, we complement the structured literature review with a statistical meta-analysis of the empirical studies. First, we renew the SLR, resulting in a total of 33 articles. Next, we sample 9 articles from the 33 that meet the requirements for meta-analysis. The articles comprised multiple experimental studies with multiple treatments, which led to a collection of 44 treatments. Subsequently, all necessary performance metrics are extracted from the articles. Based on this data foundation, a meta-analysis (Higgins et al., 2019) is conducted, which allows a statistical comparison between human, AI-, and XAI-assisted task performance. The statistical analysis confirms the previous structured literature review. The majority of studies in the field merely demonstrate that humans teaming with AI *may* achieve higher team performance than conducting the decision task alone. However, we find no effect of explanations on users' performance compared to AI assistance. Some studies report positive XAI assistance performance effects, whereas others find no or slightly adverse effects. This statistical analysis further highlights the ambiguity in harnessing complementarity potential.

In **Part III**, our objective is to gain a deeper understanding of the ambiguous empirical findings and consequently explore **RQ2**, i.e., identifying the key factors that influence effective collaboration between humans and AI agents.

Therefore, in **Chapter 5**, we derive a concept for understanding and developing effective human-AI collaboration, i.e., achieving CTP. Essentially, CTP depends on sufficient complementarity potential and the ability to harness it. Based on existing justificatory knowledge of Fügener et al. (2021), we derive a formalization of complementarity potential. In detail, we argue that complementarity potential has an inherent and a collaborative component. Whereas the first captures the idea that humans and AI possess different inherently present capabilities in the form of unique human and AI knowledge, the second component captures a new type of knowledge that only emerges through human-AI interaction. In the formalization, for both components that together result in the total complementarity potential, we distinguish between the realized amount that has materialized and a theoretical amount that serves as an upper boundary. Next, we outline possible sources of complementarity potential and introduce a classification of mechanisms for integrating human and AI agent decisions during collaboration. We call the triad of formalization, sources of complementarity potential, and classification of integration mechanisms the human-AI complementarity concept. To illustrate our concept, we apply it in an empirical study: we focus on information asymmetry as a promising

source of complementarity potential and demonstrate, for a real estate appraisal use case, that humans can indeed leverage their unique information to achieve CTP. To sum it up, CTP essentially depends on complementarity potential and the ability to harness it. In this thesis, we focus in the next chapters on how to harness existing complementarity potential.

As a starting point, in **Part IV**, to analyze how to harness complementarity potential, we focus on the most common form of human-AI collaboration—AI-assisted decision-making (Lai et al., 2023), thereby addressing **RQ3**.

AI-assisted decision-making describes a setting in which a human receives AI advice and is asked to act upon it. Human agents should not always rely on AI advice but should differentiate when to rely on AI advice and when to rely on their own, i.e., they should display appropriate reliance (AR) (Bansal et al., 2021; Wang & Yin, 2021; Yang et al., 2020a; Zhang et al., 2020). Despite being a necessary condition for effective human-AI collaboration, current research on AR on AI advice is still ambiguous with regard to definition, measurement, and impact factors (Bansal et al., 2021). Therefore, we derive a measurement concept and a research model in **Chapter 6**.

The term “appropriate reliance” is currently used inconsistently in research, referring to both a binary target state (where AR is either achieved or not achieved) and a metric indicating varying degrees of appropriateness. To address this ambiguity, we propose a two-dimensional metric, termed appropriateness of reliance (AoR), to define and quantify reliance behavior. This metric takes into account the relative frequency of accurate overrides of incorrect AI suggestions (referred to as relative self-reliance—RSR) and adherence to correct AI suggestions (referred to as relative AI reliance—RAIR). AoR embodies a metric understanding of AR. Using this metric, we can define different levels of AR that represent the achievement of specific goals, such as meeting legal, ethical, and performance standards.

In addition, we aim to analyze how the provision of explanations of AI influences AoR. Existing literature is ambiguous with regard to the effects of explanations (Alufaisan et al., 2021; Bansal et al., 2021; Wang & Yin, 2021): while in some experiments, explanations support AR (Wang & Yin, 2021; Yang et al., 2020a), in others they cause “blind trust” (Alufaisan et al., 2021; Bansal et al., 2021) in AI advice. To better understand and reconcile conflicting results, we consider additional constructs that may mediate the effect of explanations. More specifically, we hypothesize that explanations do not only influence the information available to the decision-maker but also have an impact on trust toward AI and self-confidence. Based on those hypotheses, we derive an initial research model on AoR. Our results

show that in certain situations, explanations can improve the relative AI reliance and that this effect is partially mediated by a change in self-confidence and trust.

Next, in **Chapter 7**, we further explore **RQ3** and extend our previously derived research model. One of the core influence factors of reliance behavior in human-AI collaboration seems to be the expertise of decision-maker (Nourani et al., 2020a; Wang & Yin, 2021). We follow this line of thought and hypothesize that learning in AI-assisted decision-making (decision-makers gradually gaining expertise) could be a relevant mediator of AoR. Therefore, we extend our research model on AoR, including theory-driven hypotheses, and subsequently conduct a behavioral experiment to evaluate the model. We use example-based explanations (Fahse et al., 2022) to design a human-AI collaboration scenario with a high potential for learning. Our results reveal several interrelated findings. First, we see that example-based explanations enhance human learning during the process of human-AI collaboration. Furthermore, this enhanced learning provides individuals with a better ability to determine when to rely on their own judgment. Finally, when a significant amount of learning is already present, it can effectively help determine the optimal times to rely on the AI agent's advice.

Our insights address how to harness complementarity in AI-assisted decision-making. However, many different forms of human-AI collaboration exist beyond AI-assisted decision-making, e.g., anomaly detection and investigation, generative AI, etc. Thus, next, in **Part V**, we analyze how to harness complementarity beyond AI-assisted decision-making addressing **RQ4**. Many different forms of human-AI collaboration and tasks are possible to investigate. Therefore, we briefly outline the reasoning for the focus of this thesis, followed by our methodological approach.

AI-assisted decision-making, per definition, comprises classification and regression tasks which are both parts of so-called supervised ML (Kühl et al., 2022). Supervised ML refers to techniques that allow a system to learn a particular task from a set of given instances (Mitchell, 1997). In the learning process, no manual adjustment or programming of rules or strategies to solve a problem is required (Kühl et al., 2022). Beyond supervised ML, literature usually refers to two other types of ML, unsupervised and reinforcement (Kühl et al., 2022). Unsupervised ML comprises methods that reveal previously unknown patterns in data. Reinforcement learning refers to methods that are concerned with teaching intelligent agents to take those kinds of actions that increase their cumulative reward (Kaelbling et al., 1996). To broaden our perspective, we want to study harnessing complementarity in unsupervised ML, an approach used in many practical use cases, such as cancer detection (Haq et al., 2021), predictive maintenance (Amruthnath & Gupta, 2018),

or intrusion detection (Verkerken et al., 2022). Therefore in **Chapter 8**, we shift the focus from supervised ML to unsupervised ML. Further, we selected the case of anomaly detection due to its significant relevance in practical applications of unsupervised ML (Casolla et al., 2019; Roohi et al., 2020).

Anomaly detection is a critical task in many domains, including cybersecurity (Blazquez-Garcia et al., 2021) and maintenance (Ren et al., 2018), where human experts often rely on anomaly detection systems based on unsupervised ML due to the tedious nature of continuous data monitoring. However, these systems may flag events that are unusual, such as a scheduled machine shutdown, but not events of interest, such as unexpected machine failures. This requires human expertise to investigate the relevance of the detected anomalies, revealing a potential area of complementarity potential. To harness this complementarity, we propose a novel method that supports anomaly investigation by providing explanations of the unsupervised anomaly detection. To evaluate the effectiveness of our method, we conduct a behavioral experiment using New York City taxi records as a testbed. Participants are tasked with distinguishing between anomalies due to extreme weather conditions and those due to other unusual events. The results of the experiment show that the inclusion of counterfactual explanations improves the examination of anomalies. Our results show that providing counterfactual explanations does improve the investigation of anomalies, indicating the potential for explainable anomaly detection to harness complementarity potential in general.

In addition, to analyzing complementarity in unsupervised ML, we aim to address one of the core challenges in AI-assisted decision-making, presuming the theoretical inherent complementarity over time (Fügener et al., 2021). Despite the many benefits that recent breakthroughs in AI-assisted decision-making have brought to business and society, there are also some drawbacks. It has long been known that AI-assisted decision-making can lead to various drawbacks, such as automation bias and deskilling (Goddard et al., 2012). In particular, the deskilling of knowledge workers is a major issue, as they are the same people who should also train, challenge and evolve AI. Therefore, in **Chapter 9**, we further explore **RQ4** and address deskilling in AI-assisted decision-making. To this end, based on a literature review of two different research streams—decision support system (DSS) and automation—we conceptualize a new form of human-AI collaboration, which we call intelligent decision assistance (IDA). IDA supports human agents without influencing them through explicit AI advice. Specifically, we propose to use techniques of XAI while withholding concrete AI recommendations. To test this conceptualization, we develop hypotheses on the impacts of IDA and provide first evidence for their validity based on empirical studies in the literature.

Finally, in **Part VI**, we summarize the results of this thesis, including theoretical contributions, and discuss its managerial implications as well as limitations and a research outlook.

1.4 Integration of Articles

The foundation of this thesis is built upon eight separate articles (Chapter 2 through Chapter 9), out of which five have already received acceptance for publication in scholarly forums. The other three are included as working papers. Table 1.2 on page 15 summarizes the integration of articles. The subsequent tables provide a summarization of these original publications, arranged in the order of their presentation within this thesis (refer to Table 1.4 – Table 1.10). Each table enumerates the title, authors, the outlet, the outlet type (conference vs journal paper), and year of publication, along with their abstract. To give an intuition about the quality of the work, we additionally report ranking measures. Since we publish in different research communities (IS, HCI, CS) depending on the focus of our research, we report the VHB-Jourqual3 (Verband der Hochschullehrerinnen und Hochschullehrer für Betriebswirtschaft e.V., 2022) ranking, the CORE conference or journal rating, and the H5 index of google scholar.

Part I	Introduction and Foundations	
	Chapter 2 <i>Foundations</i>	Integrated Article: Kühl, N., Schemmer, M., Goutier, M., & Satzger, G. (2022). <i>Artificial Intelligence and Machine Learning. Electronic Markets</i> , 32(4), 2235–2244.
Part II - RQ1 -	Analysis of the Current State of Empirical Work on Human-AI Collaboration	
	Chapter 3 <i>Structured Literature Review of Empirical Studies on Human-AI Collaboration.</i>	Integrated Article: Hemmer, P., Schemmer, M., Vössing, M., & Kühl, N. (2021). <i>Human-AI Complementarity in Hybrid Intelligence Systems: A Structured Literature Review. Proceedings of the 17th Pacific Asia Conference on Information Systems</i> , 78–92.
	Chapter 4 <i>A Meta-Analysis of the Impact of Explainable AI on Decision Performance</i>	Integrated Article: Schemmer, M., Hemmer, P., Nitsche, M., Kühl, N., & Vössing, M. (2022). <i>A Meta-Analysis of the Utility of Explainable Artificial Intelligence in Human-AI Decision-Making. Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society</i> , 617–626.
Part III - RQ2 -	Conceptualization of Human-AI Complementarity	
	Chapter 5 <i>Conceptualization of Human-AI Complementarity and the Influence of Information Asymmetry</i>	Integrated Article: Schemmer, M., Hemmer, P., Kühl, N., Vössing, M., & Satzger, G. (2023). <i>Human-AI Complementarity: Conceptualization and the Effect of Information Asymmetry. Working Paper.</i>
Part IV - RQ3 -	Harnessing Complementarity in AI-assisted Decision-Making	
	Chapter 6 <i>Harnessing Complementarity: The Role of Appropriate Reliance</i>	Integrated Article: Schemmer, M., Kuehl, N., Benz, C., Bartos, A., & Satzger, G. (2023). <i>Appropriate Reliance on AI Advice: Conceptualization and the Effect of Explanations. Proceedings of the 28th International Conference on Intelligent User Interfaces</i> , 410–422.
	Chapter 7 <i>Harnessing Complementarity: The Influence of Human Learning on Appropriate Reliance</i>	Integrated Article: Schemmer, M., Bartos, A., Spitzer, P., Hemmer, P., Kühl, N., Liebschner, J., & Satzger, G. (2023) <i>Towards Effective Human-AI Decision-Making: The Role of Human Learning in Appropriate Reliance on AI Advice. Working Paper.</i>
Part V - RQ4 -	Harnessing Complementarity Beyond AI-assisted Decision-Making	
	Chapter 8 <i>Harnessing Complementarity in Anomaly Detection</i>	Integrated Article: Schemmer, M., Holstein, J., Kühl, N., Vössing, M., & Satzger, G. (2023) <i>From Anomaly Detection to Anomaly Investigation: Support by Explainable AI. Working Paper.</i>
	Chapter 9 <i>Harnessing Complementarity in Intelligent Decision Assistance</i>	Integrated Article: Schemmer, M., Kühl, N., & Satzger, G. (2022). <i>Intelligent Decision Assistance Versus Automated Decision-Making: Enhancing Knowledge Work Through Explainable Artificial Intelligence. Proceedings of the Hawaii International Conference on Systems Sciences</i> , 617–626.

Table 1.2.: Overview of the integrated articles in the structure of the thesis.

Table 1.3.: Summary of the paper included in Chapter 2

Title	Artificial Intelligence and Machine Learning
Author(s)	Kühl, N., Schemmer, M., Goutier, M., Satzger, G.
Outlet	Electronic Markets
Type	Journal Paper
Year	2022
Ranking	VHB-JQ3: B CORE Conference Ranking: A H5-Index: 41
Abstract	Within the last decade, the application of “artificial intelligence” and “machine learning” has become popular across multiple disciplines, especially in information systems. The two terms are still used inconsistently in academia and industry—sometimes as synonyms, sometimes with different meanings. With this work, we try to clarify the relationship between these concepts. We review the relevant literature and develop a conceptual framework to specify the role of machine learning in building (artificial) intelligent agents. Additionally, we propose a consistent typology for AI-based information systems. We contribute to a deeper understanding of the nature of both concepts and to more terminological clarity and guidance—as a starting point for interdisciplinary discussions and future research.

Table 1.4.: Summary of the paper included in Chapter 3

Title	Human-AI Complementarity in Hybrid Intelligence Systems: A Structured Literature Review
Author(s)	Hemmer, P., Schemmer, M., Vössing, M., Köhl, N. (Shared first authorship)
Outlet	Proceedings of the 28th Pacific Asia Conference on Information Systems
Type	Conference paper
Year	2021
Ranking	VHB-JQ3: C CORE Conference Ranking: N/A H5-Index: 24
Abstract	Hybrid Intelligence is an emerging concept that emphasizes the complementary nature of human intelligence and artificial intelligence (AI). One key requirement for collaboration between humans and AI is the interpretability of the decisions provided by the AI to enable humans to assess whether to comply with the presented decisions. Due to the black-box nature of state-of-the-art AI, the explainable AI (XAI) research community has developed various means to increase interpretability. However, many studies show that increased interpretability through XAI does not necessarily result in complementary team performance (CTP). Through a structured literature review, we identify relevant factors that influence collaboration between humans and AI. Additionally, as we collect relevant research articles and synthesize their findings, we develop a research agenda with relevant hypotheses to lay the foundation for future research on human-AI complementarity in Hybrid Intelligence systems.

Table 1.5.: Summary of the paper included in Chapter 4

Title	A Meta-Analysis on the Utility of Explainable Artificial Intelligence in human-AI collaboration
Author(s)	Schemmer, M., Hemmer, P., Nitsche, M., Kühl, N., Vössing, M. (Shared first authorship)
Outlet	Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society
Type	Conference paper
Year	2022
Ranking	VHB-JQ3: N/A CORE Conference Ranking: N/A H5-Index: 44
Abstract	Research in artificial intelligence (AI)-assisted decision-making is experiencing tremendous growth with a constantly rising number of studies evaluating the effect of AI with and without techniques from the field of explainable AI (XAI) on human decision-making performance. However, as tasks and experimental setups vary due to different objectives, some studies report improved user decision-making performance through XAI, while others report only negligible effects. Therefore, in this article, we present an initial synthesis of existing research on XAI studies using a statistical meta-analysis to derive implications across existing research. We observe a statistically positive impact of XAI on users' performance. Additionally, the first results indicate that human-AI collaboration tends to yield better task performance on text data. However, we find no effect of explanations on users' performance compared to sole AI predictions. Our initial synthesis gives rise to future research investigating the underlying causes and contributes to further developing algorithms that effectively benefit human decision-makers by providing meaningful explanations.

Table 1.6.: Summary of the paper included in Chapter 5

Title	Human-AI Complementarity: Conceptualization and the Effect of Information Asymmetry
Author(s)	Schemmer, M., Hemmer, P., Kühl, N., Vössing, M., Satzger, G. (Shared first authorship)
Outlet	Working Paper (Under review)
Type	Journal paper
Year	2023
Ranking	VHB-JQ3: A CORE Conference Ranking: N/A H5-Index: 47
Abstract	Artificial intelligence (AI) can improve human decision-making in various application areas. Ideally collaboration between humans and AI systems should lead to complementary team performance (CTP), i.e., a level of performance that none of them can reach individually. However, CTP has rarely been observed, suggesting an insufficient understanding of the complementary constituents within human-AI collaboration that can contribute to CTP in decision-making. Therefore, this work aims at a holistic theoretical foundation for understanding and developing human-AI complementarity: Based on existing IS justificatory knowledge, we conceptualize complementarity which consists of formalizing the notion of complementarity potential, outlining possible sources, and introducing a classification of mechanisms for the integration of human and AI decisions during collaboration. To illustrate our conceptualization, we apply it in an empirical study: We focus on information asymmetry as a promising source of complementarity potential, and, for a real estate appraisal use case, demonstrate that humans can in fact leverage contextual information to achieve CTP. Our work provides researchers with a theoretical foundation of complementarity for human-AI collaboration and demonstrates that information asymmetry can constitute a promising source of inherent complementarity potential that can be turned into CTP.

Table 1.7.: Summary of the paper included in Chapter 6

Title	Appropriate Reliance on AI Advice: Conceptualization and the Effect of Explanations
Author(s)	Schemmer, M., Kühl, N., Benz, C., Bartos, A., Satzger, G.
Outlet	Proceedings of the 28th International Conference on Intelligent User Interfaces
Type	Conference paper
Year	2023
Ranking	VHB-JQ3: N/A CORE Conference Ranking: A H5-Index: 40
Abstract	AI advice is becoming increasingly popular, e.g., in investment and medical treatment decisions. As this advice is typically imperfect, decision-makers have to exert discretion as to whether actually follow that advice: they have to “appropriately” rely on correct and turn down incorrect advice. However, current research on appropriate reliance still lacks a common definition as well as an operational measurement concept. Additionally, no in-depth behavioral experiments have been conducted that help understand the factors influencing this behavior. In this paper, we propose Appropriateness of Reliance (AoR) as an underlying, quantifiable two-dimensional measurement concept. We develop a research model that analyzes the effect of providing explanations for AI advice. In an experiment with 200 participants, we demonstrate how these explanations influence the AoR, and, thus, the effectiveness of AI advice. Our work contributes fundamental concepts for the analysis of reliance behavior and the purposeful design of AI advisors.

Table 1.8.: Summary of the paper included in Chapter 7

Title	Towards Effective Human-AI Collaboration: The Role of Human Learning in Appropriate Reliance on AI Advice
Author(s)	Schemmer, M., Bartos, A., Spitzer, P., Hemmer, P., Köhl, N., Liebschner, J, Satzger, G.
Outlet	Working Paper (Under review)
Type	Conference paper
Year	2023
Ranking	VHB-JQ3: A CORE Conference Ranking: N/A H5-Index: 34
Abstract	<p>The true potential of human-AI collaboration lies in exploiting the complementary capabilities of humans and AI to achieve a joint performance superior to that of the individual AI or human, i.e., to achieve Complementary Team Performance (CTP). To realize this complementarity potential, humans need to exert discretion in following an AI's advice, i.e., they need to appropriately rely on the AI's advice. While previous work has focused on building a mental model of the AI to assess an AI recommendation, recent research has shown that the mental model alone cannot explain appropriate reliance. We hypothesize that, in addition to the mental model, human learning is a key mediator of appropriate reliance and, thus, CTP. In this study, we demonstrate the relationship between learning and appropriate reliance in an experiment with 100 participants. This work provides fundamental concepts for analyzing reliance and derives implications for the effective design of human-AI collaboration.</p>

Table 1.9.: Summary of the paper included in Chapter 8

Title	From Anomaly Detection to Anomaly Investigation: Support by Explainable AI
Author(s)	Schemmer, M., Holstein, J., Kühl, N., Satzger, G.
Outlet	Working Paper (under review)
Type	Journal paper
Year	2023
Ranking	VHB-JQ3: C CORE Journal Ranking: N/A H5-Index: N/A
Abstract	<p>Fast and reliable anomaly detection is critical in many areas, including cybersecurity and maintenance. However, continuously monitoring data streams for anomalies is an error-prone and tedious task that requires innovative solutions. One approach is the use of machine learning to identify anomalous patterns. Given the inherent rarity and potential diversity of anomalies, the availability of labels for training is often limited. For this reason, unsupervised machine learning methods are typically used that do not require any labels. However, the anomalies detected by unsupervised approaches may include rare events, such as a scheduled machine shutdown, but not the actual event of interest, such as a machine failure. Therefore, human experts are generally needed to investigate the relevance of the detected anomalies. Yet, the high dimensionality of data sets and the potential abundance of detected anomalies often exceed the human capabilities to investigate anomalies. Our results show that incorporating these explanations improves the accuracy of human anomaly investigation, providing a novel empirical link between anomaly detection explanations and anomaly investigation. Our work has the potential to significantly impact the design and use of anomaly detection systems in various domains.</p>

Table 1.10.: Summary of the paper included in Chapter 9

Title	Intelligent Decision Assistance Versus Automated Decision-Making: Enhancing Knowledge Workers Through Explainable Artificial Intelligence
Author(s)	Schemmer, M., Kühl, N., Satzger, G.
Outlet	Proceedings of the 55th Hawaii International Conference on System Sciences
Type	Conference paper
Year	2021
Ranking	VHB-JQ3: C Core Conference Ranking: N/A H5-Index: 55
Abstract	Detecting rare events is essential in various fields, e.g., in cyber security or maintenance. Often, human experts are supported by anomaly detection systems as continuously monitoring the data is error-prone and tedious task. However, among the anomalies detected may be events that are rare, e.g., in industrial maintenance a planned shutdown of a machine, but are not the actual event of interest, e.g., breakdowns of a machine. Therefore, human experts are needed to validate whether the detected anomalies are relevant. Related work neglects this human anomaly investigation and instead, solely focuses on the technical implementation of anomaly detection. To close this gap, we introduce a human-centered workflow linking anomaly detection and investigation. We propose to complement the anomaly investigation through explanations of the automated anomaly detection. To evaluate the utility of the workflow, we conduct a behavioral experiment using records of taxi rides in New York City as a testbed. Our results show that providing counterfactual explanations does improve the investigation of anomalies, indicating potential for explainable anomaly detection in general.

To provide a common understanding, we outline the theoretical foundations relevant to this thesis. In Section 2.1, we discuss the relationship between AI and ML to provide a common understanding of AI agents and underlying methods. Next, Section 2.2 outlines basic terminology in the area of human-AI collaboration and clarifies the scope of the thesis. Finally, Section 2.3 provides an overview of XAI, which is explored in this thesis as a central means of designing effective human-AI collaboration.

2.1 Artificial Intelligence and Machine Learning

This chapter comprises an article that was published as: Kühl, N., Schemmer, M., Goutier, M., & Satzger, G. (2022). Artificial Intelligence and Machine Learning. *Electronic Markets*, 32(4), 2235–2244. Note: The abstract has been removed. Tables and figures were reformatted, and newly referenced to fit the structure of the thesis. The terminology was standardized with the dissertation. Chapter, and section numbering and respective cross-references were modified. Formatting and reference style was adapted and references were integrated into the overall references section of this thesis.

AI has been named as one of the most recent, fundamental developments of the convergence in electronic markets (Alt, 2021) and has become an increasingly relevant topic for IS research (Abdel-Karim et al., 2021; Alt, 2018). While a large body of literature is concerned with designing AI to mimic and replace humans (Dunin-Barkowski, 2020; Fukuda et al., 2001), IS research in general, and DSS research in particular, emphasize the support of humans with AI (Arnott & Pervan, 2015). Recent research in hybrid intelligence (HI) and human-AI collaboration offers a promising path in synthesizing AI research across different fields (Dellermann et al., 2019b): The ultimate goal of HI is to leverage the individual advantages of both human and AI to enable synergy effects (Wilson & Daugherty, 2018) and to achieve CTP (Hemmer et al., 2021).

However, in many cases in both research and practice, AI is simply equated with the concept of ML—negatively impacting terminological precision and effective communication. Ågerfalk et al. (2020, p. 2) emphasizes that differentiating between AI and ML is especially important for IS research: “Is it not our responsibility as IS scholars to bring clarity to the discourse rather than contributing to its decline? (...) It would mean to distinguish between different types of AI and not talk of AI as synonymous with ML, which in itself is far from a monolithic concept.”

The practical relevance of a clear understanding is underlined by observing confusion and misuse of the terms AI and ML: During Mark Zuckerberg’s U.S. senate hearing in April 2018, he stressed that Facebook had “AI tools to identify hate speech” as well as “terrorist propaganda” (The Washington Post, 2018). Researchers, however, would usually describe tasks identifying specific social media platform instances as classification tasks in the field of (supervised) ML (Waseem & Hovy, 2016). The increasing popularity of AI (Hidemichi & Shunsuke, 2017) has led to the term often being used interchangeably with ML. This does not only hold true for the statement of Facebook’s CEO above, but also across various theoretical and application-oriented contributions in recent literature (Brink, 2017; Nawrocki et al., 2018; Shirazi et al., 2017). Camerer (2018) even mentions that he still uses AI as a synonym for ML despite knowing it is inaccurate.

As the remainder of this paper shows, both concepts are not identical—although in many cases both terms will appear in the same context. Such ambiguity might lead to multiple imprecisions in both research and practice when conversing about the relevant concepts, methods, and results. This is especially important in IS research—being interdisciplinary by nature (D’Atri et al., 2008). Ultimately, misuse can either lead to fundamental misunderstandings (Carnap, 1955) or to research that ought to be undertaken not being conducted (Davey & Cope, 2008). After all, misunderstandings can potentially lead to low perceived trustworthiness of AI (Thiebes et al., 2021).

It seems surprising that despite the frequent use of the terms, there is hardly any helpful academic delineation—apart from the notion that ML is a (not well-defined) subset of AI (Campeato, 2020), comparable to other possible subdisciplines of AI: Expert systems, robotics, natural language processing, machine vision, and speech recognition (Léon & Dejoux, 2018; Vickers, 2017). Consequently, this paper aims to shed light on the relationship between the two concepts: We analyze the role of ML in AI and, more precisely, in intelligent agents, which are defined by their capability to sense and act in an environment (Schleiffer, 2005). We do so by taking an ML perspective on intelligent agents’ capabilities and their relevant

implementation—with IS research in mind. To this end, we review the relevant literature for both terms and synthesize and conceptualize the results.

Our article’s contributions are twofold: First, we identify different contributions of ML to intelligent agents as specific AI instantiations. We base this on an expansion of the existing AI framework by (Russell, 2010) — explicitly breaking down intelligent agents’ capabilities into separate “execution” and “learning” capabilities. Second, we develop a typology to provide a common terminology for AI-based information systems, where we conceptualize which systems employ ML—and which do not. The result should provide guidance when designing and analyzing systems.

Next, in Section 2.1.1, we review relevant literature in the fields of AI and ML. In Section 2.1.2, we then analyze the capabilities of intelligent agents in more depth and examine the role of ML in them. Section 2.1.3 develops a framework and typology to differentiate the terms AI and ML and to explain their relationship.

2.1.1 Terminology

Over the last decade, both terms, artificial intelligence (AI) and machine learning (ML), have enjoyed increasing popularity in IS research. An analysis of the “AIS Senior Scholars’ Basket¹” journals since 2000², illustrates how the occurrences of both terms increased in titles, abstracts, and keywords (cf. Figure 2.1 on page 28). While over the last 21 years, we observe a small but constant number of publications covering AI-related topics, ML only gained relevance in the literature after 2017: The late reflection of ML—despite of the earlier adoption and spread in industry (Brynjolfsson & McAfee, 2017)—may raise questions about whether IS has picked up the topic early enough.

As the analysis demonstrates, the two terms do exist for quite some time, while their related subjects are highly and increasingly topical now. In this section, we will elaborate on the meaning of the terms.

Artificial Intelligence

In 1956, a Dartmouth workshop, led by Minsky and McCarthy, coined the term “Artificial Intelligence” (McDonald et al., 2017)—later taking in contributions from a

¹As of March 2022, see <https://aisnet.org/page/SeniorScholarBasket>, last accessed 16.05.2022

²We start with the year 2000, as it was the last point in time when a journal (JAIS) was added to AIS Senior Scholars’ Basket.

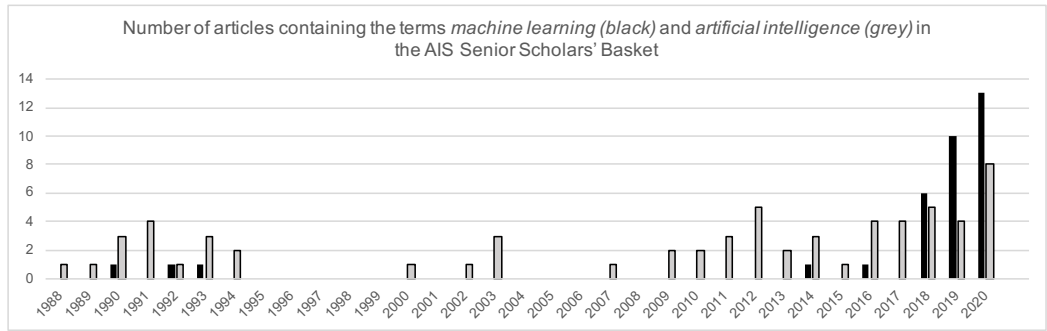


Figure 2.1.: Appearance of the terms “artificial intelligence” and “machine learning” in AIS Senior Scholars’ Basket journals.

variety of different research disciplines, such as computer science (He et al., 2016) and programming (Newell & Simon, 1961), neuroscience (Ullman, 2019), robotics (Brady, 1985), linguistics (Clark et al., 2012), philosophy (Smith, 2013), and futurology (Krawczyk, 2016). While the terminology is not well defined across disciplines, even within the IS domain definitions do vary widely; (Vickers, 2017) provide a comprehensive overview. Recent AI definitions transfer the human intelligence concept to machines in its entirety as “the ability of a machine to perform cognitive functions that we associate with human minds, such as perceiving, reasoning, learning, interacting with the environment, problem solving, decision-making, and even demonstrating creativity” (Rai et al., 2019). Still, over the last decades various debates have been raging on the *depth* and *objectives* of AI. These two dimensions span the space for different AI research streams in computer science and IS that were categorized by Russell (2010): On the one hand (depth dimension), it may target either the thought process or a concrete action (thinking vs. acting); on the other hand (objective dimension), it may try to either replicate human decision-making or to provide an ideal, “most rational” decision (human-like vs. rational decision). The resulting research streams are depicted in Table 2.1.

Table 2.1.: AI research streams (based on Russell (2010))

Depth \ Objective	Human	Rational
	Thinking	Cognitive modeling
Acting	Turing Test	Rational agent <i>(Perspective of this work)</i>

According to the *cognitive modeling* (i.e., thinking humanly) stream, AI instantiations must be “machines with a mind” (Haugeland, 1989) that perform human thinking

(Bellman, 1978). Not only should they arrive at the same output as a human when given the same input, but also apply the same reasoning steps leading to this conclusion (Newell & Simon, 1961). The laws of thought stream (i.e., thinking rationally) requires AI instantiations to arrive at a rational decision despite what a human might come up with. AI must therefore adhere to the laws of thought by using logic-based computational models (Charniak & McDermott, 1985). The Turing test stream (i.e., acting humanly) implies that AI must act intelligently when interacting with humans. To accomplish such tasks, AI instantiations must perform human tasks at least as well as humans (Rich & Knight, 1991), which can be tested via the Turing test (Turing, 2012). Finally, the rational agent stream considers AI as a rational (Russell, 2010) or intelligent (Poole et al., 1998) agent³. This agent does not only act autonomously, but also with the objective of achieving the rationally ideal outcome.

Machine Learning

Many researchers perceive ML as an (exclusive) part of AI (Copeland, 2016; Ongsulee, 2017; Vickers, 2017). In general, learning is a key facet of human cognition (Neisser, 2014). Humans process a vast amount of information by utilizing abstract knowledge that helps them to better understand incoming input. Owing to their adaptive nature, ML models can mimic a human being's cognitive abilities (Janiesch et al., 2021) : ML describes a set of methods commonly used to solve a variety of real-world problems with the help of computer systems, which can learn to solve a problem instead of being explicitly programmed to do so (Koza et al., 1996). For instance, instead of explicitly telling a computer system which words within an tweet would indicate it to contain a customer need, the system (given a sufficient set of training samples) learns the typical patterns of words and their combination which results in a need classification (Kühl et al., 2020).

In general, we differentiate between unsupervised, supervised, and reinforcement ML. Unsupervised ML comprises methods that reveal previously unknown patterns in data. Consequently, unsupervised learning tasks do not necessarily have a “correct” solution, as there is no ground truth (Wang et al., 2009).

Supervised ML refers to methods that allow the building of knowledge about a given task from a series of examples representing “past experience” (Dietterich, 2009). In the learning process, no manual adjustment or programming of rules or strategies to

³In this case, the terms rational and intelligent are used interchangeably in related work (Gama et al., 2014; Koza et al., 1996; Russell, 2010).

solve a problem is required, i.e., the model is capable to learn “by itself”. In more detail, supervised ML methods always aim to build a model by applying an algorithm to a set of known data points to gain insight into an unknown set of data (Hastie et al., 2009): Known data points are semantically labeled to create a target for the ML model. So-called semi-supervised learning combines elements from supervised and unsupervised ML by jointly using labeled and unlabeled data (Zhu, 2005).

Reinforcement learning refers to methods that are concerned with teaching intelligent agents to take those kinds of actions that increase their cumulative reward (Kaelbling et al., 1996). It differs from supervised learning in that no correctly matched features and targets are required for training. Instead, rewards and penalties allow the model to continuously learn over time. The focus is on a trade-off between the exploration of the uncharted environment and the exploitation of the existing knowledge base.

2.1.2 The Role of Rational Agents in Information Systems

To further elaborate on the role of ML within AI, we need to take a clear perspective on the different definitions of AI to be beneficial to IS research. IS traditionally utilizes ML in predictive analytics tasks within (intelligent) decision support systems (DSS) (Arnott & Pervan, 2015; Müller et al., 2016) where the goal is to generate the best possible outcome (Arnott & Pervan, 2015; Hunke et al., 2022; Power et al., 2019). As Phillips-Wren et al. (2019, p. 63) emphasize, DSS “should help the decision-maker think rationally”. The perspective of rationality is also endorsed by other researchers in the field (Bakos & Treacy, 1986; Dellermann et al., 2019b; Klör et al., 2018; Power et al., 2019; Schuetz & Venkatesh, 2020). Thus, in the following we will explore the relationship between ML and AI in IS from the lens of the rational agent stream as discussed above. Furthermore, we will focus on supervised ML as it is the most common type of ML (Jordan & Mitchell, 2015). In the remainder of this section, we will first distinguish different types of (rational) agents and then use the insights to differentiate between the necessary layers when designing them as part of information systems.

Types of Rational Agents

According to the selected research stream, intelligence manifests itself in how rational agents act. Five features characterize agents in general: they “operate autonomously, perceive their environment, persist over a prolonged time period, adapt to change, and create and pursue goals” (Russell, 2010). An agent defines its

action, not for itself, but within the environment it operates and interacts with. It recognizes the environment through its sensors, relies on an agent program to handle and digest input data, and performs actions via actuators. A rational agent targets to achieve the highest expected outcome according to one or multiple objective performance measures—which are based on current and past knowledge of the environment and possible actions. For example, a rational agent within a medical diagnosis system aims to maximize the health of a patient measured via blood pressure, heart rate, and blood oxygen (potentially while minimizing the financial costs of a treatment as a secondary condition) (Grosu, 2022).

The agent’s conceptualization and surroundings are summarized in the *agent-environment* framework. It consists of three components: an agent, an environment, and a goal. Intelligence is the measurement of the “agent’s ability to achieve goals in a wide range of environments” (Legg & Hutter, 2007). The agent obtains input through perceptions that the environment generates. Observations of the environment are one type of perception, while others are reward signals that indicate how well the agents’ goals have been achieved. Based on these input signals, the agent decides to perform actions, which are subsequently communicated back as signals to the environment.

Rational Agents in Information System Architectures

As we investigate the role of ML in AI for IS research, we also need—apart from the theoretical and definitory aspects of agents—to consider how the functionality of a rational agent is reflected in an IS architecture. The implementation of agents is a key step to embed their functionality into practical, real-world (intelligent) information systems in general or into DSS specifically (Gao & Xu, 2009). Any rational agent needs to be capable of at two least two tasks: cognition (Lieto et al., 2018) and (inter)action with the environment (Russell, 2010). If we map these capabilities to system design terms, then acting capabilities are the ones built into a frontend, while the cognitive capabilities are embedded in a *backend*.

The frontend as the interface to the environment may take various forms; it may be designed as a very abstract, machine-readable web interface (Kühl et al., 2020), a human-readable application (Engel et al., 2022; Hirt et al., 2019), or even a humanoid template with elaborated expression capabilities (Guizzo, 2014). For the frontend to interact with the environment, two technical components are required: sensors and actuators. Sensors detect events or changes in the environment and forward the information via the frontend to the backend. They can, for instance,

read the signals within an industrial process network (Hein et al., 2019), read visuals of an interaction with a human (Geller, 2014), but also perceive a keystroke input (Russell, 2010). Actuators, on the other hand, are components responsible for moving, controlling, or displaying content. While sensors merely process information, actuators act, for instance, by automatically making bookings (Neuhofer et al., 2015) or changing a humanoid’s facial expressions (Berns & Hirth, 2006). One could argue that the Turing test (Turing, 2012) takes place at the environment’s interaction with the frontend, or, more precisely, when sensors and actuators are combined in a way to test the agent’s AI for *acting humanly*.

The backend provides the required functionalities to depict an intelligent agent’s cognitive capabilities. More precisely, this executing backend allows the agent to draw on its built-in knowledge. The backend translates signals from the frontend and transforms them into signals sent back to the frontend as a response by executing actions. In some cases, there is an additional component modifying this response function over time, and thus modifying the execution part of the backend. We call this the learning part of the backend as depicted in Figure 2.2. Within the next subsections, we will further elaborate this framework and its components.

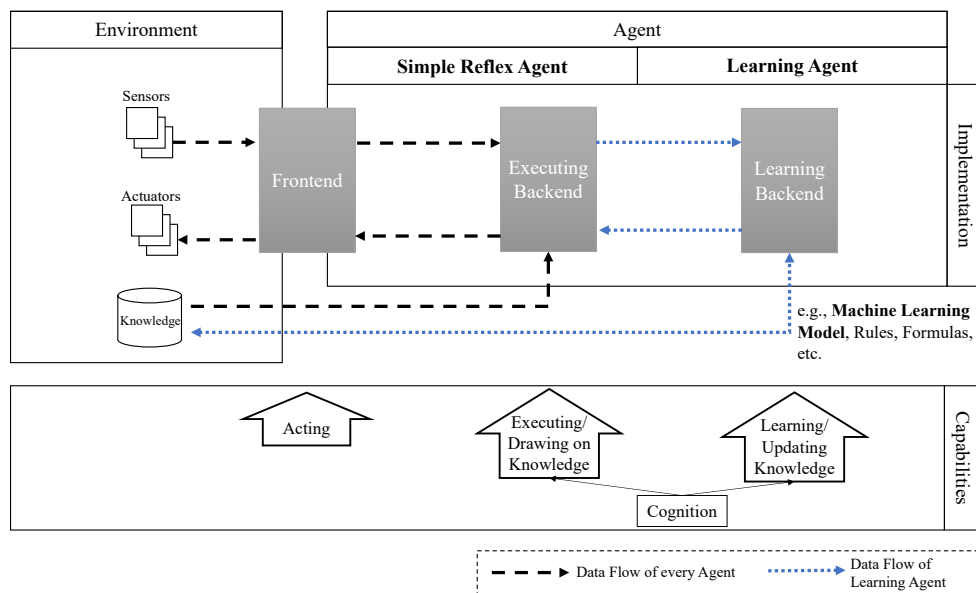


Figure 2.2.: Conceptual framework describing the general architecture for intelligent agents in AI-based information systems.

The Role of Machine Learning in Rational Agents

In terms of supervised ML, we need to further differentiate between the process *task of building* (training) adequate ML models (Witten et al., 2011) and the one of executing the deployed models (Chapman et al., 2000). To further understand ML's role in intelligent agents, we partition the agent's cognition layer into a learning sublayer (model building) and an executing sublayer (model execution). We, therefore, regard the implementation required by the learning sublayer as the learning backend, while the executing backend denotes the executing sublayer.

The learning backend first dictates *if* the intelligent agent is able to learn, and, second, *how* it does so—with respect to the algorithms it actually uses, the type of data processing it applies, and the handling of concept drift (Gama et al., 2014), etc. Using the terminology of Russell (2010), we distinguish two different types of intelligent agents: *simple-reflex agents* and *learning agents*. This differentiation holds explicitly in terms of a ML perspective on AI because it considers whether the underlying models in the cognition layer are trained just once and after that never touched (simple-reflex), or whether they are continuously updated to be adaptive (learning). Related work provides suitable examples of both. Kitts and Leblanc (2004) build a bidding agent for digital auctions as a simple-reflex agent: While building and testing the model for the agent may show convincing results, the system's adaptive learning after deployment could be critical. Other examples of agents with models trained just once are common in different areas, for example, in terms pneumonia warning for hospitals (Oroszi & Ruhland, 2010), the (re)identification of pedestrians (Zheng et al., 2018), and object annotation (Jorge et al., 2014). On the other hand, recent literature also provides examples of learning agents. Jordan and Mitchell (2015) present the concept of “never-ending learning” agents that strongly focus on continuously building and updating models in agents. Neuhofer et al. (2015) suggest an agent capable of personalization through a continuous learning processes of guest information for digital platforms, which an example of such an agent. Other examples include agents capable of making recommendations on music platforms (Liebman et al., 2014), regulating heat pump thermostats (Ruelens et al., 2015), acquiring collective knowledge across different tasks (Rostami et al., 2017), and learning the meanings of words (Yu et al., 2017). The choice of the learning type in agents (simple-reflex vs. learning agent) influences the agent's general overall design and the contribution of ML.

As a result from the layers of agents and types of learning, our conceptual framework combining both is shown in Figure 2.2 on page 32. Regarding the previously mentioned ML methods, supervised ML can be the basis for either simple-reflex

or learning agents, depending on whether the learning backend exists and on its feedback to the agent’s knowledge base. In terms of reinforcement learning, the agent, by definition, is a learning agent. However, there are also examples of where an agent functions without the utilization of ML—because the execution is based on rules (Wang et al., 2009), formulas (Billings et al., 2002) or other methods (Abasolo & Gomez, 2000). From this perspective, this means there can be AI without ML.

2.1.3 A Typology for Machine Learning in AI Systems

Based on the differentiation between simple-reflex and learning agents, we can now derive a typology for IS research. We refer to IS systems as static AI-based systems if they employ simple reflex agents that may be based on a model trained with ML. Adaptive AI-based systems, though, use learning agents, i.e., do have a learning backend—that may be based on ML, but alternatively also could be based, e.g., on rule-based knowledge representation. We, thus, propose the typology (as depicted in Figure 2.3) for AI-based IS along the two dimensions: the existence of an ML-base for the executing backend and the existence of a learning backend.

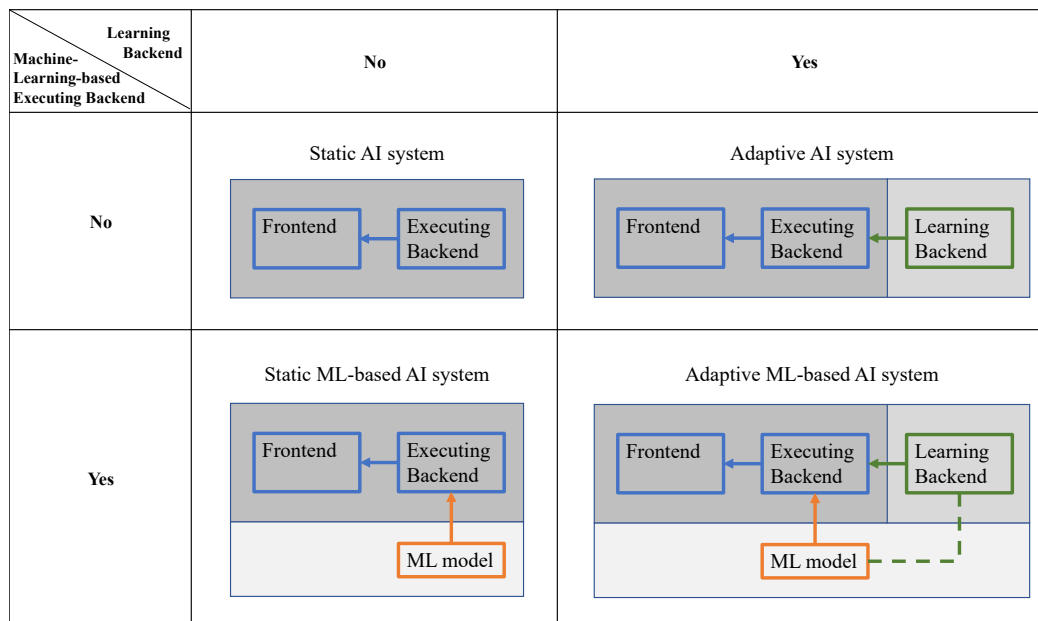


Figure 2.3.: Typology of AI-based information systems.

We illustrate these findings in concrete IS research examples: *Static AI systems* are characterized by an executing backend which is based on algorithms not classified as ML and they lack a learning backend, i.e. they have a fixed response model (Chuang & Yadav, 1997). The executing backend of such systems is based on rules

(like nested if-else statements), formulas (like mathematic equations describing a phenomena) or algorithms (like individual formal solution descriptions for specific problems). As an example for such systems, Hegazy et al. (2005) build a static AI system based on a self-developed algorithm and evaluate its performance within a cybersecurity context by simulating multiple attacks. Another example is provided by Ritchie (1990) who has developed an architecture and an instantiation of a static AI system for a traffic management platform.

In contrast, a static ML-based AI system has an executing backend which is based on ML. An example is provided in He et al. (2018). The authors develop an artifact to classify marketing on Twitter in either defensive or offensive marketing and show convincing prediction results. While their work did not aim at designing a productive artifact and is rather focused on showing the general feasibility of the approach, they choose a static ML-based AI system—which, however, might not be sufficient for permanent use: After the release of the article in 2018, Twitter changed its tweet size from 140 to 280 characters, thus changing the environment. It would be interesting to see how the developed model would need to adapt to this change. As another example, Samtani et al. (2017) build a model to identify harmful code snippets, typically utilized by hackers. They show how to design an artifact that can detect these code assets accurately for a proactive cyber threat intelligence. However, also in this case the environment and the assets of the hackers could and will change over time.

Adaptive AI systems, which are not based on ML, do comprise an executing backend with the flexibility to dynamically adapt the model to changing environments. This type of system is oftentimes enabled through the interaction between humans and AI systems. Most of the times, the system provides means and triggers for updates, while the human provides “manually encoded” knowledge updates. For example, Zhou et al. (2009) implement an adaptive AI system for pipeline leak detection which is based on a rule-based expert system and offers means to update the system online. In another example, Hatzilygeroudis and Prentzas (2004) develop an adaptive AI system to support the teaching process which has a specific component for knowledge updates. Both examples are inherently knowledge-based, but are explicitly designed to allow and force updates—although not on the basis of ML.

Finally, *adaptive ML-based AI-systems* implement learning in both sublayers of the cognition layer. For example, Zheng et al. (2013) design a reinforcement-learning-based artifact to obtain information from hidden parts (“deep web”) of the internet. As their developed system perceives its current state and selects an action to submit to the environment (the deep web), the system continuously learns and builds up

experience. In another example, Ghavamipoor and Hashemi Golpayegani (2020) build an adaptive ML-based AI system to predict the necessary service quality level and adapt an e-commerce system accordingly. As their system is continuously learning, their results show the total profits improve through effective cost reduction and revenue enhancement.

In this section, we clarified the relationship of machine learning (ML) in artificial intelligence (AI), particularly in intelligent agents, for the field of information systems research. Based on a rational agent view, we differentiated between AI agents capable of continuously improving as well as those who are static. Within these agents as instantiations of artificial intelligence, (supervised) ML can serve to support in different ways: either to contribute a once-trained model to define a static response pattern or to provide an adaptive model to realize dynamic behavior. As we point out, both could also be realized without the application of ML. Thus, “ML” and “AI” are not terms that should be used interchangeably—but as a conscious choice. Without question, ML is an important driver of AI, and the majority of modern AI cases will utilize ML. However, as we illustrate, there can be cases of AI without ML (e.g., based on rules or formulas).

This distinction enables our proposed framework to apply an intelligent agent’s perspective on AI-based information systems, enabling researchers to differentiate the existence and function of ML in them. Interestingly, as of today, many AI-based information systems remain static, i.e. employ once-trained ML models (Kühl et al., 2021). With increasing focus on deployment and life cycle management, we will see more adaptive AI-based systems that sense changes in the environment and use ML to learn continuously (Baier et al., 2019). Our framework and the resulting typology should allow IS researchers and practitioners to be more precise when referring to ML and AI, as it highlights the importance of not using the terms interchangeably but clarifying the role ML plays in AI’s system design.

2.2 Human-AI Collaboration

In this section, we elaborate on key concepts in human-AI collaboration and the scope of the thesis.

In the past, AI was mainly used as an input provider to human decision-making. Advances in AI are leading to an increase in its ability to operate autonomously, essentially changing the role of AI from a provider of decision-making input to an autonomous *AI agent* (Qian & Qian, 2020). *Human-AI teams* refer to the idea of

combining the unique strengths of both human and AI agents in repetitive tasks (Seeber et al., 2020). In this thesis, we focus on human-AI teams comprising of a single AI agent and one human agent. *Human-AI collaboration* is the process by which a human-AI team works together in a synergistic manner to achieve shared goals, e.g., with the AI agent providing recommendations or insights and humans guiding and refining the AI-generated outputs (Terveen, 1995; Vössing et al., 2022). *Hybrid intelligence* is an emerging paradigm with the idea of leveraging complementary heterogeneous intelligence in the form of a socio-technical ensemble to resolve current AI limitations. We refer to the work of Dellermann et al. (2019a, p. 640), who define hybrid intelligence as “the ability to achieve complex goals by combining human and AI, thereby achieving better results than what each of them could have accomplished separately, and continuously improve by learning from each other.”

The predominant form of human-AI collaboration is typically characterized by a human receiving advice from an AI agent while retaining the freedom to adhere to, modify, or disregard that advice (Lai et al., 2023). The term most commonly used in research to describe this setting is *AI-assisted Decision-Making*, see for example Bućinca et al. (2021), Wang and Yin (2021), and Zhang et al. (2020). In this thesis, we follow this naming convention. The term AI-assisted decision-making refers to applying supervised ML and, more specifically, to classification and regression tasks. The types of AI-assisted decision tasks vary widely, from sentiment classification to house price prediction to cancer classification (Lai et al., 2023). Other forms of collaboration beyond AI-assisted decision-making range from different collaboration mechanisms, e.g., delegation (Hemmer et al., 2023), to different ML approaches, e.g., unsupervised ML, to different tasks, e.g., clustering or content generation.

In addition to providing advice, AI agents have an array of other features to assist humans in understanding and evaluating AI predictions for superior final decision-making. In this thesis, these features are defined as AI assistance elements following the terminology of Lai et al. (2023). AI assistance elements can be grouped in information about prediction, information about models and training data, and other AI agent elements (Lai et al., 2023). Information about the prediction includes the provision of model performance metrics, uncertainty indicators, and explanations for predictions. Since the main AI assistance elements used in practice are explanations (XAI) (Bansal et al., 2021; Hemmer et al., 2021; Lai et al., 2023), in the following, we provide an overview of XAI.

2.3 Explainable Artificial Intelligence

In 2004, Van Lent et al. first used the term XAI to describe the ability of their system to explain the behavior of agents in simulation games. The recent rise in popularity of XAI is largely due to the increasing demand for improving the understandability of complex models (Wanner et al., 2020). Although highly complex models can produce superior results compared to linear models, their internal operations can often pose interpretability challenges for humans.

The field of XAI encompasses a wide range of techniques. According to Adadi and Berrada (2018), these techniques can be broadly organized based on their complexity, scope, and degree of dependence.

The complexity of a model is inextricably linked to its interpretability. Wanner et al. (2020) have divided the levels of complexity into white-box, gray-box, and black-box models. White-box models, such as linear regressions, are those with complete transparency and are therefore inherently explicable without the need for additional explanatory techniques. On the other hand, black-box models, despite their superior performance, often struggle with the problem of limited interpretability. Gray box models bridge the gap between the two—they are not interpretable out of the box, but their interpretability can be enhanced with the help of additional explanation techniques.

Explanation techniques can be further characterized by their scope (Adadi & Berrada, 2018): global or local explanations. Global XAI techniques provide thorough explanations for the entire model, while local explanations focus on individual instances. In addition to scope, these XAI techniques can be classified according to whether they are model agnostic, i.e., they can work with any type of model, or model specific.

The most common explanation techniques are feature importance, example-based explanations, and counterfactual explanations (Lai et al., 2023). Feature importance is a model-agnostic technique that provides the decision maker with information about the importance of certain features. Two well-known feature importance algorithms are LIME (Ribeiro et al., 2016a) and SHAP (Lundberg & Lee, 2017). Example-based explanations provide historical data similar to the current instance (Van der Waa et al., 2021). Thus, example-based explanations are essentially a form of information retrieval. Research in psychology suggests that people prefer explanations that use examples (Cai et al., 2019). Counterfactual explanations

provide information about what the smallest change would be to get a different AI decision (Wachter et al., 2017).

Several studies have evaluated whether different types of explanations can support humans' understanding of the AI agent with the goal of better relying on recommendations in the correct cases (Alufaisan et al., 2021; Buçinca et al., 2021; Carton et al., 2020; Van der Waa et al., 2021). However, it has also been shown that some types of explanations can lead people to rely too much on the AI agents' recommendation, especially in cases where the advice is wrong (Bansal et al., 2021; Poursabzi-Sangdeh et al., 2021; Schemmer et al., 2022c). Overall, we find mixed results regarding the effect of explanations on human decision-making. Therefore, in Part II, we begin this thesis with a rigorous analysis of the state of the art of empirical studies in human-AI collaboration in general and in XAI in particular.

Part II

Analysis of the Current State of Empirical
Work on Human-AI Collaboration

Structured Literature Review of Empirical Studies on Human-AI Collaboration

This chapter comprises an article that was published as: Hemmer, P., Schemmer, M., Vössing, M., & Kühn, N. (2021). Human-AI Complementarity in Hybrid Intelligence Systems: A Structured Literature Review. *Proceedings of the 17th Pacific Asia Conference on Information Systems*, 7–39. Note: To improve the structure of the work, the title was changed. The abstract has been removed. Tables and figures were reformatted, and newly referenced to fit the structure of the thesis. The terminology was standardized with the dissertation. Chapter, section and research question numbering and respective cross-references were modified. Formatting and reference style was adapted and references were integrated into the overall references section of this thesis.

3.1 Introduction

Over the last years, an unprecedented development in the field of artificial intelligence (AI) has contributed to an improvement in prediction accuracy of modern AIs—even exceeding the capabilities of domain experts in an increasing number of fields (He et al., 2015). These advancements have fueled the ongoing discussion of whether AI will replace domain experts in the foreseeable future (Schuetz & Venkatesh, 2020). However, in many application domains, reducing human autonomy might not be desirable. For example, the cost of errors in situations in which perfect algorithmic accuracy is not attainable might not be acceptable. Moreover, legal regulations and ethical considerations might make full algorithmic automation undesirable from a societal perspective. Additionally, the capabilities of AI are often limited to narrowly defined application contexts as the utilized algorithms often struggle to handle instances that differ from the patterns learned during training (D’Amour et al., 2022). In these cases, humans can leverage capabilities not

possessed even by state-of-the-art AI—for example, intuition, creativity, and also common sense.

This line of thought gives rise to the vision of so-called hybrid intelligence (HI) (Dellermann et al., 2019a). The HI concept proposes to combine the complementary capabilities of humans and AI by facilitating collaboration to achieve superior results in comparison to the isolated entities operating independently (Dellermann et al., 2019a; Liu et al., 2021). In this context, humans and AI are regarded as equal team members that solve tasks in cooperation (Siemon et al., 2020). Hereby, we understand complementary team performance (CTP) as the desired outcome of human-AI collaboration, i.e., the team performance exceeds the maximum performance of both individual entities. One of the key requirements for the success of human-AI collaboration lies in the fact that they enable humans to understand the decisions provided by the AI and allow them to draw conclusions about when and to what extent they can rely on the AI's prediction. This concept can be traced back to the early research on expert systems (Nunes & Jannach, 2017; Swartout & Moore, 1993).

Nowadays, with the rise of AI, algorithms emerging from the field of explainable AI (XAI) offer a rich fundus of explainability techniques to be applied within human-AI collaboration (Bansal et al., 2019a; Chu et al., 2020; Liu et al., 2021). Since XAI techniques are a means to explain the decision-making process of black-box models, we focus on human-AI collaboration that leverage XAI as a collaboration mechanism (Zschech et al., 2021). In this context, Doshi-Velez and Kim (2017) propose a taxonomy of evaluation approaches for interpretability. Their results emphasize the need for the rigorous empirical evaluation of XAI algorithms. Although the body of literature dedicated to the evaluation of these XAI approaches is steadily increasing, their utility in terms of CTP remains largely unexplored, as the research community has initially focused on the study of constructs such as system trust (Davis et al., 2020). To date there exists no structured literature review on relevant factors impacting CTP of human-AI collaboration. Therefore, in this paper, we conduct a structured literature review (SLR) on user studies analyzing the task performance of humans and AI separately as well as in the form of an HI system to answer the following research question:

RQ1: *What factors have been analyzed in user studies regarding the design of human-AI collaboration that impact CTP?*

As the identified factors depend on the human, the AI, and the task, we cluster them from a socio-technical perspective (Maedche et al., 2019). A subsequent in-depth analysis of each perspective reveals further factors that have not been taken into

consideration by existing user studies. Thus, we formulate the following second research question:

RQ2: *What factors have not been analyzed in user studies regarding the design of human-AI collaboration that impact CTP?*

To answer the second research question, we derive and discuss, based on the SLR, neglected but relevant factors of CTP. Additionally, we propose testable hypotheses future research needs to address to realize the full potential of human-AI collaboration.

The contributions of this paper are twofold: First, we collect the existing body of knowledge for human-AI collaboration and describe possible relevant factors of CTP. Second, we discuss yet neglected but relevant factors of CTP and formulate respective hypotheses for future work.

The remaining work is structured as follows: In the next section, we explain the conceptual foundations with regard to XAI and human-AI collaboration. Consecutively, we outline the methodology applied to conduct the SLR and present our findings in the results section. Following, we derive and discuss yet neglected factors that might be taken into consideration to achieve CTP. Lastly, the conclusion summarizes our work by stressing its importance for the IS research discipline.

3.2 Conceptual Foundations

In the following, we provide a short overview of the two key concepts addressed in our work—HI and explainable AI (XAI).

3.2.1 Hybrid Intelligence

Dellermann et al. (2019a, p. 640) define HI as “the ability to achieve complex goals by combining human and AI, thereby reaching superior results to those each of them could have accomplished separately, and continuously improve by learning from each other.”. We focus on the first part of the definition. On the one hand, humans can rely on their senses, perceptions, emotional intelligence, and social skills (Braga & Logan, 2017). On the other hand, AI excels at detecting patterns or calculating probabilities (Dellermann et al., 2019a). These complementary skill sets allow for superior performance in specific tasks through collaboration. For example, managers can use emotional intelligence to build relationships and motivate employees to work

for the company (Davenport & Kirby, 2016). In contrast, repetitive and monotonous work can be conducted by AI.

The goals of human-AI collaboration are manifold. Among them are increasing the effectiveness and efficiency of the outcome of a specific task (Dellermann et al., 2019a). In this work, we focus on task effectiveness with regard to CTP. We follow the definition of Liu et al. (2021) and define CTP as the performance of teams consisting of humans and AI with the goal of achieving superior performance than AI or humans could have accomplished alone. The performance can be measured by different metrics, depending on the particular task, e.g., accuracy, recall, or the f1-score. and define CTP as the performance of teams consisting of humans and AI with the goal of achieving superior performance than AI or humans could have accomplished alone. The performance can be measured by different metrics, depending on the particular task, e.g., accuracy, recall, or the F1 score.

To enable CTP, humans need insights into AI decision-making. An emerging research stream that enables interpretability of AI decisions is the field of XAI.

3.2.2 Explainable Artificial Intelligence

Explainability is a concept with a long tradition in the information system (IS) research community. With the rise of knowledge-based systems, expert systems, and intelligent agents in the 1980s and 1990s, the IS community laid the foundations for research on explainability (Meske et al., 2022). In this context, Gregor and Benbasat (1999) provide a comprehensive overview of explanations in IS research.

XAI encompasses a wide spectrum of algorithms. A comprehensive survey on many existing explanation techniques can be found in Burkart and Huber (2021). In general, they can be differentiated by their complexity, their scope and their level of dependency (Adadi & Berrada, 2018). Interpretability of a model directly depends on the complexity of the model. (Wanner et al., 2020) cluster different types of complexity in white-, grey-, and black-box models. They define white-box models as models with perfect transparency, such as linear regressions. These models do not need additional explainability techniques but are intrinsically explainable. Black-box models, on the other hand, tend to achieve higher performance but lack interpretability. Lastly, grey-box models are not intrinsically interpretable but are made interpretable with the help of additional explanation techniques. These techniques can be differentiated in terms of their scope, i.e., being global or local explanations. Global XAI techniques address holistic explanations of the models as

a whole. In contrast, local explanations function on an individual instance basis. Besides the scoop, XAI techniques can also be differentiated whether they are model agnostic, i.e., can be used with all kinds of models, or model specific.

3.3 Methodology

To answer our research questions, we conducted a structured literature review (SLR) based on the methodology outlined by vom Brocke et al. (2009). We developed our search string consisting of two main areas. The first was XAI, including relevant synonyms, such as “explainable AI” or “interpretability” combined with “artificial intelligence”. The second part comprised synonyms of behavioral experiments, e.g., “user study” or “user evaluation”. To find the synonyms, we initiated our SLR with an explorative search. The search string was iteratively extended resulting in the following final search string:

TITLE-ABS-KEY(“explainable artificial intelligence” OR XAI OR “explainable AI” OR ((interpretability OR explanation) AND (“artificial intelligence” OR ai OR “machine learning”))) AND (“human performance” OR “human accuracy” OR “user study” OR “empirical study” OR “online experiment” OR “human experiment” OR “behavioral experiment” OR “human evaluation” OR “user evaluation”)

Next, we selected an appropriate database. Our exploratory search revealed that relevant work is dispersed across multiple publishers, conferences, and journals. Thus, we chose the SCOPUS database, to ensure comprehensive coverage. Following that, we defined our inclusion criteria, i.e., articles that were in scope of this SLR. We included every article that (a) did conduct empirical research, (b) did report performance measures and (c) did focus on an application context where humans and AI perform the same task.

With our search string defined, we conducted the SLR from January to March 2021. We identified 256 articles through the keyword-based search. As a next step, we analyzed the abstract of each article and filtered based on our inclusion criteria, leading to 61 articles. Afterwards, two independent researchers read all articles in detail and applied the inclusion criteria again. This led to a total of 14 remaining studies. Based on these, we conducted forward and backward search. With the forward and backward search, we identified 15 additional articles leading to a final set of 29 articles that were consequently analyzed in-depth to collect data about

each experiment. The increase in articles can be attributed to a large number of yet unpublished papers.

The data collection process was conducted by two independent researchers. Differences were discussed and corrected. The main focus of the SLR was to extract the treatments and outcomes of each experiment reported in the studies. For example, if two XAI techniques were used and compared as separate experimental treatments we added two entries into our database.

We clustered the extracted treatments from a socio-technical view, which is in line with other research (Buçinca et al., 2020). From a socio-technical view, human-AI collaboration can be divided in the following three key elements that are connected through a collaboration mechanism (Goodhue & Thompson, 1995) human(s) with a specific goal, the task that needs to be accomplished, and the technology—in our case the AI (Maedche et al., 2019). The collaboration mechanism enables teamwork between humans and AI regarding the task to be done. Figure 3.1 depicts the relationship between all relevant elements of human-AI collaboration.

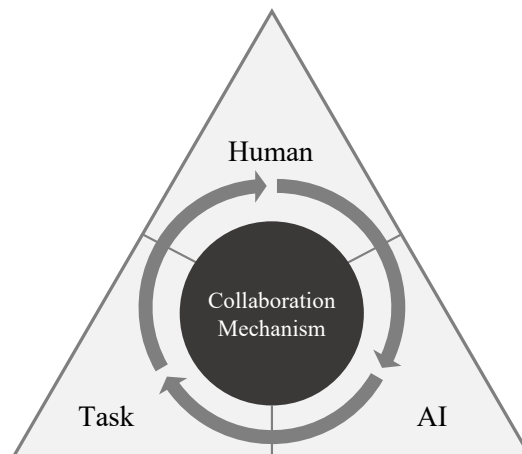


Figure 3.1.: Key elements of human-AI collaboration.

3.4 Results

In this section, we present the results of our SLR. The final data set consists of 29 articles in which 93 XAI-related experimental conditions are described. We start by providing an overview of the subset of articles that report CTP. Subsequently, we analyze the experimental conditions and cluster them according to our introduced socio-technical approach.

3.4.1 Overview

We extract, whenever possible, four different performance metrics from the articles—human, AI, AI-assisted and XAI-assisted performance. Human and AI performance refer to the performance achieved by humans or the AI when conducting the task individually. AI-assisted performance refers to the human performance when provided with the AI prediction. Finally, XAI-assisted refers to the performance of humans when provided with the AI’s recommendation as well as a supplementary explanation. Both, AI-assisted and XAI-assisted are measures of team performance. CTP is reached if this team performance exceeds both human and AI performance.

Figure 3.2 on page 50 displays the number of studies in which AI or XAI has a positive impact on HI-system performance considering different constraints. As not all studies report all four performance metrics, the following observations apply to different subsets of experiments. In general, 72 experiments measure the AI-assisted performance and XAI-assisted performance, but not necessarily human or AI performance. Of these, in 46 experiments XAI-assisted performance exceeds AI-assisted performance. Moreover, 59 out of 63 experiments report that providing either the AI’s prediction or a supplementary explanation (i.e., XAI) has a positive effect on human performance. In 53 experiments, all necessary information to analyze CTP are given—human and AI performance and either AI-assisted or XAI-assisted performance. Just 16 out of 53 experiments achieve CTP. Of these 16 experiments, five times, the XAI-assisted performance exceeds the AI-assisted performance. The 16 experiments are reported in two articles conducted by Bansal et al. (2019a) and Chu et al. (2020).

Bansal et al. (2019a) report 11 different experiments in which CTP is achieved meaning the team performance exceeds both the individual AI and human performance. All experiments focus on textual data. It is important to highlight that while CTP is reached, XAI does not yield a significant improvement over pure AI-assisted recommendations.

Further, in 5 experiments reported by Chu et al. (2020), CTP is achieved. The experiments focus on the task of predicting the age of a human based on facial images. In contrast to Bansal et al. (2019a), XAI has a significant effect on performance.

In addition to identifying existing studies that demonstrate CTP in human-AI collaboration, in the following, we provide an overview of all experimental conditions examined in the 93 studies. While most of these conditions have currently not led to CTP, they still have shown some effect on team performance or on relevant behavioral constructs, such as trust or cognitive load. We group the experimental

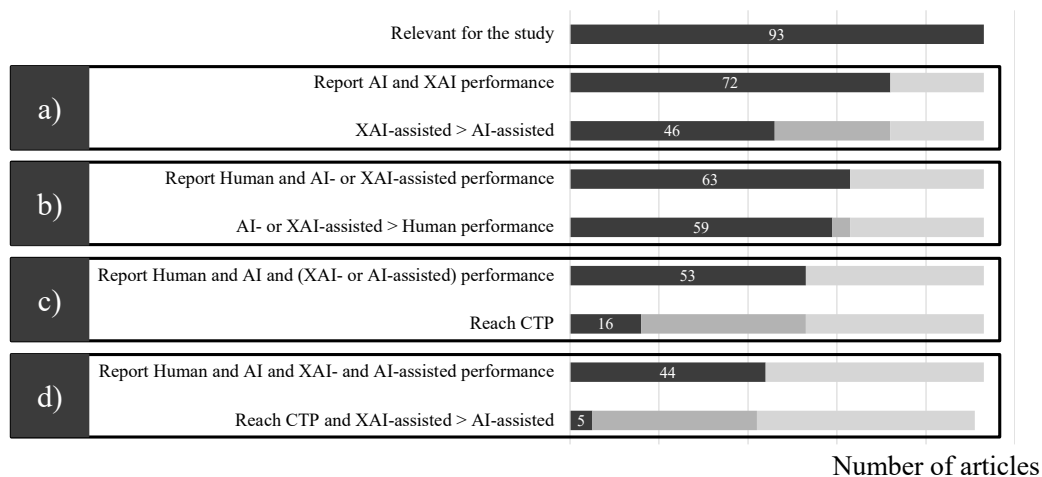


Figure 3.2.: Overview of the number of studies in which a) XAI-assisted performance exceeds AI-assisted performance, b) AI- or XAI-assisted performance exceeds human performance, c) CTP is achieved, and d) CTP is reached by using XAI.

conditions into four groups: collaboration characteristics, task characteristics, AI characteristics, and human characteristics (see Figure 3.1 on page 48).

3.4.2 Collaboration Characteristics

One important factor regarding the collaboration is the order in which the AI's predictions and explanations are made available to the human. Green and Chen (2019) find empirical evidence that asking participants to make a prediction before providing them the AI prediction or explanation leads to a better XAI-assisted performance. One possible reason might be that users are encouraged to invest cognitive capability in an active way instead of passively accepting the AI's suggestion (Green & Chen, 2019).

Another important factor is the interactivity in the human-AI collaboration. Liu et al. (2021) test this experimental condition and find an improvement in terms of human perception of AI-assistance. Bansal et al. (2021) implement adaptive explanations based on the confidence of the AI decision. If the AI prediction has a low confidence, more explanations are given. Their reasoning behind this approach is that a low confidence indicates a higher probability of a wrong AI prediction. By displaying more explanations, the human is encouraged to reflect more about the prediction.

Lastly, some researchers modify the degree of automation. One factor is whether the AI's prediction should be revealed at all (Lai & Tan, 2019). Lai et al. (2020) test

various XAI techniques without displaying the actual AI prediction. For example, they highlight all words that are relevant for the AI decision. Another condition is to colorize these highlights differently depending on the influence of the words. Their results show that differently colorized highlights result in a significant increase in XAI-assisted performance (70.7% accuracy for colorized highlights compared to 60.4% accuracy for human performance).

3.4.3 Task Characteristics

The tasks conducted in the studies play a decisive role in terms of individual and team performance. In general, in order to allow for generalizability, many studies utilize multiple tasks and data sets. In this context, Liu et al. (2021, p. 23) state that “[...] it is important to explore the diverse space and understand how the choice of tasks may induce different results in the emerging area of human-AI interaction”. For example, Alufaisan et al. (2021) study the influence of explanations on an income prediction and a recidivism task. In terms of concrete experimental conditions, our SLR shows that existing research focuses on skill sets and data types.

As discussed in the conceptual foundations, the skill set of AI and humans are complementary. Liu et al. (2021) find initial empirical evidence consistent with this line of thought by demonstrating enhanced team performance in correctly classifying out-of-distribution examples as the AI struggles to deal with instances that are beyond the patterns learned during training (D’Amour et al., 2022). However, it remains unclear how strong the distribution shift between in- and out-of-distribution data must be to discern a positive influence on CTP.

Concerning the data type, Lai et al. (2020, p. 744) state that “there may also exist significant variation between understanding text and interpreting images, because the former depends on culture and life experience, while the latter relies on basic visual cognition.”. The experimental studies analyzed in this SLR deal with human-AI tasks that are performed on either image ($n = 24$), tabular ($n = 28$), text ($n = 39$), or video ($n = 2$) data. In total, the analysis reveals the existence of only one study achieving CTP on textual data (Bansal et al., 2019a) and one on image data (Chu et al., 2020), respectively. Regarding tabular data, we are not aware of any study demonstrating comparable results. Hase and Bansal (2020) explicitly compare tabular and textual data. In this context, they observe that users rate explanations from tabular data higher than from textual data. Moreover, they find that displaying feature importance improves team performance on tabular data and that displaying examples helps for both data types.

3.4.4 Artificial Intelligence Characteristics

In addition to the first two components, the AI plays a central role in the human-AI collaboration. It can be divided into two different components—backend and frontend, i.e., the AI and XAI techniques and how explanations and predictions are displayed to humans.

Generally, most studies do not vary their AI techniques. An exception is the work of Lai et al. (2020) who test the influence of AI complexity on the XAI results. Their results show that explanations from simple models lead to better XAI-assisted performance.

In general, the XAI techniques vary from simple white-box models (Poursabzi-Sangdeh et al., 2021) over confidence scores (Zhang et al., 2020) to complex counterfactual explanations (Liu et al., 2021). As specific experimental conditions mostly confidence scores, feature importance, examples, and rules are used. Hereby, AI confidence refers to a probabilistic value that quantifies the certainty of the model with respect to a specific prediction (Zhang et al., 2020). Feature importance quantifies the contribution of each input feature to the prediction, and example-based explanations select and display most similar instances of a knowledge base (Adadi & Berrada, 2018). Lastly, rules refer to explanations based on if-then statements that are either extracted from more complex models or directly generated by the model (Van der Waa et al., 2021).

Our SLR highlights that the included studies report contradicting results with regards to the effect of these XAI techniques. For example, while Bansal et al. (2019a), Lai and Tan (2019), and Zhang et al. (2020), find that communicating the confidence of an AI's prediction is more effective than providing the importance of individual features, Chandrasekaran et al. (2017) report exactly the opposite. This indicates that focusing purely on XAI conditions might not be sufficient to extract generalizable and valuable insights. Table 3.1 on page 53 highlights all XAI-technique-related experimental conditions identified in the SLR. We want to emphasize that these results should not be interpreted quantitatively, since a lot of control variables and even measurements differ between the experiments.

In addition to the actual XAI technical category there are also differences in terms of the concrete implementations of the XAI algorithm. For example, Schmidt and Biessmann (2019) as well as Chandrasekaran et al. (2017) compare different technical implementations of feature importance algorithms. In particular, Schmidt and Biessmann (2019) compare LIME Ribeiro et al. (2016a) and covariance-based explanations. They find significant XAI-assisted performance differences (81.72%

Table 3.1.: Experimental conditions of XAI and their comparisons in existing studies.

Comparison of different XAI-assisted decision performances	References
Feature Importance outperforms Confidence	Chandrasekaran et al. (2017)
Feature Importance outperforms Examples	Lai and Tan (2019)
Feature Importance outperforms Prototypes	Hase and Bansal (2020) and Yeung et al. (2020)
Feature Importance outperforms Rules	Hase and Bansal (2020)
Confidence outperforms Feature Importance	Bansal et al. (2021), Lai and Tan (2019), and Zhang et al. (2020)
Confidence outperforms Examples	Lai and Tan (2019)
Confidence outperforms Prototypes	None
Confidence outperforms Rules	None
Examples outperform Feature Importance	Adhikari et al. (2019)
Examples outperform Confidence	None
Examples outperform Prototypes	None
Examples outperform Rules	Van der Waa et al. (2021)
Prototypes outperform Feature Importance	Hase and Bansal (2020)
Prototypes outperform Confidence	None
Prototypes outperform Examples	None
Prototypes outperform Rules	Hase and Bansal (2020)
Rules outperform Feature Importance	Hase and Bansal (2020) and Ribeiro et al. (2016a)
Rules outperform Confidence	None
Rules outperform Examples	None
Rules outperform Prototypes	Hase and Bansal (2020)

for LIME and 84.52% for covariance) indicating that the size of the improvement also depends on the selection of the specific XAI algorithm.

Besides these rather algorithmic experimental conditions our SLR reveals also more frontend-specific conditions. Lai and Tan (2019) show that the way the performance of the AI is displayed has an impact on team performance. Their results illustrate that independent of the AI's performance, sharing information about the performance increases trust. However, lower displayed performance relatively decreased XAI-assisted performance and trust. Carton et al. (2020) test the influence of the information amount communicated through the explanation. They evaluate the performance of detecting misclassification of online toxicity with full and sparse explanations. Full explanations highlight all words that have an influence on the prediction. Sparse explanations highlight just the most important words. This condition analyzes whether explanations need sufficient or comprehensive information. Their results indicate that there is no significant difference between both conditions (52.4% for full and 52.6% accuracy for sparse explanations).

3.4.5 Human Characteristics

Lastly, human characteristics influence the effectiveness of human-AI collaboration. One human-specific factor considered in the experiments is the human knowledge of the XAI techniques. Lai and Tan (2019) examine the usage of human-centered tutorials to build up essential knowledge and report a positive effect.

Another experimental condition tested is the self-assessment capability. Green and Chen (2019) explicitly test this experimental condition by asking participants how confident they are in their own decisions on a 5-point Likert scale. In their study they find that participants can not determine their own or the model's accuracy and fail to calibrate their use of AI.

3.5 Research Implications for Human-AI Collaboration

The SLR revealed relevant factors impacting the collaboration of human and AI. However, due to the small number of studies achieving CTP, it becomes evident that further research is needed on the design of human-AI collaboration. Therefore, based on the findings of the previous section, we derive and discuss possible factors beyond those tested in existing studies, which should be taken into consideration when studying CTP. Again, we structure this discussion from a socio-technical perspective. Finally, for each identified characteristic with a possible effect on CTP, we formulate multiple testable hypotheses with CTP as a dependent variable that future research should address.

3.5.1 Collaboration Characteristics

In the collaboration scenarios analyzed in existing studies, AI typically assists the human in form of recommendations. However, this must not lead to the human any longer questioning the AI's prediction. Against this background, it has been demonstrated that it is beneficial to actively involve the human in the decision-making process—either through encouragement by asking the human to make an informed decision before receiving the AI's prediction (Green & Chen, 2019), or by having the possibility to dynamically interact with the system (Liu et al., 2021). Following Lai and Tan (2019), who find a positive effect of human training for the interpretation of static explanations, we hypothesize a similar effect when humans are being trained to dynamically interact with the AI.

H1: *Training humans to dynamically interact with the AI and interpret its recommendations has a positive effect on CTP.*

In addition to training for dynamic interaction and interpretation, it could also be beneficial to visualize the AI's error boundary which highlights for each input if the model output is the correct action for that input feature combination—potentially enabling the human to predict when the AI will err and decide when to override the prediction. An improved understanding of the AI's error boundary in turn might positively contribute to CTP (Hase & Bansal, 2020).

H2: *Visualizing the AI's error boundary depending on provided input features has a positive effect on CTP.*

Moreover, to balance the amount of required cognitive load invested by the human, we suggest that it could also be viable to let the human decide case by case whether the recommendation or explanation should be revealed.

H3: *Allowing the human to decide whether the AI's prediction should be revealed has a positive effect on CTP.*

One aspect all studies considered have in common is the fact that the responsibility of the final team decision is with the human. A new perspective on collaboration could be that the AI distributes a priori who will have the final responsibility of the team decision depending on who has the higher expected probability of correctly executing the task taking the individual strengths of both team members into consideration (Mozannar & Sontag, 2020; Wilder et al., 2020).

H4: *Assigning the final decision dynamically to either the human or the AI has a positive effect on CTP.*

3.5.2 Task Characteristics

Even though existing literature discusses various facets of task characteristics, we see multiple directions for future research in terms of task complexity as well as performance differences between humans and AI.

In terms of task difficulty, previous studies found humans tend to rely more on heuristics the harder a task becomes (Goddard et al., 2014). The presence of AI support within human-AI collaboration might create the risk that people will rely more strongly on the proposed AI recommendation as the difficulty increases (Xu

et al., 2007). In this context, we hypothesize that increasing task difficulty for the human might be counterproductive in terms of CTP.

H5: *Increasing task difficulty has a negative effect on CTP.*

When working together on the same task, the absolute performance difference of humans and AI might play a significant role. There are situations in which the AI outperforms humans (Alufaisan et al., 2021) and vice versa (Chu et al., 2020). In this context, a very low performance of the AI, independent of the human performance, could result in algorithmic aversion (Manzey et al., 2012). Contrary, a very high performing AI could induce over-reliance in human action (Skitka et al., 2000). Following this line of reasoning, one might assume that the potential for CTP might be leveraged when humans and AI have comparable performance. In this context, Bansal et al. (2019a) hypothesize that a similar performance level of human and AI may contribute to achieving a significant effect on CTP. However, a small performance gap alone might not be sufficient for achieving CTP. In this context, we want to highlight that comparable performance does not inherently increase the probability of achieving the threshold to CTP. The important point is not the comparable performance but no positive correlation of human and AI errors. We hypothesize that even within the same task no positive correlation of human and AI errors contributes to reaching CTP.

H6: *No positive correlation of human and AI errors has a positive effect on CTP.*

3.5.3 Artificial Intelligence Characteristics

An important component to ensure the collaboration between human and AI is the communication capability of the AI. In this context, the frontend serves as a means for communicating the AI's prediction including explanations to the human. Its design has a significant influence on how well the user will interpret and use the recommendations derived by the AI. In general, it is crucial to balance the amount of information in order to prevent information overload (Bederson & Shneiderman, 2003). For instance, Klapp (1986) states that high volumes of information can have the same effect as noise, distraction or stress resulting in erroneous judgement. Besides the amount of information, the visualization quality also plays a significant role. For this reason, the question emerges whether a more user-centered design regarding the information presentation resulting from existing explanation algorithms could result in better human understanding. In this context, Suresh et al. (2021) propose a framework for characterizing the stakeholders of interpretable AI

including their needs. According to them not only the knowledge of humans but also the context in which human-AI interaction occurs plays a decisive role. For this reason, in-line with Kühl et al. (2020), we hypothesize that tailoring the information presentation particularly to the user while considering its knowledge and application context yields potential for improved CTP.

H7: *Personalized information presentation considering the humans' knowledge and the application context has a positive effect on CTP.*

Besides the information presented to the user, the accuracy of the inferred explanations plays a decisive role. Researchers have proposed evaluations to assess the performance of explanations, which is also known as fidelity of explanations (Shen & Huang, 2020). It can be interpreted as the capability of the explanation to reflect the AI's behavior (Alvarez Melis & Jaakkola, 2018). For example, Papenmeier et al. (2019) find that humans could lose trust in the AI when exposed to low fidelity explanations. In this context, Shen and Huang (2020, p. 172) mention “the representational power—including the correctness, sensitivity, etc., of the interpretation model—might not be sufficient to augment human reasoning about errors.”. Therefore, we hypothesize a strong positive correlation between the fidelity of explanations and the ability of humans to detect when the AI errs.

H8: *Increasing explanation fidelity has a positive effect on CTP.*

Even if established explanation techniques suffice the fidelity criterion, a further notable aspect should be their robustness. In this context, a large body of research found that explanations can vary significantly even for instances that are nearly identical and have the same classification Alvarez Melis and Jaakkola (2018), Ghorbani et al. (2019), and Tomsett et al. (2020) test the consistency of saliency maps and state their statistical unreliability. For this reason, we formulate the following hypothesis:

H9: *Increasing explanation robustness has a positive effect on CTP.*

3.5.4 Human Characteristics

Intensive research has been conducted across multiple disciplines over the past decades on characteristics that allow individuals to succeed in team settings (Zhao & Feng, 2019). For example, Morgeson et al. (2005) emphasize the importance of social skills, personality characteristics as well as team knowledge in the context of team member selection. Similar to human teams, we suggest that individual human

characteristics also play a crucial role in human-AI collaboration. We focus on a small subset of characteristics that have been examined in the context of human teams, which we believe will also have a major impact on human-AI collaboration. In this context, multiple meta-analytic studies found an influence of personality characteristics such as conscientiousness, agreeableness, or emotional stability (Hogan & Holland, 2003; Hertz & Donovan, 2000). Even though human-AI teams differ significantly from human teams, we assume these characteristics to impact CTP for the following reasons (Riefle & Benz, 2021). Conscientious people tend to demonstrate willingness to contribute to team performance regardless of their designated role (Barrick et al., 1998; Neuman & Wright, 1999). Moreover, they stand out by being especially concerned with performing their required behaviors towards achieving defined team goals (LePine et al., 1997). Furthermore, the characteristic of agreeableness encompasses traits such as cooperativeness and flexibility (John M., 1990). We hypothesize that these traits might be beneficial in the human-AI setting as well, since those individuals might be more willing to contemplate the AI's opinion. In addition, emotional stability might play a decisive role in this context as people with this trait tend to be more stress-resistant allowing them to more sovereignly manage demanding and ambiguous situations (Barrick et al., 1998). For the reasons mentioned above, we formulate the following hypotheses:

H10: *Personality characteristics (e.g., conscientiousness, agreeableness, emotional stability) have an effect on CTP.*

A further interesting direction of future research might be the influence of human cognitive capacity on CTP. In addition to various studies that find empirical evidence for cognitive ability being a strong predictor of individual performance (Hunter & Hunter, 1984; Wagner, 1997), a similar relationship can be shown at the team level (Devine & Philips, 2001). Thus, we suspect a similar relationship with regard to human-AI teams in the context of human-AI collaboration.

H11: *Cognitive ability has a positive effect on CTP.*

Lastly, human decision-making is heavily influenced by human biases (Kahneman, 2011). Human-AI collaboration is not spared of these biases. A particularly serious bias is the automation bias, i.e., “the tendency to use automated cues as a heuristic replacement for vigilant information seeking and processing” (Skitka et al., 2000, p. 344) that may lead to an overreliance on AI recommendations. Therefore, we formulate the following hypothesis:

H11: *Automation bias has a negative effect on CTP.*

3.6 Conclusion

The main goal of this study was to determine the current state of human-AI collaboration with regard to CTP. Therefore, we conducted a SLR. Subsequently, we provided an overview of the proportion of articles that reached CTP and presented experimental conditions that were tested in the articles. Based on the SLR and supplementary work we derived and discussed further experimental conditions and formulated testable hypotheses.

Unleashing the potential of human-AI collaboration that leverage the complementary capabilities of humans and AI to achieve CTP requires a multidimensional design process. For this reason, we see IS as the predestined research discipline to advance research in this field. We hope to motivate IS researchers and practitioners to actively participate in the exploration of understanding of the factors contributing to the design of human-AI collaboration for CTP.

Considerably more work needs to be done to determine the underlying patterns of effective Human-AI collaboration. Therefore, rigorous research models based on behavioral constructs need to be developed. Constructs such as mental model, cognitive load, and trust need to be measured to understand and enable CTP. Future work needs to address the testable hypotheses outlined in this work in behavioral experiments. We invite researchers to support and join us on the path to CTP.

A Meta-Analysis of the Impact of Explainable AI on Decision Performance

This chapter comprises an article that was published as: Schemmer, M., Hemmer, P., Nitsche, M., Kühl, N., & Vössing, M. (2022b). A Meta-Analysis on the Utility of Explainable Artificial Intelligence in Human-AI Decision-Making. *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, 617–626. Note: To improve the structure of the work, the title was changed. The abstract has been removed. Tables and figures were reformatted, and newly referenced to fit the structure of the thesis. The terminology was standardized with the dissertation. Chapter, section and research question numbering and respective cross-references were modified. Formatting and reference style was adapted and references were integrated into the overall references section of this thesis.

4.1 Introduction

Over the last years, the rapid developments in artificial intelligence (AI) have increased its use in many application domains. In this context, AI's continuously rising capabilities have surpassed human performance in an increasing number of tasks, such as playing poker (Brown & Sandholm, 2019), go (Silver et al., 2018), or correctly recognizing various categories of interest in images (He et al., 2015). Due to these remarkable developments, AI is increasingly applied to support decision-makers in an increasing number of domains, such as medicine (McKinney et al., 2020; Wu et al., 2020), finance (Day et al., 2018), law (Kleinberg et al., 2018) or manufacturing (Stauder & Kühl, 2022).

To offer decision-makers meaningful support, AI models are expected to provide accurate predictions and a notion of how a particular decision has been derived. In particular, explaining the rationale behind an algorithmic decision should enable domain experts to learn when to trust the recommendations of the AI and when to

question it (Zhang et al., 2020). This requirement fueled the continuous development of explainability techniques from the field of explainable AI (XAI), intending to make the decision-making process of black-box AI models more transparent and, thus, comprehensible for domain experts (Adadi & Berrada, 2018). Common approaches include among others feature importance-based (Ribeiro et al., 2016a), example-based (Cai et al., 2019), or rule-based methods (Ribeiro et al., 2018). A better understanding of how the AI's decision was derived should subsequently enable the user to appropriately rely on the AI's suggestions on a case-by-case basis (Bansal et al., 2021; Lee & See, 2004). For instance, explanations contradicting the AI's prediction could signify the user to become skeptical, consequently considering the AI prediction less in the final decision-making process.

With the ongoing development of XAI techniques, researchers have started to evaluate AI with and without explanations to assess whether their utility for better decision-making can be quantified (Bansal et al., 2021; Buçinca et al., 2020; Buçinca et al., 2021; Carton et al., 2020; Chu et al., 2020; Green & Chen, 2019; Hase & Bansal, 2020; Hemmer et al., 2022b; Lai et al., 2020; Lai & Tan, 2019; Liu et al., 2021; Van der Waa et al., 2021; Yeung et al., 2020; Zhang et al., 2020). Whereas some researchers identify a benefit of XAI-based decision support in user studies (Buçinca et al., 2020; Lai et al., 2020), others find only negligible evidence (Carton et al., 2020; Liu et al., 2021), with the underlying causes remaining partly unexplored (Hemmer et al., 2021; Schoeffler et al., 2022b). Therefore, in this article, we aim to clarify the current “snapshot” of the utility of XAI-based decision support. We conduct a meta-analysis of user studies identified in a structured literature review to shed light on the effect of XAI-assisted decision-making on user performance. In detail, our analysis encompasses studies that allow a comparison between human, AI-, and XAI-assisted task performance. Our initial findings are the following: First, on average, XAI-assisted decision-making enhances human task performance compared to no assistance at all. However, we find no additional effect of explanations on users' performance in XAI-assisted decision-making compared to isolated AI predictions, which raises questions on how to further develop current XAI methods that improve users' task performance. Second, we find that distinct data types affect user performance differently. In this context, human-AI collaboration turns out to be more effective on text data compared to tabular data.

The remainder of this article is structured as follows. In Section 4.2, we first outline related work in the context of human-AI collaboration. In Section 4.3, we describe the methodological approach of our meta-study. Subsequently, we present the results of the meta-study, including a subgroup analysis in Section 4.4. In this context, we provide additional qualitative insights on an individual level. We outline

the current limitations of our work in Section 4.5, followed by a discussion on relevant implications that result from these findings for the future development of XAI algorithms in Section 4.6. Finally, Section 4.7 concludes our work.

4.2 Related Work

Over the last years, research has focused on developing algorithms that provide explanations for AI predictions (Adadi & Berrada, 2018; Das & Rad, 2020). By now, these algorithms are increasingly employed in a growing number of practical use cases such as in manufacturing (Senoner et al., 2022; Treiss et al., 2020), medicine (Pennisi et al., 2021), or the hospitality industry (Vössing et al., 2022). Usually, XAI is utilized in scenarios that involve humans-in-the-loop processes. The underlying idea is that humans will benefit from the AI's suggestion if it is accompanied by an explanation. Therefore, a constantly rising number of studies has started to analyze the effects of explanations in behavioral experiments (Hemmer et al., 2021). In these experiments, many different target variables are taken into consideration, e.g., whether humans are capable of predicting what a model would recommend (proxy tasks) (Arjun et al., 2018; Buçinca et al., 2020; Hase & Bansal, 2020) or whether explanations support them in model debugging (Adebayo et al., 2020; Kaur et al., 2020).

In the scope of this study, we explicitly focus on AI-assisted decision-making—a setting in which an AI supports a human with the goal of improving the decision-making quality. The AI's prediction might be accompanied by additional information, e.g., about its prediction uncertainty or different types of explanations. After receiving the AI's advice, the human decision-maker is responsible for making the final decision—a scenario which is often also required from a legal perspective (Bauer et al., 2021). By providing either additional information on the AI's prediction uncertainty (Nguyen et al., 2021; Zhang et al., 2020) or explanations on how a decision was derived (Bansal et al., 2021; Lai & Tan, 2019), humans shall be enabled to better question the AI's decision. To develop a deeper understanding of this assumption, research has evaluated the effect of explanations on users' trust and how reliance on AI decisions can be appropriately calibrated (Bansal et al., 2019b; Buçinca et al., 2021; Kunkel et al., 2019; Schemmer et al., 2022c; Yu et al., 2019; Zhang et al., 2020). In this context, providing humans not only with the AI's prediction and respective explanations but also with a notion about its global performance can influence the overall team performance (Lai & Tan, 2019). Additional benefits can be found when humans are provided with model-driven tutorials about AI functionality and the task

itself (Lai et al., 2020). Further work has investigated the influence of AI advice in the out-of-distribution setting—instances differing from the distribution used for AI training—on the final human decision (Liu et al., 2021).

Besides these factors, the explanation type of an AI prediction can play a decisive role. In this context, research has developed various explainability techniques (Adadi & Berrada, 2018) ranging from feature importance methods (Ribeiro et al., 2016a) over example-based approaches (Cai et al., 2019) to rule-based explanations (Ribeiro et al., 2018) that have been evaluated in user studies accordingly (Hemmer et al., 2021). However, the current picture emerging from the results of different studies regarding the effects of XAI methods on AI-assisted decision-making performance is not unambiguous. For example, whereas Carton et al. (2020) conclude that feature-based explanations do not help users in classification tasks, Hase and Bansal (2020) find some of them to be effective in model simulatability, which refers to the ability to predict the model behavior given an input and an explanation. In this context, further studies demonstrate the utility of explanations (Buçinca et al., 2020), whereas others find that they can convince humans to follow incorrect suggestions more easily (Bansal et al., 2021; Van der Waa et al., 2021).

Of course, ambiguous findings can also be attributed to the specific setups of each study and the different goals pursued by the researchers. We aim to shed light on this ambiguity by conducting a meta-analysis of human-AI collaboration—particularly on the influence of explainability.

4.3 Methodology

We elaborate on our data collection approach to identify relevant articles, followed by the statistical analysis conducted on the final set of user studies.

4.3.1 Data Collection

For the collection of empirical user studies in the field of XAI, we conducted a structured literature review based on the methodology outlined by vom Brocke et al. (2009). In detail, we developed a search string focusing on XAI and behavioral experiments. For both topics, several synonyms were included after an explorative search. Subsequently, the search string was iteratively refined, resulting in the following final search string:

TITLE-ABS-KEY("explainable artificial intelligence" OR XAI OR "explainable AI" OR ((interpretability OR explanation) AND ("artificial intelligence" OR AI OR "machine learning"))) AND ("human performance" OR "human accuracy" OR "user study" OR "empirical study" OR "online experiment" OR "human experiment" OR "behavioral experiment" OR "human evaluation" OR "user evaluation")

To ensure comprehensive coverage of relevant articles, we chose the SCOPUS database for our initial search (Schotten et al., 2017). We filtered identified articles according to the following three criteria: an article identified with the search string was included if it (a) conducted at least one empirical user study and (b) reported the task performance as a performance measure for humans and AI- or XAI-assisted decision-making on the same task.

Additionally, we conducted a forward and backward search starting from the articles that fulfill our inclusion criteria. We extracted all individual treatments and outcomes for each article. For instance, if an experiment compared AI- with XAI-assisted decision-making in a between-subject design in two separate treatments, each of them was registered as a separate record in our database. If an article includes multiple experiments, we performed the data extraction process for each experiment separately. We contacted authors by email in case of missing or not reported information in the articles regarding the conducted user studies.

The collected studies vary considerably in terms of tasks, problem settings, and reported performance metrics. Accordingly, we filtered our set of studies in the following way: First, we focused on studies assessing classification tasks as they account for the largest subset across all entries in our database. Second, we restricted the subset of relevant studies to those reporting the mean accuracy as the performance measurement in each study since we require a common metric across multiple studies. This ensures that we base our meta-analysis on comparable and interpretable effect sizes. Third, we only included studies that have been conducted as a between-subject design. By excluding studies conducted in a within-subject design, we avoid taking into account the learning effect of participants between treatments that might distort the effect sizes of our analysis.

Subsequently, we extracted all necessary performance metrics from the articles. We define the case in which the human performs the task without any AI support as human performance. If the human is additionally equipped with AI advice, but without explanations, we call the performance AI-assisted performance. AI assistance with explanations is called XAI assistance, and the resulting performance measure is denoted as XAI-assisted performance. Based on these definitions, we excluded all

studies that do not report human performance and either AI-assisted or XAI-assisted performance. Based on the resulting sample, we conducted the following statistical analysis.

4.3.2 Statistical Analysis

For each study, we calculate the effect size as the between-group standardized mean difference (SMD) of the task performance. Furthermore, we report Hedges' g (Hedges & Olkin, 1985) to correct the SMD for a possible upward bias of the effect size when the sample size of a study is small ($n \leq 20$). Thus, Hedges' g is smaller for $n \leq 20$ than the uncorrected SMD but approximately the same for larger sample sizes. We obtain the standard deviations from standard errors and confidence intervals for group means reported for each treatment following the procedure outlined in Higgins et al. (2019). In case an article encompasses multiple studies with a single control group, we divide the size of this control group by the number of studies to avoid multiple comparisons against the same group (Higgins et al., 2019).

For our meta-analytic model and the pooling of effect sizes, we estimate a random-effects model as the setups and populations are considerably heterogeneous between studies. Hence, we calculate the distribution mean of effect sizes instead of estimating and assuming one single true effect size underlying the studies (fixed-effect model (Borenstein et al., 2010)). To assess the between-study heterogeneity variance τ^2 and its confidence intervals we use the DerSimonian-Laird estimator (DerSimonian & Laird, 1986) and Jackson's method (Jackson & Bowden, 2016), respectively.

Additionally, we conduct a subgroup analysis to provide further insights across current XAI studies. Several studies have discussed that task choice has a strong influence on the experimental outcome (Fügener et al., 2021; Lai et al., 2023). This article focuses on the influence of the task's data type. In this context, many researchers have argued about the importance of data types in human-AI collaboration. For example, Fügener et al. (2021) reason that image recognition, in general, is well suited for human-AI collaboration since it is an intuitive task for humans.

4.4 Results

We start by presenting the final set of included studies, then outline the meta-study results, including the respective subgroup analyses. Finally, we provide additional qualitative insights on an article level.

4.4.1 Data Collection

As of February 2022, we identified a total number of 393 articles. After applying our inclusion criteria and conducting a forward and backward search, the number of relevant articles is reduced to 33.

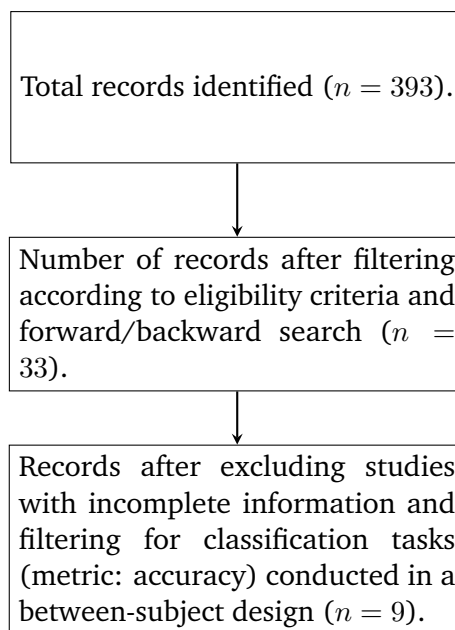


Figure 4.1.: Flowchart describing the data collection and article selection procedure.

As classification tasks form the largest subset, we focus on this particular prediction problem. After filtering for accuracy as a common metric and removing articles with missing information, e.g., sample size or dispersion measures, we include 9 articles in the meta-analysis and the respective subgroup analyses. Figure 4.1 visualizes the entire filtering process. Moreover, Table 4.1 on page 68 provides an overview of all included articles together with information about each dataset, datatype, and treatments extracted from the articles. Each article contains at least one behavioral experiment conducted with a particular dataset. Each experiment consists of several experimental treatments. The treatment in which humans conducted a task on their

Table 4.1.: Overview of articles that were identified in the structured literature review and are analyzed in this work. All articles are peer-reviewed at the time of the meta-study.

Source	Dataset	Datatype	Studies
Alufaisan et al. (2021)	COMPAS (ProPublica, 2016) Census (Dua & Graff, 2019)	Tabular	AI-, XAI-assisted (Anchor) AI-, XAI-assisted (Anchor)
Bansal et al. (2021)	LSAT (Team, 2017) Book reviews (He & McAuley, 2016) Beer reviews (McAuley et al., 2012)	Text	XAI-assisted (Confidence, Explain Top-1 Expert, Explain Top-2 Expert, Adaptive Expert) XAI-assisted (Confidence, Explain Top-1 AI, Explain Top-2 AI, Adaptive AI, Adaptive Expert) XAI-assisted (Confidence, Explain Top-1 AI, Explain Top-2 AI, Adaptive AI, Adaptive Expert)
Buçinca et al. (2020)	Fat content prediction (Buçinca et al., 2020)	Image	AI-, XAI-assisted (Inductive & Deductive Explanation)
Carton et al. (2020)	Online toxicity (Wulezyn et al., 2017)	Text	AI-, XAI-assisted (Keyword, Partial, Full Explanation)
Figgener et al. (2021)	Dog breed ImageNet (Russakovsky et al., 2015a)	Image	AI-, XAI-assisted (Certainty)
Lai et al. (2022)	Deception detection (Ott et al., 2013; Ott et al., 2011)	Text	XAI-assisted (Signed & Predicted Label, Signed & Predicted Label & Guidelines, Signed & Predicted Label & Guidelines & Accuracy)
Liu et al. (2021)	COMPAS (ProPublica, 2016) ICPSR (United States Department of Justice, 2014) BIOS (De-Arteaga et al., 2019)	Tabular Text	XAI-assisted (Static/Static, Interactive/Static, Interactive/Interactive) XAI-assisted (Static/Static, Interactive/Static, Interactive/Interactive) XAI-assisted (Static/Static, Interactive/Static, Interactive/Interactive)
Van der Waa et al. (2021)	Diabetes mellitus type 1 (Van der Waa et al., 2021)	Tabular	AI-, XAI-assisted (Rule-based, Example-based)
Zhang et al. (2020)	Census (Dua & Graff, 2019)	Tabular	AI-, XAI-assisted (Confidence, Feature Importance)

own without AI assistance is referred to as a control group. In the following, we denote each treatment as an individual study. Overall, we thereby have a sample size of 44 studies.

4.4.2 AI Assistance Versus XAI Assistance

We start our meta-analysis by investigating AI- and XAI-assisted performance. For this reason, we first focus on all studies that report AI- and XAI-assisted performance, which leads us to a sample of 11 studies and a total number of 999 observations.

Figure 4.2 on page 70 displays the forest plot of the SMD between AI-assisted and XAI-assisted performance. The results of the analysis reveal that, on average, the SMD of all studies that reported AI- and XAI-assisted performance is 0.07 with a 95% confidence interval (CI) [-0.15, 0.30]. A z-test against the null-hypothesis that the effect size is 0 cannot be rejected ($z = 0.63, p = 0.53$). This means we do not find a significant difference between AI-assisted and XAI-assisted performance in our current sample of studies.

Regarding heterogeneity, we find an I^2 of 57.00% (95% CI [15.70%, 78.00%]), which can be considered moderate (Higgins et al., 2019). The τ^2 is 0.07 (95% CI [0.01, 0.54]) and Q is significantly different from 0 ($Q = 23.24, df = 10, p < 0.01$). Thus, we can reject the null hypothesis that the true effect size is identical in all studies. To provide an intuitive understanding of the heterogeneity, we also report the prediction interval that represents the expected range of true effects in other studies (IntHout et al., 2016). The prediction interval ranges from -0.59 to 0.74. That means we can expect negative as well as positive effects of XAI assistance in comparison to AI assistance. In summary, on average, XAI-assisted decision-making does not significantly influence the performance of human-AI collaboration in our sample. The highest improvement was measured by Van der Waa et al. (2021). Contrary, the highest negative impact of XAI is measured by Alufaisan et al. (2021).

4.4.3 Human Versus XAI Assistance

We analyze the overall effect of XAI in comparison with human performance. Therefore, we filter all studies that report human and XAI-assisted performance. This results in a sample of 33 studies and a total number of 5,083 participants. Based on this sample, we analyze whether XAI-assisted decision-making improves performance compared to humans conducting a task alone.

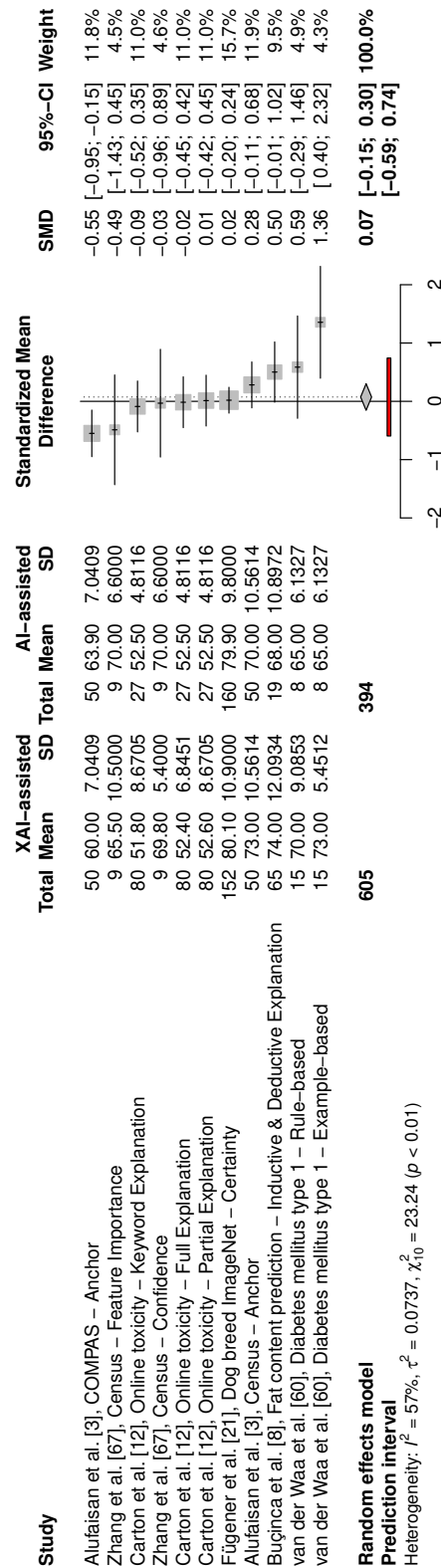


Figure 4.2.: Forest plot of the standardized mean difference between AI-assisted and XAI-assisted performance.

Figure 4.3 on page 72 visualizes the forest plot of the standardized mean difference between human and XAI-assisted performance. The meta-analysis indicates that, on average, XAI assistance increases task performance by 0.59 SMD as compared to humans conducting the tasks alone. The 95% CI of the SMD ranges from 0.39 to 0.79. As this range does not include an effect size of 0 and a z-test is significant ($z = 5.73, p < 0.0001$), we can reject the null hypothesis concluding that, on average, XAI-assisted decision-making improves human task performance.

Looking at heterogeneity, we can reject the null hypothesis that the true effect size is identical in all studies ($Q = 219.24, df = 32, p < 0.0001$). Moreover, I^2 is 85.40% with a 95% CI of 80.50% to 89.10%. The estimated τ^2 is 0.29 (95% CI [0.18, 0.59]). Thus, the level of heterogeneity can be considered substantial (Higgins et al., 2019). The prediction interval in the analysis is -0.54 to 1.71. This means that we cannot say with certainty that XAI always has a positive impact on human decision-making as the prediction interval is not exclusively larger than 0. Even though most studies report a performance improvement through XAI assistance, we also find studies that report a performance decline. In this context, one of the two studies conducted by Alufaisan et al. (2021) encountered the most negative effects with a human performance decline. Participants are asked to predict whether a defendant will recidivate in two years and receive AI predictions with decision rules aiming to support users' understanding of the AI's decisions. However, this reduction is not statistically significant due to a high level of dispersion. The most considerable improvement can be found in a study by Buçinca et al. (2020). Here, participants have to decide based on an image of a meal whether the fat content of this meal on a food plate exceeds a certain threshold.

It is important to highlight that the significant SMD does not imply that including explanations will improve performance over simply providing AI advice without any form of explainability, as we did not find a significant difference between AI-assisted and XAI-assisted performance in Section 4.4.2. Instead, it can be interpreted as a positive effect of *some* form of AI advice.

4.4.4 Tabular Versus Text Data

Additionally, we conduct a subgroup analysis based on three data types used in our sample—tabular, text, and image data. As only two articles report experiments using image data, the interpretation of this data type might not be conclusive. Thus, we focus on tabular and text data types for the subgroup analysis resulting in a total sample size of 31 studies and 4,702 participants. Figure 4.4 on page 74 displays the

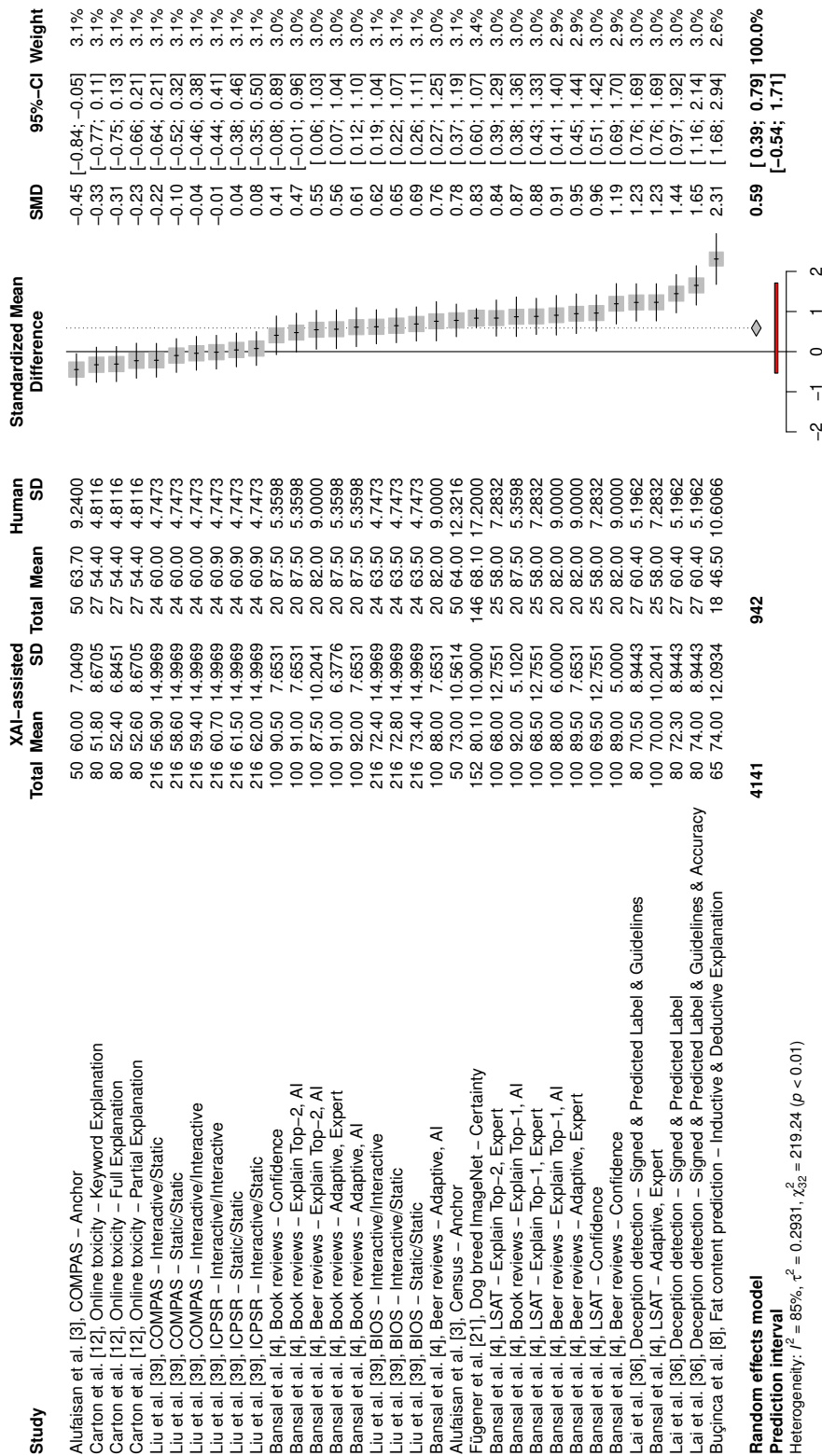


Figure 4.3.: Forest plot of the standardized mean difference between human and XAI-assisted performance.

forest plot of the standardized mean difference between human and XAI-assisted performance with regard to both data types.

The SMD of the tabular subgroup is 0.01 (95% CI [-0.24, 0.26]). As this range does include an effect size of 0 and the z-value is 0.08 with a corresponding p-value of 0.94, we cannot reject the null hypothesis. That means we cannot conclude that the SMD is significantly different from 0. Regarding heterogeneity, we can reject the null hypothesis that the true effect size is identical in all studies ($Q = 20.30, df = 7, p < 0.005$). The I^2 is 65.50% with a 95% CI ranging from 26.70% to 83.80% and τ^2 has a value of 0.09 (95% CI [0.01, 0.48]). Finally, the prediction interval ranges from -0.77 to 0.79. That means we can expect future negative impacts of XAI assistance on human decision performance in certain situations.

The text data subgroup has a higher SMD of 0.72 (95% CI [0.50, 0.93]). This range does not include an effect size of 0. Additionally, the z-value is 6.65 with $p < 0.0001$ denoting that the SMD is significantly different from 0. In terms of heterogeneity, we can reject the null hypothesis that the true effect size is identical in all studies ($Q = 103.57, df = 22, p < 0.0001$). In this context, τ^2 is 0.21 (95% CI [0.10, 0.47]) and I^2 is 78.80% (95% CI [68.70%; 85.60%]). Thus, the heterogeneity of the text data subgroup can be considered higher than the tabular data subgroup. The prediction interval ranges from -0.26 to 1.69, which means we can also expect some negative XAI effects with text data.

Lastly, comparing both subgroups, we observe significant performance differences between tabular and text data suggesting that the data type in our sample influences the effect of XAI assistance on the performance ($Q = 17.81, df = 1, p < 0.0001$).

4.4.5 Summary of the Articles

Having analyzed the collected studies in the form of a meta-analysis, we pursue a discussion and summarization of the individual articles from a qualitative perspective. We focus on extracting further insights that can be derived from comparing human performance without any assistance and AI or XAI assistance. In particular, we are interested in the question of why XAI assistance did or did not improve AI assistance.

Alufaisan et al. (2021) draw upon rule-based explanations generated by anchor LIME (Ribeiro et al., 2018) for two real-world tabular datasets—an income prediction task using the Census dataset (Dua & Graff, 2019) and a recidivism prediction task using the COMPAS dataset (ProPublica, 2016). The XAI algorithm provides rules

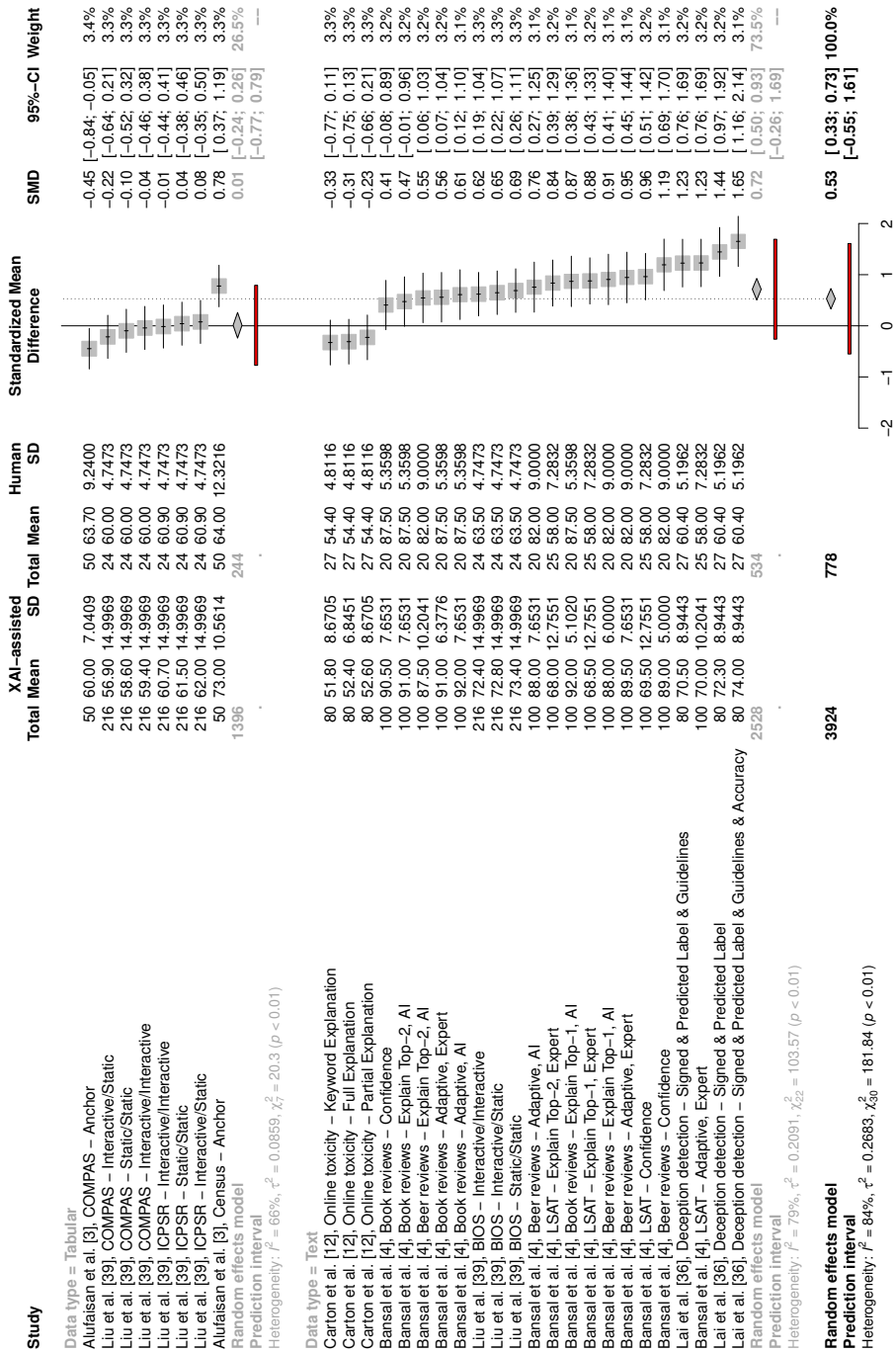


Figure 4.4.: Forest plot of the standardized mean difference between human and XAI-assisted performance considering the subgroups tabular and text data.

which are denoted as anchors to explain a prediction on an instance level. A rule is considered an anchor if any changes in the features that are not included in the anchor do not impact the AI decision (Alufaisan et al., 2021). The authors find that AI assistance improves human performance. However, they find no additional benefit of XAI assistance regarding decision-making performance. They hypothesize that one reason for the negligible effects of XAI could be information overload. When taking a closer look at the individual decision level, humans tend to follow AI predictions more often when the AI makes a correct prediction than when it makes an incorrect one. In this context, explanations did not alter this observation.

Bansal et al. (2021) compare multiple forms of explanations of AI predictions on three tasks. Two of them are about sentiment analysis of book (He & McAuley, 2016), and beer reviews (McAuley et al., 2012). The third task consists of a set of law school admission test questions requiring logical reasoning (Team, 2017). All three datasets consist of text data. As explanations, they use feature importance generated by LIME (Ribeiro et al., 2016a) and feature importance generated by human experts. Additionally, the authors evaluate the effect of providing only confidence ratings of the AI's prediction, which we consider as a form of explanation as well to ensure that AI assistance encompasses only the predictions of the AI in all studies. Therefore, we are restricted to the comparison of human performance with XAI assistance in the meta-analysis. Interestingly, the authors find no performance improvements of explanations over confidence ratings. However, they find that providing explanations tends to increase the chance that humans accept the AI's prediction regardless of its correctness—which shows contrary results to the observations of Alufaisan et al. (2021). Moreover, participants receiving XAI assistance outperform both participants and AI alone.

Buçinca et al. (2020) analyze two different techniques for evaluating XAI systems. First, a proxy task, e.g., asking humans to predict the AI's decision. Second, an actual decision-making task, e.g., asking humans to make a decision with XAI assistance. We consider the actual decision-making task for the meta-analysis. The authors show participants images of different food plates and ask them to predict the fat content of the meal. The participants are requested to decide whether the fat content in percent is higher than a threshold. Users are provided with inductive (i.e., example-based) and deductive (i.e., rule-based) explanations. With XAI assistance, the users significantly outperform users with AI assistance, which yields better performance than users conducting the task without any support. Interestingly, example-based explanations enable users to identify erroneous AI predictions better. However, they prefer and trust the rule-based explanations more.

Carton et al. (2020) consider a social media comment toxicity prediction task for their experiment by sampling comments from the dataset provided by Wulczyn et al. (2017). They utilize feature importance by highlighting passages and words pointing towards a toxic comment. In detail, three different forms of explanations are analyzed—full, partial, and keyword explanations. The full explanations are generated by an attention model (Carton et al., 2018) that produces a discrete attention mask over the input text for toxic content. Partial explanations refer to reducing the mask to the most toxic passage. Keyword explanations are derived from a bag-of-words logistic regression classifier that highlights only the most toxic single words instead of considering the context. Their study shows a marginal negative trend of AI assistance performance in comparison to users with no assistance at all. However, the effect does not vary significantly regardless of explanations being present or not. In this context, users tend to follow the AI prediction without an effect of the different explanations provided. A detailed analysis of false negative and false positive rates reveals that explanations tend to increase false negatives and reduce false positives. The authors hypothesize that this might indicate a reduced cognitive engagement with the social media comments by focusing on the highlighted text without considering the not highlighted passages.

Fügner et al. (2021) use a dog breed image classification task (Russakovsky et al., 2015a). They evaluate the effect of AI assistance with and without additional confidence ratings and compare both treatments with the performance of users that do not receive any AI support. While the authors find that AI assistance significantly improves human performance, no additional effect of providing confidence ratings can be identified. In this context, both AI-assisted performance and the one with additional confidence ratings outperform humans and AI when conducting the task alone. A detailed analysis of the confidence rating treatment reveals that providing AI certainty decreases AI adherence. The authors hypothesize that a decrease in users' trust might explain this effect as it also declines in this treatment.

Lai et al. (2020) consider a deception detection task which is about identifying fake hotel reviews (Ott et al., 2013; Ott et al., 2011). The authors analyze the effect of three types of XAI assistance provided together with the AI prediction. First, they provide users with signed feature importance, i.e., they highlight the most important words that indicate real and fake reviews. Word importance is derived from the absolute value of the coefficients using a linear SVM with a unigram bag-of-words. Second, they additionally provide the participants with supporting guidelines derived from related research and observations of the model, which are paraphrased by the authors. Third, they provide an additional AI accuracy performance statement on top. Before the actual decision-making task, participants had to undergo a task-

specific training phase. In a prior experiment, the authors demonstrated that these so-called tutorials enhance participants' performance without any further assistance from the AI. In general, they find an improvement in XAI-assisted performance over human performance in all three treatments. However, no statistical significance between the different forms of XAI assistance can be detected. Moreover, the authors also emphasize that XAI-assisted performance remains inferior to the AI performing the task alone, even though a small proportion of participants who were able to outperform the AI can be identified. The authors do not consider an AI-assisted treatment in their experiment.

Liu et al. (2021) explore the effect of out-of-distribution data instances and interactive explanations on human-AI collaboration. For the meta-analysis, we focus on the in-distribution setting, including a comparison of static with interactive explanations, as out-of-distribution data might impede the comparability with the other studies. The authors draw upon three tasks in their article. In the first, participants predict whether arrested defendants will violate the terms of pretrial release using the ICPSR dataset (United States Department of Justice, 2014). The second task is about predicting whether defendants will recidivate in two years using the COMPAS dataset (ProPublica, 2016). Both are tabular datasets. The third task requires participants to predict a person's profession given a textual biography using the BIOS dataset (De-Arteaga et al., 2019). Regarding the AI model, a linear SVM classifier with unigram bag-of-words for BIOS and one-hot encoded features for ICPSR and COMPAS is employed. The static explanations consist of feature importance by coloring features that contribute to the AI prediction. Interactive explanations offer users the possibility to explore what-if scenarios. In general, the authors find that XAI assistance improves users' performance in predicting a person's profession compared to users without any assistance. However, no significant performance difference between static and interactive explanations can be found. For the two recidivism prediction datasets, they find no significant difference between the performance of XAI assistance and human alone. The authors hypothesize that the complexity of both recidivism tasks might have prevented noticeable performance improvements. Even though no performance difference can be observed, users rate interactive explanations more useful in the recidivism prediction tasks. The authors do not consider sole AI-assisted decision-making in their article.

Van der Waa et al. (2021) evaluate the effect of example- and rule-based explanations in the context of a diabetes self-management task where participants are requested to select the appropriate dose of insulin. The authors compare AI assistance with both types of explanations but do not report sole human performance. In this context, the presence of either example- or rule-based explanations does not result

in a significant performance difference compared to pure AI assistance. A closer analysis of explanations' "persuasive power", i.e., how often humans agree with the AI recommendation regardless of being correct or not, reveals that users without explanations follow the AI prediction significantly less than with example- or rule-based explanations.

Zhang et al. (2020) compare the effect of additional confidence ratings with feature importance explanations using Shapley values (Lundberg & Lee, 2017) in the context of human-AI collaboration. They utilize the Census dataset (Dua & Graff, 2019) for asking participants to predict whether a person's income would exceed \$50,000. In the experiment, the authors find no significant difference between additional confidence ratings and feature importance explanations in terms of task performance. Moreover, both treatments do not differ significantly from sole AI assistance. Even though displaying confidence scores does not affect task performance, it can be found that it improves overall trust and contributes to a calibration over different confidence levels. The authors explain this phenomenon with a high correlation between human and AI confidence, showing a large overlap of instances with low AI and human confidence. In contrast, explanations do not affect users' trust in their experiment.

In summary, most studies, except for *Buçinca et al. (2020)*, found no significant differences when comparing XAI and AI assistance on an individual study level. *Alufaisan et al. (2021)* argue that the negligible effect of XAI might stem from information overload. *Carton et al. (2020)* discuss that XAI, in the case of feature importance, might reduce the amount of data humans process as they just focus on the features relevant to the AI. *Liu et al. (2021)* discuss whether task complexity could be a reason for no significant improvements of XAI assistance over AI assistance. We also observe some contrasting results that require further investigation in future work. *Fügener et al. (2021)* find a general decrease in AI adherence in the context of confidence ratings. In contrast, *Van der Waa et al. (2021)* and *Bansal et al. (2021)* find that explanations increase the general probability of accepting AI advice. Based on the qualitative review, we derive two potential factors influencing the utility of XAI: First, XAI could improve decision-making performance by increasing the acceptance of AI advice. Note that the performance improvement will just happen if the AI, on average, performs significantly better than the human (*Zhang et al., 2020*). Second, XAI assistance could influence appropriate trust and reliance, which means humans can discriminate between correct and incorrect AI advice (*Bansal et al., 2021*). The idea is that humans will be better able to distinguish between correct and incorrect advice if the AI conveys its reasoning. Our qualitative review showed that while there is evidence for increased acceptance of AI advice due to XAI

(Bansal et al., 2021; Van der Waa et al., 2021; Zhang et al., 2020), just one article reports some form of appropriate reliance (Buçinca et al., 2020).

4.5 Limitations

As XAI is a relatively new field of research, at least in comparison to research domains where meta-analyses are more common, e.g., medical research (Eaden et al., 2001), we encounter some major limitations that form around the current existing sample of XAI studies.

First, the current existing sample of XAI studies contains just online studies. In these online studies, people are recruited via online platforms such as Mechanical Turk. They conduct a task not in a controlled lab environment inducing higher variability. Second, the studies use different XAI algorithms ranging from providing an additional confidence score to personalized explanations. We also considered a subgroup analysis for the XAI algorithm category but were limited due to the current sample size. Therefore, interpretable findings could not be derived yet. Third, also task design differs between the studies. Some studies use more intuitive tasks for humans, such as sentiment analysis of reviews, while others consider more complicated ones, such as recidivism prediction. In this context, future studies should evaluate other task-related factors beyond data type. Fourth, in the data type subgroup, the tabular subgroup contains just two articles (Alufaisan et al., 2021; Liu et al., 2021). Even though it contains 8 studies, this poses a possible limitation. Moreover, as many studies do not report dispersion metrics numerically, we had to extract them from the plots. We conducted a multi-step approach. Two researchers extracted the values individually and afterward discussed the differences. Furthermore, we want to highlight that the meta-analysis is limited with regard to drawing causal conclusions. Our analysis should instead be considered as a synthesis of existing research. Lastly, the main limitation of the comparison of AI-assisted and XAI-assisted performance is the small sample size due to other studies not reporting the dispersion of their AI-assisted condition.

4.6 Discussion and Future Work

In this work, we conducted the first meta-analysis of XAI-assisted decision-making. Based on a structured data collection process, we collected 393 XAI-related articles.

After applying our inclusion criteria, we identified a set of 9 articles encompassing 44 studies for the meta-analysis.

In the current sample, we find no statistically significant difference between XAI and AI-assisted performance. However, some studies report positive XAI assistance performance effects, whereas others find no or slightly negative effects. Additionally, we find a positive effect of XAI assistance on human task performance. Since we do not identify a difference between XAI and AI assistance, the results need to be interpreted carefully. Therefore, we cannot conclude that XAI will lead to an overall performance superior to AI assistance. However, we observe a positive tendency of AI to support humans in decision-making, either with explanations or without. A promising avenue for future research is to investigate the factors that determine consistent performance gains in human-AI collaboration. In this context, the work of Lai et al. (2023) could be a foundation for the data collection. Furthermore, our subgroup analysis indicates a stronger positive effect of explanations on task performance with text data than tabular data. If this effect can be confirmed in future studies, more work would be required regarding human-AI collaboration with tabular data. Reasons for this difference could be that text data is a more intuitive data type for humans. Additionally, analyzing the utility of explanations and AI predictions concerning image data requires more attention, as we identified only 2 articles using image data. Therefore, future work should explicitly investigate performance differences induced by different data types. Moreover, due to the currently high heterogeneity of the studies, future analyses could consider not only a distinction by data type but also by task type, e.g., complexity and users' task-specific knowledge. For example, in easy tasks, humans might be able to evaluate better whether AI advice is correct or not.

Our qualitative review of the collected studies highlights that explanations can easily lead to increased acceptance of AI advice. In the scenario in which AI performs, on average, better than humans, its performance might bound the maximum joint performance of both. However, if the goal is that human-AI collaboration ideally results in superior team performance (Bansal et al., 2021), appropriate reliance on the AI's decision becomes indispensable. Therefore, future research should investigate design mechanisms that enable appropriate reliance. Additionally, from this observation emerges the need to discuss the ethical implications of pure AI decision acceptance increase in future work as it can be understood as a form of manipulating people to follow AI advice blindly.

4.7 Conclusion

This article presents the results of a meta-analysis of the utility of explanations in human-AI collaboration. We identify a total sample of 9 articles that report all necessary information as a prerequisite through a structured literature review. We analyze whether humans' decision-making can benefit from AI support with and without explanations and derive three major findings: First, we do not find a significant effect of state-of-the-art explainability techniques on AI-assisted performance. Second, we observe a significant positive effect of XAI assistance on human performance. Third, our analysis indicates that XAI assistance is more effective on text than tabular data. We hope that our work will motivate scholars to pursue meta-analyses in future human-AI research to systematically assess previous studies to derive conclusions about the current body of research.

Part III

Conceptualization of Human-AI
Complementarity

Human-AI Complementarity: Conceptualization and the Influence of Information Asymmetry

This chapter comprises a working paper that is currently under review as Schemmer, M., Hemmer, P., Köhl, N., Vössing, M., & Satzger, G. (2023b). Human-AI Complementarity: Conceptualization and the Effect of Information Asymmetry [Working paper]. Note: To improve the structure of the work, the title was changed. The abstract has been removed. Tables and figures were reformatted, and newly referenced to fit the structure of the thesis. The terminology was standardized with the dissertation. Chapter, section and research question numbering and respective cross-references were modified. Formatting and reference style was adapted and references were integrated into the overall references section of this thesis.

5.1 Introduction

The increasing capabilities of artificial intelligence (AI) have paved the way for supporting human decision-making in a wide range of application domains. Examples include decision support for humans in application areas such as customer service (Adam et al., 2021), medicine (Wu et al., 2020), law (Mallari et al., 2020), finance (Day et al., 2018), and industry (Stauder & Köhl, 2022). While increasingly accurate decisions of AI systems can tempt the automation of specific tasks, automation generally overlooks the potential of combining individual team members' strengths in a human-AI team (Seeber et al., 2020) to achieve even better performance.

The recent emergence of large language models illustrates this (Malone et al., 2023). Applications like ChatGPT often provide helpful results but cannot be relied upon completely, and a human decision-maker has to collaborate with the system,

for example to override erroneous answers, to achieve superior task performance (Malone et al., 2023). Similarly, in the medical domain, AI as well as humans are able to conduct a disease diagnosis on their own. It has however also been demonstrated that humans and AI can make different errors on individual task instances (Geirhos et al., 2021; Steyvers et al., 2022), such as ensuring that CT images complement each other (Jussupow et al., 2021). In this context, the AI may detect patterns in large amounts of data that humans will find difficult to discover, while humans in turn excel at the causal interpretation and intuition required to interpret these patterns (Lake et al., 2017; Li et al., 2019a).

This potential for complementarity has led researchers to investigate how the individual capabilities of humans and AI can be leveraged to achieve superior team performance, compared to either one performing the decision task independently. Such an outcome is defined as *complementary team performance* (CTP) (Bansal et al., 2021).

Various studies have demonstrated that human-AI teams can outperform human individuals, but they often do not exceed the AI's individual performance (Bansal et al., 2021; Hemmer et al., 2021). The observation that CTP, meaning a performance beyond that of an isolated human or AI, is often not attained in these studies raises questions about which factors could contribute to achieving this phenomenon. It illustrates that the current knowledge of how the respective capabilities of humans and AI can be utilized to create joint decision-making synergies has not yet been sufficiently developed and that there is a need for additional concepts that foster an in-depth understanding of complementarity in human-AI collaboration.

To address this gap, we build on the theoretical work of Fügener et al. (2021) who analyze how AI advice affects unique human knowledge. In particular, we propose a conceptualization of human-AI complementarity by introducing the notion of complementarity potential (CP). The conceptualization consists of formalizing the complementarity potential and its constituent components, delineating relevant sources of complementarity potential, and classifying integration mechanisms to infer the final team decision and realize the human-AI complementarity potential. In detail, we argue that complementarity potential has an inherent and a collaborative component. Whereas the first captures the idea that humans and AI possess different inherently present capabilities in the form of unique human and AI knowledge, the second component captures a new type of knowledge that only emerges through human-AI interaction.

To demonstrate the utility of the proposed conceptualization, we choose the domain of real estate valuation and instantiate a relevant source of complementarity poten-

tial together with an integration mechanism. We train and deploy an AI model to predict real estate prices based on tabular data. Human decision-makers have additional access to a real estate photograph that represents unique human contextual information - a relevant source of inherent complementarity potential - that they can use. The human decision-maker constitutes the final integrator of their own and the AI decision.

Our results highlight the usefulness of our conceptualization, allowing us to develop a more nuanced understanding of the factors that influence whether synergies in decision-making emerge during human-AI collaboration. Our experiment shows that the distribution of instances for which the human or the AI performs better on (indicating the inherent complementarity potential) changes when the human sees an additional photograph of the house. More specifically, providing the photograph increases the number of house price estimates where the human performs better than the AI. We find that providing an additional photograph of the house has a surprising effect on the human's estimate after receiving advice from the AI. Intuitively, the awareness of having more information than the AI could lead to algorithm aversion (Jussupow et al., 2020), which prevents humans from appropriately adjusting the house prices the AI recommended. Our research however indicates the opposite, as our results show that providing the photo improves human adjustment of house prices suggested by the AI.

In summary, our contribution is threefold: First, we conceptualize human-AI complementarity by introducing and formalizing the notion of complementarity potential, delineating sources of complementarity potential, and providing a classification of existing integration mechanisms to realize the complementarity potential. Second, we demonstrate the applicability and usefulness of the conceptualization for a human-AI collaboration setting by drawing on information asymmetry as a source of complementarity potential. Third, through our proposed conceptualization, we find a new and surprising insight that unique human contextual information can result in human decision-makers better adjusting AI advice.

Our work should be guiding further theoretical and empirical work on human-AI collaboration. The concepts we study in this article can be applied to various other application domains, such as medical diagnosis, weather forecasting, or co-programming.

In the remainder of this work we start by outlining the relevant background and related work in Section 5.2. In Section 5.3, we derive the proposed conceptualization of human-AI complementarity by formalizing the notion of complementarity potential, identifying relevant sources of complementarity potential as well as classifying

integration mechanisms to realize the complementarity potential. In Section 5.4, we illustrate the utility of our conceptualization in an experimental study in the context of real estate valuation and study the effect of information asymmetry as one source of complementarity potential. We discuss our results in Section 5.5, before Section 5.6 concludes the work.

5.2 Theoretical Foundations and Related Work

In this section, we elaborate on key concepts and existing work on human-AI collaboration, human-AI complementarity, and information asymmetry.

Many terms have been used to describe the interaction between humans and AI. Common terms are human-AI team (Seeber et al., 2020), human-AI collaboration (Vössing et al., 2022), or human-AI collaboration (Lai et al., 2023). Human-AI teams, human-AI collaboration, and human-AI collaboration are interrelated concepts that emphasize the integration of human and AI systems to achieve optimal outcomes.

Human-AI teams refer to the idea of combining the unique strengths of humans and AI, such as human creativity, empathy, and contextual understanding with AI's data processing, complex calculations, and efficiency in repetitive tasks (Seeber et al., 2020). *Human-AI collaboration* is the process by which these teams work together in a synergistic manner to achieve shared goals, for example with the AI system providing recommendations or insights and humans guiding and refining the AI-generated outputs (Terveen, 1995; Vössing et al., 2022). *human-AI collaboration* refers to the combined input of humans and AI in making choices or judgments, where it is a decision-making task and the AI offers data-driven decision suggestions while humans leverage their domain expertise, emotional intelligence, and ethical considerations to reach a final decision (Lai et al., 2023).

By focusing on these integrated approaches, organizations can capitalize on the complementary strengths of humans and AI systems, fostering improved performance and decision-making in various domains.

5.2.1 Human-AI Collaboration

In recent years, research on human-AI collaboration has experienced tremendous growth as a particular field of human-AI collaboration (Bansal et al., 2021; Buçinca et al., 2020; Lai et al., 2020; Liu et al., 2021). An increasing number of studies have

conducted behavioral experiments to understand how humans make decisions with AI support (Alufaisan et al., 2021; Bansal et al., 2021; Buçinca et al., 2020; Carton et al., 2020; Fügenger et al., 2021; Lai et al., 2020; Liu et al., 2021; Malone et al., 2023; Reverberi et al., 2022; Van der Waa et al., 2021; Zhang et al., 2022; Zhang et al., 2020).

In this context, a growing body of work examines how human reliance on AI decisions can be appropriately calibrated to establish effective decision-making (Buçinca et al., 2020; He et al., 2023; Kunkel et al., 2019; Yu et al., 2019; Zhang et al., 2020). One idea is to enable humans to judge the quality of an AI decision by revealing insights about the uncertainty of the prediction (Zhang et al., 2022) or by providing explanations that shed light on the AI's decision-making (Alufaisan et al., 2021; Bansal et al., 2021; Buçinca et al., 2020; Lai et al., 2020; Liu et al., 2021).

For this purpose, explainable artificial intelligence (XAI) research has developed various approaches to enable humans to understand the underlying mechanisms that contribute to an AI's prediction (Adadi & Berrada, 2018). Several studies investigate how human-AI collaboration benefits from different XAI techniques. Examples range from feature-based (Ribeiro et al., 2016a) and example-based (Van der Waa et al., 2021) to rule-based (Ribeiro et al., 2018) techniques. Even though experimental evidence has demonstrated the benefits of explanations (Buçinca et al., 2020), it has also revealed that explanations can convince humans to follow an incorrect AI decision (Bansal et al., 2021) which is referred to as automation bias (Schemmer et al., 2022c).

A closer look at the studies that analyze human-AI collaboration quantitatively reveals that, in general, human performance increases when supported by a high-performing AI. In the vast majority of cases, the team performance however still remains inferior to that of the AI when it had performed the task alone (Hemmer et al., 2021; Malone et al., 2023).

5.2.2 Human-AI Complementarity

Complementarity between humans and AI is discussed as part of three closely related paradigms: intelligence augmentation, human-machine-symbiosis, and hybrid intelligence.

Intelligence augmentation is defined as “enhancing and elevating human ability, intelligence, and performance with the help of information technology” Zhou et al. (2021), p. 245. It is a form of human-AI collaboration in which machines use their

complementary strengths to assist humans, not necessarily with the goal of achieving CTP, but to improve human objectives.

Human-machine symbiosis is a paradigm that envisions deepening the collaborative human-AI connection. It is based on the notion of a symbiotic relationship between humans and AI, which implies considering both as a common system rather than two separate entities, aiming to become more effective together compared to working separately (Licklider, 1960). This leads to overcoming human restrictions by extending their abilities and reducing the time needed to solve problems (Gerber et al., 2020). Another aspect that emphasizes the symbiotic nature of humans and AI is the focus on human-like communication and interaction between both team members. Researchers argue that the machine should be able to understand verbal and non-verbal communication to exchange information with a human (Sanchez & Principe, 2009; Sandini et al., 2018).

Hybrid intelligence is an emerging paradigm with the idea of combining human and artificial team members in the form of a socio-technical ensemble to resolve current AI limitations. We refer to the work of Dellermann et al. (2019a), p. 640, who define hybrid intelligence as “the ability to achieve complex goals by combining human and AI, thereby achieving better results than what each of them could have accomplished separately, and continuously improve by learning from each other.”

Human-machine symbiosis and hybrid intelligence share our view that humans and AI should complement each other to achieve superior results. Existing studies under both labels however do not provide theoretical foundations or classify sources of complementarity potential and integration mechanisms. The only work of which we are aware that contains a theoretical view of human-AI complementarity are articles by Donahue et al. (2022), Rastogi et al. (2022), and Steyvers et al. (2022), develop a taxonomy that characterizes differences between human and AI decision-making. The authors formalize a particular integration mechanism in detail—the technical integration of human and AI decisions. However, they neither provide a formalization of complementarity nor a classification of possible integration mechanisms beyond their instantiated approach. Donahue et al. (2022) also focus on a single integration mechanism and formalize an integration algorithm that assigns weights to human and AI predictions. Lastly, Steyvers et al. (2022) derive a framework for combining the decisions and different types of confidence scores from humans and AI. They focus on a particular integration mechanism where a Bayesian model combines human and AI decisions.

In summary, to our knowledge, there is no work that holistically formalizes the complementarity of humans and AI, delineates relevant sources of complementarity, and provides a classification of integration mechanisms.

5.2.3 Information Asymmetry

In many application domains, a possible source of complementarity potential resides in the information asymmetry between humans and AI—as additional contextual information is often available to humans only. While AI requires digitally available and sufficient training data to detect patterns that can subsequently be used for decision recommendations (LeCun et al., 2015), relevant information may not have been digitized, for technical or economic reasons (Ibrahim et al., 2021). Contextual information may also not be sufficiently available to be included in AI model training. Humans can however use their expertise and additionally consider non-digitized information as well as information about rare events to create a holistic picture for decision-making. We hypothesize that this information asymmetry represents a promising source of harnessing human-AI complementary capabilities.

This hypothesis is strengthened by forecasting theory (Sanders & Ritzman, 1991, 1995, 2001). (Sanders & Ritzman, 1995) analyze the effects of averaging the predictions of a statistical forecast and a human with access to contextual information. They find a positive effect when giving higher weights to the predictions of humans with more contextual information. However, they do not provide a deeper analysis of why and how unique human contextual information improves team performance. While this study indicates promising effects of unique human contextual information on team performance in forecasting, it lacks a theoretical foundation.

Systematic analyses of leveraging unique human contextual information to improve human-AI collaboration are scarce, with a notable exception being the work of Zhang et al. (2020). The authors focus on the psychological effect of humans knowing to have more information than an AI, independent of the possible performance gain by including unique human contextual information. To do so, they provide the participants in their behavioral experiment with an additional feature that contains no relevant information for the task. Their study shows that this treatment has no significant effect on performance.

We therefore conclude that the current knowledge of how the complementarity potential between humans and AI determines whether CTP can be achieved during joint decision-making is not sufficiently developed. Research work currently lacks a

theoretical conceptualization supported by empirical evidence of human-AI complementarity. In this work, we aim to provide both a theoretical foundation as well as empirical insights.

5.3 Conceptualization of Human-AI Complementarity

In this section, we first elaborate on the purpose and scope of human-AI complementarity. Subsequently, we introduce our conceptualization by formalizing the notion of complementarity potential, followed by delineating possible sources of complementarity potential and their realization by classifying different existing integration mechanisms.

5.3.1 Purpose and Scope of Human-AI Complementarity

We initiate the development of the conceptualization by defining the purpose and scope of human-AI complementarity (Jones & Gregor, 2007). Researchers have discussed different perspectives of human-AI collaboration. For example, humans are required for the training and debugging process of AI systems (Dellermann et al., 2019a), take over the final decision in high-stakes decision-making scenarios due to ethical and legal considerations (Lai & Tan, 2019), or are involved in realizing human-AI performance synergies (Bansal et al., 2021; Reverberi et al., 2022).

In this work, we focus on the performance perspective, particularly decision-making tasks where humans and AI can perform the task independently. With the recent rise of large language models, there is an increasing number of tasks that both humans and AI can conduct. Examples include accurately diagnosing diseases in medicine (Goldenberg et al., 2019), conducting loan decisions in finance (Turiel & Aste, 2020), or writing entire programs with AI code generation systems (Ross et al., 2023).

This differentiates human-AI complementarity from classical decision support, where an algorithm only provides selective input for more comprehensive downstream decisions that only a human can make. For example, in credit allowance decisions, a technical system may only provide an aggregated probability for default that the lender uses in the downstream task of making the final loan decision. Nowadays, AI systems have become increasingly capable of conducting such decisions in isolation, meaning that more and more tasks can be automated. Recent studies in the medical domain demonstrate that AI can detect diabetic retinopathy as accurately as highly trained experts (Gulshan et al., 2019). Therefore, AI's role increasingly changes

from being a pure input provider to becoming an equitable team member for the human.

5.3.2 Formalization of Complementarity Potential

In this subsection, we introduce and formalize the notion of complementarity potential. With the formalization we aim to provide the means to quantify the synergetic potential between humans and AI to foster a more detailed understanding of their individual and joint decision-making capabilities. This is because the sole comparison of performance metrics does not capture the underlying mechanisms that drive the resulting performance outcomes.

In detail, the formalization pursues the idea that human-AI complementarity potential is composed of an inherent and collaborative component. The first component captures any complementarity potential that can be attributed to the individual capabilities of both with respect to a decision-making task and is inherently present in both team members. We denote it as inherent complementarity potential. The second component captures the complementarity potential that emerges only through any form of collaboration during joint decision-making and is therefore ex-ante not existent. We denote it as collaborative complementarity potential. In the formalization, for both components that together result in the total complementarity potential, we distinguish between the realized amount that has materialized and a theoretical amount that serves as an upper boundary.

Figure 5.1 on page 94 provides a high-level conceptual overview of the notion of complementarity potential, including its components in human-AI collaboration. In addition to the components, it depicts the average decision-making performance for a task conducted by a human and AI together as a team as well as both individually, expressed by an arbitrary loss function that quantifies the prediction error with respect to a given ground truth. The difference between the best individual team member and the joint human-AI team performance materializes in realized complementarity potential, whereas its theoretical upper boundary for potential further improvement is given by the performance of the best team member. Viewing human-AI collaboration from this granular perspective allows a deeper understanding of the joint decision-making behavior to develop. In the next step, we introduce and develop the proposed formalization.

Let us consider a decision task $T = \{x_i, y_i\}$ as a set of N instances $x_i \in X$ with corresponding ground truth labels $y_i \in Y$. The ground truth may not be known at

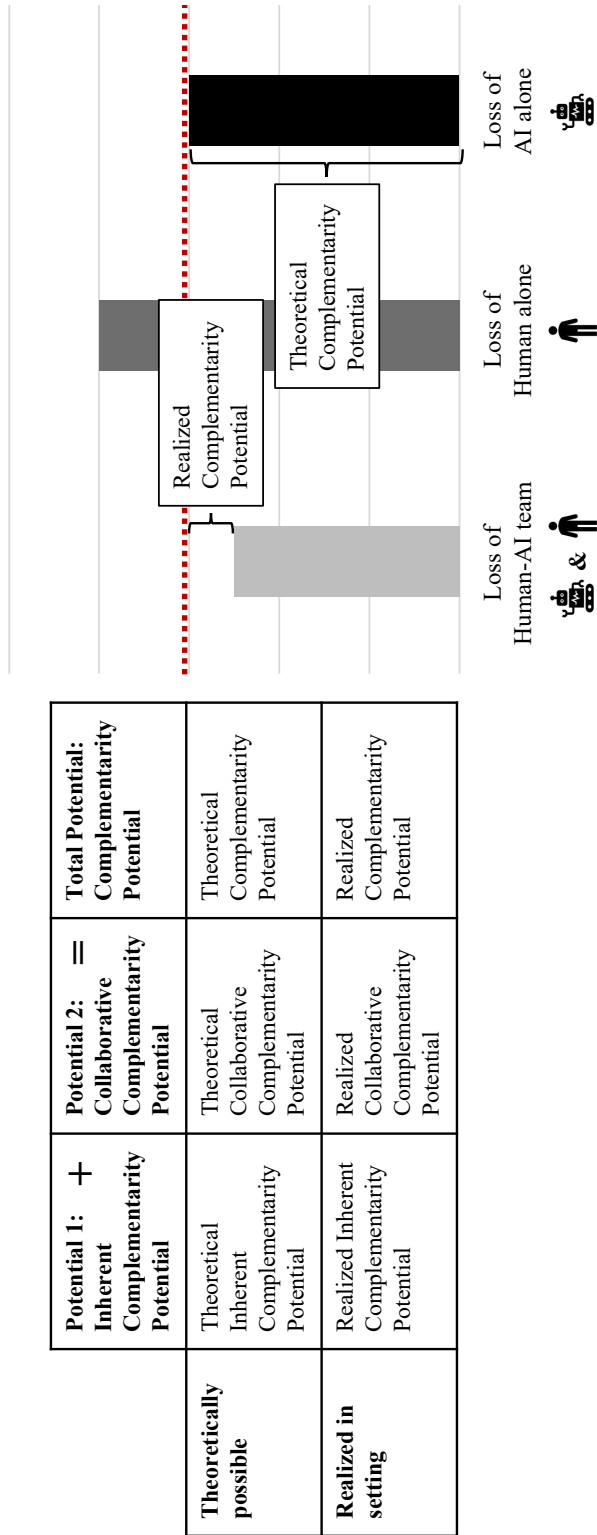


Figure 5.1.1: Conceptual overview of the notion of complementarity potential including its components inherent and collaborative complementarity potential. Note that in the bar chart performance is represented by a loss that can be interpreted as an error measure. This means a lower value represents a better performance

the time of the decision but can be determined and revealed later. Both a human decision-maker and an AI model are capable of independently inferring a prediction \hat{y}_i^H and \hat{y}_i^{AI} for a given instance x_i . Additionally, let us consider some loss function l with its loss bounded in R^+ . A loss function determines the error between a single prediction of a human or an AI model and the corresponding ground truth label. In this context, it can be understood as a generic measure of task performance. In our formalization, it can encompass classification as well as regression tasks. For a given prediction task, we denote the instance-specific human loss as l_H and the average loss over all available instances as $L_H = \frac{1}{N} \sum_{i=1}^N l_H(\hat{y}_i^H, y_i)$. Likewise, we denote the instance-specific AI loss as l_{AI} and the average loss considering all available instances as $L_{AI} = \frac{1}{N} \sum_{i=1}^N l_{AI}(\hat{y}_i^{AI}, y_i)$. For both the human and the AI model, we assume their decisions are made independently.

In addition, we represent any way of collaboration between the human and the AI model by an integration mechanism $I(\hat{y}_i^H, \hat{y}_i^{AI})$ producing a joint prediction \hat{y}_i^I . This decision also incurs an instance-specific loss l_I . Similarly, we define the average loss of the integrated decision as $L_I = \frac{1}{N} \sum_{i=1}^N l_I(\hat{y}_i^I, y_i)$. Complementary team performance (CTP) exists once the average loss of the integrated decision is lower than both the individual average losses of the human and the AI model (Bansal et al., 2021):

$$CTP = \begin{cases} 1, & L_I < \min(L_H, L_{AI}), \\ 0, & otherwise \end{cases}$$

In addition to this binary outcome, we quantify the difference between the average loss of the best individual team member and the team performance and denote this amount as realized complementarity potential:

$$CP_{realized} = \min(L_H, L_{AI}) - L_I$$

In this context, a positive value of the realized complementarity potential ($CP_{realized}$) denotes that synergies between human and AI could be realized during the collaboration, whereas a negative value means that the collaboration resulted in a worse outcome compared to the best team member alone. It also allows us to interpret the average loss of the best individual team member as an upper boundary for further improvement. A reduction by this amount through collaboration would mean perfect human-AI collaboration behavior. We therefore denote it as theoretical complementarity potential:

$$CP_{theoretical} = \min(L_H, L_{AI}) - L_I$$

The notion of complementarity potential forms the basis for developing a deeper understanding of the factors that can lead to CTP in human-AI collaboration. Specifically, we argue that it is essentially composed of two components - inherent human-AI complementarity potential and collaborative complementarity potential, which does not exist inherently, but can only emerge through the collaboration itself.

The first component can be understood as the potential performance increase resulting from the exploitation of unique human or AI knowledge. In accordance with Fügener et al. (2021b), we denote it as unique human knowledge (*UHK*) and unique AI knowledge (*UAIK*). When a human can contribute unique human knowledge to particular instances of a decision task, it is reflected in a lower loss compared to that of the AI. Vice versa, when the AI can contribute its unique knowledge, it is reflected in a lower loss compared to that of a human. These knowledge-based performances can be defined as the sum of the differences between the instance-specific AI and the instance-specific human losses:

$$UHK = \sum_{i=1}^N \max(0, l_{AI}^{(i)} - l_H^{(i)})$$

$$UAIK = \sum_{i=1}^N \max(0, l_H^{(i)} - l_{AI}^{(i)})$$

From the perspective of the team member with the lower average individual loss (e.g. $L_{AI} < L_H$ or $L_H < L_{AI}$) the other team member's unique knowledge constitutes the theoretically existing unique knowledge that can materialize in improved team performance, for example by selecting the individual decision of either human or AI that is more accurate on the task instance level than the team decision. Alternatively, it can be understood as the existing additional potential for team performance improvement that the lower-performing team member possesses uniquely and can contribute to the better-performing team member's capabilities. As this unique knowledge roots in the individual team members' capabilities and therefore exists inherently ex-ante when human and AI team up, we denote it as theoretical inherent complementarity potential. It can be defined as

$$CP_{theoretical}^{inh} = \begin{cases} UHK, & L_{AI} \leq L_H, \\ UAIK, & L_{AI} > L_H. \end{cases}$$

If the average loss of a human and AI is equal $L_{AI} = L_H$, both team members can contribute quantitatively the same amount of unique knowledge ($UHK = UAIK$) which then corresponds to the theoretical inherent complementarity potential $CP_{theoretical}^{inh}$.

In practice, it is unlikely that the integrated decision, resulting from human-AI collaboration, will always fully exploit the theoretical inherent complementarity potential $CP_{theoretical}^{inh}$. It rather represents an upper boundary of the first component of the realized complementarity potential $CP_{theoretical}^{inh}$. We are however interested in measuring the amount that has been exploited by integrating the human and AI decisions. We therefore denote the realized amount of the theoretical inherent complementarity potential $CP_{theoretical}^{inh}$ as realized inherent complementarity potential $CP_{theoretical}^{inh}$. To determine this amount, we must distinguish which of the two team members has the lower average individual loss, as the lower-performing team member's unique knowledge contributes to the realized inherent complementarity potential.

When we consider $L_{AI} < L_H$, any integrated decision with $l_{AI} > l_I \geq l_H$ means that unique human knowledge is present but not fully exploited. In this case, $CP_{theoretical}^{inh}$ is the difference between the instance-specific loss of the AI model l_{AI} and the instance-specific loss of the integrated decision l_I . Figure 5.2 case a on page 98 exemplifies this scenario. It displays the absolute difference of the instance-specific loss of each team member individually and of the integrated decision with respect to the ground truth for a particular instance. As shown in Figure 5.2 case b on page 98, the loss of the integrated decision l_I may even fall below the loss of the human l_H ($l_{AI} > l_H > l_I$). Then, $CP_{theoretical}^{inh}$ is the difference between the losses of the AI model l_{AI} and the loss of the human l_H . Any remaining improvement beyond the smaller loss (in Figure 5.2 case b on page 98 the human one) cannot be attributed to the “upfront” inherent knowledge asymmetry. We will interpret this when we elaborate the concept of collaborative complementarity potential.

Now, let us consider the case $L_H < L_{AI}$. Any integrated decision with l_H, l_I, L_{AI} can be interpreted that unique AI knowledge is present but not fully exploited. In this case, $CP_{theoretical}^{inh}$ is the difference between the instance-specific loss of the human l_H and the instance-specific loss of the integrated decision l_I . Figure 5.3 case a exemplifies this scenario. Again, if the instance-specific loss of the integrated



Figure 5.2.: Deviation of human prediction, AI prediction and integrated prediction from the ground truth for a single instance. Possible cases of how realized inherent complementarity potential $CP_{theoretical}^{inh}$ emerges, considering the scenario $L_{AI} < L_H$

decision l_I even falls below the instance-specific loss of the AI model l_{AI} ($l_H > l_{AI} > l_I$), $CP_{theoretical}^{inh}$ can only be the difference between the instance-specific loss of the human l_H and the instance-specific loss of the AI model l_{AI} . We visualize this scenario in Figure 5.3 case b on page 98.

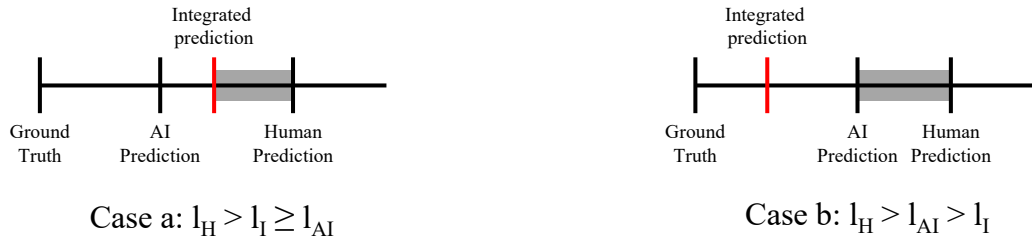


Figure 5.3.: Deviation of human prediction, AI prediction and integrated prediction from the ground truth for a single instance. Possible cases of how realized inherent complementarity potential $CP_{theoretical}^{inh}$ emerges, considering the scenario $L_H < L_{AI}$

Finally, we can summarize each component of the realized inherent complementarity potential in the following formula¹ :

$$CP_{theoretical}^{inh} = \sum_{i=1}^N \begin{cases} l_{AI}^{(i)} - l_I^{(i)}, & L_{AI} \leq L_H \text{ and } l_{AI} > l_I \geq l_H \\ l_{AI}^{(i)} - l_H^{(i)}, & L_{AI} \leq L_H \text{ and } l_{AI} > l_H \geq l_I \\ l_H^{(i)} - l_I^{(i)}, & L_H \leq L_{AI} \text{ and } l_H > l_I \geq l_{AI} \\ l_H^{(i)} - l_{AI}^{(i)}, & L_H \leq L_{AI} \text{ and } l_H > l_{AI} \geq l_I \end{cases}$$

¹If both team members have the same average loss ($L_{AI} = L_H$), it is valid to calculate the realized inherent complementarity potential $CP_{theoretical}^{inh}$ either according to the case $L_{AI} < L_H$ (used in this work) or the case $L_H < L_{AI}$.

We denote the second component of the complementarity potential as collaborative complementarity potential. It refers to the observation that collaboration between a human and an AI model can also result in an integrated decision l_I that improves or deteriorates for a specific instance beyond the better or worse individual human or AI decision. This may be driven by effects that happen due to the collaboration itself, such as (un)learning. Given the unique knowledge of the human-AI team determined by the theoretical inherent complementarity potential $CP_{theoretical}^{inh}$ and the amount of the overall theoretical complementarity potential $CP_{theoretical}^{inh}$, we can calculate an upper boundary for potential further improvement through collaboration. We refer to it as theoretical collaborative complementarity potential:

$$CP_{theoretical}^{coll} = CP_{theoretical} - CP_{theoretical}^{inh}$$

There might be task instances where improvement beyond the best individual decision is attainable through joint decision-making. Likewise, there will also be situations where joint decision-making leads to deterioration. In practice, we are interested in capturing both effects that contribute to the realized collaborative complementarity potential ($CP_{realized}^{coll}$). We reflect this idea in the formalization by distinguishing positive realized collaborative complementarity potential ($CP_{realized}^{coll, positive}$) and negative realized collaborative complementarity potential ($CP_{realized}^{coll, negative}$) in the following way:

$$CP_{theoretical}^{coll} = CP_{theoretical}^{coll, positive} - CP_{theoretical}^{coll, negative}$$

Let us first focus on the positive realized collaborative complementarity potential ($CP_{realized}^{coll, positive}$). For a better understanding, let us again consider the scenario with $L_{AI} < L_H$. If the instance-specific loss of the integrated decision l_I falls below the lower individual instance-specific loss of the human $l_H (l_{AI} \geq l_H > l_I)$ or the AI model ($l_H \geq l_{AI} > l_I$), the difference between the lower individual instance-specific loss of either the human l_H or the AI model l_{AI} and the integrated loss l_I refers to positive realized collaborative complementarity potential (see Figure 5.4 case a and b), as this improvement can only be driven by collaboration and not by the inherently present unique knowledge of one team member.

The same phenomenon applies to the scenario with $L_H < L_{AI}$. If the instance-specific loss of the integrated decision l_I falls below the individual instance-specific loss of the AI model $l_{AI} (l_H \geq l_{AI} > l_I)$ or the human $l_H (l_{AI} \geq l_H > l_I)$ the difference between the lower individual instance-specific loss of either the human l_H

or the AI model l_{AI} and the integrated loss l_I refers to positive realized collaborative complementarity potential (see Figure 5.4 case a and b). We summarize positive realized collaborative complementarity potential in the following formula²

$$CP_{realized}^{coll, positive} = \sum_{i=1}^N \max \left(0, \min \left(l_H^{(i)}, l_{AI}^{(i)} \right) - l_I^{(i)} \right)$$

Lastly, we consider scenarios in which negative realized collaborative complementarity potential ($CP_{realized}^{coll, positive}$) can occur.



Figure 5.4.: Deviation of human prediction, AI prediction and integrated prediction from the ground truth for a single instance. Possible cases of how positive realized collaborative complementarity potential $CP_{theoretical}^{coll, positive}$ emerges.

In the scenario with $L_{AI} < L_H$ the AI model on average outperforms the human. Negative realized collaborative complementarity potential $CP_{realized}^{coll, positive}$ incurs, once the instance-specific loss of the integrated decision l_I is larger than the instance-specific loss of the AI model l_{AI} ($l_I > l_{AI}$) independent of the instance-specific loss of the human l_H (see Figure 5.5 case a on page 101). Negative realized collaborative complementarity potential $CP_{realized}^{coll, positive}$ is the difference between the instance-specific loss of the integrated decision l_I and the instance-specific loss of the AI model l_{AI} .

Similarly, in the scenario with $L_H < L_{AI}$ the human on average outperforms the AI model. We therefore incur negative realized collaborative complementarity potential ($CP_{realized}^{coll, negative}$) once the instance-specific loss of the integrated decision l_I is larger than the instance-specific loss of the human l_H ($l_I > l_H$) independent of the instance-specific loss of the AI model l_{AI} (see Figure 5.5 case b). Negative realized collaborative complementarity potential ($CP_{realized}^{coll, negative}$) is the difference between the instance-specific loss of the integrated decision l_I and the instance-

²Note that the formula also applies to a case where both team members have the same average loss ($L_{AI} = L_H$)



Figure 5.5.: Deviation of human prediction, AI prediction and integrated prediction from the ground truth for a single instance. Possible cases of how negative realized collaborative complementarity potential $CP_{realized}^{coll, negative}$ emerges, with case a considering the scenario $L_{AI} < L_H$ and case b considering the scenario $L_H < L_{AI}$.

specific loss of the human l_H . We summarize negative collaborative knowledge in the following formula ³ :

$$CP_{realized}^{coll, negative} = \sum_{i=1}^N \begin{cases} l_I^{(i)} - l_{AI}^{(i)} & L_{AI} \leq \text{and } l_I > l_{AI} \\ l_I^{(i)} - l_H^{(i)} & L_H < L_{AI} \text{ and } l_I < l_H \end{cases}$$

Based on both introduced components, we can formulate the realized complementarity potential ($CP_{realized}$) as the sum of the realized inherent complementarity potential ($CP_{realized}^{inh}$) and the realized collaborative complementarity potential ($CP_{realized}^{coll}$)

$$CP_{realized} = CP_{realized}^{inh} + CP_{realized}^{coll}$$

This alternative perspective on the realized complementarity potential ($CP_{realized}$) allows us to analyze human-AI collaboration on a more granular level that can help uncover new insights about the joint decision-making behavior.

5.3.3 Sources of Theoretical Complementarity Potential

A continuously growing body of research on human-AI collaboration assumes complementary capabilities between humans and AI (Bansal et al., 2021; Dellermann et al., 2019a). The discourse about complementary capabilities usually remains

³If both team members have the same average loss ($L_{AI} = L_H$) it is valid to calculate the negative realized collaborative complementarity potential $CP_{realized}^{coll, negative}$ either according to the case $L_{AI} < L_H$ (used in this work) or the case $L_H < L_{AI}$

superficial, for example hypothesizing that humans excel in creativity, whereas AI better identifies patterns in large amounts of data (Dellermann et al., 2019a). A conceptualization together with an empirical analysis of complementarity is however still lacking, therefore we aim to provide a more nuanced understanding of theoretical complementarity potential by identifying its sources. To that end, we conceptually distinguish three phases in human and AI decision-making. The first is the *learning or training phase* that encompasses AI training and the human learning process. Note that AI models are typically developed in a short period of time by learning from aggregated training data, while human learning is a lifelong process. The second is the inference phase, in which the human and AI infer a decision on a particular instance. The third is the integration phase, in which the human and AI collaborate. Based on this simplified decision-making process, we can discuss sources of theoretical complementarity potential (inherent and collaborative) in a more granular fashion.

Training phase: First, training data as input of the training process differs between humans and AI. Humans may have seen a considerable *number of training* instances, whereas the AI is typically trained on a limited, customized training data set. Second, their *inherent capabilities*, which are stimulated through training, differ from each other. For instance, the AI can efficiently identify patterns in high-dimensional data or infer decisions from probabilistic reasoning (Dellermann et al., 2019a; Jarrahi, 2018). In contrast, humans can already learn abstract concepts from a small number of samples (Zheng et al., 2017). Third, during the training process, humans and AI learn *different decision boundaries* from which their final decision is inferred (Geirhos et al., 2021). In this context, human decision-making can often be purely heuristic or intuitive without considering all available information (Jarrahi, 2018). These differences provide the potential for complementary capabilities; their presence results in theoretical inherent complementarity potential.

Inference phase: After the training process, the AI can be used to infer a decision for a particular instance. Even assuming identical decision boundaries of humans and AI, different available input data can constitute a source of theoretical inherent complementarity potential - as in real-world settings, AI and humans often have access to different features (Bansal et al., 2021; Sanders & Ritzman, 2001). A famous example is the “broken leg” scenario, which refers to side information that is known to humans (Meehl, 1957). This information could not be incorporated as features in the model due to its rare occurrence. In the majority of application domains in which AI is applied to support human decision-making, additional information exists beyond the data used to train the AI model (Ibrahim et al., 2021). In practice, due to technical or economic reasons, this discretionary data may often

not be digitally available at all, or only in a small quantity that is insufficient for model training (Ibrahim et al., 2021). Nevertheless, in human-AI collaboration, human team members may leverage their expertise to use the additional information (Ibrahim et al., 2021). On the other hand, researchers highlight the consistency of AI decision-making at inference (Blattberg & Hoch, 2010; Dellermann et al., 2019a). To summarize, theoretical inherent complementarity potential can be present during training or inference and is typically either based on information asymmetries or complementary skills.

Integration phase: The integration process adds a third phase to the overall decision-making process. In the integration phase, theoretical collaborative complementarity potential can arise as an additional part of theoretical complementarity potential beyond the theoretical inherent complementarity potential. For the collaborative aspect of theoretical complementarity potential, it is essential for the human-AI team to possess the capability to further improve the best individual decision achieved either by the AI or human alone on a given task instance. Figure 5.6 summarizes all three phases.

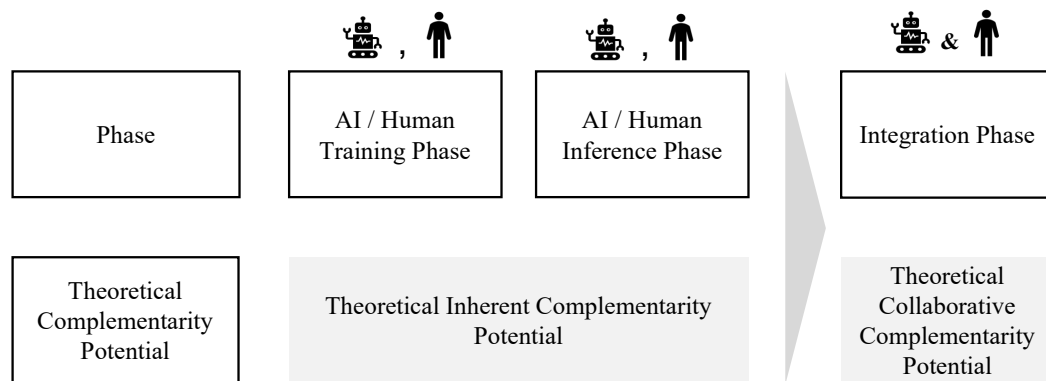


Figure 5.6.: Conceptual overview of the human-AI collaboration process.

5.3.4 Integration Mechanisms: Realizing the Theoretical Complementarity Potential

Human as well as AI team members can make decisions individually for any instance of a particular task. However, from a team perspective, there is a final team decision. We denote a generic representation of inferring the final team decision as an integration mechanism. We consider the integration mechanism as the collaborative element of the human-AI team that can be instantiated arbitrarily. In this subsection, we aim to derive a classification of possible integration mechanisms.

We differentiate integration mechanisms along two dimensions. First, different *timings* of integration are conceivable. We may ex-ante assign the decision either to the human or AI. This means that only one of the team members would have to solve the task without interactive collaboration. Alternatively, we may have both propose a decision and integrate them ex-post. Second, we may differentiate who performs the integration – the human or the technical system (*agency*). Whereas literature on decision support systems usually considers the human as the only decision-maker, research in forecasting and human-AI collaboration proposes several technical delegation and aggregation methods (Bondi et al., 2022; Hemmer et al., 2021; Sanders & Ritzman, 2001). Note that in some instantiations of the integration mechanisms, the integrator can also be one of the team members. Table 5.1 on page 105 depicts the resulting four classes of integration mechanisms that we describe in the following subsections.

Technical delegation: Analogously to human delegation, a technical delegation of task instances is also possible (Bondi et al., 2022; Hemmer et al., 2021; Mozannar & Sontag, 2020). Recent work proposes to specifically train the AI to consider the capabilities of human team members and accordingly decide who to assign the task to (Hemmer et al., 2021; Mozannar & Sontag, 2020; Wilder et al., 2020). A simpler technical delegation might consist of distributing task instances solely based on the confidence of the AI (Fügener et al., 2021). In the case of the ex-post integration, a decision is made by both the human and AI respectively. They are the basis for the final decision integrated either by the human or a technical system.

Human aggregation: Human aggregation refers to the human who receives both the human and an AI prediction and integrates them. The most common integration mechanism might be one that allows humans to make judgmental adjustments to AI decisions (Hemmer et al., 2021; Ibrahim et al., 2021; Lai et al., 2023; Sanders & Ritzman, 1995).

Technical aggregation: Similar to the human aggregation, a technical aggregation is also conceivable. One of the most common approaches is stacked generalization (Wolpert, 1992) where typically a secondary AI is trained that aggregates the decisions of multiple AI models. Similarly, this secondary AI may also aggregate human and AI decisions. Alternative technical, but non-AI, approaches consider a weighting mechanism for the decisions of both team members (Blattberg & Hoch, 2010).

The theoretical complementarity potential realized by ex-ante and ex-post integration mechanisms is different. Ex-ante integration mechanisms are able to realize inherent theoretical complementarity potential, while ex-post integration mechanisms can

realize both inherent and collaborative theoretical complementarity potential. The reason is that to realize collaborative complementarity potential, humans and AI must jointly make a better decision than individually at the task instance level. This can only happen through collaboration in ex-post integration, because in ex-ante integration the task instance is delegated to one of the team members, and therefore the maximum performance is that of the better team member on that particular task instance.

		Timing	
		Ex-ante	Ex-post
Agency	Human	Human delegation	Human aggregation
	Technical system	Technical delegation	Technical aggregation

Table 5.1.: Conceptual overview of the human-AI collaboration process.

In this subsection, we developed a conceptualization of human-AI complementarity by introducing and formalizing the notion of complementarity potential, identified sources of theoretical complementarity potential, and classified different integration mechanisms to realize the theoretical complementarity potential. We focus on settings where a human and an AI form a team with the overarching goal of reaching CTP. Our conceptualization allows an analysis of the contributing factors, including their magnitude, to foster an in-depth understanding through which factors CTP was achieved during human-AI collaboration or not.

In Table 5.2 on page 106, we summarize our conceptualization. It highlights starting points for the design of human-AI complementarity. System designers can develop mechanisms that influence the realized inherent and the realized collaborative complementarity potential. In the following subsections, to demonstrate the utility of our conceptualization, we apply it in a behavioral experiment.

5.4 Experimental Design

To illustrate our conceptualization, we conduct a behavioral experiment. As discussed in the previous section, the key drivers of CTP are the complementarity potential, the presence of its sources and the integration mechanism to realize the theoretical complementarity potential. Several previous empirical studies focused

Isolated performance (better team member)

	Potential 1: Inherent complementarity potential +	Potential 2: Collaborative complementarity potential =	Total potential: Complementarity potential
Theoretically possible	Theoretical inherent complementarity potential $CP_{theoretical}^{inh}$	Theoretical collaborative complementarity potential $CP_{theoretical}^{coll}$	Theoretical complementarity potential $CP_{theoretical}$
	Source: Training phase Inference phase	Source: Integration phase	
- Realized in setting	Realized inherent complementarity potential $CP_{realized}^{inh}$	Realized collaborative complementarity potential $CP_{realized}^{coll}$	Realized complementarity potential $CP_{realized}$
	Integration mechanism: Ex-ante Ex-post	Integration mechanism: Ex-post	

= Team performance

Table 5.2.: Conceptualization of human-AI complementarity consisting of the notion of complementarity potential, sources of complementarity potential, and integration mechanisms. The realized complementarity potential subtracted from the isolated performance of the better team member results in the final team performance after collaboration. Note that performance is represented by a loss that can be interpreted as an error measure. This means a lower value represents a better performance.

on investigating the effect of the integration mechanisms on the effectiveness of human-AI collaboration (Bondi et al., 2022; Hemmer et al., 2021; Lai et al., 2023). In this behavioral experiment, we use our conceptualization to study the effect of designing for the theoretical inherent complementarity potential by ensuring the presence of a relevant source. To induce complementarity potential, we generate a setting with asymmetric information between humans and AI. More specifically, we equip the human with unique human contextual information and choose a human aggregation setting in which the human adjusts the AI suggestions in the best

possible way from the ex-post integration mechanisms. In the following subsection, we denote this setting as human adjustment. It allows us to investigate the realized inherent as well as the realized collaborative complementarity potential. The latter does not arise in the ex-ante integration classes. More specifically, we instantiate an AI-assisted decision-making setting in which the human integrates their own and the AI decision.

5.4.1 Task and AI Model

In this subsection, we explain our chosen task and AI model. We draw upon a real estate appraisal task provided on the data science website Kaggle (2019). As housing is a basic need, and because it is ubiquitous in everyone's life, all people have to some degree the ability to assess the value of a house based on relevant factors such as size or appearance. The data set encompasses 15,474 houses and contains information about the street, city, number of bedrooms, number of bathrooms, and size (in square feet). House prices in the data set denote their listing price. The average house price is \$703,120, with a minimum of \$195,000 and a maximum of \$2,000,000. An image of each house's exterior is provided.

For the house price prediction task, we use a random forest regression model (Breiman, 2001). We draw upon the individual trees in the random forest to generate a predictive distribution for each instance and provide the 5% and 95% quantiles as indicators of the AI's prediction uncertainty. We use 80% of the data as the training set and 20% as the test set. We randomly draw a hold-out set of 15 houses from the test set, serving as samples for our behavioral experiment. We train the random forest on the features of street, city, number of bedrooms, number of bathrooms, and square feet of the house, while the image of the house is withheld from the AI model. The AI model achieves a performance measured in terms of the mean absolute error (MAE) of \$163,080 on the hold-out set, which is representative of its performance on the test set.

For the condition with unique human contextual information, we create sufficient complementarity potential between human and AI. From prior studies in real estate appraisal, we know that AI usually outperforms humans Viriato, 2019. We can therefore assume a certain amount of unique AI knowledge, also in our task setup. To ensure a certain amount of unique human knowledge, we give the humans access to additional contextual information. To positively affect CTP, this information must be useful for humans. For the selected use case of real estate appraisal, a valuable piece of information may be an image of the house, as humans can use their general

understanding to form an overall assessment based on the house’s features, the visible neighborhood, and its appearance. To verify this assumption, we conduct an initial pilot study (Appendix A.1.1 contains the detail).

5.4.2 Study Design and Procedure

We conduct an online experiment with a between-subject design. We recruit participants via prolific.co. The study includes two treatments and randomly assigns each participant to one of these treatments. We do not allow any repeated participation. Each participant passes through the steps displayed in Figure 5.7 and described below.

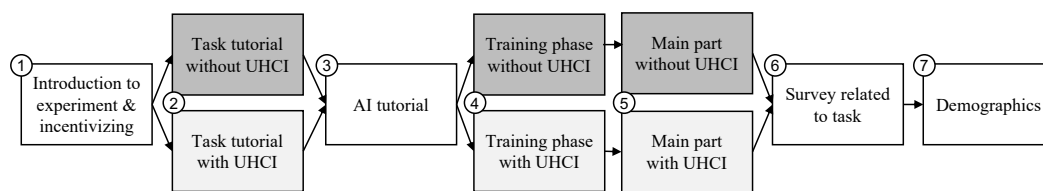


Figure 5.7.: Sequence of the individual steps in the experiment (UHCI = unique human contextual information).

Step 1: After accepting the task, participants are transferred to our experimental website. They are asked to give their consent and read the instructions. The study is initiated by a control question. *Step 2:* As prior work highlights the importance of providing training about the task that the participants must conduct, we include a mandatory tutorial (Grootswagers, 2020). To familiarize participants with the task and data, both treatments receive an identical in-depth introduction about the data set, including summary statistics about the entire data set, like mean, maximum and minimum house prices as reference points (Figure A.1 and Figure A.2 in Appendix A.1.2 contain more information). Participants in the treatment without unique human contextual information (without UHCI) are only introduced to the tabular data of the houses, while the participants in the treatment with unique human contextual information (with UHCI) are also provided with the house images.

Step 3: As part of the tutorial, we introduce the participants to the AI prediction. We highlight that the AI did not have access to the images during training. We show the AI prediction in the context of minimum and maximum house prices. The participants also receive information about the AI’s uncertainty in the form of the 5% and 95% quantiles (see Figure 5.8). We explain the interpretation of the AI’s advice, including all data points mentioned above. The participants are then asked to answer a control question to verify their understanding.

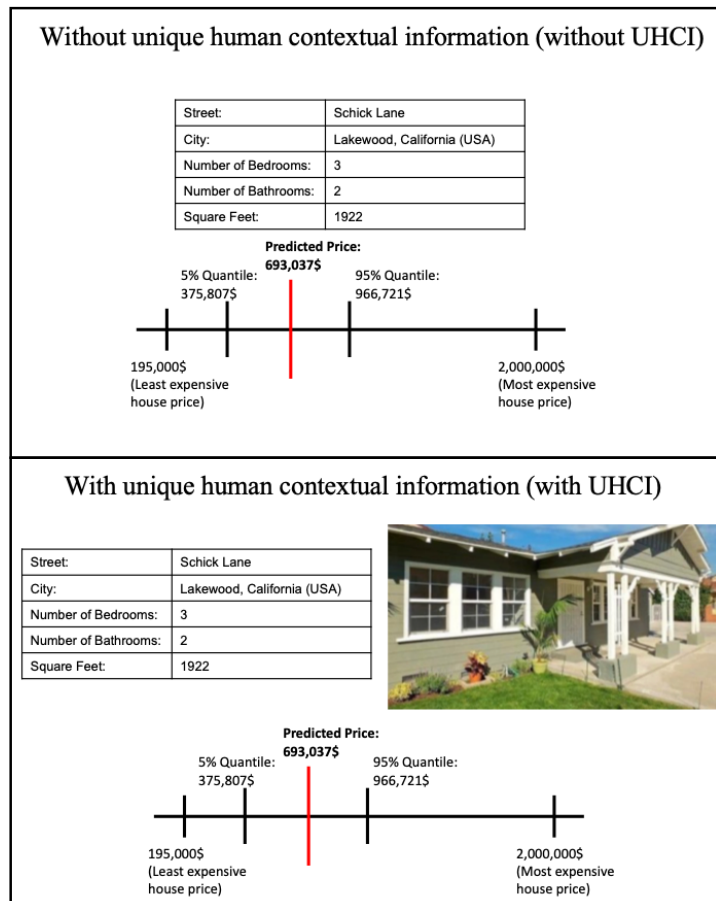


Figure 5.8.: An overview of the interfaces containing the information that the participants are provided with in the respective treatments of the behavioral experiment

Step 4: The participants conduct two training tasks. For each task, we initially let the participants provide a prediction on their own, before we reveal the AI’s recommendation. They are asked to adjust the AI’s prediction in the best possible way. After each training example, participants receive feedback in the form of the true house price. After completing the two training tasks, they are informed about the start of the study.

Step 5: Each participant completes 15 house price prediction tasks presented in a randomized order in the same procedure as described in *Step 4*. During the tasks, participants are not informed about the true house price. After completion of all tasks, we ask participants to complete a questionnaire to collect qualitative feedback (*Step 6*) and demographic information (*Step 7*).

The overall task lasts approximately 30 minutes. We recruit a total of 120 participants, 60 per condition. They receive a base payment of £5 and are additionally incentivized following the approach of Kvaløy et al., 2015. Note that the two

training tasks are not included in the final evaluation. To ensure the quality of the collected data, we remove participants by entering house prices that are higher than the communicated maximum house price in the data set of \$2,000,000. We also identify outliers for removal using the median absolute deviation Leys et al., 2013; Rousseeuw and Croux, 1993. After applying these criteria, we continue with the data from 101 participants over both conditions - 53 in the treatment without UHCI and 48 in the treatment with UHCI (Table C1 in Appendix A.1.3 contains the detail).

5.4.3 Evaluation Measures

We measure the human's and AI's individual performances (L_H and L_{AI}) as well as the joint team performance resulting from the instantiated integration mechanism (L_I). The loss per participant is measured by the mean absolute error (MAE):

$$MAE = \frac{1}{N} \sum_{i=1}^N |y - x|$$

In addition to the respective losses, we calculate the specific components of the formalization of complementarity potential as defined in subsection 5.3.2. It includes complementary team performance (CTP) as a binary task outcome, unique human knowledge (UHK), unique AI knowledge (UAIK), complementarity potential (theoretical: $CP_{theoretical}$ realized $CP_{realized}$) including its components inherent complementarity potential (theoretical: $CP_{theoretical}^{inh}$ realized $CP_{realized}^{inh}$) and collaborative complementarity potential (theoretical: $CP_{theoretical}^{coll}$ realized $CP_{realized}^{coll}$). Regarding the realized collaborative complementarity potential, we report its positive ($CP_{realized}^{coll,positive}$) as well as its negative components ($CP_{realized}^{coll,negative}$).

5.5 Results

In this section, we analyze the impact of unique human contextual information on team performance and complementarity potential. We evaluate the significance of the results using the Student's T-test and the Mann-Whitney U-test depending on the fulfillment of the prerequisites while applying the Bonferroni correction. First, we focus on the impact of contextual information on performance, followed by an

in-depth analysis of the effect on complementarity potential and its constituting components.

Isolated performance and team performance. Figure 5.9 displays the sole human and joint AI-assisted performance for both conditions. It also includes the performance of the AI alone. We first evaluate the impact of unique human contextual information without any AI assistance. Participants in the treatment without UHCI achieve an MAE of \$251,282, while those in the treatment with UHCI yield an MAE of \$200,510 - an improvement of \$50,772 (20.21%), which is significant ($p < 0.001$, two-sample, two-tailed T-test). This result confirms the general usefulness of the provided unique human contextual information (the house images) for the human decision-making.

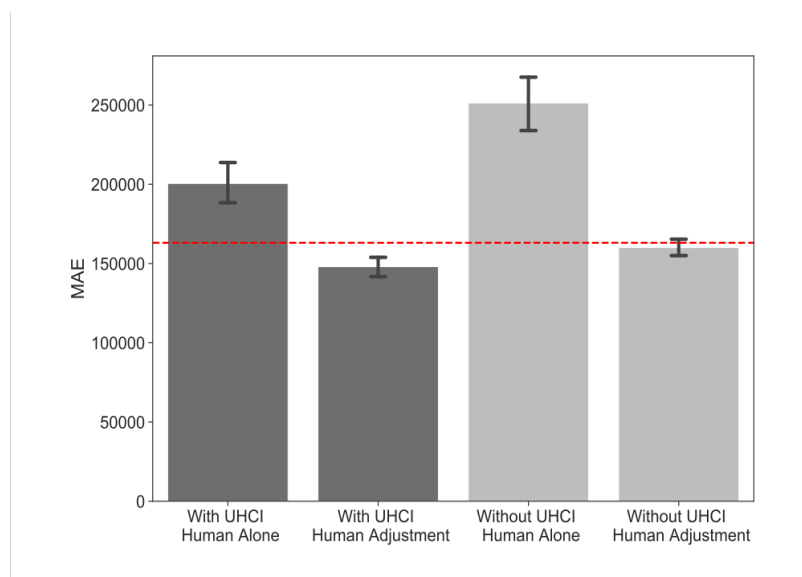


Figure 5.9.: Performance results as MAE of the integration mechanism “human adjustment” across conditions (UHCI = unique human contextual information), including 95% confidence intervals. The red horizontal line denotes the AI performance.

Next, we evaluate the impact of unique human contextual information when the human is teamed up with the AI (human adjustment). Looking at the team performance after adjusting the AI’s prediction, the treatment without UHCI results in an MAE of \$160,095 versus an MAE of \$148,009 in the treatment with UHCI - an improvement of \$12,086 (7.55%), which is significant ($p = 0.04$, two-sample, two-tailed T-test). In both treatments, the human-AI team outperforms the AI (MAE: \$163,080). Whereas the difference between the performance of the human-AI team in the treatment with UHCI is significant ($p < 0.001$, one-sample, two-tailed T-test), the difference in the treatment without UHCI does not constitute a significant improvement ($p = 0.999$, one-sample, two-tailed T-test).

In the next step, we demonstrate the usefulness of the formalization by analyzing the theoretically and realized complementarity potential, including its components, as this allows a more in-depth understanding of the factors that contribute to CTP.

Theoretical complementarity potential. The theoretical complementarity potential ($CP_{theoretical}$) consists of the theoretical inherent ($CP_{theoretical}^{inh}$) as well as the theoretical collaborative ($CP_{theoretical}^{coll}$) complementarity potential. The theoretical inherent complementarity potential refers to the unique knowledge the lower-performing team member can contribute. We observe an increase in unique human knowledge (UHK) in the presence of unique human contextual information that is significantly different. It is \$42,995 in the condition without UHCI and \$61,970 in the condition with UHCI ($p < 0.001$, two-tailed Mann-Whitney U test). This finding can be interpreted as that the images contain useful contextual information for humans that is not accessible to the AI. At the same time, unique AI knowledge decreases (UAIK without UHCI: \$131,196; UAIK with UHCI: \$99,399), which is significant ($p = 0.008$ two-sample, two-tailed T-test). This can be explained by the observation that unique human contextual information also increases human performance on tasks where the AI performs better, which consequentially reduces the unique AI knowledge. To determine the theoretical inherent complementarity potential ($CP_{theoretical}^{inh}$), the unique knowledge attributable to the lower-performing team member must be determined. Since the AI performs better than the human, the theoretical inherent complementarity potential ($CP_{theoretical}^{inh}$) is defined by the unique human knowledge (UHK).

Next, we calculate the theoretical collaborative complementarity potential ($CP_{theoretical}^{coll}$) as the difference between the performance of the better team member (L_{AI}), meaning the total theoretical complementarity potential ($CP_{theoretical}$): without UHCI: \$163,080; with UHCI: \$163,080), minus the theoretical inherent complementarity potential ($CP_{theoretical}^{inh}$). Whereas in the condition without UHCI it results in \$120,085, in the condition with UHCI it takes on a value of \$101,110. The difference is statistically significant ($p < 0.001$, two-tailed Mann-Whitney U test).

Note that it is highly unlikely that the total theoretical complementarity potential will be fully realized, as this would mean that each prediction is exactly correct.

Realized complementarity potential. Now that we have analyzed the impact of unique human contextual information on the theoretical complementarity potential ($CP_{theoretical}$), we focus on the realized complementarity potential ($CP_{realized}$).

We find a significant difference between the realized inherent complementarity potential ($CP_{theoretical}^{inh}$) in both conditions (without UHCI: \$14,468; with UHCI:

\$27,860; $p < 0.001$, two-tailed Mann-Whitney U test), which highlights the potential of contextual information. This absolute increase may be due to the increase in complementarity potential and/or improved integration. Therefore, to investigate this further, we also calculate the relative amount of realized inherent complementarity potential $\left(\frac{CP_{theoretical}^{inh,realized}}{CP_{theoretical}^{inh}}\right)$. This analysis reveals that unique human contextual information not only enhances the theoretically available inherent complementarity potential, but participants can also use significantly more of it (without UHCI: 34%; with UHCI: 45%; $p < 0.001$, two-tailed Mann-Whitney U test). This is a surprising result, as having more information available than the AI could also have detrimental psychological effects. For example, Jussupow et al. (2020) find that perceived AI capabilities and expertise influence aversion towards AI. This can lead to humans taking less account of the AI's suggestions when making their final decision (Longoni et al., 2019), which could result in CTP not being achieved. Our results show that unique human contextual information can however not only increase the potential but also the overall effectiveness of the integration.

Lastly, we analyze the impact of unique human contextual information on realized collaborative complementarity potential ($CP_{theoretical}^{coll}$). The realized collaborative complementarity potential is not significant between both treatments ($CP_{theoretical}^{coll}$): without UHCI: \$-11,483; with UHCI: \$-12,789; $p = 0.999$, two-tailed Mann-Whitney U test). In this context, the positive realized collaborative complementarity potential ($CP_{theoretical}^{coll,positive}$) is \$6,770 in the treatment without UHCI and \$7,731 in the treatment with UHCI. The negative realized collaborative complementarity potential ($CP_{theoretical}^{coll,negative}$) takes on larger values in both treatments (without UHCI: \$18,253; with UHCI: \$20,520), explaining the overall negative realized collaborative complementarity potential. In our experiment, it was expectable that the negative exceeds the positive realized collaborative complementarity potential. Given that the AI in our setup outperforms its human team member, the occurrence of positive realized collaborative complementarity potential can only occur on task instances where the team performance surpasses that of the AI and the human. Conversely, each instance where the human underperforms in comparison to the AI and fails to exactly adapt the AI's decision, amplifies the negative realized collaborative complementarity potential. As humans tend to choose decisions between two boundaries (in our case their own and the AI decision), our setup naturally fosters a negative realized collaborative complementarity potential. The more important point is however that the human can realize the theoretical inherent complementarity potential and even improves in the presence of unique human contextual information, which finally even results in CTP.

Summing the realized inherent complementarity potential and the realized collaborative complementarity potential results in the total realized complementarity potential ($CP_{realized}^{coll}$), which is equivalent to the performance difference between the best individual team member and the joint human-AI team performance (without UHCI: \$2,985; with UHCI: \$15,071).

To summarize, we have shown that unique human contextual information increases the complementarity potential and also improves the human ability to integrate their own and the AI decision. This constitutes a new empirical insight that was only possible to measure due to the granular formalization. Table 5.3 on page 115 summarizes the results of our experiment.

5.6 Discussion

Our research highlights CTP as an increasingly relevant goal of human-AI collaboration. As long as AI still had limited capabilities and was used to merely augment the human in low-stakes decision-making tasks, such as calculating decision-making inputs, CTP had not been a focus. Nowadays AI can however conduct a rising number of tasks independently and is even entering high-stakes decision-making domains such as medicine (McKinney et al., 2020) and law (Hillman, 2019). It also offers general-purpose support fueled through the recent advances in large language models that enable applications like ChatGPT (Bubeck et al., 2023). While this opens up the potential for automating tasks, it also calls for a new form of human-AI collaboration in which AI models and humans are team members that can complement each other. In this work, we provide guidance to reach CTP more consistently.

5.6.1 Contributions

Current research lacks a concise picture of human-AI complementarity, as the majority of studies focusing on human-AI collaboration do not achieve CTP (Bansal et al., 2021; Hemmer et al., 2021). The contributions of our research are threefold and provide guidance on the endeavor to reach CTP. First, we conceptualize human-AI complementarity to facilitate the understanding of the inner workings in human-AI teams that are relevant for achieving CTP. Our conceptualization consists of three parts. We start with a formalization of human-AI complementarity and introduce associated metrics. In addition, we systematically identify information asymmetry

Isolated performance: AI performance
\$163,080

	Potential 1: Inherent complementarity potential		Potential 2: Collaborative complementarity potential		Total potential: Complementarity potential	
Theoretically possible	Theoretical inherent complementarity potential ($CP_{theoretical}^{inh}$)		Theoretical collaborative complementarity potential ($CP_{theoretical}^{coll}$)		Theoretical complementarity potential ($CP_{theoretical}$)	
	without UHCI	with UHCI	without UHCI	with UHCI	without UHCI	with UHCI
	\$42,995	\$61,970	\$120,085	\$101,110	\$163,080	\$163,080
Realized in setting	Realized inherent complementarity potential ($CP_{realized}^{inh}$)		Realized collaborative complementarity potential ($CP_{realized}^{coll}$)		Realized complementarity potential ($CP_{realized}$)	
	without UHCI	with UHCI	without UHCI	with UHCI	without UHCI	with UHCI
	\$14,468	\$27,860	\$-11,483	\$-12,789	\$2,985	\$15,071
Percentage	+34%	+45%	-10%	-13%	+2%	+9%

=	Team performance:	
	without UHCI	with UHCI
	\$160,095	\$148,009

Table 5.3.: Summary of human-AI complementarity results.

and complementary skills as the key sources of complementarity potential. Finally, we structure the options for designing effective human-AI integration mechanisms that do not only exploit the inherent complementarity potential but also allow the creation of collaboratively complementarity potential.

Second, for one instantiation of an integration mechanism—human adjustments—we empirically show that asymmetric information between human and AI can result

in CTP. In this context, we illustrate the application of the developed conceptualization and metrics.

Third, our formalization allows us to reveal an interesting empirical insight about the effect of unique human contextual information on human integration. Our experiment highlights that providing humans with unique contextual information not only affects the theoretical inherent complementarity potential, but also significantly increases the realized inherent complementarity potential. This is an interesting finding because the perception of having more information than the team member could intuitively also worsen collaboration.

5.6.2 Theoretical Implications

Our work advances the discourse on human-AI collaboration (Bansal et al., 2021; Dellermann et al., 2019a) and in particular expands the theoretical groundwork of Fügener et al. (2021). Our conceptualization - the formalization, including operational metrics, sources of complementarity, and the classification of integration mechanisms - provide a basis for future research in human-AI collaboration. In this context, we do not only provide the research community with a formalization for human-AI collaboration but also supply concrete measures that allow investigation into the inner workings of human-AI collaboration on an in-depth level in behavioral experiments.

Providing sources of complementarity potential in a structured way may inspire researchers to look for complementarity beyond arbitrary statements about complementarity capabilities such as human creativity and AI's computational power. The classification of integration mechanisms aids comprehension of the appropriate application of specific instantiations for different scenarios. By employing our proposed concepts, researchers can conduct a more thorough analysis of distinct aspects of human-AI collaboration, namely the theoretical complementarity potential and its realization.

The most important implication of our research is the need to design for CTP, which is influenced by two factors - the complementarity potential and the integration mechanism. Both can and should be designed. The inherent complementarity potential can be influenced by increasing the unique knowledge. From an AI perspective, this could for example be realized by designing “complementary” AIs that are particularly trained in areas of the feature space where humans do not perform well (Hemmer et al., 2021; Mozannar & Sontag, 2020; Wilder et al., 2020).

From a human perspective, humans could be trained to focus on their unique capabilities and build awareness to use unique human contextual information. The integration mechanism needs to be consciously designed to maximize the realized inherent and collaborative complementarity potential.

Our work prepares the ground for the rigorous and fruitful development of design knowledge and artifacts for human-AI complementarity.

5.6.3 Managerial and Political Implications

Our work also has major implications for decision-makers in management and politics. In application areas suitable for human-AI collaboration tasks, such as high-stakes decision-making, responsible managers should focus on developing AI systems that enable the realization of CTP. Competitors might otherwise realize competitive advantages. They need to start collecting data to train complementary AI and invest in training that upskills their workers to improve the inherent complementarity potential of human-AI teams.

Our work provides guidance on where to identify complementarity potential. Our two classifications, one for sources of complementarity potential and one for integration mechanisms, allow an intuitive and easy evaluation of use cases' eligibility for CTP. Current AI endeavors often result in blindly adopting human-AI collaboration due to decision-makers' concerns regarding full automation (Jussupow et al., 2021). It is however worth purposefully designing human-AI collaboration, as it can potentially achieve CTP. Rather than fearing automation, decision-makers should explore the benefits of working with AI. Our research provides valuable guidance to managers, assisting them in determining when and how to collaborate with AI to optimize decision-making and achieve CTP.

Many politicians and researchers warn of the ethical implications of giving AI the final authority in decision-making. Our research shows that in addition to ethical considerations, employing AI without human input does not allow leveraging the potential of human-AI collaboration to reach CTP. Politicians and managers should therefore consider supporting human-AI collaboration as commercially driven companies might only see the short-term benefit in cost reduction through automation while neglecting the long-term potential value of CTP.

5.6.4 Limitations

Our current research design has several limitations that need to be addressed in future work. We conducted our behavioral experiment based on a regression task. Many design elements of our study are therefore closely related to regressions, such as providing quantile information as confidence proxies in AI decisions. In future work, classification as well as generative tasks should be evaluated, especially due to the increase of large language models that enable applications like ChatGPT. In this study, we also focus on a single task, namely real estate appraisal.

Another limitation is the way in which we measure the counterfactual human decision if the user did not receive AI advice. In this work, we conducted a sequential decision-making setup to first measure the human prediction and afterwards the AI-assisted decision. The sequential setup alone might however influence human decision-making. The timing of revealing the AI's recommendation is therefore a critical aspect in the experimental design (Jussupow et al., 2021). By immediately presenting the AI recommendation to the participants, humans might invest reduced cognitive capacities in the task, as the AI already provided a possible answer (Green & Chen, 2019).

5.6.5 Future Research

We see several potential areas of future research regarding human-AI complementarity, which we structure along the formalization, the sources of complementarity potential, and the integration mechanisms.

Future work could apply our concepts and metrics to other domains. Our formalization could also be extended further, for example to a team setting with more than two team members, considering multiple AIs or multiple humans.

Additionally, future work should use the conceptualization and extend the collection of sources. In this work, we highlighted the existence of unique human contextual information as a particular source of complementarity potential. Besides identifying possible further sources, it is worthwhile to identify general criteria when certain information can be considered contextual information. To extend the current understanding of complementarity potential, future work could derive design principles that ensure a sufficient amount of complementarity potential in human-AI collaboration. AI and humans can be trained to increase their complementarity potential. From a human perspective, researchers should investigate how to specifically train people to use unique human contextual information and actively integrate it into

human-AI collaboration. From an AI perspective, research should explore how to train AI that maximizes complementarity potential.

In addition, future work needs to address the realization of theoretical complementarity potential, inherent as well as collaborative. Our work shows that even with unique human contextual information humans only capture 9% of the theoretical complementarity potential. Future research needs to investigate how to improve the integration. With human aggregation, researchers could derive research models and evaluate relevant constructs, such as trust (Söllner et al., 2014) and mental models (Rouse & Morris, 1986), to improve the appropriate reliance of humans in AI decisions.

5.7 Conclusion

human-AI collaboration has predominantly been concerned with AI supporting human users. With an increasing number of tasks that can be automated, meaning they can be solved by AI alone, the focus has however shifted to the purposeful design of the collaboration between humans and AI as team members. The ultimate objective of these teams must be the achievement of complementary team performance (CTP), with the team outperforming each individual team member. The IS community is predestined to drive the development of appropriate theories and lay the foundation for practical applications. We hope that the conceptual base developed in this paper with regard to formalization, sources, and integration mechanisms will provide fruitful ground for future research, and that the empirical studies will illustrate the validity and potential of the human-AI complementarity paradigm.

Part IV

Harnessing Complementarity Potential in
AI-Assisted Decision-Making

Harnessing Complementarity: The Role of Appropriate Reliance

This chapter comprises an article that was published as: Schemmer, M., Kuehl, N., Benz, C., Bartos, A., & Satzger, G. (2023d). Appropriate Reliance on AI Advice: Conceptualization and the Effect of Explanations. *Proceedings of the 28th International Conference on Intelligent User Interfaces*, 410–422. Note: To improve the structure of the work, the title was changed. The abstract has been removed. Tables and figures were reformatted, and newly referenced to fit the structure of the thesis. The terminology was standardized with the dissertation. Chapter, section and research question numbering and respective cross-references were modified. Formatting and reference style was adapted and references were integrated into the overall references section of this thesis.

6.1 Introduction

Most important decisions are made by calling upon advisors. While in the past advice was typically obtained from human experts, nowadays advisors based on artificial intelligence (AI) are becoming increasingly frequent in research and practice (Jung et al., 2018; Jussupow et al., 2021). For example, AI advises medical professionals in breast cancer screening (McKinney et al., 2020), or in loan decisions (Demajo et al., 2020).

In the past, research has predominantly focused on maximizing the reliance (Kerasidou et al., 2021), trust (Siau & Wang, 2018), utilization (Alnowami et al., 2022), compliance (Kühl et al., 2019), or acceptance (Shin, 2021) of AI advice (Jussupow et al., 2021). With certain nuances, all these terms basically describe a concept that aims at maximizing the amount of AI advice that a human decision-maker eventually follows, i.e. maximizing the AI reliance. Recently, a new line of research has emerged that argues that maximising AI reliance does not fully exploit the potential

of state-of-the-art human-AI collaboration (Bansal et al., 2021; Buçinca et al., 2021; Schemmer et al., 2022a; Zhang et al., 2020). We summarize the reasons for this line of thought in three main categories:

Increasing usage of imperfect AI advisors. First, prior research on AI advice often assumed “perfect” advice (Jussupow et al., 2021). This makes sense if one considers narrow application spaces, e.g. performing deterministic algebra. But AI nowadays is used for more complex tasks (Frey & Osborne, 2017) which increases both the number and severity of AI errors. Therefore, generally accepting AI advice would also include the acceptance of incorrect AI advice: Assume that without an AI advisor’s intervention, a physician would have diagnosed cancer, yet in the setting with AI advice had been misled by an incorrect AI advice and, thus, had failed to detect the cancer.

Increasing alignment of the objectives of human decision-makers and AI advisors. Second, in previous research, the goal of human or AI advisors was often inconsistent with the goal of the decision-maker, e.g., financial advisors wanted to persuade a client to purchase certain financial products that would yield the highest financial benefit for their own bonus. In this situation, AI designers do not want customers to differentiate good from bad advice but simply increase acceptance of advice. Today, however, AI advisors are often specifically designed to enhance human decision-making (Lai et al., 2023). The advice seekers can design the advisor based on individual goals and with desired features such as “honest” explainability. This honesty might be missing if the inherent goal of the advisor is different from the advice seeker’s. If the AI is not designed in alignment with human goals, it is often not in the AI designer’s interest to enable humans to critically question the advice they receive.

Increasing potential for complementary team performance. Third, as modern AI is not only more performant and offers more application areas, but also complements humans (Hemmer et al., 2022a; Hemmer et al., 2022b; Lai et al., 2022), there is an increasing potential to achieve complementary team performance (CTP), i.e., a performance that exceeds both—individual AI and human performance (Bansal et al., 2021; Fügener et al., 2021; Hemmer et al., 2022b; Lai et al., 2022; Nguyen et al., 2022). However, this level of performance can only be achieved by exploiting complementary capabilities. Even in the case of superior AI advice, it cannot be achieved simply by accepting AI advice, as it is then tied to AI performance without considering the potential additional strengths of the human.

In conclusion, human decision-makers should not simply rely on AI advice, but should be empowered to differentiate when to rely on AI advice and when to rely

on their own, i.e. they should display *appropriate reliance (AR)* (Bansal et al., 2021; Wang & Yin, 2021; Yang et al., 2020a; Zhang et al., 2020). Despite being a necessary condition for the effective use of AI, current research on AR on AI advice is still very ambiguous with regard to definition, measurement, and impact factors.¹

First, we deal with the ambiguous concept of AR: Current research inconsistently uses the term both for a binary target state (“appropriate reliance is either achieved or not”) and a metric indicating a degree of appropriateness. To clarify this, we introduce a two-dimensional metric—the appropriateness of reliance (AoR)—to describe and measure reliance behavior. It is based on relative frequencies of correctly overriding wrong AI suggestions (self-reliance) and following correct AI suggestions (AI reliance), and reflects a metric understanding of AR. This metric can then be used to define different levels as target states of AR that mark the achievement of particular objectives like certain legal, ethical and performance requirements.

Second, we analyze how the provision of explanations influences AoR and the achievement of AR states. Existing literature is ambiguous with regard to effects of explanations (Alufaisan et al., 2021; Bansal et al., 2021; Wang & Yin, 2021): While in some experiments, explanations support AR (Wang & Yin, 2021; Yang et al., 2020a), in others they cause “blind trust” (Alufaisan et al., 2021; Bansal et al., 2021) in AI advice. To better understand and reconcile conflicting results, we consider additional constructs that may mediate the effect of explanations. More specifically, we hypothesize that explanations do not only influence the information available to the decision-maker, but also have an impact on trust toward AI and on self-confidence.

To test our hypotheses, we conduct a behavioral experiment with 200 participants. Our experiment underscores the advantages of AoR as a metric to examine in detail the factors that lead to changes in overall performance. Moreover, our results help explain the relationship between explanations and AoR by assessing the role of reliance and confidence as mediators, thus mitigating the ambiguity of previous research.

Our work provides researchers with a theoretical foundation of AR within human-AI collaboration research and provides guidance on how to design AI advisors. More specifically, our research contributes to research and practice by defining AR, developing a measurement concept (AoR), and analyzing how explainable

¹It is worth to mention that several studies have examined AR in automation and robotics research (Lee & See, 2004; Wang et al., 2008), but an agreed-upon definition of AR and a respective metric are still missing (Lai et al., 2023; Wang et al., 2008).

AI influences the AoR. Our definition should help researchers to more accurately describe whether they have achieved AR in their experiments. The AoR metric allows to precisely steer the development towards AR. Lastly, our experimental insights can be seen as a starting point for in-depth experimental evaluations of factors impacting AoR.

The remainder of this article is structured as follows: In Section 6.2, we first outline related work on AR in the context of human-AI collaboration. In Section 6.3, we define AR and develop a measurement concept, the AoR, to isolate different possible effects. Subsequently, we derive impact factors on AoR in Section 6.4. In Section 6.5, we describe the design of our behavioral experiment and summarize the results in Section 6.6. In Section 6.7, we discuss our results and provide ideas for future work. Section 6.8 concludes our work.

6.2 Related Work

In the following, we introduce the related work of this article, structured along the topics of appropriate reliance in human advice, automation, and human-AI collaboration as well as the role of explainability.

Appropriate reliance in human advice. Historically, the use of (human) advice is generally discussed in the so-called judge-advisor system (JAS) research stream (Harvey & Fischer, 1997; Sniezek & Van Swol, 2001; Yaniv, 2004). The term “judge” refers to the decision-maker who receives the advice and must decide what to do with it (Bonaccio & Dalal, 2006). The judge is the person responsible for making the final decision. The “advisor” is the source of the advice (Bonaccio & Dalal, 2006). The research stream mainly focuses on advice acceptance.²

Appropriate reliance in automation. In contrast, many researchers have worked on AR with regard to automation (Lee & See, 2004) and robotics (Talone, 2019). In the following, we will provide an overview of the most common definitions. Fundamental work in the context of AR in automation has been laid by Lee and See (2004). The authors outline the relationship between “appropriate trust” and AR in their work. However, they do not define AR explicitly but provide examples of inappropriate reliance, such as “misuse and disuse are two examples of inappropriate reliance on automation that can compromise safety and profitability” (Lee & See, 2004, p. 50). Wang et al. (2008) define appropriate reliance as the impact of reliance

²In this article, we use the term advice acceptance as a generic term to describe the behavior of following AI advice regardless of its quality.

on performance. For example, they discuss the situation in which automation reaches a reliability of 99%, and the human performance is 50%. In their opinion, it would be appropriate to always rely on automation as this would increase performance. Talone (2019) follows the work by Wang et al. (2008) and defines AR as “the pattern of reliance behavior(s) that is most likely to result in the best human-automation team performance” (Talone, 2019, p. 13). Both see AR as a function of human-automation team performance.

Appropriate reliance in Human-AI Collaboration. Recent work in human-AI collaboration has started to discuss AR in the context of AI advice. Lai et al. (2023) give an overview of empirical studies that analyze AI advice considering AR. For example, Arjun et al. (2018) analyze whether humans can learn to predict the AI’s behavior. This ability is associated with an improved ability to appropriately rely on AI predictions. Moreover, Gonzalez et al. (2020) measure the acceptance of incorrect and correct AI advice. Similarly, Poursabzi-Sangdeh et al. (2021, p. 1) point out the idea of AR in the form of “making people more closely follow a model’s predictions when it is beneficial for them to do so or enabling them to detect when a model has made a mistake”. However, the authors do not explicitly relate this idea to the concept of AR. In this context, additional work uses the term “appropriate trust” with a similar interpretation as the behavior to follow “the fraction of tasks where participants used the model’s prediction when the model was correct and did not use the model’s prediction when the model was wrong” (Wang & Yin, 2021, p. 323). Finally, also Yang et al. (2020a, p. 190) define “appropriate trust is to [not] follow an [in]correct recommendation”. All these articles have in common that they consider AR or appropriate trust on a case-by-case basis. Bussone et al. (2015) assess how explanations impact trust and reliance on clinical decision support systems. The authors divide reliance into over- and self-reliance as part of their study. They use a qualitative approach to answer their research questions. Chiang and Yin (2021) evaluate the impact of tutorials on AR and measure AR through team performance.

The related work highlights that current research inconsistently uses the term both for a binary target state (“AR is either achieved or not”) and a metric indicating a degree of appropriateness. Additionally, previous research does not provide a unified measurement concept that allows measuring the degree of appropriateness (Lai et al., 2023).

Explainable AI and appropriate reliance. Most researchers that studied AR so far have proposed to use explanations of AI as a means for AR (Adadi & Berrada, 2018; Lai & Tan, 2019; Zhang et al., 2020). We refer to AI that generates explanations as

explainable AI (XAI). Explanations can be differentiated in terms of their scope, i.e., being global or local explanations (Adadi & Berrada, 2018). Global XAI techniques address holistic explanations of the models as a whole. In contrast, local explanations work on an individual instance basis. Besides the scope, XAI techniques can also be differentiated with regard to being model specific or model agnostic, i.e., whether they can be used with all kinds of models (Adadi & Berrada, 2018). The most commonly used model agnostic technique is feature importance (Lundberg et al., 2018; Lundberg & Lee, 2017). Feature importance can be used to generate saliency maps for computer vision tasks or highlight important words for text classification.

Table 6.1.: Related work on explainable AI (XAI) and appropriate reliance (AR).

Study	AR Metric	Independent variable	Effect of XAI on AR
Bansal et al. (2021)	Accuracy on correct or incorrect AI advice	Local feature importance	Negative
		Adaptive explanations based on AI confidence	Positive
Buçinca et al. (2021)	Ratio of reliance on incorrect AI advice	Local feature importance	Negative
Jakubik et al. (2023)	Ratio of reliance on correct or incorrect AI advice	Local feature importance	No effect
		Predictive outcomes	Negative
Wang and Yin (2021)	Accuracy on correct or incorrect AI advice	Global feature importance	No effect
		Local feature importance	Positive
		Examples	No effect
		Counterfactuals	No effect
Yang et al. (2020a)	Ratio of reliance on correct or incorrect AI advice	Local feature importance	Positive

Several studies have evaluated whether different types of explanations can support humans' understanding of the AI model with the goal of better relying on recommendations in the correct cases (Alufaisan et al., 2021; Buçinca et al., 2021; Carton et al., 2020; Van der Waa et al., 2021). However, it has also been shown that some types of explanations can lead people to rely too much on the AI's recommendation, especially in cases where the AI advice is wrong (Bansal et al., 2021; Poursabzi-Sangdeh et al., 2021; Schemmer et al., 2022c). In Table 6.1, we provide a comprehensive overview of the results that were found in the current studies on AR in XAI-assisted decision-making. Overall, we find mixed results regarding the effect of explanations.

To sum it up, related work is missing a precise definition of AR, a unified measurement concept, and a precise understanding of when and why explanations of AI advisors influence AR.

6.3 Conceptualization of Appropriate Reliance

Although several studies have examined AR in automation and robotics research (Lee & See, 2004; Wang et al., 2008), an agreed-upon definition of AR and a respective metric are still missing (Lai et al., 2023; Wang et al., 2008). We, therefore, initiate our research by deriving a definition of AR and a corresponding metric. To do so, we first discuss the terms reliance and appropriateness. Following that, we derive our metric and lastly define AR.

6.3.1 Reliance and Appropriateness

Reliance itself is defined as a behavior (Dzindolet et al., 2003; Lee & See, 2004). This means it is neither a feeling nor an attitude but the actual action conducted. This means reliance is directly observable. Scharowski et al. define reliance in the AI advisor context as “user’s behavior that follows from the advice of the system” (Scharowski et al., 2022, p. 3). Defining reliance as behavior also clarifies the role of trust in this context, which is defined as “the attitude that an agent will help achieve an individual’s goals in a situation characterized by uncertainty and vulnerability” (Lee and See, 2004, p. 51). In general, research has shown that trust in AI increases reliance, but reliance can also take place without trust being present (Lee & See, 2004). For example, we might not trust the bank advisor but consciously decide that following the advice is still the best possible decision. The other way around, we could also generally trust an advisor, but consciously decide that the given advice is not correct in a particular situation. Finally, reliance is influenced beyond trust by other attitudes such as perceived risk or self-confidence (Riley, 2018).

After establishing a common understanding of reliance, we proceed by defining “appropriateness”. Appropriateness depends on different types of AI errors. Current AI is imperfect, i.e., it may provide erroneous advice. This erroneous advice can be divided into systematic errors and random errors (Talone, 2019). While humans can identify systematic errors, random errors have no identifiable patterns and can not be distinguished. These different types of errors allow differentiation between two cases. If all errors are random and cannot be detected, then, from a performance

perspective, humans should always rely on the AI's advice when the AI performs better on average, and never when the AI performs worse on average (Talone, 2019). However, suppose there are some systematic errors. In that case, humans might be able to differentiate between correct and incorrect advice, which may even result in superior performance, i.e. complementary team performance (CTP) (Hemmer et al., 2021), compared to a scenario in which AI and humans conduct the task independently of each other.

This changes the overall discrimination to a case-by-case discrimination. In the presence of systematic errors, humans should evaluate each case individually. Since the solution approach in the presence of just random errors is relatively simple, as pointed out above, in this article, we focus on the more complicated setting when a significant proportion of task instances exhibit systematic errors. After having defined the terms reliance and appropriateness in the following, we derive a metric.

6.3.2 Towards a Measurement Concept—Appropriateness of Reliance

Appropriateness is often measured by the percentage in which the decision-maker relies on correct AI advice and the percentage in which the decision-maker does not rely on incorrect AI advice (Bansal et al., 2021; Gonzalez et al., 2020). The major disadvantage of this measurement is that we cannot know whether the reliance on correct AI advice stems from a correct discrimination or simply an overlap of the human and the AI's decision, i.e. instances where the AI advisor just confirms the human (Tejeda et al., 2022). Especially from an ethical point of view, it makes a major difference whether the final decision is incorrect because an AI advisor "convinced" a human decision-maker to accept an incorrect AI advice or whether the human decision-maker would also not have been competent to solve it alone.

Therefore, to measure the degree of appropriateness in a more narrow sense, we follow the approach of the JAS paradigm and include an initial human decision (Sniezek & Van Swol, 2001). This approach requires participants to make a decision, receive advice, and then make a second, potentially revised decision. In general, if we do not consider the initial human decision, information about the human discrimination ability, including the consequent action, gets lost—it is not traceable how the human would have decided without the AI advice. Nevertheless, especially this interaction needs to be documented to research AR holistically.

We use a simple discrete decision case to highlight the different possible outcomes of reliance. Note that for simplicity, we refer to classification problems. However, the measurement concept can be extended to regression problems as well (see for example Petropoulos et al. (2016)). We focus on a single task instance perspective and consider a sequential decision process which can be described as follows: First, the human makes a decision, then receives AI advice. Second, the human is asked to update the initial decision, i.e., either adopt or overwrite the AI advice. This allows measuring appropriateness in a fine-granular way. Figure 6.1 highlights the different combinations. Four of the eight combinations are cases where the human’s initial decision and the AI’s advice are the same, i.e. the AI confirms the human’s decision. In our reliance measure, we exclude these confirmation cases for two reasons: First, if the same decision is made in all three steps, it is impossible to objectively measure whether AI or self-reliance was present. Second, if the final decision differs from the advice and the initial human decision, it is questionable whether we can speak of a reliance outcome. For example, if both the human and the AI are initially incorrect, then arriving at a correct final decision is less a matter of reliance than of human-AI collaboration. While these cases of collaboration are relevant to CTP, they are beyond the scope of our work. Therefore, we arrive at four different reliance outcomes, which we present below.

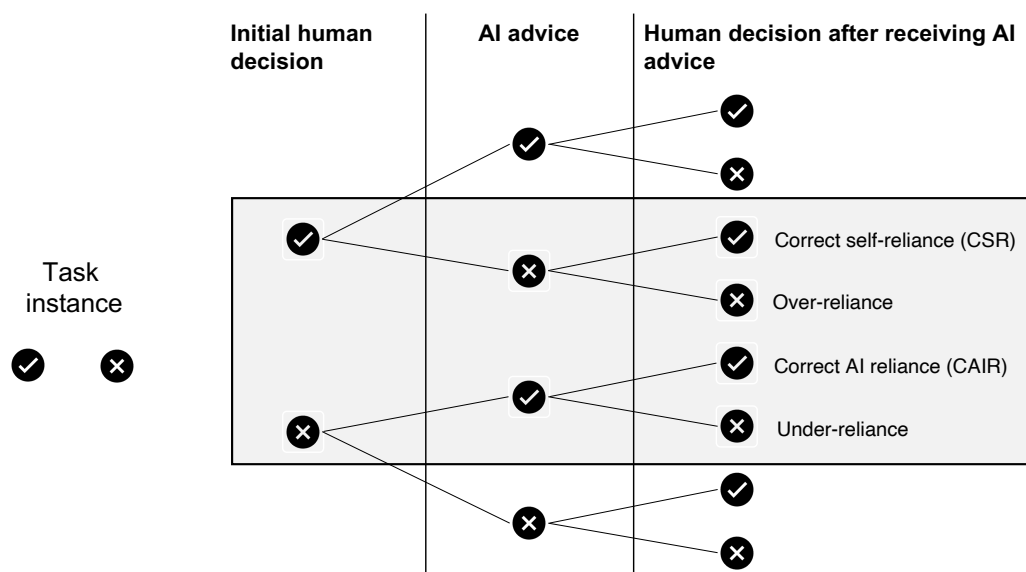


Figure 6.1.: Combinatorics of initial human decisions, AI advice and human reliance for a single task instance in a sequential task setting.

On a high level, we can cluster those four outcomes into either AI or self-reliance. *AI reliance* refers to cases in which the initial decision-maker’s decision is different from the AI advisor and the decision-maker relies on the AI’s advice. Likewise, *self-reliance*

refers to cases in which the decision-maker is different from the AI advisor but finally relies on themselves. On a more detailed level, we can further differentiate whether the final decision is correct or incorrect which leaves us with the following four reliance outcomes: First, *correct AI reliance (CAIR)*, which describes the case when the human is initially incorrect, receives correct advice, and relies on that advice. Second, the case in which the human relies on the initial incorrect decision and neglects correct AI advice. This is denoted as *incorrect self-reliance* or under-reliance. Third, if the human is initially correct and receives incorrect advice, this can either result in *correct self-reliance (correct self-reliance (CSR))*, i.e., neglecting the incorrect AI advice, or relying on it, which is denoted as *incorrect AI reliance* or over-reliance.³ Based on these cases, we propose a two-dimensional metric that transfers the instance perspective on a measurement for multiple task instances.

Let us consider a prediction task $T = \{X_i, y_i\}_i^N$ as a set of N instances $x_i \in X$ with a corresponding ground truth label $y_i \in Y$. On the first dimension, we calculate the ratio of the number of cases in which humans rely on correct AI advice and the decision was initially not correct, i.e., in which humans rightfully change their mind to follow the correct advice.

$$\text{Relative AI reliance (RAIR)} = \frac{\sum_{i=0}^N CAIR_i}{\sum_{i=0}^N CA_i}$$

$CAIR_i$ is one if, in this particular case, the original human decision was wrong, the AI recommendation was correct and the human decision after receiving the AI recommendation is correct, and zero otherwise. CA_i is one if the original human decision was wrong and the AI advice was correct, regardless of the final human decision, and zero otherwise. On the second dimension, we propose to measure the relative amount of correct self-reliance in the presence of incorrect AI advice.

$$\text{Relative self-reliance (RSR)} = \frac{\sum_{i=0}^N CSR_i}{\sum_{i=0}^N IA_i}$$

CSR_i is one if on this particular instance the initial human decision was correct, the AI advice was incorrect and the human decision after receiving AI advice is correct. IA_i is one, if the initial human decision for a task instance i was correct and the AI advice was incorrect.

³This also clarifies our definition of over- and under-reliance. Both are errors on a task instance level, when humans do not rely appropriately.

Figure 6.2 highlights both dimensions. On the x-axis, we depict the relative AI reliance (*RAIR*), and on the y-axis, the relative self-reliance (*RSR*). The figure highlights the properties of the measurement concept. It ranges between 0 and 1 along both dimensions. We call the tuple of *RAIR* and *RSR* *appropriateness of reliance* (AoR).

$$\text{Appropriateness of Reliance (AoR)} = (RSR; RAIR)$$

We refer to the theoretical goal of having a *RSR* and a *RAIR* metric of “1” as *optimal AoR*. Most likely, this theoretical goal will not be reached in any practical context as humans will not always be able to perfectly discriminate on a case-by-case basis whether they should rely on AI advice. Furthermore, random errors will reduce AoR as they cannot be discriminated against. Therefore, optimal AoR will most likely be a theoretical goal.

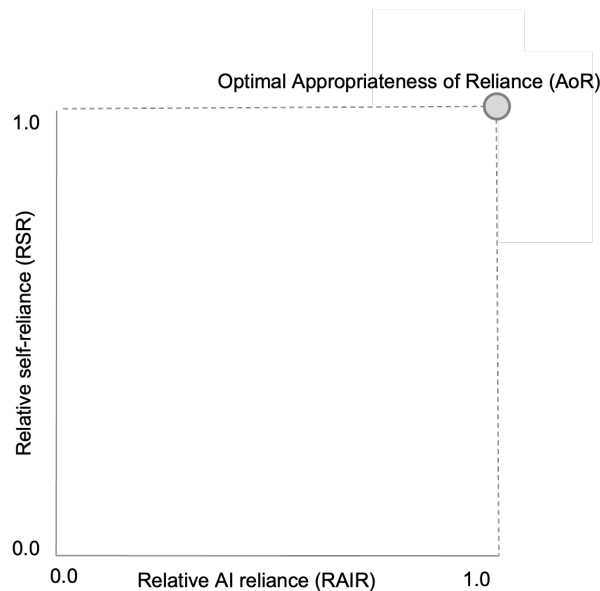


Figure 6.2.: Two-dimensional depiction of appropriateness of reliance (AoR).

6.3.3 Definition of Appropriate Reliance

So far, we have discussed how to measure AoR and the theoretical upper boundaries of AoR. The challenge, however, is to define the level of *RAIR* and *RSR* that constitutes AR. In our work, we take an objective-oriented perspective and propose to define AR individually in the context of the task. Dependent on the task, different levels of *RAIR* and *RSR* might be appropriate. In this work, we focus on AR with

respect to performance (P) following Talone (2019) and Wang and Yin (2021).⁴ Thus, we define AR with regard to CTP:

$$\text{Appropriate Reliance}(AR) = \begin{cases} 1, & \text{if } P_{H\&AI} > \max(P_H, P_{AI}) \\ 0, & \text{otherwise} \end{cases}$$

With $P_{H\&AI}$ being the performance after receiving AI advice, P_H the individual human performance and P_{AI} the individual AI performance. Essentially, this means any tuple of $RAIR$ and RSR that leads to CTP is considered AR.

In summary, we derived a metric (AoR) and defined AR. In the next section, we will use this foundation to derive a research model to analyze the impact of AI recommendation explanations on AoR.

6.4 Theory Development And Hypotheses

With the measurement concept (AoR) at hand, we now develop hypotheses on the effect of explanations on AoR. Research has already investigated in-depth the effect of explanations on AI advice acceptance (Shin, 2021). However, research is missing theoretical and empirical evidence on how explanations influence AoR (Bansal et al., 2022). Thus, our work contributes a research model that is evaluated applying the AoR concept.

As a dependent variable, we use the before-defined two dimensions of AoR—namely $RAIR$ and RSR . We believe that both dimensions need to be treated differently as there are inherent differences between the $RAIR$ and the RSR . The $RAIR$ essentially deals with cases where the human is initially incorrect, gets correct advice, and relies on this advice. In contrast, RSR focuses on cases where the human is initially correct and receives incorrect advice, which is then rightfully ignored. Thus, we formulate different hypotheses for the $RAIR$ and the RSR .

The most central impact factor in the assessment of the AI advisor might be the explanations of its recommendations. These explanations should provide insights into the AI's thought process. In the presence of incorrect advice, explanations might enable the human to detect whether the advice is incorrect, for example, by validating if the AI advisor violates some universal axioms of the task. Similarly,

⁴Note that besides this performance perspective, appropriateness could also be discussed from an ethical perspective. Since in high-stake decision-making humans have an oversight responsibility (Schoeffler et al., 2022a).

(Bansal et al., 2021) hypothesize that if explanations do not “make sense”, humans will reject the AI advice. However, on the other hand, sometimes explanations are rather interpreted as a general sign of competence and thereby increase over-reliance (Buçinca et al., 2021). Which effect exceeds the other is unclear. Therefore, we hypothesize, without specifying a direction of the effect, that the explanations have a general effect on the *RSR*:

H1a: *Providing explanations of the AI advisor influences the relative self-reliance (RSR).*

The second effect of explanations on AoR is through the *RAIR*. Essentially, the *RAIR* measures the percentage of times decision-makers follow the correct advice after initially being wrong about the task instance. This means they do not have enough domain knowledge to solve the task on their own. Thus, to increase the *RAIR*, human decision-makers need to extend their knowledge and simultaneously validate whether this knowledge extension makes sense. Here, explanations are needed to first get inspired to derive new knowledge and second to validate the knowledge. Figure 6.3 on page 136 shows an illustrative example based on an animal classification task. Imagine that a child has just seen big dogs and then sees a very small dog. It might think that this animal is something else, like a rat. Next, the child receives AI advice that says the animal it sees is a dog, and provides additional justification by highlighting the part of the image that led to the AI’s decision. Now, the child’s first task is to figure out whether this advice makes sense in general, while building the knowledge base. In this illustrative case, it might understand that the animal has characteristics of a dog, but is only smaller and therefore relies on the AI, thereby increasing the *RAIR*. In the presence of correct advice, explanations might point humans towards new patterns they have not seen before and help discriminate these knowledge extensions. In the presence of correct advice, the convincing element of explanations would not have a negative effect as a higher overall reliance on the AI advice would simply increase the *RAIR*. We therefore hypothesize:

H1b: *Providing explanations of the AI advisor increases the relative AI reliance (RAIR).*

Beyond the provisioning of additional information, explanations might change the attitude toward the AI advisor. In 1992, Lee and Moray (1992), already discussed the influence of self-confidence and trust as predominant attitudes for reliance decisions in the context of automation. Therefore, in the following, we discuss potential impacts on trust and the change in self-confidence induced through explanations and their impact on AoR.

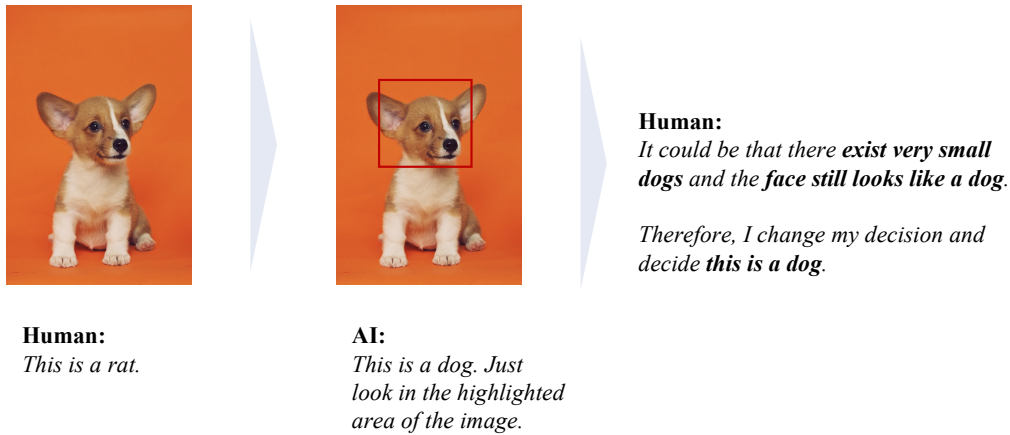


Figure 6.3.: Illustrative example of how humans can increase their relative AI reliance (*RAIR*) based on explanations of the AI advice.

Confidence is defined as a person's degree of belief that their own decision is correct (Peterson & Pitz, 1988; Zarnoth & Snizek, 1997). Confidence in one's own decision is a key mechanism underlying advice acceptance (Chong et al., 2022; Wang & Du, 2018). So far, most research has discussed the influence of static human self-confidence on AoR, i.e. a confidence in doing the task instance without any advisor (Chong et al., 2022). However, we hypothesize that the absolute level of human confidence actually plays a minor role in comparison to the change of confidence after seeing the AI advice as it essentially reflects the combination of a self-assessment and the AI advisor's assessment. Explanations should not influence the initial human self-confidence but the confidence level after seeing the AI advice. Thus, we hypothesize:

H2 *Providing explanations of the AI advisors increases the change in self-confidence.*

We hypothesize that this change in self-confidence positively correlates with the discrimination capability and thus should increase the *RSR* and *RAIR*.

H3a *An increased change in human self-confidence increases the relative self-reliance (*RSR*).*

H3b: *An increased change in human self-confidence increases the relative AI reliance (*RAIR*).*

We hypothesize that also trust in the AI advisor influences AoR. There are different levels of trust, e.g. trust in AI in general, trust in a specific AI advisor (Jacovi et al., 2021) and some researchers even refer to the case-by-case discrimination as trust on a task instance level (Wang & Yin, 2021). In this work, we focus on the specific trust in our developed AI advisor. In general, trust is a complex, multidisciplinary construct with roots in diverse fields such as psychology, management and information systems

(Niese & Adya, 2022). In our study, we define trust as a belief in the integrity, benevolence, trustworthiness, and predictability of the AI advisor following (Crosby et al., 1990; Doney & Cannon, 1997; Ganesan, 1994; McKnight et al., 2002). Understanding is crucial in building trust (Gilpin et al., 2019). Psychological research shows that in general explanations of humans increase trust (Koehler, 1991). Thus, we hypothesize:

H4 *Providing explanations of the AI advisor increases trust in the AI advisor.*

Trust influences reliance, but does not fully determine it (Lee & See, 2004). When people show a high level of trust in the advisor, they consider the advice to be high-quality advice from an advisor with good intentions, and they will give more weight to that advice (Wang & Du, 2018). In general, this should increase the acceptance of AI consulting. For systems' most effective use, however, users must appropriately trust AI advisors (Lee & See, 2004). Trust should be calibrated and match the AI's capabilities (Lee & See, 2004). Insufficient trust is called distrust and when trust exceeds capability it is called over-trust (Lee & See, 2004). Research on automation has shown that over-trust can result in over-reliance on automation (Bailey & Scerbo, 2007; Goddard et al., 2012; Parasuraman et al., 1993) and, therefore, should decrease the *RSR*. Thus, we hypothesize:

H5a: *Trust decreases the relative self-reliance (RSR).*

Since trust increases reliance it should also increase the *RAIR*. Therefore, we hypothesize:

H5b: *Trust increases the relative AI reliance (RAIR).*

Figure 6.4 highlights all our hypotheses and combines them into one integrated research model.

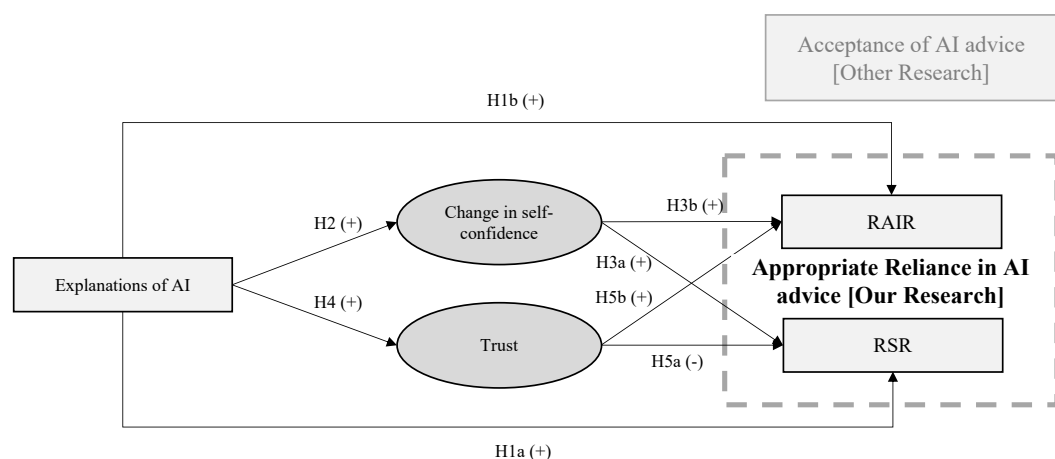


Figure 6.4.: Research model on the effect of explanations on appropriateness of reliance (AoR).

6.5 Experimental Design

In this section, we present our study task, the AI model, and the corresponding explanations. We then explain the study procedure and measurements.

6.5.1 Task, Model, and Explanations

As an experimental task, we have chosen a deceptive hotel review classification. Humans have to differentiate whether a given hotel review is deceptive or genuine. Ott et al. (2013), Ott et al. (2011) provide the research community with a data set of 400 deceptive and 400 genuine hotel reviews. The deceptive ones were created by crowd-workers, resulting in corresponding ground truth labels.

The implemented AI advisor is based on a Support Vector Machine with an accuracy of 86%, which is a performance that is similar to the performance in related literature (Lai et al., 2020). For the explanations, we use a state-of-the-art explanation technique, LIME feature importance explanations (Ribeiro et al., 2016b), as it is the most common one for textual data. Feature importance aims to explain the influence of an independent variable on the AI's decision in the form of a numerical value. Since we deal with textual data, a common technique to display the values is to highlight the respective words according to their computed influence on the AI's decision (Lai et al., 2020). We additionally provide information on the direction of the effect and differentiate the values into three effect sizes following the implementation of Lai et al. (2020) (see step 2 in Figure 6.5).

6.5.2 Study Design and Procedure

The research model is tested in an online experiment with a between-subject design. We tested two different conditions. First, a *control* condition in which the human receives AI advice without feature importance explanations and second, a *feature importance* condition.

In each condition, participants are provided with 16 reviews. We incorporate an advanced sampling strategy to isolate the effects of discriminating AI advice. We have a test set of 32 reviews to which we apply stratified sampling and select four reviews of each class of a confusion matrix (True Positive, False Positive, True Negative, False Negative), two with a positive sentiment and two with a negative

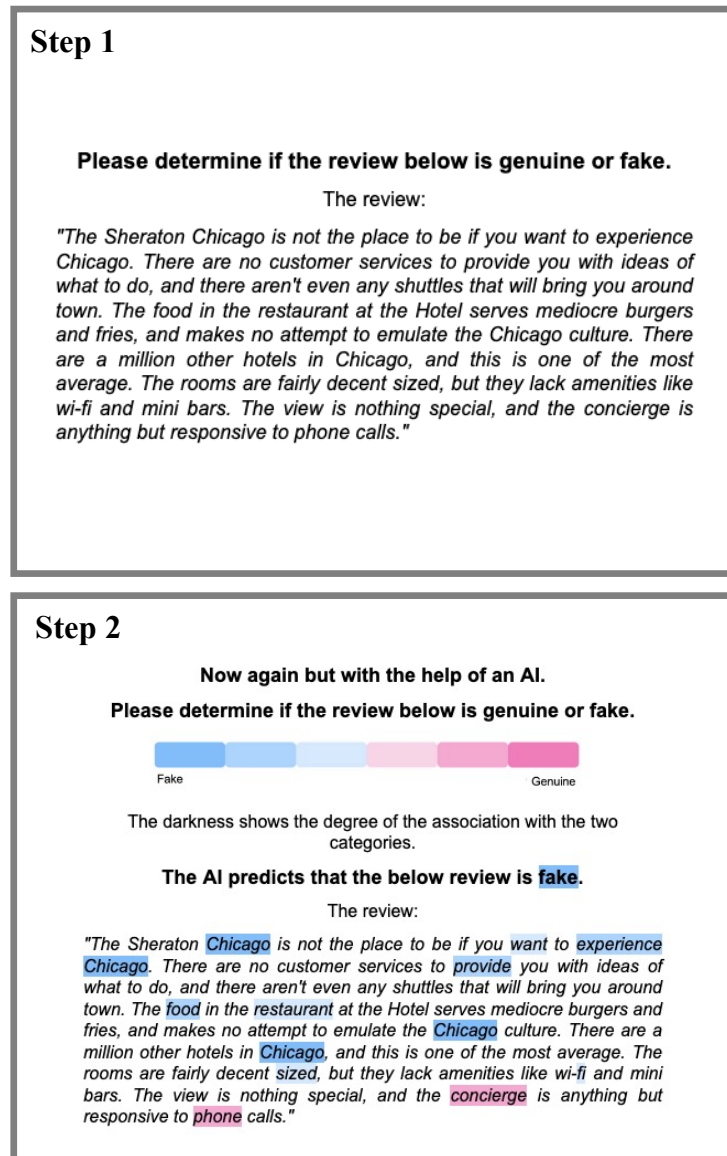


Figure 6.5.: Online experiment graphical user interface for the feature importance condition. The ground truth of the exemplarily shown hotel review is “fake”. The design of the interface is adapted from Lai et al. (2020).

sentiment. This approach allows us to ensure a high-performing AI that should provide good explanations but also the potential for incorrect AI advice.

Task flow. The online experiment is initiated with an attention control question that asks participants to state the color of grass. To control for internal validity, participants are randomly assigned to the condition groups. Then, both condition groups receive an introduction to the task and either AI alone or AI including feature importance. We provide the participants with a general intuition of the AI but not

with specific performance information. Then, the participants conduct two training tasks, to familiarize the participants with the task and the AI and, depending on the condition, with its explanations. Additionally, the participants receive feedback on the training tasks. After the two training reviews, the participants are provided with the 16 main tasks. For the AoR measurement concept, sequential task processing is essential. In our study, this means the human first receives a review without any AI advice, i.e., just plain text, and classifies whether the review is deceptive or genuine (see step 1 in Figure 6.5 on page 139). Then the participant is asked to classify the review and provide a confidence rating. Following that, the human either receives a simple AI advice statement, e.g. “the AI predicts that the review is fake” or the AI advice and additional explanations (see step 2 in Figure 6.5 on page 139). After receiving the AI advice the participant is able to change the initial decision and provide a new self-confidence assessment. This sequential two-step decision-making allows us to measure AoR. During the main tasks, the participants do not receive feedback on their performance. After classifying the hotel reviews, we collect data on trust and demographic variables.

Reward. To incentivize the participants, they were informed that for every correct decision, they get an additional 12 Cents in addition to a base payment of 5.83 Euro. Hereby, the two training classifications do not count for the final evaluation.

Participant information. The participants are recruited using the platform “Prolific.co”. We note that crowd workers might limit the generalizability of our results. However, deception detection of digital information is often done in online communities. Future work could analyze the effects in professional deception detection screening services. In total, we conducted the experiment with 200 participants. We excluded one participant in the feature importance condition because of a failed manipulation check. Table 6.2 shows the age, gender, and education distribution of the participants.

6.5.3 Evaluation Measures

To measure AoR, we use the upfront derived measurements *RAIR* and *RSR*. We measure the *change in self-confidence* for all task instances per participant:

$$\text{Change in self-confidence} = \sum_{i=1}^{16} \text{Conf}_{H2,i} - \text{Conf}_{H,i}$$

Table 6.2.: Summary of participants' characteristics.

Number per condition	Control = 100 Feature importance = 99
Age	$\mu = 27.5, \sigma = 8.5$
Gender	46 % Male 54 % Female
Education	32 % High school 38 % Bachelor 14 % Master 16 % Other

$Conf_H$ refers to the human self-confidence when doing the task instance alone and $Conf_{H2}$ to the human self-confidence after receiving AI advice. Both are measured with a 7-point Likert scale (“How confident do you feel in your decision?”). We measure the trust in the AI advisor as a subjective latent construct with a 7-point Likert scale. We use four items based on (Crosby et al., 1990; Doney & Cannon, 1997; Ganesan, 1994; Gefen et al., 2003). The items were: “I think I can trust the AI.”, “The AI can be trusted to provide reliable support.”, “I trust the AI to keep my best interests in mind.” and “In my opinion, the AI is trustworthy.”. Cronbach’s Alpha was 0.89 (high). The original scales were validated. As both classes are equally distributed, task performance was measured by the percentage of correctly classified images, i.e., accuracy. To measure the *human accuracy*, we calculated this measure for both conditions based on the initial human decision across all 16 task instances. Furthermore, we calculate the *AI-assisted accuracy* based on the revised human decision after receiving AI advice.

6.6 Results

In the following, we present the results of our behavioral experiment. We start by presenting descriptive results, followed by the results with respect to AoR and AR. Following that, we analyze our full research model, including mediations, by applying structural equation modelling (SEM).

6.6.1 Descriptive Analysis

Descriptive results of our study can be found in Table 6.3. They are split according to the experimental condition. We evaluate the significance of the results using t-tests after controlling for normality. The participants' *RAIR* is significantly higher in the explanation group compared to the control group ($t = -1.95, p = 0.05$). The change in confidence is statistically significant ($t = -2.33, p = 0.02$) which means that on average people feel more confident after receiving AI advice including explanations. Neither AI-assisted nor human accuracy is significantly different between conditions. However, the difference between the human and AI-assisted performance is significant (feature importance condition = 2.45 pp; control condition = -1.56 pp ; $t = 2.29, p = 0.02$) which means that explanations not only improve the *RAIR* but also as a consequence the overall performance.

Table 6.3.: Descriptive results.

Condition	<i>RAIR</i>	<i>RSR</i>	Trust (SD)	Change in Self- Confidence (SD)	AI-assisted Accuracy	Human Accuracy
Control	29.59 %	71.87%	4.45 (1.17)	0.19 (0.42)	53.94 %	55.50 %
Feature Importance	38.87%	69.45 %	4.4 (1.18)	0.06 (0.35)	56.30 %	53.85 %

6.6.2 Appropriateness of Reliance & Appropriate Reliance

We depict our AoR results of the experiment in Figure 6.6 on page 143. They highlight in the control condition a high *RSR* of 71.87% ($\pm 3pp$) and a relatively low *RAIR* of 29.59% ($\pm 3pp$). This indicates that humans in the setting were able to differentiate between wrong AI advice and self-rely to a high degree. The *RAIR* of 29.59% shows that we can observe a severe share of under-reliance on AI.

In the XAI condition, we can observe a significant increase ($t = -1.95, p = 0.05$) in *RAIR* from 29.59% ($\pm 3pp$) to 38.87% ($\pm 3pp$) while the *RSR* does not change significantly (71.87% $\pm 3pp$ for the control condition and 69.45% $\pm 3pp$ for the feature importance condition, $t = 0.61, p = 0.54$). This means explanations of AI decisions can reduce the share of under-reliance. It is important to highlight that *RAIR* is not increased simply by relying more often on AI advice, as this would have also reduced

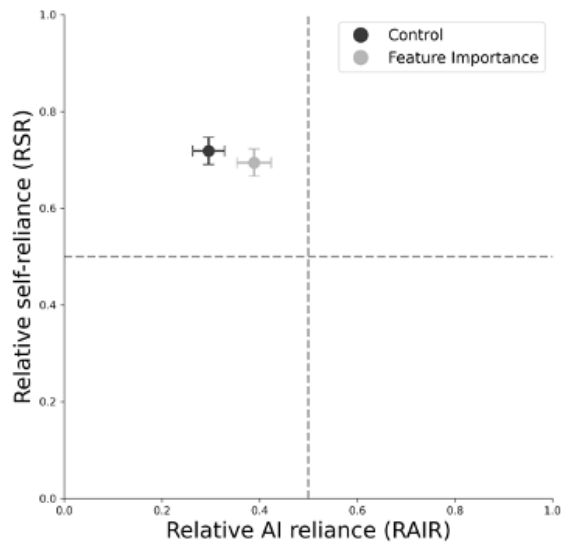


Figure 6.6.: Illustration of appropriateness of reliance (AoR) including standard errors. Explanations increase the *RAIR* significantly. Differences in *RSR* are not significant.

the *RSR* significantly. Thus, our experiment indicates that feature importance on textual data can have a positive effect on human-AI collaboration.

Following our AR definition, to evaluate whether the participants display AR, we need to calculate whether we reached CTP. Therefore, we compare the individual human and AI performance with the human-AI team performance. The down-sampled AI performance is 50 % for both conditions. The human accuracy varies depending on the condition. The human-AI team performance is not significantly different from the human accuracy which means we do not reach CTP and therefore AR is not displayed. Further means would be necessary to reach AR.

6.6.3 Structural Equation Model

In addition to analyzing the direct effect of explanations on the *RAIR* and *RSR*, we use SEM analysis to test our hypothesized research model. Before fitting our SEM, we conducted missing data analysis, outlier detection, a test for normality, and the selection of an appropriate estimator. We observe no missing data and no outliers. However, one participant failed our attention check leaving us with a final sample size of 199 for both conditions. Shapiro’s test for normality indicates that several variables of interest deviate significantly from normal distributions. As a result, we

conducted the analysis with an estimator that allows for robust standard errors and scaled test statistics (Kunkel et al., 2019). Therefore, we use the MLR estimator (Lai, 2018).

Table 6.4.: Structural equation model fitting index using a chi-squared test (χ^2), root mean square error of approximation (RMSEA), Comparative Fit Index (CFI), Tucker-Lewis Index (TLI), and Standardized Root Mean Squared Residual (SRMR).

	χ^2	RMSEA	CFI	TLI	SRMR
Measurement criteria (Bentler, 1990; Hu & Bentler, 1999)	> 0.05	< 0.05	> 0.96	> 0.95	< 0.08
Value	0.083	0	1	1.02	0.02

Our dependent variables are the *RAIR* and *RSR*. Since these dependent variables are between 0 and 1, we employed a logistic model in the Lavaan package, version 0.6-9, in R (Rosseel, 2012). This model has an excellent overall fit (see Table 6.4). The results for each independent variable are discussed below and visualized in Figure 6.7.

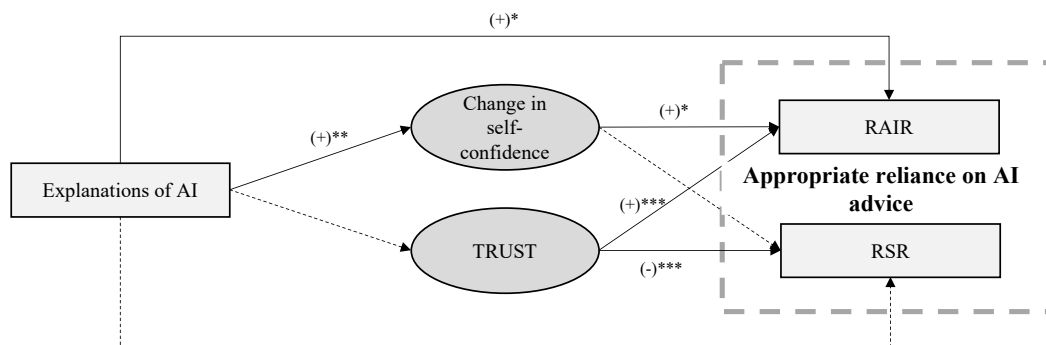


Figure 6.7.: Structural equation modeling results. Significance: $*** p < 0.01$, $** p < 0.05$, $* p < 0.1$.

We do not find any significant relationship between providing explanations of the AI advisor and the *RSR* (**H1a**). However, also in the SEM the effect of explanations on the *RAIR* is significant (**H1b**). Additionally, we find a significant positive relationship between providing explanations of the AI advisor and the change in self-confidence which confirms **H2**. We also observe a positive correlation between the change in self-confidence and the *RAIR* (**H3b**). Since the absolute strength of the relationship also decreased with including the change in confidence, we can conclude that the positive effect of explanations on the *RAIR* is partially mediated by the change in confidence. Interestingly, we do not find a relationship between the change in self-confidence and the *RSR* (**H3a**). Additionally, we find no effect

of our explanations on trust (**H4**). However, both **H5a** and **H5b** are confirmed, i.e. increasing trust increases *RAIR* but also decreases *RSR*. We display a summary of the hypotheses results in Table 6.5. In the following Section, we discuss our results.

Table 6.5.: Analysis results of the structural equation model (*** $p < .01$, ** $p < .05$, * $p < .1$)

X	Y	z-value	Standardized regression coefficient	Result
H1a: Explanations ->	<i>RSR</i>	0.04	-0.03	not supported
H1b: Explanations ->	<i>RAIR</i>	1.73	0.08 *	supported
H2: Explanations ->	Change in self- confidence	2.30	0.13 *	supported
H3a: Change in confidence ->	<i>RSR</i>	0.02	0.00	not supported
H3b: Change in confidence ->	<i>RAIR</i>	1.84	0.11 *	supported
H4: Explanations ->	Trust	-0.21	-0.04	not supported
H5a: Trust ->	<i>RSR</i>	-3.08	-0.05 ***	supported
H5b: Trust ->	<i>RAIR</i>	3.19	0.06 ***	supported

6.7 Discussion

In this article, we first defined AR and conceptualized a measurement concept (AoR). Following that, we derived a research model regarding the impact of explanations on AoR which we subsequently tested on a deception detection task. Though conducted in a limited scope, our findings should help to guide future work on AR.

Theoretical foundation of appropriate reliance. The main contribution of our work is the theoretical development of AR. So far, terms like “appropriate trust”, “calibrated trust” and AR were often used interchangeably in prior research. We provide clarity by defining AR and putting the terms in perspective. Second, we derive a granular, two-dimensional measurement concept—appropriateness of reliance (AoR). Most prior work neglected the initial human decision that would have been made without any AI advice.⁵ Taking this initial human decision into account allows to differentiate between the effects of advice (correct and incorrect) and confirmation. Without the human initial decision, we cannot say for sure whether a

⁵The one exception is the work of (Buçinca et al., 2021) who measure over-reliance in the same way that we do but do not consider under-reliance.

final wrong human decision is due to over-reliance on a wrong AI advice or due to the human and the AI being wrong.

Implications for appropriate reliance and explainable AI. We further have investigated the effect of AI explanations on AoR. We confirm the results of prior research (Bansal et al., 2021; Gonzalez et al., 2020; Wang & Yin, 2021) by finding an effect of explanations on the *RAIR*. We believe that the reason for this could be that the AI’s explanations increase people’s knowledge of the task (Gonzalez et al., 2020; Spitzer et al., 2022). Maybe in such cases, the human should be seen less as a “judge” but instead more as a student of the AI. On the other hand, our results show that explanations do not influence the *RSR*. While this may sound disappointing at first, it also shows that the claim that explanations would reduce overreliance (Bansal et al., 2021; Bućinca et al., 2021) does not seem to hold for all kinds of tasks.⁶ It further suggests that new techniques must be developed to distinguish incorrect AI advice. Moreover, our study is the first one that analyzed mediators of the effect of explanations on AR. Interestingly, in our study, we find no effect of explanations on trust. However, prior research has shown that it depends on a lot of confounding factors. We find significant effects of trust on *RAIR* as well as *RSR*. Additionally, we show that the effect of explanations on *RAIR* partially depends on the change in confidence after receiving AI advice.

Appropriate reliance and complementary team performance (CTP). Lastly, we want to elaborate on the relationship between AoR and CTP. To reach CTP, the task needs to have instances where the AI is better than the human and vice versa, i.e. a certain amount of complementarity potential needs to be present (Hemmer et al., 2022b). CTP essentially depends on the relationship between *RSR* and *RAIR* and this complementarity potential.

The human impact on CTP is given by multiplying the *RSR* and the share of incorrect AI advice. However, simultaneously involving a human might reduce the AI performance ($1 - RAIR$ multiplied by the share of correct advice). That means to reach CTP, the gain through human involvement needs to be larger than the loss through discounting correct AI advice or more formally⁷:

$$CTP = \begin{cases} 1, & \text{if } RSR * IA > (1 - RAIR) * CA \\ 0, & \text{otherwise} \end{cases}$$

⁶An earlier study by this research team (Schemmer et al., 2022c) found initial signs of a reduced *RSR* in a pretest. However, this study with more participants shows that the effect is not significant in a larger sample.

⁷Provided that nothing changes in the cases where the initial decision of the human and the AI advice are the same.

Here IA is the total number of task instances in a test set where the human is initially correct and receives incorrect advice. CA refers to the number of task instances where the human is initially incorrect and receives correct advice. If this condition is not fulfilled, AR depends on the relationship between human and AI performance. If the human performs worse than the AI advisor and has a low RSR and $RAIR$, one should always favor AI advice. If the human decision-maker performs better on average than the AI, one can argue that from a performance perspective the AI should not be used.

Limitations. No research is without limitations. We would like to emphasize that deception detection is a difficult task for humans (Lai et al., 2020; Lai & Tan, 2019). Humans on average just perform slightly better than by chance on this particular hotel review task (Lai et al., 2020). This makes AR difficult as the task of discrimination requires human domain knowledge. Additionally, the generalizability of our experimental findings is limited due to the choice of explanations. However, we have deliberately chosen the most modern form of explanation to maximize the impact of our results. Our concept is limited to classification tasks but will be extended in future work. First approaches can be found in the work of Petropoulos et al. (2016).

Furthermore, the sequential task setup necessary for our measurement concept has some disadvantages as it changes the task itself. Since conducting the same task initially alone before receiving AI advice, the human is already mentally prepared and might react differently than after directly receiving AI advice. More specifically, research has shown that letting humans conduct the task alone before receiving AI advice might reduce over-reliance (Buçinca et al., 2021). The sequential task setup could induce an anchoring effect which prevents the human to more actively take the AI into account (Buçinca et al., 2021). This could have led to the overall low $RAIR$ in our experiment. Moreover, sequentially conducted tasks with AI advice might not always be possible or desired in real-world settings. Therefore, the measurement should be seen as an approximation of real human behavior. Instead of having a sequential task setup, one alternative option could be to simulate a human model based on a data set of task instances solved by humans without AI advice. This simulation model could approximate the initial human decision within a non-sequential task setting. However, this approach is also an approximation of real human behavior. In other work, a latent construct has been derived to measure reliance behaviour (Tejeda et al., 2022). Future work should compare the approaches.

Future Work. Essentially, the improvement in *RAIR* depends on the knowledge gain of the human. Future work could therefore extend our research model by adding newly learned knowledge as a mediator. Empirically, this could be measured by asking humans before collaborating with AI to do a couple of task instances on their own and afterwards (Spitzer et al., 2022). The performance improvement can be interpreted as learned knowledge.

Most importantly, future research needs to investigate the impact of different design features of AR. We initiate our research with state-of-the-art feature importance but many other ones can be thought of, e.g. counterfactuals, global explanations. Future research should evaluate these potential design features to provide practitioners with a toolkit for effective use of AI.

6.8 Conclusion

Appropriate reliance in AI advice is the next milestone after a decade of research focused on AI adoption and acceptance. Nowadays, many AI applications are deployed and used on a daily basis. While adoption and acceptance remain important, we argue that a perspective shift is necessary. In the use phase of AI, researchers need to find ways to ensure appropriate reliance and, thus, effective use of AI. In this article, we provide guidance for future research on appropriate reliance by providing a definition and a measurement concept—appropriateness of reliance (AoR). Furthermore, we generate initial insights how explanations influence the appropriateness of reliance. We hope that our research will inspire researchers and practitioners for future research on appropriate reliance, resulting in effective human-AI collaboration.

Harnessing Complementarity: The Influence of Human Learning on Appropriate Reliance

This chapter comprises a working paper that is currently under review as Schemmer, M., Bartos, A., Spitzer, P., Hemmer, P., Kühl, N., Liebschner, J., & Satzger, G. (2023a). Towards Effective Human-AI Decision-Making: The Role of Human Learning in Appropriate Reliance on AI Advice [Working paper]. Note: To improve the consistency of the thesis, the title has been changed. The abstract has been removed. Tables and figures were reformatted, and newly referenced to fit the structure of the thesis. The terminology was standardized with the dissertation. Chapter, section and research question numbering and respective cross-references were modified. Formatting and reference style was adapted and references were integrated into the overall references section of this thesis.

7.1 Introduction

Over the past years, Artificial Intelligence (AI) systems have entered a wide range of areas, even high-stake decision domains. For instance, AI applications support doctors in their diagnoses (Leibig et al., 2022), help recruiters in the hiring process (Peng et al., 2022), and support legal decisions in court (Kleinberg et al., 2018). This proliferation is driven by the continuous development of AI systems, which results in advanced capabilities and increased performance (Ren et al., 2015). Self-supervised models and generative AI have further fueled a new era of human-AI collaboration, e.g., Notion AI, GitHub Co-pilot, DeepL, and ChatGPT.

Modern AI is not only more performant and versatile in its applications, but also bears the potential to enhance humans through complementary capabilities (Dellermann et al., 2019a; Fügener et al., 2021; Hemmer et al., 2021) reaching performance

levels beyond the ones humans or AI can reach on their own. This desired superior performance in human-AI collaboration is referred to as complementary team performance (CTP) (Bansal et al., 2021; Hemmer et al., 2021).

Despite the tremendous advances in performance and capabilities, it is essential to bear in mind that every AI application has inherent uncertainty. AI models and their recommendations are based on probabilities. So, no matter how good a model is, it will not always be accurate. Since AI advice is imperfect, general acceptance by a human decision-maker would also comprise incorrect advice. For example, physicians would blindly follow AI advice on cancer diagnosis—although AI advice might be wrong, and the physicians might have known better. Thus, it is important for human decision-makers to have the ability to discern when to rely on AI advice and when to rely on their judgment, i.e., they should display a high level of appropriateness of reliance (AoR) (Schemmer et al., 2023d).

So far, research has focused on driving AoR by enabling humans to build an accurate mental model of AI (Bansal et al., 2019a; Kloker et al., 2022; Kühl et al., 2022; Taudien et al., 2022). The term mental model refers to the human's knowledge about various aspects of the AI system's capabilities. Bansal et al. (2019a) particularly stress the importance of recognizing the AI's error boundaries to develop a realistic mental model. The mental model research stream focuses on enabling decision-makers to assess the quality of an AI prediction. Researchers aim to build this mental model by providing explanations of the AI's decision process (Bansal et al., 2019a; Zhang et al., 2020).

However, recent review articles have shown that the current focus on influencing the mental model through the provisioning of explanations is not sufficient to consistently reach CTP (Hemmer et al., 2021; Schemmer et al., 2022b). One of the core influence factors of reliance behavior in human-AI collaboration seems to be whether the decision-maker is an expert (high domain knowledge) or a lay worker (low domain knowledge) (Nourani et al., 2020b; Wang & Yin, 2021). We follow this line of thought and hypothesize that learning during human-AI collaboration (decision-makers gradually gaining expertise) could be a relevant mediator of AoR.

Figure 7.1 on page 151 illustrates this line of thought based on a brain cancer classification example. Previous work has focused on using AI explanations to enable the decision-maker to determine a classification's quality. Although we do not want to abandon this line of research, we believe that an additional facilitator could be that a decision-maker learns new patterns through AI's explanations that can diagnose cancer.

However, the specific effect of learning on AoR is ambiguous. Learning during collaboration could also lead to unwanted effects, such as aversion—as humans might think they have learned enough to solve the task alone and no longer need the AI advice. Therefore, we formulate the following research question.

RQ1: *How does human learning during human-AI collaboration influence the Appropriateness of Reliance on AI advice?*

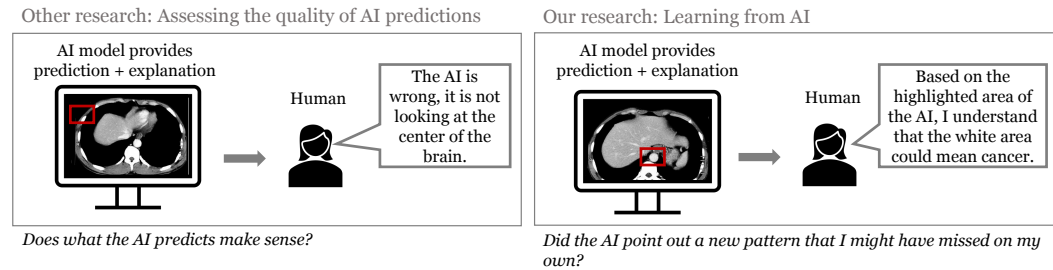


Figure 7.1.: Distinction between related work and our contribution.

Even if learning improves AoR, we must find ways to enable and improve learning during human-AI collaboration. Recent research on learning systems has seen first positive results of using explanations of AI to improve learning (Goyal et al., 2019; Wang & Vasconcelos, 2020). However, it is an open question if the promising result in learning systems can be replicated in human-AI collaboration. Therefore, we formulate our second research question.

RQ2: *Can explanations of AI increase human learning in human-AI collaboration?*

To gain first answers to our broader research questions, we derive a research model including theory-driven hypotheses and subsequently conduct a behavioral experiment with 100 participants using an image classification task as a testbed to evaluate the model. We use example-based explanations (Fahse et al., 2022) to design a human-AI collaboration scenario with a high potential for learning.

Our results show that a) example-based explanations can improve human learning during human-AI collaboration, b) learning improves the human ability to assess when to rely on themselves, and c) if sufficient learning is present, human learning helps to assess better when to rely on AI.

We contribute to the body of knowledge on human-AI collaboration in general and on AoR and learning from AI in particular. To the best of our knowledge, this research depicts the first study covering the effect of explanations on learning and the mediating effect on AoR. We thereby extend the research model of AoR

developed by Schemmer et al. (2023d) by a learning construct. Our work provides a new perspective on AoR and the design of human-AI collaboration systems.

7.2 Theoretical Foundations and Related Work

In the following, we introduce the related work of this article, structured along the topics of human-AI collaboration, explainable AI, appropriate reliance on AI advice, and learning from AI.

7.2.1 Human-AI Collaboration

In recent years, there has been a surge of research in human-AI collaboration, with a growing number of studies conducting behavioral experiments to gain a better understanding of how humans form decisions in the presence of AI (Alufaisan et al., 2021; Buçinca et al., 2020; Carton et al., 2020; Lai et al., 2020; Lai & Tan, 2019; Liu et al., 2021; Zhang et al., 2020). This research has focused on improving human-AI collaboration to optimize team performance (Buçinca et al., 2020; Zhang et al., 2020).

The idea behind human-AI collaboration is to be more effective than both human and AI individually (Dellermann et al., 2019a). This improvement through collaboration stems from the idea that both the AI and the human possess a unique set of skills that can enrich each other in specific tasks. Thus, the true potential of human-AI collaboration lies in leveraging these complementary capabilities to reach the desired state of superior performance, i.e., complementary team performance (CTP) (Bansal et al., 2021; Hemmer et al., 2021).

The question of how to realize this complementarity potential and thus reach the desired superior performance remains an active area of research. In most empirical studies, the performance of human-AI collaboration is still inferior to that of individual AI, and thus CTP is not achieved (Bansal et al., 2021; Hemmer et al., 2021; Schemmer et al., 2022b). Current research identifies the missing appropriateness of reliance as a main cause preventing the achievement of CTP (Bansal et al., 2021; Hemmer et al., 2021; Schemmer et al., 2022b).

7.2.2 Explainable Artificial Intelligence

Explainability has a long-standing history in information system (IS), dating back to the emergence of knowledge-based systems, expert systems, and intelligent agents in the 1980s and 1990s (Meske et al., 2022). The term “Explainable Artificial Intelligence” (XAI) was first introduced by Van Lent et al. (2004), referring to the capacity of their system to clarify agent behavior.

XAI techniques can be differentiated in terms of their scope, i.e., global or local explanations (Adadi & Berrada, 2018): Global XAI techniques deal with holistic explanations of the models as a whole. In contrast, local explanations work based on individual task instances. Local approaches can be based on examples, features, or rules. Example-based explanations provide examples from historical data that are either from the AI predicted class (normative examples) or from a different class (comparative examples) (Cai et al., 2019). In an image classification task, a normative example would be an image from the AI predicted class. A comparative example would be the most similar images from a different class. It is important to note that example-based explanations have a link to ground truth, as they have historically validated labels.

7.2.3 Appropriate Reliance in Artificial Intelligence

The concept of appropriate reliance has gained attention in the field of human-AI collaboration (Bansal et al., 2021; Schemmer et al., 2022b). In general, appropriate reliance refers to desirable behavior where humans override incorrect AI advice and follow correct advice (Bansal et al., 2021; Schemmer et al., 2023d).

First work on the conceptualization and measurement of appropriate reliance was done by Schemmer et al. (2023d). The authors differentiate between appropriate reliance as a binary target state (“appropriate reliance is either achieved or not”) and a metric indicating a degree of appropriateness. They introduce a two-dimensional metric—the appropriateness of reliance (AoR)¹—to describe and measure reliance behavior. It is based on relative frequencies of correctly overriding wrong AI suggestions (correct self-reliance) and following correct AI suggestions (correct AI reliance) and reflects a metric understanding of appropriate reliance (see Figure 7.2 on page 154). This metric can then be used to define different levels as target states of appropriate reliance that mark the achievement of objectives like certain legal, ethical

¹Note that the measurement of AoR requires a sequential task setup as visualized in Figure 7.2 on page 154 and described in Schemmer et al. (2023d)

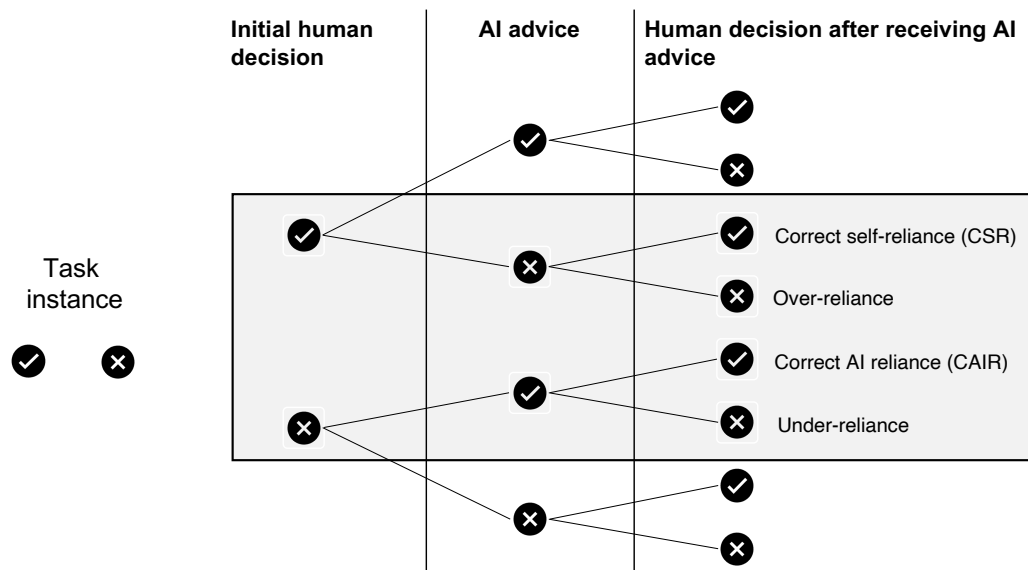


Figure 7.2.: Combinatorics of initial human decisions, AI advice and human reliance for a single task instance in a sequential task setting (Schemmer et al., 2023d).

and performance requirements. From an effectiveness perspective they argue that appropriate reliance is achieved if CTP is present.

Several studies have explored the impact of different explanation techniques on AoR, including feature-based (Ribeiro et al., 2018), example-based (Van der Waa et al., 2021) and rule-based (Ribeiro et al., 2018) explanations. Some empirical evidence has shown that XAI explanations can help humans differentiate better between correct and incorrect AI predictions (Buçinca et al., 2020), but they can also be misleading, convincing humans to follow incorrect AI advice and leading to poorer team performance (Bansal et al., 2021; Poursabzi-Sangdeh et al., 2021). This ambiguous prospect raises the question of what constitutes good explanations in the context of human-AI collaboration and highlights the importance of further research into suitable explanations to support AoR and its underlying factors. Moreover, Nourani et al. (2020b) shows, that the effect of XAI explanations on humans' reliance is dependent on their initial domain knowledge. In this work, we explore the potential impact of leaning from AI as a mediator between explanations and AoR.

7.2.4 Learning from Artificial Intelligence

In recent years, researchers have been exploring the potential for AI to augment human learning (Cakmak & Lopes, 2012; Edwards et al., 2018). One field of

research that has emerged is machine teaching (Zhu et al., 2018). In this field, instead of assisting humans in decision-making, AI systems are set up as learning systems to teach humans. One example are AI systems that train crowd-sourcing workers to correctly annotate images (Wang & Vasconcelos, 2020). The concept involves selecting the optimal teaching set (Zhu et al., 2018) to achieve the best learning performance (Singla et al., 2014).

In recent research, the focus in this interaction between humans and AI has been on explainable AI (XAI). Recent studies have utilized XAI to generate explanations in such learning systems. The AI system provides additional explanations to the human to improve their knowledge in a specific domain (Alipour et al., 2021). For example, Goyal et al. (2019) use a convolutional neural network for various image classification tasks and apply comparative examples to provide visual explanations. The authors select images with minor changes from another class to generate comparative explanations.

With the advent of research to facilitate XAI in learning systems, to the best of our knowledge, there are no studies that investigate how explanations affects human learning in human-AI collaboration. In this study, we investigate how XAI affects human learning without explicitly developing a learning system but by observing its use in human-AI collaboration to improve AoR. Dellermann et al. (2019a) argue that humans and AI systems can learn from each other when they collaborate. We follow this line of thought and following, derive a research model on the potential mediating effect of learning on AoR.

7.3 Theoretical Development

In this work, we postulate that human learning plays a crucial role in AoR and that it can be influenced by providing explanations. In this section, we derive a corresponding research model that establishes the link between explanations and human learning and its mediating role on AoR.

As a dependent variable, we use the previously introduced tuple of AoR, which comprises the two dimensions of relative self-reliance textit(RSR) and relative AI-reliance (*RAIR*). RSR encompasses the cases where the human is initially correct, receives wrong advice, and rightly dismisses it. In this case, the humans' complementary knowledge is leveraged by correcting an AI's wrong output on an instance level. In contrast, RAIR encompasses the cases where the human is initially incorrect, gets correct advice, and rightly follows. In this case, complementarity potential from an

AI can be exploited as the human would not have been able to correctly solve the task instance without the help of the AI advisor.

We now first derive hypotheses related to the impact of providing explanations on learning and thereafter focus on the potential impact of learning on AoR and potential direct effects.

Prior research at the intersection of IS and human-computer interaction has leveraged learning systems that are supported by AI to teach humans in an example-based manner. The objective of learning systems is the user's knowledge extension. In human-AI collaboration, the goal is to improve performance (Hemmer et al., 2021). In learning system research, the effect of explanations generated by XAI on humans' learning performance is an evolving research stream (Alipour et al., 2021; Goyal et al., 2019). However, previous research has only examined how explanations can be utilized in learning systems. To the best of our knowledge, there are no studies investigating how XAI affects learning in human-AI collaboration.

In general, explanations hold the potential to stimulate new ways of thinking, which can lead to the generation of new knowledge (Saeed & Omlin, 2023). Prior research indicates that learning performance among humans can be enhanced in learning systems through example-based learning (Basu & Christensen, 2013; Martin-de-Castro et al., 2008; Stark et al., 2004). In this study, we examine example-based explanations as both normative and comparative instances. Normative examples embody instances of the predicted class, while comparative examples illustrate instances from a different class. As Cai et al. (2019) explain, normative examples aim to set a standard for the intended class by displaying training instances from that class, while comparative explanations provide a contrast between the AI prediction and the most similar instances from a different class. Example-based explanations are anticipated to be easily comprehensible and prompt causal thinking, enabling individuals to deduce cause-and-effect relationships. Fahse et al. (2022) suggest that the efficacy of example-based explanations can be attributed to their compatibility with human reasoning processes and the minimal cognitive burden they impose on users. Yang et al. (2020b) further argue that these explanations align with people's inductive (i.e., bottom-up logic) and analogical reasoning (i.e., drawing comparisons from one instance to another), which helps users understand why certain objects are deemed similar or dissimilar.

The main difference between the use of Explainable Artificial Intelligence (XAI) in a learning system and in human-AI collaboration may be the level of accuracy. In the learning system, the AI is expected to achieve perfect accuracy, whereas in human AI decision contexts, such perfect accuracy may not be achievable. This is because

learning system designers can use historical data and select a training set where the AI prediction is known to be correct. In human-AI decision-making, it is unclear whether an AI's advice is right or wrong. AI is inherently imperfect in human-AI collaboration. However, we argue that example-based explanations are actually not entirely dependent on the performance of the AI. Example-based explanations have a clear ground truth because they are drawn from the training data set. Therefore, example-based explanations have the potential to increase knowledge even if the AI is wrong.

To sum it up, we hypothesize that the benefits of normative and comparative examples on learning that are present in learning systems will also be present during human-AI collaboration. Therefore, we formulate:

H1: *Example-based explanations have a positive effect on human learning during human-AI collaboration.*

Next, we hypothesize the effect of human learning on RSR and RAIR. Distinguishing these two dimensions allows for a deeper understanding of the underlying mechanisms. We base our hypotheses on theories of domain knowledge (Nourani et al., 2020b). Much of the latter work in appropriate reliance has examined the differences between experts (high domain knowledge) and lay workers (low domain knowledge) in terms of their reliance behavior in human-AI collaboration. We argue that the differences between experts and lay workers can be seen as an analogy for learning.

RSR essentially refers to the ability to override incorrect AI advice. In other words, if humans can correctly solve a task instance and receive an incorrect AI recommendation, the RSR tells us how well they can reject that incorrect advice. Nourani et al. (2020b) have shown that experts are better at correcting AI errors than lay workers. Consequently, learning should increase the effectiveness of validation, and thus increase the RSR. Therefore, we hypothesize:

H2a: *Learning increases relative self-reliance (RSR).*

Improving RAIR reveals a more complex process. In this setting, the human does not have enough domain knowledge to solve the task instances independently. Using our analogy, they could be considered “lay workers” in these cases. Becoming an “expert” can have two advantages. First, at the task instance level, learning a new pattern based on the explanations received could allow the decision-maker to solve the task independently, which in turn would increase RAIR. Second, an increase in overall knowledge over time may help the decision-maker to better recognize and

follow correct AI advice (following RSR logic). Both mechanisms would essentially lead to an increase in RAIR through learning.

H2b: *Learning increases relative AI-reliance (RAIR).*

In addition to our hypothesized mediation effects, we also follow the line of thought of previous work and consider that explanations might influence AoR by improving the mental model. For this reason, we formulate additional direct path hypotheses between explanations and AoR that represent additional effects of explanations beyond learning².

Explanations allow insights into the reasoning and decision-making of AI models. In the case of inaccurate advice, these insights might help the human decision-maker to evaluate the validity of such reasoning by checking for its alignment with universal axioms of the task. This process might, in turn, enhance their knowledge regarding the underlying AI model (mental model) and thus improve validation capability. As RSR is increased by the correction of wrong AI advice, explanations that enhance this kind of knowledge would increase the RSR.

H3a: *Example-based explanations have a positive effect on relative self-reliance (RSR).*

Additionally, it could also be possible to calibrate the own mental model of the AI in such a way that without learning, it is still possible to detect a good recommendation of an AI. Therefore, we hypothesize:

H3b: *Example-based explanations have a positive effect on relative AI-reliance (RSR).*

Figure 7.3 on page 159 summarizes all hypotheses in one integrated research model. Lastly, any improvement in AoR should in turn improve team performance and at a certain level enable CTP.

7.4 Methodology

In this section, we present our design for a behavioral experiment to test our research model.

²The inclusion of both indirect and direct effect hypotheses in a research model is well known in IS research. For example, see Tereschenko et al. (2022)

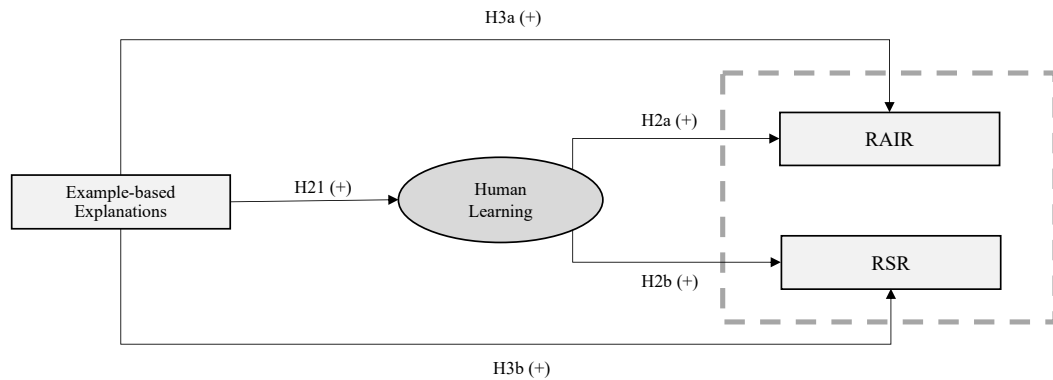


Figure 7.3.: Research model on the effect of human learning on appropriateness of reliance (AoR).

7.4.1 Task, Model and Explanations

As an experimental task, we choose a bird species classification task based on image data. We chose the context of image recognition following the reasoning of Fügenger et al. (2021): Firstly, image recognition is a broad task that all humans should be capable of executing without requiring specialized skills or training. In behavioral research, the objective is often to establish a context where findings are applicable to various situations. It is believed that observations in general tasks can transfer to more specialized tasks, while contexts that necessitate specific training result in less generalizable outcomes. Secondly, image classification is an area where contemporary AI systems excel (Szegedy et al., 2015), performing at least on par with human abilities (Russakovsky et al., 2015b). Thirdly, prior studies have shown the high complementarity potential of image classification (Fügenger et al., 2021; Nguyen et al., 2022).

The bird species classification task is based on the Caltech-UCSD Birds-200-2011 dataset (Wah et al., 2011). This dataset has been used extensively in high-profile publications and includes 11,788 images of 200 different bird categories (Goyal et al., 2019; Nguyen et al., 2022). Four black-colored bird classes (American Crow, Groove billed Ani, Shiny Cowbird, and Boat tailed Grackle) are chosen for the experiment. In a preliminary study, we tried different combinations of bird classes and the number of classes to get a solvable but still difficult task. Only the consistently black-colored birds are selected from each of these classes, resulting in 216 images in the dataset used. We filter out no-black birds as the preliminary studies have shown that participants tend to choose color as an important factor during learning which, however, was not an unambiguous feature of the bird classes.

As a model, we use a pre-trained ResNet50 (He et al., 2016) and fine-tune it on our data set. For training, we use Adam as an optimizer with a StepLR rate scheduler with step size 5 and gamma of 0.1. The training data is augmented with random rotation, change in sharpness and contrast. The model is trained for 9 epochs and achieves an accuracy of 87.96%.

To sample the normative and comparative examples, we follow the approach of (Cai et al., 2019). The comparative examples are searched within the second most probable classes according to the model. Next, we calculate the cosine similarity of the feature vectors of the flattened layer of the model between the target image and all images of the searched class (Chen, 2020). Then, the two most similar images are selected. The normative examples are randomly selected from the class predicted by the model. The images which are to be classified in the experiment were omitted from the examples.

7.4.2 Experimental Design

The experiment is conducted online with a between-subject design where two different conditions are tested (in the following, these conditions are referred to as baseline condition and XAI condition). Depending on the condition, either a baseline or an example-based explanation is provided. The study is approved by the University IRB.

Our experimental design is influenced by the requirements to measure learning and AoR. To measure learning, we use standard methods from the organizational learning research stream (Spitzer et al., 2022) and conduct two knowledge tests in the experiment. The difference between the two tests then constitutes as learning (Spitzer et al., 2022). Additionally, to measure AoR, a sequential task setup is necessary as it is necessary to measure an initial human decision (Schemmer et al., 2023d). These two requirements shape the design of our experiment.

Participants are randomly assigned to the condition groups to control for internal validity. The online experiment is initiated with an attention control question. Then, both condition groups receive an introduction to the task. The participants are not provided with any specific performance information about the AI. Then, the participants conduct a tutorial where we show them one example per bird class. After the tutorial, we conduct a knowledge test by asking the participants to classify 8 pictures. Hereby, the 8 pictures are drawn stratified from a sample of 16 images in total.

Once the initial tutorial and first knowledge assessment have been completed, participants move to the main task consisting of 16 individual task instances. We use an advanced sampling strategy to ensure that RSR as well as RAIR is possible in our study. We stratified sample each bird class and control for 50% correct predictions and 50% incorrect predictions. For the AoR measurement concept, sequential task processing is essential. In our study, this means the human first receives an image without any AI advice (see step 1 in Figure 7.4). Then the participant is asked to classify the image. Following that, the human either receives a simple AI advice statement, e.g., “the AI predicts that the image below shows Class 4” or the AI advice and additional example-based explanations (see step 2 in Figure 7.4). After receiving the AI advice, the participant can change the initial decision. This sequential two-step decision-making allows us to measure AoR. During the main tasks, the participants do not receive feedback on their performance. After finishing the main task, again, their task-specific knowledge is assessed. Additionally, data on demographic variables are collected.

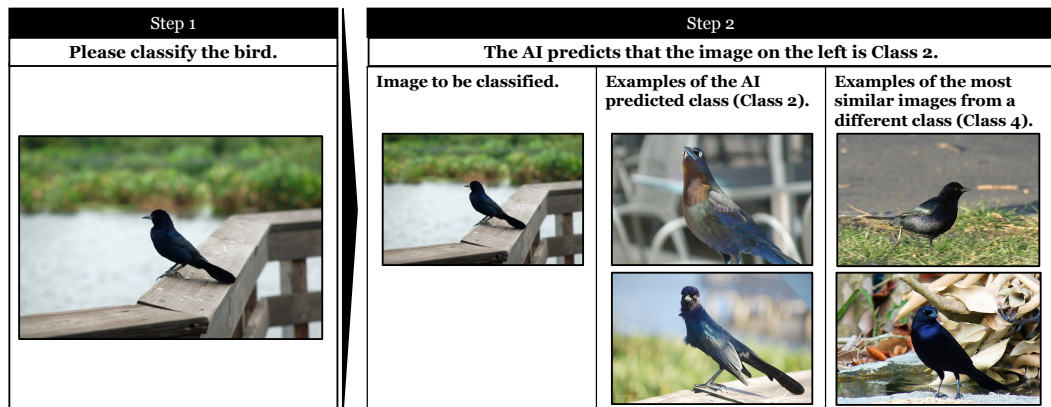


Figure 7.4.: Online experiment graphical user interface for the example-based explanation condition. The ground truth is class 2.

7.4.3 Measurements

In this work, we measure initial task knowledge by counting the number of correctly classified images in the first knowledge test. Learning is measured as the difference between the number of correctly classified images in the second knowledge test and the first knowledge test.

We measure the AoR following the work of Schemmer et al. (2023d) as a tuple of RSR and RAIR. RSR is hereby measured as the number of cases of correct self-reliance divided by the total number of cases in which a previously correct decision-maker

receives incorrect AI advice. Correct self-reliance (CSR) hereby is “1” if, on this particular instance i , the initial human decision was correct, the AI advice was incorrect, and the human decision after receiving AI advice is correct. Incorrect advice (IA) is “1” if the initial human decision for a task instance i was correct and the AI advice was incorrect.

$$\text{Relative self-reliance (RSR)} = \frac{\sum_{i=0}^N CSR_i}{\sum_{i=0}^N IA_i}$$

RAIR is the ratio of the number of cases in which humans rely on correct AI advice, and the decision was initially not correct, i.e., in which humans rightfully change their minds to follow the correct advice. Correct AI reliance (CAIR) hereby is “1” if, in this particular case i , the original human decision was wrong, the AI recommendation was correct, and the human decision after receiving the AI recommendation is correct, and “0” otherwise. Correct advice (CA) is “1” if the original human decision is wrong and the AI advice is correct, regardless of the final human decision, and “0” otherwise.

$$\text{Relative AI reliance (RAIR)} = \frac{\sum_{i=0}^N CAIR_i}{\sum_{i=0}^N CA_i}$$

The human performance is measured as the number of correctly classified images divided by the total number of images before receiving AI advice. Likewise, the human performance after AI advice refers to the number of correctly classified images divided by the total number of images after receiving AI advice. Finally, CTP is defined as a binary state that is achieved if the human performance after AI advice is significantly higher than the human as well as the AI performance.

7.4.4 Participants

The experiment was performed in April 2023 on the platform “Prolific”. Image classification is often done as crowd work and previous IS research has shown the validity of using online studies for image classification (Fügener et al., 2021).

Overall, 100 participants were recruited, 50 per condition. All of them passed our attention check. To incentivize the participants, they were informed that for every correct decision, they get an additional 5 Pennies in addition to a base payment of 1.5 Pounds for an estimated study time of 15 min. Average duration of the experiment was 9:00 minutes in the base line condition and 12:36 minutes in the XAI condition.

Roughly 64% of the final sample identified as being female, almost 36% identified as being male, and one participant preferred not to report. Participants' age ranges from 18 years to the age group of 76 years with an average age of approximate 32 years.

We excluded one participant in the baseline condition because of a failed manipulation check. In addition, some participants had no cases where the AI was correct, and they were previously wrong leading to an undefined RAIR. We assigned those participants the average value of the condition (1 in the baseline condition and 2 in the XAI condition). Next, after removing missing values, we remove outliers from our data using the z-score method (2 sigma). We find one outlier in the XAI condition with a learning value of -5. Additionally, we find 4 outliers with zero correct classifications in the first knowledge test. Which leaves us with a final data set of 48 participants in the baseline condition and 46 participants in the XAI condition.

7.5 Results

In this section we report the results of our behavioral study. First, we provide an overview of the descriptive findings, followed by the results in terms of AoR and CTP. We then examine our comprehensive research model, which includes mediations, using structural equation modelling (SEM) and conduct an exploratory subgroup analysis.

7.5.1 Descriptive Results and Appropriateness of Reliance

The descriptive results of our study are presented in Table 7.1 on page 164. They are divided according to the experimental condition. We evaluated the significance of the results using t-tests after controlling for normality. Otherwise, we use Mann-Whitney U tests. In order to control for multiple comparisons and reduce the likelihood of Type I errors, we applied Bonferroni corrections to our statistical analyses. First, we report descriptive measures of initial knowledge and learning. Next, we report correlation measures and then our analysis of AoR.

Our control variable initial task knowledge is not significantly different between groups, meaning that participants start the experiment with the same knowledge on average (baseline mean = 4.54, XAI mean = 4.26; two-tailed t-test: $T = -0.8$, $p = 0.43$). On average, participants correctly classify about 4 of the 8 birds in the initial knowledge test. Note that this does not mean that they perform randomly

on average, as we perform a multiclass classification with 4 classes, i.e., random guessing would lead, on average, to a performance of 2 out of 8. Testing against randomness shows that participants in both groups have significant initial knowledge (one-sample t-test: baseline: $T = 9.55$, $p < 0.01$; XAI: $T = 9.86$, $p < 0.01$). In summary, initial knowledge is not significantly different between the two groups, and participants have sufficient initial knowledge to avoid guessing at random.

Learning is significantly different between the groups (two-tailed t-test: $T = 2.09$, $p = 0.04$), which means that, on average people, learned more in the XAI condition. In addition, we test with a one-sample t-test whether the learning is significantly different from 0. In the baseline condition, we find no significant difference ($T = -0.69$, $p = 0.49$). In the XAI condition, we observe that learning is significantly greater than zero ($T = 2.25$, $p = 0.03$). This means that our example-based explanations not only increase learning but also bring it to a level higher than 0.

Table 7.1.: Descriptive results (** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$)

Treatment	Learning ** (SD)	RSR (SD)	RAIR (SD)
Baseline	-0.19 (1.89)	64.89% (32pp)	57.82% (39pp)
Example-based explanations	0.63 (1.9)	74.61% (25pp)	66% (33pp)

We also analyze the Pearson correlation on paths 2 and 3 of our hypothesis. The results are presented in Table 7.2. To analyze the correlation paths, we do not yet distinguish between conditions. We find a weak correlation between learning and RSR but no significant correlation between learning and RAIR.

Table 7.2.: Correlation analysis (** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$)

Treatment	Correlation	P-Value
Learning – RSR*	0.18	0.07
Learning – RAIR	0.08	0.43

Next, we analyze AoR. Participants' RAIR and RSR are not significantly different between conditions. However, both values are relatively high if we compare them with related literature (Schemmer et al., 2023d; Taudien et al., 2022). We also observe that, although not significant, there is a trend that our example-based explanations increase both RSR and RAIR. Figure 7.5 on page 165 illustrates our AoR analysis.

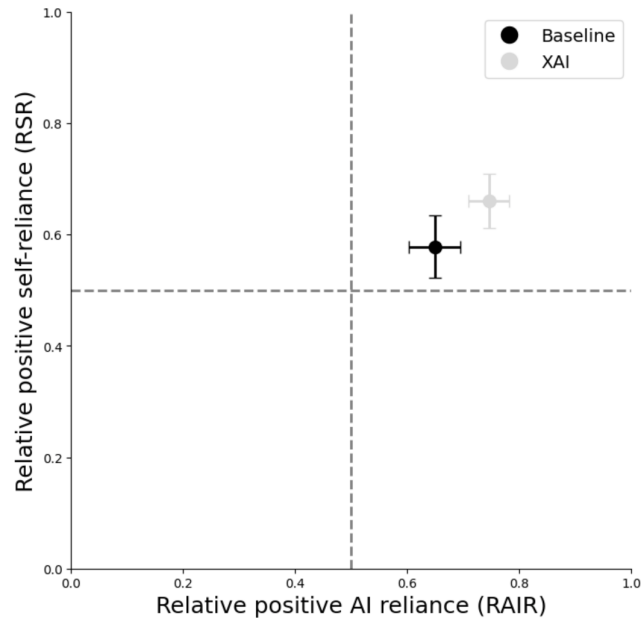


Figure 7.5.: Appropriateness of reliance analysis (including standard errors).

7.5.2 Structural Equation Modeling

In addition to analyzing the direct effect of explanations on RAIR and RSR, we use structural equation modeling (SEM) analysis to test our hypothesized research model.

Prior to fitting our SEM, we performed missing data identification, outlier detection, normality testing, and selection of an appropriate estimator. We describe how we identify missing data and remove outliers in the previous section. Shapiro’s test for normality indicates that several variables of interest deviate significantly from normal distributions. As a result, we conduct the analysis using an estimator that allows for robust standard errors and scaled test statistics (Kunkel et al., 2019). Thus, we use the MLR estimator (Lai, 2018).

Our dependent variables are RAIR and RSR. Since these dependent variables are between 0 and 1, we used a logistic model in the Lavaan package, version 0.6-9, in R (Rosseel, 2012). This model has an excellent overall fit (see Table 7.3 on page 166). The results for each independent variable are discussed below and visualized in Figure 7.6 on page 166.

First, we find a significant relationship between providing explanations and learning (**H1**). We find no direct effect of providing explanations on RSR or RAIR (**H3a &**

Table 7.3.: Structural equation model fitting index using root mean square error of approximation (RMSEA), Comparative Fit Index (CFI), Tucker-Lewis Index (TLI) and Standardized Root Mean Squared Residual (SRMR)

	RMSEA	CFI	TLI	SRMR
Measurement Criteria based on (Hu & Bentler, 1999)	< 0.05	> 0.96	> 0.95	< 0.08
Value	0	0.99	1	4.06

3b). We also find no effect of learning on RAIR (**H2b**). However, we do find a strong significant effect of learning on RSR (**H2a**). This reveals a rare but possible phenomenon where the direct effect of a mediation is not significant, but the indirect pathways are. This may occur if learning does not fully mediate the effect of examples on RSR, and a confounder reduces the overall effect. For example, this confounder could be aversion. Future studies should take this into account. We also test the influence of our control variable initial domain knowledge. We find that initial domain knowledge has a positive significant effect on both learning and RSR.

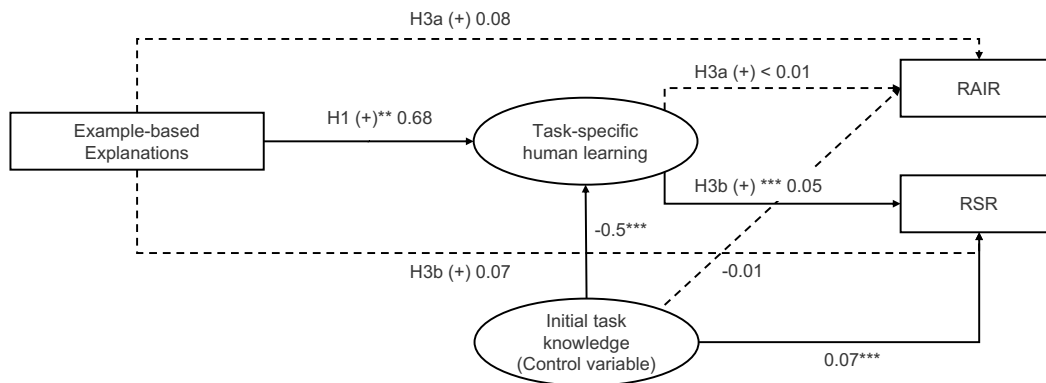


Figure 7.6.: Structural equation modeling results. Significance: (***) $p < 0.01$, (**) $p < 0.05$, (*) $p < 0.1$)

7.5.3 Explorative Sub-Group Analysis

In this section, we want to explore the interesting result that learning does not affect RAIR at all. Therefore, we analyze both conditions (baseline and XAI) separately and perform an exploratory subgroup analysis. Overall, we find significant differences between the SEMs fitted to the different conditions ($X^2 = 18.26$, $p = 0.02$). Table 7.4 on page 164 shows the path coefficients and p-values of the subgroup models.

In the XAI group, we find our expected positive significant effect of learning on RAIR (H2b). The analysis of the baseline condition reveals the reasons for the overall non-significant effect. In the baseline condition, the effect of learning on RAIR is also weakly significant, but the path coefficient is negative. This is an interesting result that deserves discussion. In order to answer this question, we need to understand what learning in the baseline condition actually means. Objectively, it is almost impossible to learn anything from the AI in the baseline condition as there is no reference to ground truth. This means that in the baseline condition, there is a tendency for humans to learn nothing or even unlearn patterns that were learned in the tutorial (which is reflected by the overall negative learning value). The limited learning potential during human-AI collaboration could then lead to automation bias (Goddard et al., 2014), i.e., blindly relying on AI advice. Automation bias, in turn, increases RAIR (Schemmer et al., 2022a). This could mean that learning in the baseline condition simply means less automation bias and, therefore less RAIR. To conduct a first validation of this reasoning, we test the impact of learning on absolute reliance³ in both subgroups. We find that learning significantly decreases reliance in the baseline condition (Path coefficient = -0.03, Z = -2.38, p= 0.02) and has no influence in the XAI condition (Path coefficient < 0.01, Z = -0.05, p= 0.96). Moreover, this reliance positively impacts RAIR in the baseline as well as in the XAI condition. This finding confirms our reasoning above. Future studies need to confirm our findings.

Table 7.4.: Subgroup analysis.

Path	Baseline Condition		XAI Condition	
	Estimate	P-Value	Estimate	P-Value
Learning – RSR	0.08	< 0.01	0.04	0.14
Learning – RAIR	-0.06	0.08	0.05	0.05

7.6 Discussion

In this work, we investigated two research questions. First, do explanations improve learning from AI during human-AI collaboration, and second, how does learning from AI influence AoR? To answer both research questions, we derived a research model and conducted a behavioral experiment.

³We measure reliance as the average number of times the participant followed the AI advice.

7.6.1 RQ1: The Effect of Learning on Appropriateness of Reliance

To answer our first research question (How does human learning during human-AI collaboration influence the Appropriateness of Reliance on AI advice?), we analyze the impact of learning on RAIR and RSR.

Applying SEM on our research model does highlight a significant effect of learning on the RSR. However, the direct path between explanations and RSR is not significant. This could mean that some confounders are prevalent that hamper the positive effect of explanations on RSR. One of those confounders could be automation bias (blindly relying on AI advice) (Bansal et al., 2021).

Interestingly, we, however, do not find a positive effect of learning on RAIR. Therefore, we conducted an exploratory sub-group analysis and found significant differences between the effect of learning in the XAI condition and the baseline condition. In the XAI condition, the expected positive effect of learning on RAIR is present. Based on our results, we deduce that a certain level of learning is necessary to observe any impact of learning on RAIR.

7.6.2 RQ2: The Effect of Explanations on Learning

To answer our second research question (Can explanations of AI increase human learning in human-AI collaboration?), we explore the impact of example-based explanations on human learning. We find a statistically significant difference between a baseline condition and the XAI condition. In the baseline condition, which may be the default configuration in many real-world applications, it may be objectively very difficult to learn anything at all from the AI, which is also confirmed by our experiment. In contrast, we observe statically significant learning in the XAI condition.

To reference our work back to IS research, we want to discuss the link to hybrid intelligence. In their seminal paper on hybrid intelligence, Dellermann et al. (2019a) discuss the impact of human-AI collaboration on mutual learning. Humans can teach AI new patterns (often referred to as human-in-the-loop systems (Dellermann et al., 2019a) or active learning (Hemmer et al., 2020)). However, the human side of mutual learning has been neglected. We complement IS research with the first work on human learning during human-AI collaboration.

Finally, we will discuss the implications of long-term human-AI collaboration for unique human and AI knowledge. Learning is only possible if the “teacher” has

some unique knowledge. The question now is whether humans and AIs can reach an equilibrium where the knowledge of both has aligned such that no further learning is possible. In the case of a static AI (and a human not continuing to learn independently of the AI), this could be the case. However, recent IS research (Kühl et al., 2022) has developed a more dynamic perspective on AI. AI models are frequently updated or being re-trained (the authors call this adaptive AI systems). Similarly, humans usually learn on the job, so they are continuously building unique knowledge. Especially in organizations, this is crucial to prevent knowledge loss and foster the distribution of knowledge (Engbom, 2020). Therefore, AI, as well as humans, are in a continuous learning process that creates unique knowledge.

7.6.3 Implications for Theory and Practice

Theory. Our research contributes to the literature on organizational learning (Levitt and March 1988) as well as appropriate reliance (Schemmer et al., 2023d). Even though learning on the job is a widely recognized approach for organizational learning, research has neglected the potential of in-process learning during human-AI collaboration. So far, learning from AI was always considered as part of a knowledge management tool. We, however, show that also in-process learning is possible and thereby opens up new research potential for the machine teaching domain. With regard to appropriate reliance, learning was so far neglected as an impact factor. With our work, we extend the research model of Schemmer et al. (2023d).

Practice. Our research implies that practitioners can, with the right design, leverage two benefits at ones from human-AI collaboration—upskilling of the workforce and better performance. With our study, we show the potential to learn from each other in a human-AI collaboration. Thus, these insights can be used to guide not only designers of AI systems but also knowledge managers within organizations to enhance the learning of humans.

7.6.4 Limitations and Future Work

Despite the contributions of this work, also limitations are present. First of all, our empirical work is limited by the choice of task as well as conducting a single study. However, we believe that image classification is a task with much potential for generalization following the arguments of Fügener et al. (2021). There are many real-world situations where humans need to classify images. Tasks can range from

low-stakes tasks, such as product quality inspection, to high-stakes tasks, such as cancer detection.

Additionally, the generalizability of our experimental findings is limited due to the choice of explanations. Both normative and comparative examples are a special type of XAI that is directly linked to ground truth. However, example-based explanations are state-of-the-art for learning systems (Goyal et al., 2019). Future work should evaluate different example techniques.

Additionally, the sequential task structure required for our measurement approach has certain drawbacks, as it alters the task itself. When humans first complete the task independently before receiving AI assistance, they are already mentally prepared and may respond differently than if they received AI advice immediately. However, this is a known challenge in research and not specific to our study (Schemmer et al., 2023d).

Most importantly, future research needs to investigate the impact of different design features. Future research should evaluate these potential design features to provide practitioners with a toolkit for the effective use of AI.

7.7 Conclusion

Over the last decade, the emphasis in research on AI adoption and acceptance by humans has facilitated the widespread use of AI in everyday life. The now widespread adoption raises the question of how to harness the potential of human-AI collaboration in the best way possible. The true potential of human-AI collaboration lies in leveraging their complementary capabilities in a way that jointly a performance is reached that is superior to individual AI or human performance. Therefore, to pave the way toward effective human-AI collaboration, it is now crucial to focus on appropriateness of reliance that realizes this potential. In this work, we study the effect of learning on appropriateness of reliance in a behavioral experiment. We find first evidence for learning as an impact factor of appropriateness of reliance and show that human learning can be influenced through explanations. Thus, this work contributes to the design of effective human-AI collaboration.

Part V

Harnessing Complementarity Potential
beyond AI-Assisted Decision-Making

Harnessing Complementarity in Anomaly Detection

This chapter comprises a working paper that is currently under review as Schemmer, M., Holstein, J., Kühl, N., Vössing, M., & Satzger, G. (2023c). From Anomaly Detection to Anomaly Investigation: Support by Explainable AI [Working paper]. Note: To improve the structure of the work, the title was changed. The abstract has been removed. Tables and figures were reformatted, and newly referenced to fit the structure of the thesis. The terminology was standardized with the dissertation. Chapter, section and research question numbering and respective cross-references were modified. Formatting and reference style was adapted and references were integrated into the overall references section of this thesis.

8.1 Introduction

Anomaly detection is essential in various domains, e.g., manufacturing (Ren et al., 2018; Susto et al., 2017), financial auditing (Debener et al., 2021), healthcare (Matschak et al., 2021; Schultz et al., 2022), and cyber security (Blazquez-Garcia et al., 2021; Gamboa, 2017). For example, engineers in manufacturing want to find early indicators of machine failures that would allow them to conduct preventive maintenance. In cyber security, experts aim to find security breaches and attacks. In financial auditing, detecting fraudulent claims is a crucial task (Schultz et al., 2022), with an estimated total of 13 billion euro in Europe in 2017 (Insurance Europe, 2019).

At the same time, manually detecting anomalies is very challenging even for human experts (Huang et al., 2022; Qian et al., 2020), primarily due to the vast amounts of data—in terms of granularity and variability—that need to be analyzed. For this reason, designers of information systems build so-called anomaly detection systems (ADSs) that aim to support human experts in identifying anomalies (Bhuyan et al., 2013; Breitenbacher et al., 2019; Moustafa et al., 2019). Information systems

research guides the development of such systems in many domains, e.g., auditing (Bhattacharya & Lindgreen, 2020), manufacturing (Ren et al., 2018), etc. Recently, more and more of these systems are based on machine learning (ML) (Garg et al., 2022) equipping them with the ability to detect complex patterns in high-dimensional datasets (Audibert et al., 2022). Due to the inherent rarity of anomalies, only limited amounts of labels are available for training ML models. Additionally, the term “anomaly” often encompasses a diverse range of underlying events (Wang et al., 2019). To address these challenges, unsupervised ML provides a practical solution by acquiring knowledge of normal data patterns and identifying anomalies as deviations from these (Cheng et al., 2021; Matschak et al., 2021; Steenwinckel et al., 2021).

Despite the capabilities of unsupervised anomaly detection techniques in identifying anomalous patterns embedded within datasets, they are insufficiently precise in the detection of *relevant* rare events and, instead, detect *any* anomalous pattern in the data. However, domain experts are often interested in a specific type of anomaly that is relevant to the business, e.g., an increase in temperature prior to a shut-in. Therefore, human experts need to carefully validate and investigate the detected anomalies to determine their business relevancy. As a result, ADS cannot perform the task of detecting relevant anomalies in an automated way. This means that human *anomaly investigation* is necessary to confirm whether an identified anomaly is indeed relevant.

While existing literature provides a comprehensive examination of anomaly *detection*, it falls notably short when it comes to providing systematic support for anomaly *investigation* (Chemweno et al., 2016; Pang et al., 2021; Steenwinckel et al., 2021). This deficit becomes increasingly apparent given that anomaly investigation often necessitates domain experts to scrutinize hundreds of distinct features (Liu et al., 2022), not all of which are necessarily relevant to a specific anomaly. Take, for example, a typical manufacturing production line that records thousands of measurements every second. Despite this volume of data, there currently exists no established systematic support to guide domain experts in their investigations. Considering the difficulty in identifying the features necessary to validate relevant anomalies, it becomes apparent that human domain experts require assistance in their anomaly investigation. Therefore we formulate our research question:

Research Question: *How can we methodologically support human experts in anomaly investigation?*

While an automated validation of anomalies may not be feasible in many cases, we hypothesize that unsupervised anomaly detection methods can still offer valuable information to human experts, enhancing the investigation process. Specifically, we hypothesize that explanations derived from anomaly detection can assist in human anomaly investigation. Therefore, our method embarks on a novel approach—it utilizes explanations generated from unsupervised anomaly detection to improve the investigation process. Our research specifically targets the investigation of anomalies in multivariate time series, which have numerous use cases for anomaly detection (Kieu et al., 2019; Malhotra et al., 2016), e.g., predictive maintenance (Choi et al., 2022), stock price movements (De Benedetti et al., 2018) or cyber security (Blazquez-Garcia et al., 2021). To evaluate the effectiveness of our proposed method, we conduct a behavioral experiment to test whether providing humans with these explanations improves the accuracy of anomaly investigation.

To instantiate our proposed method, we have chosen an LSTM-autoencoder as it is a commonly used approach for anomaly detection in time series data (Malhotra et al., 2016). Previous research has indicated that counterfactual explanations have shown potential in the realm of multivariate time series forecasting problems (Ates et al., 2021). This approach offers a nuanced means of understanding the forecasting model’s output by providing alternative scenarios that could have resulted in a different forecast. In essence, counterfactual explanations generate scenarios that are as close as possible to the actual data instance but with a different predicted outcome. This is accomplished by altering a minimal number of features, therefore highlighting the most influential factors driving the model’s decisions.

For our behavioral experiment, we leverage the public New York City taxi trip dataset (TLC, 2022) and design an autoencoder that detects anomalous patterns, i.e., events that have a significant impact on the local taxi industry. To generate explanations, we utilize the framework for counterfactual explanations for autoencoders proposed by Ates et al. (2021). As a testbed, we ask participants in the subsequent anomaly investigation task to differentiate between a specific event of interest and other events. Our experiment involves a total of 64 participants, and we find that the provided explanations can improve the accuracy of anomaly investigation.

Our contribution is twofold. First, we introduce a method to support human anomaly investigation, utilizing explanations derived from anomaly detection. Second, we demonstrate the effectiveness of these explanations through a behavioral experiment. Our study is the first to empirically investigate the impact of anomaly detection explanations on anomaly investigation. By validating their potential, we inspire

new use cases for anomaly detection which could have a significant impact on how anomaly detection systems are approached.

In the following sections, we will provide further details on our work. In Section 8.2, we will present the fundamentals and related work. Subsequently, in Section 8.3, we will conceptualize our anomaly investigation method. In Section 8.4, we will introduce our dataset, the design of our explainable autoencoder, and our experimental design. The results of our experiment will be presented in Section 8.5, followed by a discussion of the findings and a conclusion in Section 8.6. Finally, in Section 8.7, we will conclude our study.

8.2 Related Work

In this chapter, we introduce the fundamentals of our work and provide an overview of related work. First, we introduce foundations of anomalies, anomaly detection, investigation, and explainable anomaly detection. Then, we introduce the related work that covers explainable autoencoder-based anomaly detection in multivariate time series.

8.2.1 Anomaly Definition

First, it is imperative to define the term “anomaly” to establish a common ground. An anomaly is essentially a data point or a sequence of data points with substantial deviations from the majority of data points (Görnitz et al., 2013; Hawkins, 1980). The term anomaly does not describe a specific event but rather a property of those events. Those events are described as “unusual”, “rare” and simply not “normal” (Görnitz et al., 2013).

Anomalies can be categorized into different types: *Point anomalies* are the most trivial to find, as these anomalies are only single points located outside the normal value range. Next, *contextual anomalies* can consist of sequences and can only be identified as anomalous in comparison to different points with the same context. The most complex type is the collective anomaly. *Collective anomalies* always span over sequences and only gradually show a different pattern compared to normal data. Individual values within this type of anomaly may seem ordinary and only collectively raise suspicion (Braei & Wagner, 2020). Figure 8.1 on page 177 highlights the three types.

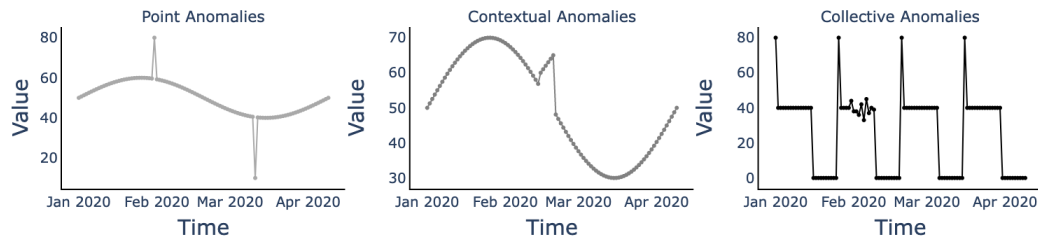


Figure 8.1.: Different anomaly types (Choi et al., 2021). The three examples depict univariate time series, with the three types of anomalies in time series.

In most use cases where anomalies are to be detected, the ultimate goal is not to detect any anomaly but relevant ones (Liu et al., 2022; Song et al., 2007). Therefore, most anomaly detection use cases essentially boil down to a rare event classification. For example, in manufacturing, the goal of operators may be to detect early indicators of machinery failures. However, not only these early indicators but also other events deviate from the “normal” operation, e.g., planned shut-downs. This means only a subset of the anomalies is actually of interest.

In the past, knowledge-based systems have been used to classify those rare events of interest, i.e., systems that explicitly store the knowledge of experts to detect the events (Steenwinckel et al., 2021). However, experts have a limited, more global view, and as data volume grows, it becomes harder for experts to explain deviations in values and their effects (Steenwinckel et al., 2021). Moreover, acquiring their knowledge is a time-consuming and challenging task (Steenwinckel et al., 2021). For this reason, more and more ADS were developed. However, ADS cannot classify the detected anomalies based on their relevancy. This means a human anomaly investigation is still imperative (Song et al., 2007).

8.2.2 Anomaly Detection

Anomaly detection approaches consist of either classification (e.g., isolation forest), nearest neighbor (e.g., distance-based), compression-based (e.g., autoencoder), or clustering methods (Muruti et al., 2018). Further, anomaly detection can be categorized into three different classes (Nassif et al., 2021). *Supervised anomaly detection* aims to build a classifier model that learns from a labeled training dataset. Here, the training dataset contains labels for normal and anomalous instances. In practice, it may be challenging to find such datasets due to anomalies being rare events and models requiring vast amounts of data. Next, *semi-supervised anomaly detection* requires only a training dataset with instances being labeled as normal. Accordingly, any instance different from the normal class is classified as

an anomaly. Finally, *unsupervised anomaly detection* is the most common type for anomaly detection as it does not require any labels in the training dataset, with autoencoders being one of the most powerful model classes.

While there are numerous methods to perform anomaly detection, the focus of this work is deep anomaly detection due to its superior performance (Chalapathy & Chawla, 2019). Deep anomaly detection describes the application of deep learning to the anomaly detection task. Autoencoders are the most frequently used deep unsupervised anomaly detection method (Chalapathy & Chawla, 2019). This architecture was already introduced in the 1980's (Rumelhart et al., 1985) and attempts to compress the input data to then reconstruct it with as little information loss as possible (Baldi, 2012). The most basic structure of an autoencoder contains an encoder that generates a compressed representation of the input data and a decoder that aims to reconstruct the input data from the compressed representation (Bank et al., 2020). Not all the information can be stored in the so-called bottleneck layer, so the model must learn statistical patterns in the training data (Bengio et al., 2009). These lower-dimensional representations are the latent space of an autoencoder (Dillon et al., 2021). For time series anomaly detection, autoencoders are usually equipped with LSTM layers that can capture temporal dependencies (Malhotra et al., 2016). To decide whether a sample is anomalous or not the autoencoder reconstruction error is used. If the reconstruction error of a sample exceeds a certain threshold, the sample is labeled as an anomaly. The threshold can be tuned manually or set by a certain percentage of the highest errors.

8.2.3 Anomaly Investigation

Only a few articles focus on anomaly investigation (Liu et al., 2022; Soldani & Brogi, 2022). Anomaly investigation can happen on multiple levels based on the necessary and available data (Soldani & Brogi, 2022). It can either be conducted based on the same data used for the anomaly detection or by taking additional data into account. This work focuses on use cases where the same data is used.

Most of the existing work deals with data visualization to improve anomaly investigation (Soldani & Brogi, 2022; Xue & Yan, 2022). Xue and Yan (2022) develop an ADS for detecting and analyzing anomalies in cloud computing performance. They provide rich visualization and interaction designs to help understand the anomalies in a spatial and temporal context. Soldani and Brogi (2022) improve the process of detecting and investigating anomalies in time series data in industrial contexts. To do so, they characterize six design elements and develop a visual ADS to support this

process. However, beyond visualization approaches, methods to improve anomaly investigation are still missing (Liu et al., 2022; Soldani & Brogi, 2022).

8.2.4 Explainable Anomaly Detection

Explanations are required to understand how specific predictions are generated (Šimić et al., 2021). On an abstract level, approaches can be divided into local and global explanations (Ates et al., 2021). Global explanations focus on the entire dataset (Ibrahim et al., 2019), whereas local explanations refer to individual observations (Plumb et al., 2018). By reviewing related work, it becomes apparent that there are many implementations of explainable anomaly detection (Choi et al., 2022; Song et al., 2018). However, none of them evaluate the effect on anomaly investigation. Overall, to the best of our knowledge, no study has ever empirically evaluated whether the explanations of the anomaly detection also provide a benefit for anomaly investigation. Therefore, we argue that this work's topic is highly relevant.

8.2.5 Explainable Autoencoder-Based Anomaly Detection in Multivariate Time Series

Within the context of multivariate time series, a lack of explainability approaches can be observed, while simultaneously, analytics for these time series are increasing in popularity (Ates et al., 2021). Counterfactuals are a promising explainability technique for time series (Filali Boubrahimi & Hamdi, 2022). While there are many counterfactual approaches in various domains, the multivariate time series domain remains mostly uncovered (Guidotti, 2022). Hereby, the work of Ates et al. (2021) is the only known framework for counterfactual explanations in time series forecasting and classification. As their approach is model agnostic, they only require class probabilities as the model's output to create explanations. To do so, they modify the input data in a way that is as close as possible to the original input while receiving a different class label. However, not all available input features are altered and, instead, only the ones with the highest deviations between the original input and the modified instance with a different label. Reducing the number of adjusted variables helps human experts as previous research has pointed out that humans are only capable of processing four variables simultaneously (Halford et al., 2005). A typical example of counterfactual explanations outside the domain of time series is a loan application scenario: An AI-based system declines a person's request, stating that

similar customers have also been declined. In contrast, a counterfactual statement can convey that the request would have been accepted if the person had slightly lowered the credit amount (Kenny & Keane, 2021).

During the review of related works that implement explainable autoencoder, it becomes apparent that most works utilize some form of feature importance as an explanation technique. Alfeo et al. (2020), Dix (2021), and Ghalehtaki et al. (2022) use the model's built-in reconstruction error to detect important features. However, Roelofs et al. (2021) argue that this methodology is not very robust as the reconstruction error does not always match the actual feature importance. Other work uses well-known frameworks such as SHAP or LIME to generate feature importance through a surrogate model (e.g., (Jakubowski et al., 2021)) or even deploy multiple SHAP explanations to capture temporal and feature interactions respectively (Hussain & Perera, 2022). Ha et al. (2022) calculates the feature importance through SHAP by applying a flattening layer on their LSTM autoencoder. The new model uses the weights from the autoencoder and generates explanations by using Gradient SHAP. Oliveira et al. (2022) designs its framework, the residual explainer, which interprets deviations of the reconstruction errors to create feature importance. In an experiment, the approach produces better results than SHAP and takes only a fraction of the time. The only work to our knowledge that uses an explanation technique besides feature importance is the work of Sulem et al. (2022), who also generate counterfactual explanations. However, none of them evaluate the impact of explanations on human anomaly investigation.

In summary, we find no study that provides methodological support for anomaly investigation and that empirically investigates the influence of explainable anomaly detection on anomaly investigation.

8.3 Conceptualization of Explainable Anomaly Detection for Anomaly Investigation

In this section, we outline our method for investigating anomalies. As discussed, we argue that the overarching goal of most anomaly detection use cases is actually to classify specific rare events—*relevant* anomalies. The small number of labels available for these rare events makes automated classification difficult. Unsupervised approaches can detect anomalies without labels (Cheng et al., 2021) and can find unknown patterns (Matschak et al., 2021). The underlying assumption of using unsupervised approaches is that the desired rare events are a subset of the

identified anomalies. While unsupervised anomaly detection methods can effectively detect anomalies, they are limited in their ability to fully automate the detection of relevant anomalies. Consequently, determining whether an identified anomaly is truly relevant involves human experts, as they can apply their domain knowledge to analyze anomalies in-depth (Liu et al., 2022). However, they are limited by their cognitive capacity, such as the patterns they can process. For example, when analyzing hundreds of sensors, humans may reach their cognitive limits. Thus, cognitive capacity may limit their ability to investigate anomalies.

To summarize, unsupervised ML-based ADSs are constrained in their capability to classify relevant anomalies, while human experts are limited by their cognitive abilities. To overcome these limitations, we adopt a human-centered perspective (Shneiderman, 2020) and develop a method that combines the strengths of automated anomaly detection with human anomaly investigation. We contend that this approach has the potential to overcome the challenges posed by both ADSs and human experts, resulting in a more effective and comprehensive approach to classifying rare events.

To reduce the required cognitive capacity for investigating anomalies, we propose to leverage recent advances in explainable AI. More specifically, we use explanations derived from the ADS to inform the anomaly investigation, which can explain the ADS's reasoning for flagging data as anomalous. Our underlying reasoning is that we hypothesize a relationship between the features influencing the automated detection of an anomalous event and the features a human needs to investigate the anomalies. For example, if the event is an early indicator for a machine failure, we hypothesize that the sensor values that behave unusual also give an insight that the event is an early indicator and not, for example, a planned shut-down. Besides giving insights for the investigation, we also hypothesize that explanations can be used to reduce the cognitive effort of human experts. For example, showing humans a filtered list of sensors relevant to anomaly detection would reduce the number of sensors a human expert needs to analyse. In summary, we argue that explanations of the anomaly detection can help humans in their anomaly investigation.

Having outlined the general idea of the method, in the following, we provide a detailed overview along with the four components of the method—data, anomaly detection, explanations of the anomaly detection, and the human investigation. Figure 8.2 on page 182 visualizes the method. The first two components are related to anomaly detection, and the last two are related to anomaly investigation. We highlight the components of the method, an exemplary instantiation, and provide

an illustration. As an example, we use the detection of early indicators for machine faults.

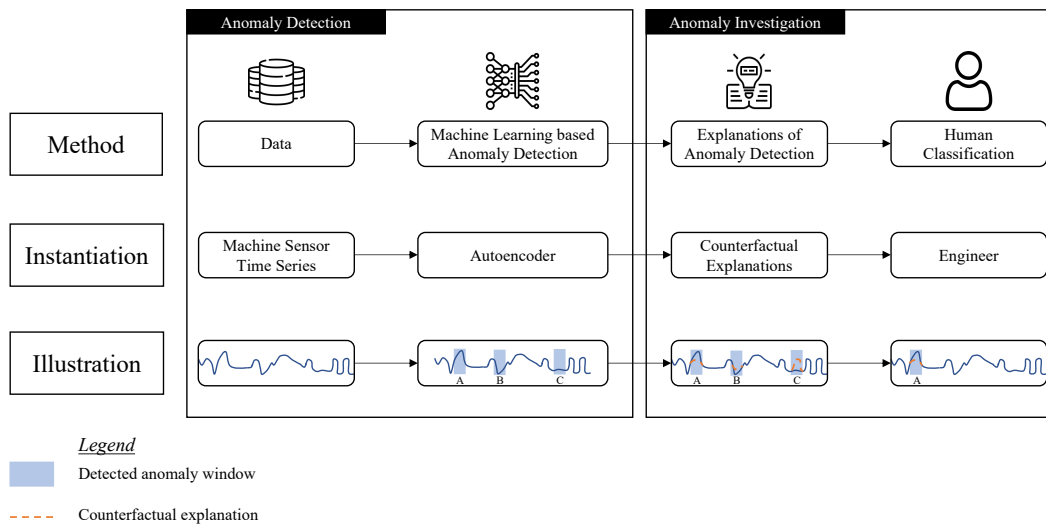


Figure 8.2.: Representation of the proposed method. Explanations generated from the anomaly detection support the human anomaly investigation.

Data. For our method, the underlying data needs to have enough events of normal behavior to be able to use an unsupervised approach. In our machine fault classification example, the input data to the method are sensor values. In the illustration in Figure 8.2, we depict a univariate time series, e.g. a single sensor value such as temperature over time.

Unsupervised Anomaly Detection. Next, an unsupervised ML-based anomaly detection flags anomalies in the input data—indicating anomalous events that require human analysis. Unsupervised anomaly detection models operate by learning what is normal or what are expected based on various characteristics of the data. Following this any data point that significantly deviates from this established norm is then considered as an anomaly. This method does pose certain challenges, primarily in determining the threshold at which a point is flagged as an anomaly. Setting this threshold usually involves a tradeoff whether to detect more anomalies (low deviations are considered as an anomaly) or less anomalies (only high deviations lead to anomalies). Commonly employed methods for unsupervised anomaly detection are, for example, auto-encoder, principal component analysis, or isolation forest. In our example, the output of this step are highlighted time windows in the sensor data that are anomalies with respect to the normal behavior of the machine. It is important to note that a human will only examine the time windows highlighted by the anomaly detection.

Explanations of Anomaly Detection. As a next step, explanations of the unsupervised anomaly detection are generated. Common approaches are ex-post explanations such as feature importance (highlighting the relevant sensor or timestamps), example-based or rule-based explanations. Research on supervised approaches has highlighted the importance of a human-centered perspective for designing explanations (Ehsan et al., 2021). Following this line of thought, we argue that explanations need to be consciously designed for the problem at hand. For example, in the time series domain, research has shown the advantages of counterfactual explanations (Ates et al., 2021). In our example, we show counterfactual explanations. This means the explanations highlight how the sensor data should have behaved to be not labeled as anomalous.

Human Anomaly Investigation. In the final step, a human expert receives the explanations and differentiates the detected anomalies in for the business problem relevant and non-relevant anomalies. This method step may be conducted by business experts, lay workers, etc. In our example, the anomalies are investigated by engineers. The outcome are those events that the engineer identifies as early indicators for a machine breakage.

Having outlined our proposed novel method, in the next chapter, we present the design of our experiment to demonstrate the value of the method.

8.4 Experiment Design

In this section, we provide information about the data and the task we use to investigate the utility of our proposed method. Then, we describe the development of the autoencoder and the counterfactual explanations, together forming our explainable ADS. Finally, we present the experimental design.

8.4.1 Dataset, Task and Data Preprocessing

Dataset. We search for a suitable task and dataset by specifying a list of requirements the dataset must fulfill. The dataset must consist of multivariate time series and must include anomalies and, ideally, external information about the respective anomalies. Since the participants are non-experts, the dataset must come from a context they can understand.

Based on these requirements, we evaluate several well-known multivariate benchmark datasets frequently used in anomaly detection on multivariate time series (e.g., (Du et al., 2017; Risdal et al., 2016)). All these datasets are multivariate and stem from a technical context. While these characteristics are desirable for a technical evaluation of a model, they conflict with our requirement to be easy enough to understand by experiment participants.

For this reason, we picked a dataset with a more familiar context. One dataset that meets all these requirements is the public New York City Taxi dataset (TLC, 2022). Currently, around one million trips are recorded every day (TLC, 2022). TLC has made this data available to the public since 2009. Each trip record contains 19 features, e.g., information about the pick-up and drop-off time and location, the trip distance, payment types, fares, and the number of passengers. Nearly 13 years of data are available - in these years, the taxi industry has changed considerably. Fares, availability of cabs, or, for example, new competitors have, among other factors, influenced the collected data and represented a considerable challenge (Baier et al., 2019) that is out of the scope of this work. We address this issue by using a shorter period of observation.

Certain days, such as holidays or days with extreme weather conditions, cause considerable deviations from the usual behavioral pattern. These days are thus suitable as anomalies because they are out-of-distribution by nature while serving as ground truth at the same time (Ferreira et al., 2013). For extreme weather events, ground truth can be found on the governmental extreme weather website ¹. All of the anomalies are collective anomalies, e.g., they are just anomalous as a sequence. During the chosen timeframe from the beginning of 2016 to the end of December 2018, several events with known large impacts on the taxi business took place, for example:

- Christmas (12/24/2018 - 12/26/2018)
- New Year's Day (01/01/2018)
- Winter storm (11/15/2018)
- Heavy snowfalls (03/21/2018)

For the training of our model, we use 2016 as the training and 2017 as the validation period. 2018 serves as our test set, of which we visualized a subset in Figure 8.3 on page 185. The colored areas shown indicate known events in New York City.

¹<https://www.weather.gov/okx/stormevents>

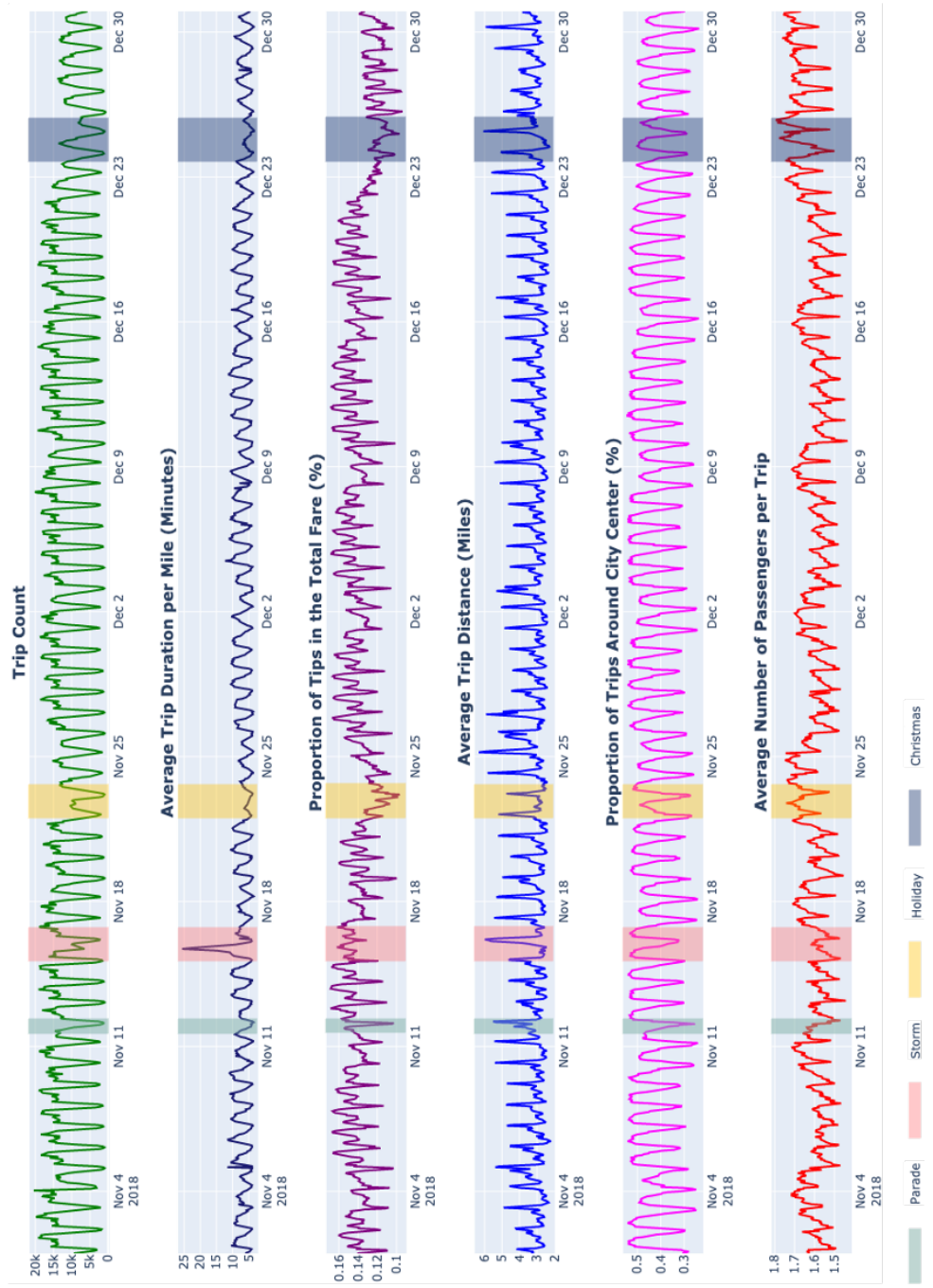


Figure 8.3.: Excerpt of test data with highlighted event days.

Task. Accordingly, we want to provide an easy-to-understand task based on the dataset that supports the classification of the identified anomalies. Our dataset lays the ideal basis for this task as the anomalies have different classes, e.g., public holidays, extreme weather events, or other events. While it may not be possible to differentiate between such events deterministically, some frequently appearing patterns can be observed, e.g., during extreme weather, fewer people use taxis and, at the same time, the share of the tips increases. Therefore, we provide participants with the task of classifying whether the shown anomaly is an extreme weather event or not. We employ a binary classification (extreme weather event or not) to be close to a realistic task. For example, in condition-based maintenance, the binary classification may be to discriminate faults from other anomalies.

Data Preprocessing. As the data is provided directly from the recording, it is necessary to clean and preprocess it. The goal of the preprocessing is to increase the data quality and, therefore, also the performance of the ADS (Frye et al., 2021). We merely make basic assumptions that ensure the validity of individual recordings while not removing any anomalies the model should detect, e.g., the trip duration should be longer than zero minutes (Baier et al., 2020). After the data cleaning, we aggregate the taxi demand by hour and perform a few preprocessing steps. To increase the comprehensibility of the dataset, we drop some of the original 19 dimensions, as they are sometimes difficult to understand and negligible for anomaly detection. Further, we create additional features that are easy to interpret and thus support the classification task. Therefore, our final dataset consists of the following features: trip count, average trip duration per mile, the proportion of tips in the total fare, average trip distance, proportion of trips starting and ending in the city center, and, finally, the average number of passengers per trip. Lastly, we scale the data to unify the magnitude of different features (Misra & Yadav, 2019), as this can have otherwise undesired results on the model's processing (Papenmeier et al., 2019).

8.4.2 Explainable Anomaly Detection System

Modeling. In the following, we present our anomaly detection modeling. Similar to related work (Ghalehtaki et al., 2022; Ha et al., 2022; Jakubowski et al., 2021), we use LSTM-layer to take intertemporal and multivariate dependencies into account. To optimize our architecture, we conduct a grid search and identify the following parameters as the best combination: window size of 8, step size of 2, hidden dimensions of the autoencoders are 8, 6 and 4, and finally the latent space with 4 dimensions. We use the reconstruction error and the detection of known anomalies as the target.

Our approach is to calculate the average reconstruction error of a window over all timesteps and features and compare this value to a threshold. The threshold must be optimized based on the results. The goal of this optimization is the recall of the model, meaning that all anomalies are identified as such by the model.

Counterfactual Explanations. The standard autoencoder architecture must be extended to enable explanations for common explanation frameworks such as SHAP (Lundberg & Lee, 2017) or CoMTE (Ates et al., 2021) that cannot handle the autoencoder output. This is because the autoencoder output has the same dimensions as the input data. Current explainable AI frameworks, however, expect outputs in the form of a classification or regression prediction. Thus, we design a new layer that manipulates the model's output to provide class probabilities (Ates et al., 2021). To calculate the necessary class probabilities, a new layer is given a threshold value in addition to the already existing sum of the reconstruction error, as proposed in (Ates et al., 2021). The reconstruction error is then, similar to (Aronsson & Bengtsson, 2021; Ates et al., 2020), converted to a binary class probability by first subtracting the threshold value (τ) from the calculated mean error. The Sigmoid function afterward projects that value to a range between 0 and 1. The layer is concatenated after training the autoencoder.

Finally, we use the CoMTE framework to generate the counterfactual explanations (Ates et al., 2021), which serve two purposes. First, they reduce the number of features that experts need to analyze (On average, our approach changes 3.2 features). Second, they highlight how the time series should have looked liked to be not flagged as anomalous. Figure 8.4 on page 188 depicts two examples of our explanations. The used approach modifies four input features in the extreme weather event and three for the public holiday.

Having introduced our explainable ADS, we now describe how we conduct our behavioral experiment.

8.4.3 Experimental Design

Pilot Study.

To obtain qualitative feedback on our ADS, we conduct a focus group interview with five experts with backgrounds in ML. The session lasts 30 minutes. First, we briefly introduce the basic information about this work and then focus on the core of the study. There, an anomalous window with the respective features is presented.

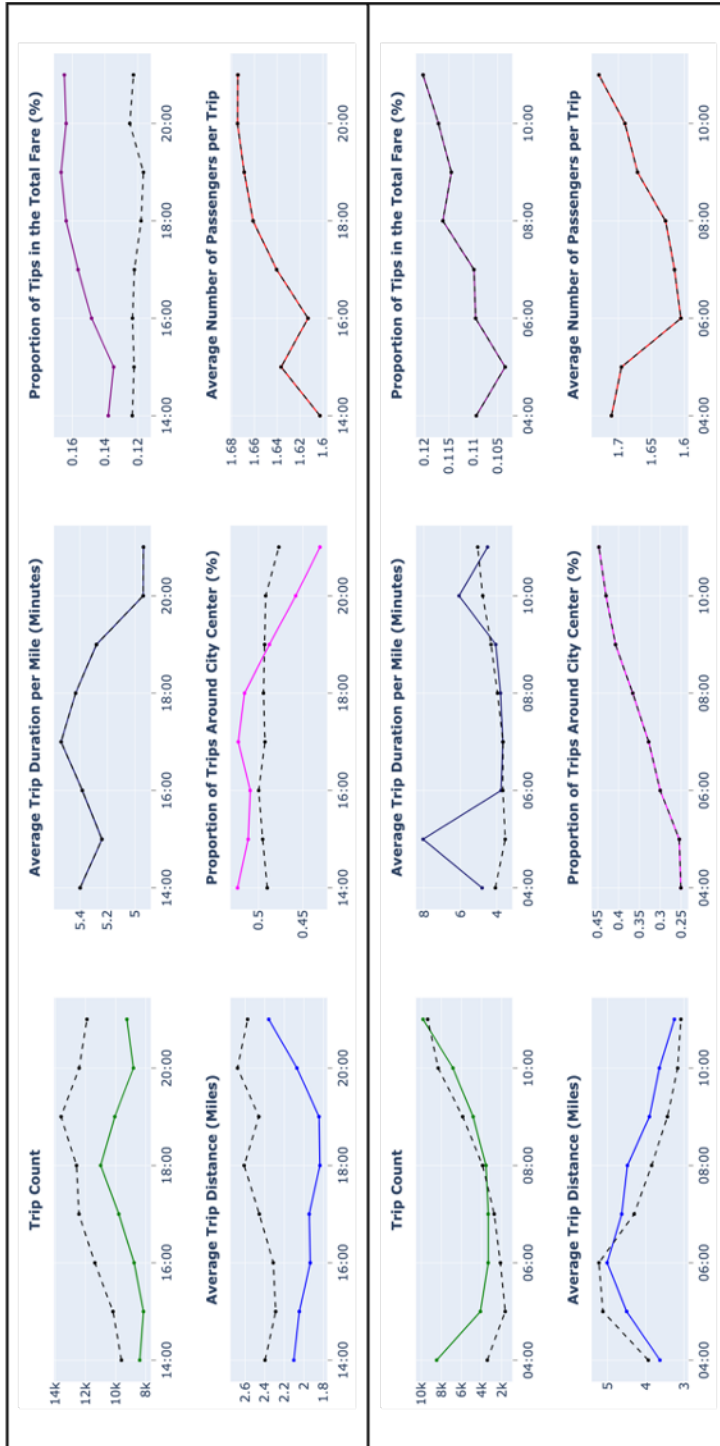


Figure 8.4.: Exemplary explanations for extreme weather event (top) and public holiday (bottom). The dotted lines visualize the counterfactual explanations.

The session is recorded and transcribed to evaluate the results more precisely in retrospect (McGrath et al., 2019).

The experts argue that the only cue generated by the explanations, the distance between the two lines of the counterfactual explanation, is not enough. We observe that it is vital to provide exemplary patterns in the pre-training of the user study to ensure that participants can understand the events being classified. We argue that this also transfers to real-world cases, as domain experts also have prior knowledge within their domain, which they incorporate into the anomaly investigation.

Study Procedure.

The research model is tested in an online experiment with a between-subject design. We test two different conditions. First, a *control* condition in which the human receives the ADS without counterfactual explanations, and second, a counterfactual explanation (*CF*) condition. The study as a whole is approved by the University IRB.

Sampling Strategy. In each condition, we provide participants with eight events. For the sampling of the eight events, we apply rules to ensure that the patterns of the underlying event are visible and prevent participants from being able to classify events based on previously seen anomalies, e.g., the same extreme weather at two different times during a day. Therefore, we first label our anomalies based on the provided dates, start and end times of extreme weather of the government storm website and public holidays². For extreme weather events, we label an anomaly as extreme weather if the identified anomaly starts at most 2 hours before the start of the extreme weather or two hours before the end of the extreme weather. Similarly, we label anomalies as a holiday if they start on the date of a public holiday. Finally, we randomly draw four extreme weather events, three holiday events, and one anomaly classified as neither. While sampling, we verify that we do not draw two anomalies on the same day.

Interface. Next, we create visualizations of the sampled anomalies (see Figure 8.5 on page 190). Similar to Liu et al. (2022), we use two views with varying information: the context view and the zoomed view. First, our context view shows past data of the last three weeks for all variables, with the anomaly being highlighted. This should support participants in understanding the behavior and interactions of the variables in non-anomalous times and thus provides context (following the requirements from the pilot study). However, we refrain from flagging additional anomalies in this

²<https://publicholidays.com/us/new-york/2018-dates/>

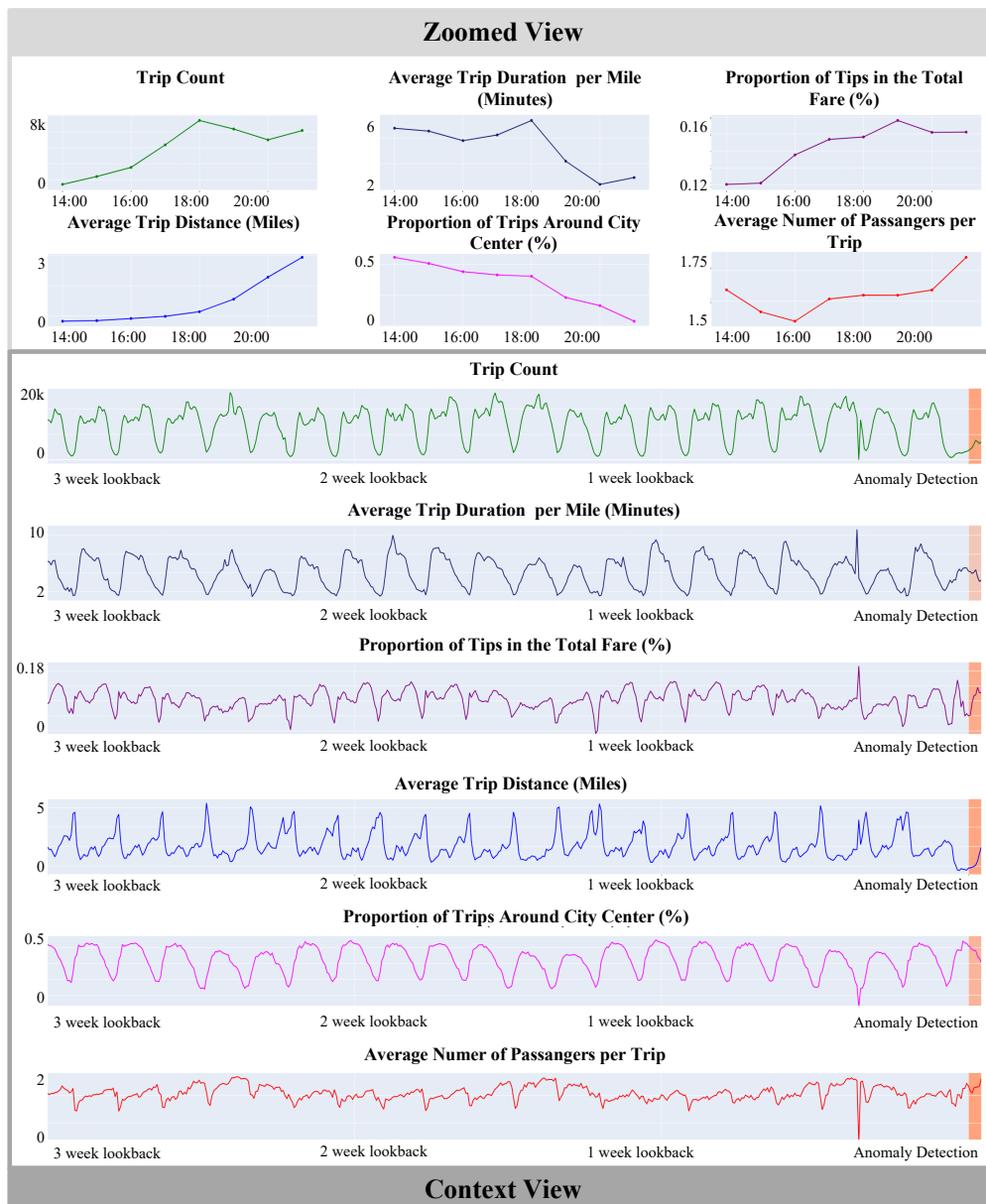


Figure 8.5.: Example of an extreme weather event with counterfactual explanations (top: zoomed view, bottom: context view).

period to avoid the possibility of inferring the date based on the position of the anomalies, e.g., New Year's Day, by the previous anomalies of Christmas. Second, the zoomed view allows a detailed look at the specific time window of the anomaly. It is also the single point in which the treatments differ. While the AI treatment solely receives the data during the anomaly, the zoomed view of the counterfactual treatment additionally displays explanations. To support the anomaly investigation, we provide supplemental information about the time of the anomaly to enable a

better understanding of the anomaly's patterns without allowing conclusions on the specific date (e.g., Christmas). For example, while an extreme weather event may result in fewer trips, this is also true for nights on regular working days. Additionally, we argue that in real-world cases of anomaly, investigation time is a feature that is also available.

Task flow. The online experiment is initiated with an attention control question that asks participants to state the color of grass. To control for internal validity, participants are randomly assigned to the condition groups. As multivariate time series are difficult to interpret for humans (Janus et al., 2021), we include multiple tutorials. First, both conditions receive an introduction to the task and are given examples of extreme weather events and other events. Following that, we explain the two views of our ADS and ask participants four comprehension questions. Afterward, we give a short tutorial on how participants can detect extreme weather events, followed by two comprehension questions. Finally, we sample event patterns based on related literature (Lee & Sohn, 2020; Qing et al., 2015). For the CF condition, we follow on with an explanation of counterfactual explanations. We provide the participants with a general intuition of the explanations rather than specific technical information. During the experiment, we neither use the terms AI and ML nor counterfactual explanations to prevent issues of AI literacy. Instead, we speak of ADS and expected values. Then, the participants conduct two training tasks to familiarize them with the task and, depending on the condition, with its explanations. Additionally, the participants receive feedback on the training tasks. After the two training reviews, the participants receive the eight main tasks. For each task, we ask them how much they agree with the statement “The anomaly is an extreme weather event” on a four-point Likert scale (Strongly agree, agree, disagree, strongly disagree). This allows us to get a binary classification and certainty information. After classifying the anomalies, we collect data on demographic variables.

Reward. To incentivize the participants, they were informed that for every correct decision, they get an additional 12 cents in addition to a base payment of 6 pounds per hour. However, the two training classifications do not count for the final evaluation.

Participant information. The participants are recruited using the platform “Prolific.co”. We note that crowd workers might limit the generalizability of our results. However, our sampling of the task should ensure that crowd workers are capable of doing the task. In total, we conduct the experiment with 66 participants (33 participants per condition). We exclude two participants in the CF condition and two participants in the control condition because of conducting the eight tasks in under

one minute. Apart from the attention check, we provide participants with in total of six questions that ensure that participants understand the task and underlying visualization, e.g., how many weeks of data are displayed in the context view. Based on these questions, we further exclude eight participants in the CF condition and nine participants in the control condition for incorrect answers. Even though this might seem like a high number of excluded participants, one needs to consider that multivariate time series anomaly detection is a very challenging task, and some crowd workers may not even understand what a time series is. In addition, we exclude all participants who fall outside the interquartile range of 1.5. By doing so, we exclude two outliers in the CF condition. This leaves us with 22 participants in the control group and 21 in the counterfactual group. Table 8.1 shows the age, gender, and education distribution of the participants.

Table 8.1.: Summary of participants' characteristics.

Number per condition	Control = 22 Counterfactual Explanations = 21
Age	$\mu = 27.12, \sigma = 6.29$
Gender	47 % Female 47 % Male 6 % Non-Binary
Education	26 % High school 51 % Bachelor 12 % Master 11 % Other

Evaluation Measures.

To evaluate our hypothesis, we calculate two measures based on the results of our experiment—effectiveness and efficiency.

Our first measure, effectiveness, is the participants' accuracy in the anomaly investigation, e.g., the share of correctly classified events. To calculate the share, we first binarize the result. Due to our sampling strategy, by chance, participants would be able to have an accuracy of 50 percent. In addition to the accuracy, we analyze the participants' certainty by comparing the share of agreement and disagreement with

the percentage of strong agreement and disagreement. Next, efficiency represents the time needed for the anomaly investigation. For both measures, we calculate the mean per participant for the global evaluation of our hypothesis. Additionally, we examine the effects of the explanation on each type of event more closely. Therefore, we also build the mean for the four extreme and non-extreme weather events.

8.5 Results

In the following section, we report the results of our study. First, we provide a qualitative interpretation of typical patterns of detected anomalies that could have been observed by the participants of the experiment. Finally, we present an analysis of the experiments' results.

8.5.1 Qualitative Interpretation of Detected Anomalies

As mentioned earlier, participants in the experiment must classify identified anomalies in extreme weather events. This classification is based on the intuition that each type of event has common patterns that are shared across events. However, these patterns often do not allow for a deterministic classification of anomalies. Nevertheless, in the following, we qualitatively present and interpret certain patterns derived from counterfactual explanations.

Extreme Weather. During extreme weather events, three variables in particular differ from normal days: the number of trips is lower, the percentage of tips in the total fare increases, and the average trip distance decreases. For an example of a winter storm, see the top of Figure 8.4. We interpret this pattern to mean that more people stay home on stormy days and forgo longer trips, such as visiting relatives or friends in other neighborhoods. Furthermore, people leave their homes only for urgent matters and then rely on taxis. Once they arrive at their destination, they express their gratitude to the taxi drivers with an increased tip due to the adverse circumstances.

Holidays. Compared to extreme weather events, holidays often have a distinct pattern (see the bottom of Figure 8.4). On holidays, the number of trips during the night is usually higher, and later the average number of passengers per trip is higher. People often go out the night before, so there are more trips during the night compared to regular days. Compared to regular workdays, we interpret the higher number of trips during the night as people going out, resulting in more trips.

The second observation with more people sharing taxis could be families visiting relatives or friends together.

8.5.2 Experiment Results

In Table 8.2 we highlight the descriptive results of our experiment. The results are broken down by experimental condition. Finally, we evaluate the significance of the results using Student’s T-tests or Mann-Whitney U-tests after controlling for normality using the Shapiro-Wilk test. In the following, we first present our results on effectiveness and then highlight the impact of explanations on efficiency. Figure 8.6 and Figure 8.7 visualize the results of our experiment.

Table 8.2.: Descriptive outcomes.

Condition	Effectiveness	Efficiency
Control	55.11 % (21.71 %)	30.92 s (22.2 s)
Counterfactual Explanations (CF)	70.24 % (15.04 %)	25.89 s (12.47 s)

Effectiveness. In the control condition, participants have an average accuracy of 55.11 %, or an average of 4.41 correct classifications. In the CF condition, we measure an average accuracy of 70.24 %, which corresponds to an average of 5.7 correct classifications with the help of counterfactual explanations. The analysis of the results of the experiment shows that the mean accuracy of the participants in the CF condition is significantly higher than in the control condition (Mann-Whitney U-test: $U = 137.5$, $p = 0.02$). Thus, we can conclude that *explanations improve the effectiveness of anomaly investigation*. In addition, Figure 8.6 shows that the interquartile range is lower when explanations are provided.



Figure 8.6.: Distribution of the effectiveness of participants.

Efficiency. On average, participants took 30.92 seconds to complete the eight classifications in the control condition (approximately 3.87 seconds per classification). In the CF condition, they took 25.89 seconds (approximately 3.24 seconds per classification). Figure 8.7 shows the measurements. No significant effect on classification time was observed between the CF and control conditions (Mann-Whitney U-test: $U = 255.5$, $p = 0.56$). We conclude that counterfactual explanations do not significantly increase human anomaly investigation efficiency in our setup. We discuss possible reasons for the non-significant effect in the next chapter.

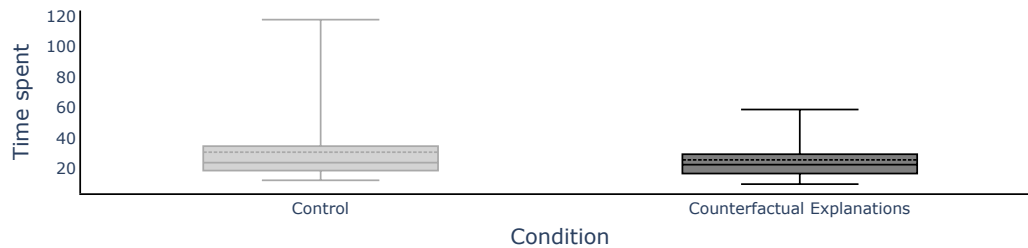


Figure 8.7.: Distribution of the efficiency of participants.

In total, our experiment highlights the potential for using counterfactual explanations to improve anomaly investigation. In the following section, we discuss our results.

8.6 Discussion

In this paper, we propose a method to support anomaly investigation. We suggest using explanations derived from anomaly detection to guide human anomaly investigation. We instantiated this method to demonstrate its usefulness and to analyze two facets of anomaly investigation—accuracy and efficiency. To this end, we conducted a behavioral experiment that showed that providing counterfactual explanations can improve the accuracy of anomaly investigation, particularly in distinguishing weather events from non-weather events. That is, humans can identify the subset of relevant anomalies from a larger set of detected anomalies. However, providing counterfactual explanations does not significantly improve efficiency in our current setup. A possible reason for this lies in the nature of counterfactual explanations—the visualization of a second instance that must be interpreted and compared to the anomaly to be classified. This additional effort may offset effects that improve efficiency, such as easier classification.

8.6.1 Implications

To our knowledge, we are the first study to show empirically that explainable anomaly detection can improve anomaly investigation. This result has important implications for research and practice.

Implications for Research. In our research, we propose an anomaly detection and investigation method that extends the well-researched area of anomaly detection with complementary anomaly investigation. We then instantiated the method and tested it in a behavioral experiment. We show that counterfactual explanations can improve human accuracy in detecting relevant anomalies from a larger set of anomalies. By demonstrating the usefulness of our method, we extend the existing research on anomaly detection with the complementary anomaly investigation.

Implications for Practice. Our work has implications for any use case with large amounts of data and rare classes of interest. In previous work, these use cases have typically been characterized as unfavorable for ML (Pang et al., 2021). However, we argue that they require a different approach. Instead of using supervised ML with sampling strategies (e.g., SMOTE), we hypothesize that using explainable anomaly detection together with human expert-based anomaly validation—i.e., human-AI collaboration—may be superior. To maximize the potential benefits of such a system, organizations should consider training their experts in AI-derived explanations to increase their familiarity with explanations and potentially increase the efficiency of anomaly investigation. By improving the usability of these interfaces, companies could increase the efficiency and effectiveness of their decision-making processes, leading to better outcomes and increased competitiveness.

8.6.2 Limitations and Future Work

As always with behavioral experiments, there is the question of how generalizable our results are. We would like to emphasize that our research does not aim to recommend general design features (e.g., the use of counterfactual explanations for time series), but rather to show that explanation can improve the investigation of anomalies per se, i.e., it shows the existence of the effect. We further argue that showing that it works with laypeople in an online experiment shows even more potential for experts. However, more work is needed to explore design recommendations and whether our findings hold in other domains. We encourage other researchers to test our method in different settings and derive design knowledge.

We also want to discuss how realistic our chosen testbed is. Therefore, we compare the classification of taxi events with a typical use case for anomaly detection in manufacturing. In manufacturing, anomaly detection is often used to detect early indicators of machine failures (Jankauskas et al., 2023). Compared to our task, the number of features in manufacturing is usually even higher. However, based on the expertise of researchers, we argue that the number of relevant features informing classification tasks is usually similarly small.

A key open question is how our method is perceived by end users. A key criterion for anomaly detection in the past has been the reduction of “false alarms,” i.e., detected anomalies that do not belong to a class of interest, in order to reduce the workload of experts (Campos et al., 2016; Pang et al., 2019). However, our results show that explanations sometimes do not reduce the time needed to interpret the anomalies. This means that explanations may be perceived as inconvenient. Future research needs to investigate whether experts find explainable ADS useful and adopt it.

In addition to detecting infrequent events of interest—thus acting as an alert system—explainable anomaly detection has the potential to act as a data mining tool to generate new knowledge for organizations. Anomaly detection can find patterns previously unknown to experts (Chandola et al., 2009). Explanations could allow experts to validate these patterns and generate entirely new insights. Future work could provide insight into this additional use case.

8.7 Conclusion

In this work, we address the problem of investigating anomalies in terms of their relevance. We develop a method that assists human experts in anomaly investigation by providing them with explanations of unsupervised automated anomaly detection. We then conduct a behavioral experiment and show that counterfactual explanations of autoencoder-based anomaly detection improve anomaly investigation in multivariate time series. With our results, we hope to motivate researchers and practitioners to research, implement, and use explainable anomaly detection.

Harnessing Complementarity in Intelligent Decision Assistance

This chapter comprises an article that was published as: Schemmer, M., Kühl, N., & Satzger, G. (2022d). Intelligent Decision Assistance Versus Automated Decision-Making: Enhancing Knowledge Work Through Explainable Artificial Intelligence. *Proceedings of the 55th Hawaii International Conference on System Sciences*, 617–626. Intelligent decision assistance versus automated decision-making: Enhancing knowledge workers through explainable artificial intelligence. Note: To improve the structure of the work, the title was changed. The abstract has been removed. Tables and figures were reformatted, and newly referenced to fit the structure of the thesis. The terminology was standardized with the dissertation. Chapter, section and research question numbering and respective cross-references were modified. Formatting and reference style was adapted and references were integrated into the overall references section of this thesis.

9.1 Introduction

The recent advances in artificial intelligence (AI) lead to an increase in automated decision-making (Coombs et al., 2020). Decisions can be classified into unstructured, semi-structured, and structured decisions (Turban et al., 2010). Traditionally, automated decision-making was applied to structured problems, and decision support system (DSS) enhanced decision-making for unstructured problems (Gorry & Scott Morton, 1971; Turban et al., 2010). Unstructured tasks were considered too difficult to automate since they require more cognitive flexibility (Lacity & Willcocks, 2016). However, advances in AI, specifically in deep learning, now increasingly enable to automate also more complex cognitive tasks, such as driving a car (Frey & Osborne, 2017). Therefore, AI has now the potential to also address semi-structured and unstructured decisions that are far from basic back-office tasks (Asatiani et al.,

2019). For example, AI is used to automate loan approval (Infosys, 2019), or to conduct recruitment choices (Albert, 2019)—decisions that in the past were unimaginable to automate. Therefore, both the number and complexity of tasks that can be automated increase.

However, it has long been known that increasing automation of decisions can lead to various drawbacks, such as automation bias and deskilling (Mosier & Skitka, 1999; Parasuraman et al., 2000). This is especially challenging since most semi-structured and unstructured tasks are knowledge work incorporating high-stake decision-making, e.g. medical diagnosis or jurisdictional decisions. In general, AI for knowledge workers should automate routine and assist knowledge-intensive work with reasoning and other high-level functions (Adelstein, 2007). The deskilling of knowledge workers is a major problem, as they are the people who should train, challenge and evolve AI. Knowledge workers create the labels for the AI that is the foundation for its initial training. After changes in the environment of the AI knowledge workers adapt and develop new solutions based on their domain expertise (Baier et al., 2019). Furthermore, they should be able to challenge the AI's recommendation, either with regard to performance but also with respect to ethical and fairness concerns. While in many use cases these disadvantages may be negligible there are cases where they must not be ignored. Reasons include, but are not limited to, losing significant competitiveness, e.g. in asset investment strategy decisions, or even potentially harming people, e.g. in medical diagnoses.

Because DSS are explicitly designed to not automate but support decision-makers (Arnott & Pervan, 2016), the initially obvious idea emerges to address these problems by using DSS instead of fully automated systems. However, automation should not be interpreted as a binary state but instead as a continuum (Parasuraman et al., 2000). Negative impacts already occur at low automation levels (Parasuraman et al., 2000)—as positive features of human decision-making are reduced such as human engagement. Therefore, when speaking about automated decision-making, we use the broader understanding of the continuum mentioned above, also including lower automation levels. As many state-of-the-art DSS do include automated, AI-based recommendations (Turban et al., 2010), they are subject to negative impacts, like automation bias in the short, reduced engagement in the medium, and deskilling in the long term. Thus, we perceive a major research gap in supporting human decision-making without those downsides, and formulate:

RQ: *How can we design AI for decision support without introducing automation disadvantages?*

Based on automation and DSS research, we conceptualize a new class of DSS, *Intelligent Decision Assistance* (IDA), that reduces automation-induced disadvantages while still preserving decision support levels. From the automation literature, we draw the critical evaluation of potential disadvantages of automated decision-making and the awareness of a continuum between full automation and human agency (Parasuraman et al., 2000). From DSS literature, we use the concept of guidance (Morana et al., 2017). Part of guidance theory is the explainability of DSS (Morana et al., 2017) which is a traditional topic of IS research (Meske et al., 2022). We discuss various combinations of automation levels and explainability and eventually follow the idea of informative guidance as a guidance that foregoes to provide explicit recommendations (Silver, 1991). In line with this notion, we propose to withhold the AI's decision and let the human "brainstorm" together with the AI by providing techniques from the Explainable AI (XAI) knowledge base (Adadi & Berrada, 2018), such as examples, counterfactuals, or feature importance. After conceptualizing IDA and deriving hypotheses on its impact, we provide first evidence for their validity through a systematic evaluation of empirical studies in the literature. With our work, we contribute to research and practice by conceptualizing a new class of DSS—*Intelligent Decision Assistance*.

9.2 Literature Review

In general, IS are designed to support or automate human decision-making (Zuboff, 1985). These two purposes are traditionally analyzed in two different research streams: *decision support* is traditionally covered in DSS literature (Power, 2007), while *Automation* is mainly addressed in Ergonomics literature (Coombs et al., 2020).

9.2.1 Decision Support Systems

DSS represent an important class of IS that aim to provide decisional advice (Arnott & Pervan, 2016). In general, "DSS is a content-free expression, which means that there is no universally accepted definition" (Turban et al., 2010, p. 16). However, DSS can be used as an umbrella term to describe any computerized system that supports decision-making in an organization (Turban et al., 2010). Originally, DSS were defined as supportive IT-based systems, aiming at supporting and improving managerial decision-making (Arnott & Pervan, 2016; Young, 1983). Later developments in DSS opened the area for application to all levels of an organization

(Arnott & Pervan, 2016). In contrast to other IS, DSS focuses on decision-making effectiveness and decision-making efficiency rather than efficiency alone (Evans & Riha, 1989).

In general, the decision-making process consists of three phases that are supported through DSS—the intelligence, design, and choice phase (Simon, 1960). In the intelligence phase, the decision-maker searches, classifies and decomposes problems (Turban et al., 2010, p. 48-49). In the design phase, decision alternatives are derived (Turban et al., 2010, p. 50). Finally, in the choice phase, the critical phase of decision-making, the decisions are chosen (Turban et al., 2010, p. 58).

An important concept of decision support is decisional guidance that has a long-lasting history in IS literature (Morana et al., 2017). Silver (1991) differentiates in the form of guidance, which can be either suggestive, quasi-suggestive, or informative. Suggestive guidance makes judgmental recommendations that can also be a set of alliterative decisions (Silver, 1991, p. 94). Quasi-suggestive guidance is guidance “that does not explicitly make a recommendation but from which one can directly infer a recommendation or direction” (Silver, 1991, p. 109). Lastly, informative guidance provides decision-makers only with decision-relevant information without suggesting or implying how to act.

Another form of guidance is the explainability of the DSS (Morana et al., 2017). Explainability is a concept with a long tradition in IS (Gregor & Benbasat, 1999). With the rise of expert systems, knowledge-based systems, and intelligent agents in the 1980s and 1990s, the IS community has built the basis for research on explainability (Meske et al., 2022). In particular, the research stream of Explainable AI (XAI), which addresses the opaqueness of AI-based systems, is gaining momentum. The term XAI was first coined by Van Lent et al. (2004) to describe the ability of their system to explain the behavior of agents. The current rise of XAI is driven by the need to increase the interpretability of complex models (Wanner et al., 2020). In contrast to interpretable linear models, more elaborate models can achieve higher performance (Briscoe & Feldman, 2011). However, their inner workings are hard to grasp for humans. XAI encompasses a wide spectrum of algorithms. These algorithms can be differentiated by their complexity, their scope, and their level of dependency (Adadi & Berrada, 2018). The interpretability of a model directly depends on its complexity. Wanner et al. (2020) define three types of complexity—white, grey, and black-box models. They define white-box models as models with perfect transparency, such as linear regressions. These models do not need additional explainability techniques but are intrinsically explainable. Black-box models, like neural networks, on the other hand, tend to achieve higher performance

but lack interpretability. Lastly, grey-box models are not inherently interpretable but are made interpretable with the help of additional explanation techniques. These techniques can be further differentiated in terms of their scope, i.e., being global or local explanations (Adadi & Berrada, 2018): Global XAI techniques address holistic explanations of the models as a whole. In contrast, local explanations function on an individual instance basis. Besides the scope, XAI techniques can also be differentiated with regard to being model-specific or model agnostic.

9.2.2 Automation

Research on automation is an essential part of IS research (Frank, 1998) and has been around for more than a century (Lacity & Willcocks, 2016) with the overarching goal to increase the efficiency of work by using automation as a means (Hitomi, 1994). In general, humans are performing worse than machines in conducting repetitive tasks and are influenced by cognitive bias (Heer, 2019). Thereby, automation can reduce human bias-induced errors. Automated decision-making applications are designed to minimize human involvement and relieve humans from exhaustive tasks (Harris & Davenport, 2005). Additionally, automation acts as a “talent multiplier” that scales human expertise and frees up human capacity to focus on more valuable work (Harris & Davenport, 2005).

Traditionally, automation has been seen as a binary state—either none or fully automatic (Endsley & Kaber, 1999). However, Parasuraman et al. (2000, p. 287) define automation as “the full or *partial* replacement of a function previously carried out by the human operator” which implies that automation may occur on different levels. The authors propose a taxonomy of automation and develop ten levels. While humans are responsible for decision-making at the first five levels, AI has control at the last five levels up to full autonomy at level ten.

Beyond developing the 10-level taxonomy, Parasuraman et al. (2000) provide a four-stage model of automated human information processing consisting of information acquisition, information analysis, decision and action selection, and action implementation. This model allows to precisely specify which stage is automated in the decision process.

Although automation has many advantages, some authors have expressed challenges, such as automation bias or cognitive skill reduction leading possibly to deskilling (Bainbridge, 1983). In the following, we discuss these disadvantages which essentially represent the problem with current approaches that we want to solve.

In the short-term, automation might lead to *Automation bias* which is the “tendency to use automated cues as a heuristic replacement for vigilant information seeking and processing” (Mosier & Skitka, 1999)—essentially representing an over-reliance on AI recommendations. For this reason, sometimes high levels of automation are not desirable if the automation is not perfectly reliable and recommends wrong decisions (Sarter & Schroeder, 2001). These wrong recommendations then can lead to a negative switch from a previously correct human decision (Goddard et al., 2012).

Furthermore, in the long-term, automation bias can result in *deskilling*, either because of the reduction of existing skills or due to the lack of skill development in general (Meske et al., 2022; Sutton et al., 2018). This attacks the collective intellectual capital that is the key asset of many organizations (Asatiani et al., 2019). Many factors might eventually result in deskilling. One factor is the reduced amount of stored information in memory, and more importantly, the reduced mental capability to store information, when using automation, which is commonly known as the “Google effect” (Sparrow et al., 2011). Users seem to reduce investing energy into storing things that can be easily retrieved (Sutton et al., 2018).

Research shows that human *engagement* in the task is particularly important to keep up the vigilantly (Mosier & Skitka, 1999). Engagement is a psychological state that is broadly defined as an “individual’s involvement and satisfaction as well as enthusiasm for work” (Harter et al., 2002, p. 269) that could reduce potential deskilling (Asatiani et al., 2019). Exemplary, the danger of deskilling can be highlighted with an intelligent asset solution for financial markets. Thereby, the engagement of the broker in the task will reduce which may lead finally to deskilling. Therefore, within the company implementing that solution, the broker deskills—while brokers from companies not implementing the project stay skilled. In the long-term, the environment may eventually change, for example, because of new regulations. Therefore, existing AI solutions need to be built and trained. One of the most important factors in the development process is domain knowledge which may now be reduced due to deskilling. If other companies did not implement AI, they can build and adapt faster and will, therefore, have competitive advantages

This long-term disadvantage of automated decision support leads to a discussion of efficiency in the short and long-term in human-AI systems. In the short-term AI might increase performance. However, in the long-term due to deskilling, AI systems will not be effectively further trained and evolved. This potentially results in severe negative long-term effects.

9.3 Conceptualization of Intelligent Decision Assistance

In this section, we use the previously depicted research streams of DSS and automation and synthesize them to conceptualize a solution against the disadvantages of automation. Subsequently, we discuss three particular techniques of this concept

We see two main dimensions that influence the undesired effects of automation, which we discuss in more detail below: First, the general level of human control and agency (Parasuraman et al., 2000) and, second, the form and degree of explainability (Morana et al., 2017).

Which level of human agency in automation should be implemented is a notorious discussion in automation literature (Heer, 2019). Asatiani et al. (2019) have discussed that retaining control of human workers may help to sustain their skill level. Similar, Endsley et al. (1997) argued that lower automation levels, in general, can keep them cognitively engaged.

Regarding the second dimension, the literature suggests “that a seamless, collaborative interaction between human agents and automated tools, as opposed to using automation as an isolated “black box”, could help to prevent the ill effects of deskilling” (Asatiani et al., 2019, p. 6). As discussed, the research stream of XAI addresses this “black box” issue in AI-based automated decision-making. Recent examples (Lai & Tan, 2019; Ribeiro et al., 2018) demonstrate the capability of XAI to support end-users in their decision-making. By varying the “degree” of explanations, i.e. the system’s transparency (Vössing, 2020), we believe different effects on the negative aspects of automation could be influenced. On the one hand, some might argue that more explainability is always better. However, the latest research suggests that a high level of automation paired with high explainability might just result in automation bias (Hemmer et al., 2021). Furthermore, the degree of explainability should be adapted to the profession and experience of the end-user, e.g. novice users might need more intuitive and simpler explanations while data scientists can get the full degree of potential explanations (Kühl et al., 2019). These examples show that also the degree of explainability needs to be chosen thoughtfully.

As introduced, there are many forms of guidance—suggestive, quasi-suggestive, and informative guidance (Silver, 1991). Suggestive guidance provides the decision-maker with explicit recommendations and tries to increase the guidance of this recommendation. However, as Parasuraman et al. (2000) states, also partially automated systems can lead to automation bias and skill degradation. In contrast, as mentioned, informative decisional guidance is a form of guidance where users do

not receive explicit recommendations (Silver, 1991). We follow this line of reasoning and propose a system could simply withhold its recommendation—although it is aware of that recommendation. Parkes (2012) validates that suggestive guidance—which is actually a form of automated decision-making—can lead to automation bias, while informative guidance does not have such effects. Research also shows that the effects of the types of guidance vary depending on the task complexity. Montazemi et al. (1996) found that suggestive guidance is better for less complex tasks and informative guidance is better with increasing task complexity. This argument strengthens our derivation. Following this line of thought gives rise to the idea to set the degree of automation to almost zero and withhold explicit AI recommendations while keeping support through explanations up. By doing so, we can minimize the drawbacks of automation while still assisting human decision-making. We are creating intelligent systems that are fully capable of solving issues on their own but use their capabilities to inspire and support instead of automating. Based on the derivation, we name this new class of DSS, Intelligent Decision Assistance (IDA) and define it as follows:

Definition: *Intelligent Decision Assistance (IDA) is an AI that a) supports humans, b) does not recommend explicit decisions or actions, and c) explains its reasoning*

Referring to the three phases of decision-making—intelligence, design, and choice—we mainly support with this approach the intelligence and to some extent the design phase. In terms of final effects on the human, we derive three hypotheses (engagement, performance, automation disadvantages).

First, IDA provides decision-makers with options to actively engage with the task by interactively requesting explanations, interpreting them and essentially communicating with the AI. As Asatiani et al. (2019) (Asatiani et al., 2019) have discussed providing explanations instead of using automation as an isolated “blackbox” could result in an engaged human-AI collaboration. Thus, we hypothesize:

H1: *IDA increases engagement with the task.*

Beyond that, we hypothesize that IDA should increase human performance. While especially, if the automation is far better than the human, IDA will most likely not exceed automated decision-making, it should still improve the performance by providing guidance and especially insights. Therefore, we formulate:

H2: *IDA performance outperforms the human alone.*

Lastly, because IDA does not incorporate higher levels of automation it should reduce automation disadvantages and especially prevent deskilling. Therefore, we formulate the following hypothesis:

H3: *IDA reduces automation induced disadvantages.*

In the next section, we are going to test these hypotheses based on empirical studies in the literature.

Figure 9.1 depicts IDA in the continuum of both discussed dimensions. We depict different types of systems for decision-making. At a high level of automation and almost no explanations, we position automation (Turban et al., 2010). Traditional DSS come also usually with a higher level of automation, through providing explicit recommendation, but additionally provide explanations for the decision-maker. We delimit ourselves from DSS that use AI to transform unstructured data into structured data and DSS that use AI to produce a pre-decision output, e.g. a forecast. As stated, Parasuraman et al. (2000) define four stages of automation—information acquisition, information analysis, decision-making, and actions. Following this classification, we focus on the decision-making level. This classification allows us also to differentiate IDA from Advanced Analytics (Watson, 2014). While advanced analytics may incorporate AI solutions they are always on the information acquisition or analysis level. In contrast, IDA allows the decision-maker to actively engage on the decision level and is positioned in the right top corner of Figure 9.1 with high explainability and full human autonomy.

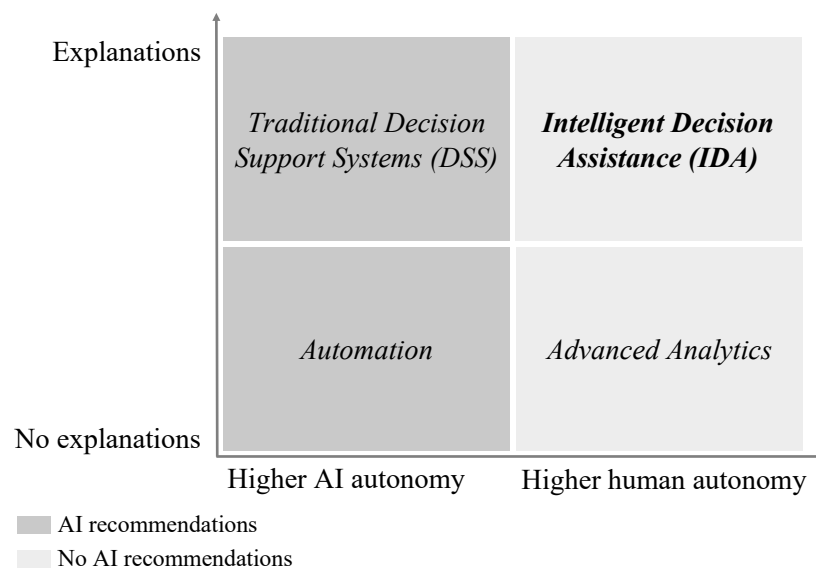


Figure 9.1.: Positioning of Intelligent Decision Assistance on the two dimensions of explainability and degree of automation

Now that we derived, defined, and delimit IDA, we discuss specific explanation techniques that support IDA and consequently pose valid implementation options. Specifically, we discuss feature importance, example-based explanations, and counterfactual explanations. We explain these features based on the example of a loan approval decision-making task.

Feature importance: Feature importance is a model-agnostic technique that gives the decision-maker information about the importance of specific data points. Two famous algorithms of feature importance are LIME (Ribeiro et al., 2016a) and SHAP (Lundberg & Lee, 2017). In a loan approval decision where the banker has information about past credits, expenses, demographics, etc., one could now train artificial intelligence to make this decision and recommend explicit decisions. In contrast, IDA would withhold the specific AI decision but provide the decision-maker, i.e. the banker, with information on which data was in particular important for the AI's decision. In an IDA this information could now be used for various use cases. Now in the time of big data, e.g. having many information on customers, one particular great use case would be to filter or sort the features in an intelligent way based on the feature importance.

Example-based explanations: Example-based explanations provide historical data that is similar to the current instance (Van der Waa et al., 2021). Example-based explanations, therefore essentially represent some form of information retrieval. Research in psychology states that humans prefer explanations that show examples (Cai et al., 2019). Furthermore, examples can be used within complex tasks (Glaser, 1986). Referring to our loan approval case, the decision-maker would receive information on past approvals that were similar. In an IDA, the decision-maker would get information about similar historical cases that are labeled. Based on these examples, the decision-maker should be able to infer differences or similarities.

Counterfactual explanation: Counterfactual explanations give information on what the smallest change would be to get a different AI decision (Wachter et al., 2017). Counterfactual explanations take a similar form to the statement (Schoeffer et al., 2021): “You were denied a loan because your annual income was 20,000. If your income had been 45,000, you would have been offered a loan.” In an IDA a counterfactual explanation would look like the following: “Your current annual income is £30,000. If your income would be £45,000, the AI's decision would change.” This type of non-intrusive explanation would lead to an increased thought process of the decision-maker.

Figure 9.2 on page 209 highlights the idea of IDA for a credit allowance example. On the left side, we display a traditional interface for automated decision-making.

On the right side, IDA is visualized. In the traditional interface, the decision-maker gets a specific recommendation. Additionally, the decision-maker gets the available information on the credit applicant, the importance of the features for the decision, and optional explanation options. In contrast, an IDA does not provide a specific recommendation, but rather various XAI techniques that allow the decision-maker to “brainstorm” with the AI.

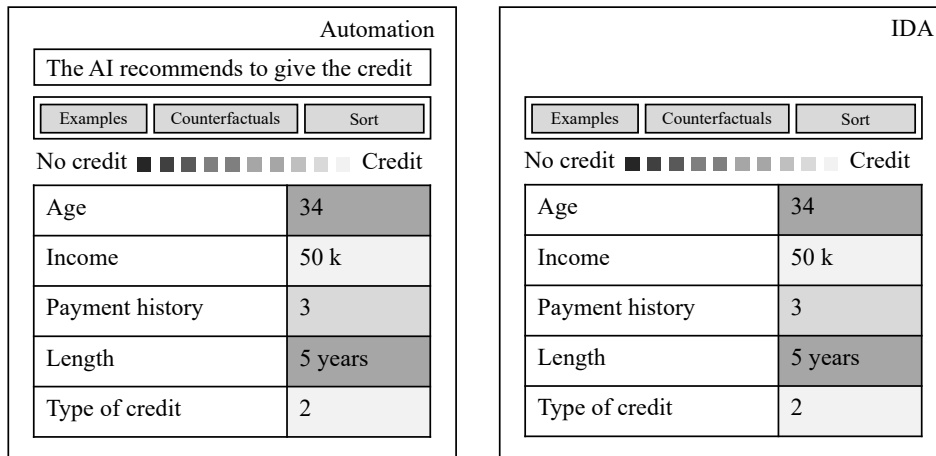


Figure 9.2.: Comparison of traditional automated decision-making and Intelligent Decision Assistance (IDA)

9.4 Validation Study

After deriving a conceptualization of IDA, we validate our concept by conducting a literature-based validation study based on the methodology outlined by vom Brocke et al. (2009). The goal of the study is to find empirical studies that tested variations of automation and explainability and to analyze whether the findings do support our hypotheses above. This means they should address the degree of automation and explainability. For this reason, our search string consists of two main parts. The first reflects XAI, including relevant synonyms, such as “explainable AI” or “interpretability” comprises of “artificial intelligence”. The second part comprised synonyms of behavioral experiments, e.g., “user study” or “user evaluation”. To find the synonyms, we initiated our SLR with an explorative search. The search string was iteratively extended resulting in the following final search string:

TITLE-ABS-KEY(“explainable artificial intelligence” OR XAI OR “explainable AI” OR (interpretability OR explanation) AND (“artificial intelligence” OR ai OR “machine learning”)) AND (“human performance” OR “human accuracy” OR “user study”

OR “empirical study” OR “online experiment” OR “human experiment” OR “behavioral experiment” OR “human evaluation” OR “user evaluation”)

Then, we selected an appropriate database. Our exploratory search indicated that relevant work is dispersed across multiple disciplines, publishers, conferences, and journals. For this reason, we chose the SCOPUS database, to ensure comprehensive coverage. Following that, we defined our inclusion criteria. We included every article that (a) conducted empirical research, (b) reported performance measures, (c) focused on an application context where AI supports humans on the decision level, and (d) provided an IDA setting. With our search string defined, we conducted the SLR from January to March 2021. We identified 256 articles through the keyword-based search. As a next step, we analyzed the abstract of each article and filtered based on our inclusion criteria, leading to 61 articles. Afterward, two independent researchers read all articles in detail and applied the inclusion criteria again. Based on these, we conducted a forward and backward search. This led to a total of five articles that were consequently analyzed in-depth to collect data about each experiment. The data collection process was conducted by two independent researchers who discussed and homogenized differences. The main focus of the validation study was to extract the treatments and outcomes of each experiment reported in the studies. For example, if two XAI techniques were used and compared as separate experimental treatments we added two entries into our database. In total, we identified five articles and 12 experiments (Carton et al., 2020; Chu et al., 2020; Lai et al., 2020; Lai & Tan, 2019; Schmidt & Biessmann, 2019). In the following, we describe the studies and their results with regard to IDA in detail.

Carton et al. (2020) conduct an experiment on online toxicity classification of social media posts. They use feature importance to highlight words that were relevant for the classification. As one condition they have the prediction presence. In their experiment, they find no significant effect of examples. However, they find signs of automation bias: “We find that the presence of a visible model prediction tends to bias subjects in favor of the prediction, whether it is correct or incorrect.” (Carton et al., 2020, p. 101)

Chu et al. (2020) conduct an experiment on age guessing supported through AI. They test three different conditions of explanations and the visibility of AI predictions. The authors found no significant effects of explanations but also signs of automation bias: “The predictions generally help whenever the human is inaccurate [...], but can hurt when the human is accurate and the model is inaccurate [...].” (Chu et al., 2020, p. 5)

Lai and Tan (2019) and Lai et al. (2020) refer in their studies also to the ten levels of automation introduced by (Parasuraman et al., 2000) and test various XAI techniques without ever displaying what the actual AI's decision is on a deception detection task. For example, they highlight all words that were relevant for the decision (unsigned) (Lai et al., 2020). Another condition was to colorize this highlight differently depending on the influence of the words (signed). Their results show that signed highlights result in a significant increase in XAI-assisted performance (70.7% for signed, and 60.4% for human performance) (Lai et al., 2020). In Lai and Tan (2019) they test additionally the influence of example-based explanations with also positive but not significant effects. However, also in Lai and Tan (2019) two highlight-based conditions showed significant positive effects in terms of short-term performance.

Lastly, Schmidt and Biessmann (2019) conduct two different tasks in their experiment—a book category classification based on their descriptions and a movie rating classification. They test two different XAI algorithms, both feature importance techniques to highlight important words. Both data sets and both XAI algorithms show an increase in IDA performance with one algorithms generating significant results on both data sets.

Table 9.1.: Validation study results

Source	Engagement	Performance	Automation
Carton et al. (2020)	No Measurement	No effect	Automation Bias
Chu et al. (2020)	No Measurement	No effect	Automation Bias
Lai and Tan (2019)	No Measurement	Improvement	No Measurement
Lai et al. (2020)	No Measurement	Improvement	No Measurement
Schmidt and Biessmann (2019)	No Measurement	Improvement	No Measurement

Table 9.1 summarizes our results of the validation study. Regarding our first hypothesis (**H1**), we can see that current research fails to provide insights into the effect of IDA on engagement. Regarding **H2**, three papers validated our hypotheses that IDA performance should exceed human performance. Lastly, regarding **H3**, two of the studies showed signs of Automation Bias in the presence of explicit AI recommendations, which is an indicator of potential long-term deskilling effects (Meske et al., 2022; Sutton et al., 2018).

9.5 Discussion

Overall, the validation study provides first support for the hypotheses on the impact of IDA and highlights the potential of IDA through five experiments with significant positive effects and none with significant negative effects. Furthermore, the study shows that current research lacks insights on the influence of IDA on engagement which should be addressed in future research.

IDA has of course also limitations. One of them might be the perceived usefulness. Telling the decision-maker that the AI would be theoretically capable of providing them with a recommendation but this recommendation is to withhold may be perceived as annoying for decision-makers, especially if they are under time pressure. Therefore, the advantages of IDA need to be highlighted. One attenuated option could be to show the explanations on default, but the recommendation just on request. Another limitation is the potential high computational costs. Some XAI techniques, e.g. SHAP values (Lundberg & Lee, 2017), are computationally inefficient. Therefore, the computational costs, especially in comparison to traditional analytics tools might be much higher. This trade-off has to be determined for individual cases.

We want to clarify that IDA should not be applied in every use case. We explicitly derive this idea for knowledge work and not for repetitive structured work. Especially for jobs where the disadvantages of automation are critical, IDA should be taken into account. Among others, in high stake decision-making such as medicine, law, or human resource. But also in knowledge-intensive areas where the competitive advantage is based on knowledge, such as finance. However, as pointed out by Endsley and Kaber (1999), for structured tasks that require low flexibility and have a high system performance, full automation can be the best option.

Additionally, we want to discuss an additional advantage that may have a temporary influence on the adoption of IDA. Paragraph 22 of the GDPR states: “The data subject shall have the right not to be subject to a decision based solely on automated processing [...]” (European Union, 2018) This means that in some cases automated decision-making is simply forbidden. Here the best possible augmentation through IDAs could be a valuable approach.

Furthermore, IDAs could have a positive influence on the fairness of AI-enhanced decision-making. AI algorithms can have biases that can lead to unfair decision-making. With IDAs, we allow people to have full control over the final decision and can thus reduce bias.

Finally, there are some open questions. Future work should empirically validate whether IDAs prevent deskilling and other automation disadvantages and in contrast increases engagement. Furthermore, one should assess the efficiency effects of IDA on human decision-making. For example, Fazlollahi et al. (1995) find that decisional guidance increases decision time. However, also direct recommendations may decrease efficiency if they lead to cognitive dissonance and consequently to an in-depth analysis of the decision-maker. The efficiency of IDAs needs to be compared to pure human and automated approaches.

9.6 Conclusion

The main goal of this study was to conceptualize a solution to automation-induced disadvantages, such as automation bias or deskilling. To do so, we initiated our research by conducting a literature review of automation and DSS literature. Based on these two research streams, we conceptualized a new class of DSS, namely *Intelligent Decision Assistance* (IDA). IDA augments human decision-making through Explainable AI (XAI) while withholding explicit AI recommendations. Thereby, IDA aims to provide insight into the data without generating automation disadvantages. Subsequently, we validated our conceptualization by searching for empirical literature which shows first evidence of our hypotheses.

Our contributions are threefold: First, we synthesize the body of knowledge in automation sciences and decision support literature. Second, we conceptualize a new class of systems—IDA—and third, we test three hypotheses regarding the potential of IDA.

Unleashing the potential of IDA requires a multidimensional design process. For this reason, we see the IS research community as the predestined research discipline to advance research in this field. We hope to motivate IS researchers and practitioners to actively participate in the exploration of IDA.

Part VI

Finale

This thesis is based on the idea that human and AI agents do not necessarily compete on all tasks but can complement each other. However, we observe that currently, human-AI collaboration leads to ineffective results (Bansal et al., 2021; Hemmer et al., 2021; Schemmer et al., 2022b). Therefore, in this study, we investigate impact factors and how to harness the complementarity potential between human and AI agents. Accordingly, our goal is to derive fundamental insights for understanding and designing *effective* human-AI collaboration. We addressed this research objective by answering four interrelated research questions.

This final chapter of this thesis is structured as follows: In Section 10.1, we summarize the results of our research and discuss the theoretical contributions. Next, in Section 10.2, we outline the managerial implications of our work. In Section 10.3, we discuss limitations and potential future research.¹

10.1 Summary and Theoretical Contributions

In this section, we highlight the results of the thesis and their theoretical contributions stemming from each study by revisiting the research questions (see Section 1.2) to structure our findings. The results presented in this thesis contribute to the domains of information systems, human-computer interaction, and computer science.

Research Question 1 (RQ1)

How effectively do human agents and AI agents collaborate?

To address RQ1, we conduct a structured literature review and a meta-analysis in Part II. Based on the structured literature review, we provide a comprehensive overview of the influence of explainable AI on human decision-making performance

¹Note that, with exceptions (artificial intelligence: AI, complementary team performance: CTP, machine learning: ML, research question: RQ), for improved readability, we do not use previously introduced abbreviations throughout this chapter.

(cf. Chapter 3). Figure 10.1 summarizes the main results of the empirical analysis of the state of the art. We collected a total of 93 experiments and further filtered them into three conditions based on available performance measures, i.e., human, AI, AI-assisted, and explainable AI-assisted performance. We observe that of all the experiments measuring AI and explainable AI-assisted performance, 46 out of 72 experiments showed an improvement in human decision performance by providing explanations of the AI agent’s prediction. Out of 63 experiments that measured AI- or explainable AI-assisted performance and human performance, 59 experiments show an improvement in performance due to AI- or explainable AI assistance. Finally, our analysis shows that in the current state of empirical research on human-AI collaboration, CTP (a joint performance superior to each agent individually) is not consistently achieved, as only 16 out of 53 experiments show CTP. In addition, we derive twelve testable hypotheses about potential influencing factors of CTP that need to be addressed in future research. The contributions of this study are twofold: First, we summarize the existing body of knowledge for empirical studies of human-AI collaboration and describe relevant researched factors of CTP. Second, we discuss neglected but relevant factors of CTP and formulate hypotheses for future work.

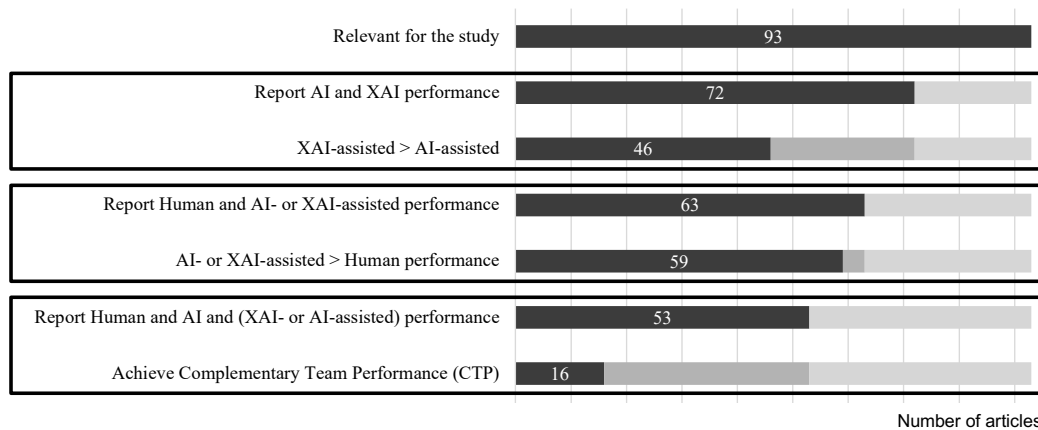


Figure 10.1.: Research contribution addressing RQ1: Empirical results of structured literature review (cf. Chapter 3).

Furthermore, in Chapter 4, we conduct a statistical meta-analysis (Borenstein et al., 2021) to complement the observations of the structured literature review. On average, explainable AI assistance improves human task performance compared to no assistance. We find that different types of data affect user performance differently. For example, human-AI collaboration is more effective for textual data than for tabular data. However, we find no additional effect of explanations on user performance in explainable AI-assisted decision-making compared to isolated

artificial intelligence (AI) assistance, raising the question of how to further develop current explainable AI methods that improve user decision-making performance. We contribute to the research by conducting, to the best of our knowledge, the first meta-analysis of empirical human-AI collaboration. As a result, we are able to derive statistically sound research gaps.

Overall, we contribute to the research by analyzing the state of empirical studies and deriving research directions. Our research shows the need for a better understanding of the mechanisms that promote effective human-AI collaboration, which is explored in RQ2.

Research Question 2 (RQ2)

What are the key factors that influence effective human-AI collaboration?

The observation that CTP is often not achieved raises the question of what factors might contribute to this observation. It illustrates that current knowledge of how to leverage the respective capabilities of humans and AI agents to create synergies in joint decision-making is not sufficiently developed and that there is a need for additional concepts that promote a deeper understanding of complementarity in human-AI collaboration.

To address this research gap and to address RQ2 (cf. Chapter 5), we build upon the theoretical work of Fügener et al. (2021) and propose a conceptualization of human-AI complementarity. The conceptualization consists of a formalization of complementarity potential and its constituting components, a differentiation of relevant sources of complementarity potential, as well as a classification of integration mechanisms to realize the complementarity potential between humans and AI. In detail, we argue that complementarity potential is composed of an inherent and collaborative component. The first component captures the idea that humans and AI agents possess different inherently present capabilities in the form of unique human and AI knowledge. The second component captures a new type of knowledge that only emerges through the interaction of humans and AI. We further differentiate between the theoretical upper limit of complementarity potential available and the realized amount during human-AI collaboration. Figure 10.2 on page 220 summarizes our conceptualization of human-AI complementarity.

We conduct a behavioral experiment to demonstrate the proposed conceptualization's usefulness.² Our results highlight the advantages of our conceptualization,

²We choose the real estate listing price prediction domain as a testbed and exemplarily instantiate a relevant source of complementarity potential together with an integration mechanism.

Isolated performance (better team member)

	Potential 1: Inherent complementarity potential	+	Potential 2: Collaborative complementarity potential	=	Total potential: Complementarity potential
Theoretically possible	Theoretical inherent complementarity potential $CP_{theoretical}^{inh}$		Theoretical collaborative complementarity potential $CP_{theoretical}^{coll}$		Theoretical complementarity potential $CP_{theoretical}$
	Source: Training phase Inference phase		Source: Integration phase		
Realized in setting	Realized inherent complementarity potential $CP_{realized}^{inh}$		Realized collaborative complementarity potential $CP_{realized}^{coll}$		Realized complementarity potential $CP_{realized}$
	Integration mechanism: Ex-ante Ex-post		Integration mechanism: Ex-post		

= Team performance

Figure 10.2.: Research contribution addressing RQ2: Conceptualization of Human-AI complementarity (cf. Chapter 5).

allowing us to develop a more nuanced understanding of the factors that influence effective human-AI collaboration. Our experiment shows that the distribution of instances in which the human or the AI agents performs better (and thus the inherent complementarity potential) changes when the human is provided with an additional house photograph. More specifically, providing the photograph increases the number of house price estimates where the human performs better than the AI agent. Moreover, we find that providing an additional photograph of the house affects the human’s estimate after receiving advice from the AI agent. Intuitively, the awareness of having more information than the AI agent could lead to algorithm aversion (Jussupow et al., 2021), which prevents humans from appropriately adjusting the house prices recommended by the AI agent. However, our research indicates the opposite, as our results show that providing the photo improves human adjustment of house prices suggested by the AI agent.

To summarize, our contributions are threefold. First, we conceptualize human-AI complementarity by introducing and formalizing the notion of complementarity potential, outlining sources of complementarity potential, and providing a classification of existing integration mechanisms for realizing complementarity potential. Second, we demonstrate the usefulness of the conceptualization to a human-AI collaboration setting by relying on information asymmetry as a source of complementarity potential. Third, through our proposed conceptualization, we find a new and surprising insight that unique human contextual information can lead human decision-makers to better adjust AI advice.

To conclude, in order to answer RQ2, we introduce the concept of human-AI complementarity, consisting of a formalization, sources of complementarity potential, and a classification of integration mechanisms. It becomes clear that to achieve CTP, it is essential to have the capabilities to harness complementarity potential, such as information asymmetry, a topic we address in RQ3.

Research Question 3 (RQ3)

How can complementarity potential in AI-assisted decision-making be harnessed?

In this thesis, we focus on harnessing the theoretical complementarity potential and, to address RQ3, analyze how this complementarity potential can be harnessed in AI-assisted decision-making, which refers to the setting where human decision-makers receive AI advice and are asked to either follow or adjust the advice. As discussed in detail in Chapter 6, human decision-makers should not simply rely on AI advice but should be empowered to differentiate when to rely on AI advice, and when to rely on their own, i.e., they should display appropriate reliance. Despite being a necessary condition for the effective use of AI, current research on appropriate reliance on AI advice still needs to be clarified with regard to definition, measurement, and impact factors (Bansal et al., 2021).

To address the current ambiguity in appropriate reliance research, we introduce a two-dimensional metric—the appropriateness of reliance (AoR)—to describe and measure reliance behavior. It is based on the relative frequency of correctly overriding incorrect AI suggestions (relative self-reliance—RSR) and following correct AI suggestions (relative AI-reliance—RAIR) and reflects a metric understanding of appropriate reliance. Figure 10.3 on page 222 visualizes the metric.

Second, we analyze how the provision of explanations influences appropriateness of reliance and the achievement of appropriate reliance states. Existing literature

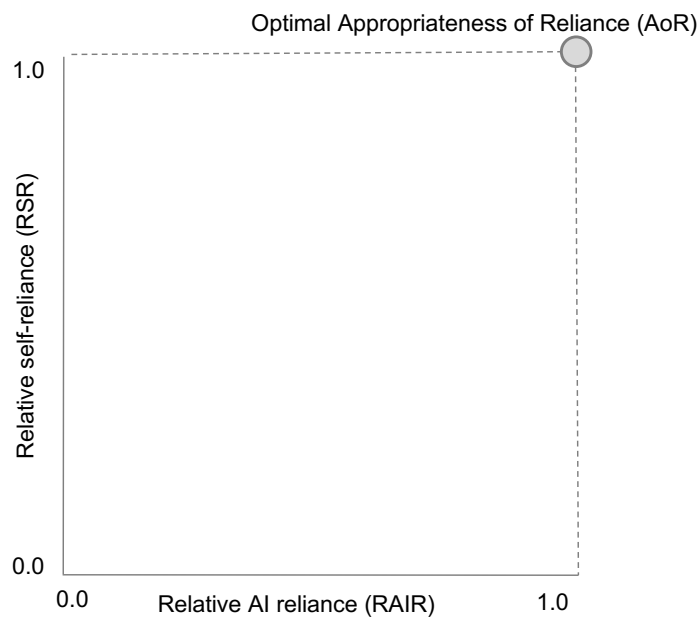


Figure 10.3.: Research contribution addressing RQ3: Measurement of appropriateness of reliance (cf. Chapter 6).

is ambiguous with regard to the effects of explanations (Alufaisan et al., 2021; Bansal et al., 2021; Wang & Yin, 2021). We consider additional constructs that may mediate the effect of explanations to better understand and reconcile conflicting results. More specifically, we hypothesize that explanations do not only influence the information available to the decision-maker but also have an impact on trust toward AI and self-confidence. Our work shows that both mediators significantly influence the appropriateness of reliance.

Our work provides researchers with a theoretical foundation of appropriate reliance on AI-assisted decision-making. More specifically, our research contributes by defining appropriate reliance, developing a measurement concept, and analyzing how explainable AI influences the appropriateness of reliance. Our definition should help researchers to more accurately describe whether they have achieved appropriate reliance in their experiments. Our measurement allows for precisely steering the development towards appropriate reliance. Lastly, our experimental insights can be seen as a starting point for in-depth experimental evaluations of factors impacting appropriateness of reliance.

Building upon the results of Chapter 6, in Chapter 7, we extend our previously derived research model by an additional mediator—the level of expertise possessed by decision-makers. Whether decision-makers are highly knowledgeable experts or

less knowledgeable laypeople appears to be an essential factor influencing reliance behavior in AI-assisted decision-making (Nourani et al., 2020a; Wang & Yin, 2021). Building on this notion, we propose that the learning process during human-AI collaboration, in which decision-makers progressively acquire expertise, may play a critical role in mediating the phenomenon of appropriateness of reliance. To explore this hypothesis, we extend the existing research model and conduct a behavioral experiment to test its validity empirically. Our results show that example-based explanations can improve human learning during human-AI collaboration, and learning improves the human ability to assess when to rely on themselves. Further, if sufficient learning is present, human learning helps to assess better when to rely on the AI agent. Figure 10.4 shows the research model, including the measured effects.

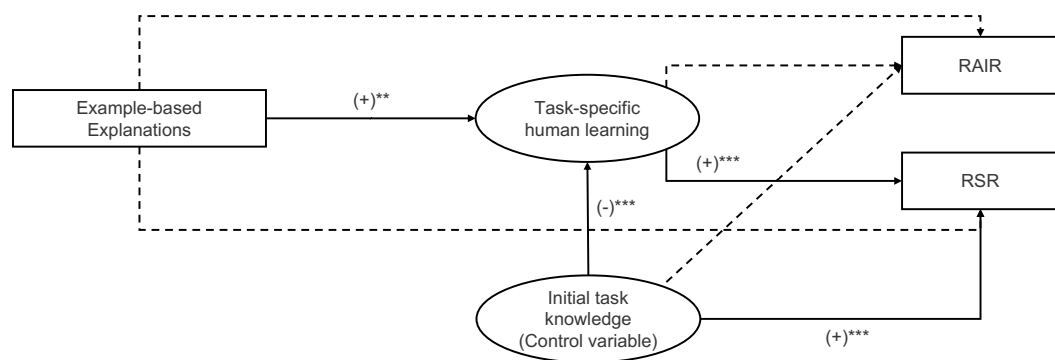


Figure 10.4.: Research contribution addressing RQ3: Research Model including task-specific human learning mediator (sub-group analysis is not included). Significance: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$ (cf. Chapter 7).

To the best of our knowledge, this research depicts the first study covering the effect of explanations on learning and the mediating effect on appropriateness of reliance. We thereby extend the research model of appropriateness of reliance in the previous chapter by a learning construct. We contribute to the body of knowledge on human-AI collaboration in general and on appropriateness of reliance and learning from AI agents in particular. Our research contributes to the literature on organizational learning (Levitt & March, 1988) as well as appropriate reliance (Bansal et al., 2021; Schemmer et al., 2023d). Even though learning on the job is a widely recognized approach for organizational learning, research has neglected the potential of in-process learning during human-AI collaboration. So far, learning from AI agents was always considered part of knowledge management. We, however, show that also in-process learning during human-AI collaboration is possible and thereby opens up new research potential for the domain of so-called AI teaching (Spitzer et al., 2022). With regard to appropriate reliance, learning was so far neglected as an impact factor. By extending the appropriateness of reliance research

model, we contribute to the understanding of harnessing complementarity potential in AI-assisted decision-making.

In summary, with regard to RQ3, we contribute to research by developing a measurement for quantifying appropriate reliance behavior, deriving a research model for appropriateness of reliance, including trust, change in confidence, and human learning, and validating this research model empirically. As these contributions are tailored to AI-assisted decision-making, in RQ4, we explore the potential to harness complementarity beyond AI-assisted decision-making.

Research Question 4 (RQ4)

How can complementarity potential beyond AI-assisted decision-making be harnessed?

Finally, to better understand how system designers can harness complementarity beyond AI-assisted decision-making, we analyze two different settings in Part V. Both studies show examples where complementarity potential exists, but it is harnessed in a different way than in AI-assisted decision-making. Complementarity potential is defined by an existing ground truth and the potential to solve the task alone by human and AI agents. AI-assisted decision-making is, per definition, bound to supervised ML tasks, i.e., classification and regression tasks. For this reason, we first analyze how to harness complementarity potential in unsupervised ML to broaden our perspective. In addition, we address one of the core challenges of human-AI collaboration presuming theoretical inherent complementarity potential over time (Fügener et al., 2021).

In the first study in Chapter 8, we shift our focus from supervised ML to unsupervised ML. We chose anomaly detection as an application case because of its relevance and prevalence in unsupervised ML applications, such as healthcare (Šabić et al., 2021), maintenance (Minarini & Decker, 2020), and cybersecurity (Karimipour et al., 2019). The anomalies detected by unsupervised ML approaches may include rare events but not the actual event of interest. Therefore, human experts are generally needed to investigate the relevance of the detected anomalies. While existing literature provides a comprehensive examination of anomaly detection, it falls notably short when it comes to providing systematic support for anomaly investigation, i.e., the classification of anomalies which are relevant to the business problem at hand (Chemweno et al., 2016; Pang et al., 2021; Steenwinckel et al., 2021). Therefore, we propose a novel method in which anomaly investigation is improved by explanations from an AI agent that conducts anomaly detection.

Figure 10.5 visualizes our proposed novel method. Subsequently, we conduct a behavioral experiment to validate the usefulness of our method. Our experiment involves a total of 64 participants, and we find that providing explanations can improve the accuracy of anomaly investigation.

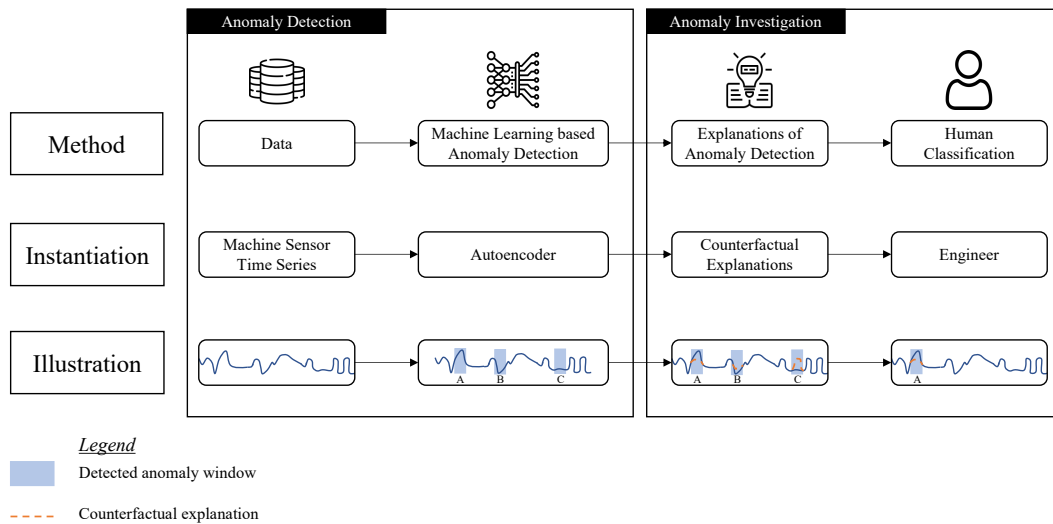


Figure 10.5.: Research contribution addressing RQ4: Our method of using explainable anomaly detection to support anomaly investigation (cf. Chapter 8).

In our research, we propose an anomaly detection and investigation method that extends the well-researched area of anomaly detection with complementary anomaly investigation. We thereby contribute to human-AI collaboration research in general as well as anomaly detection in specific. Our contribution can be summarized in two main aspects. First, we present a conceptual framework for methodological support of anomaly investigation using explanations derived from anomaly detection. Second, we demonstrate the effectiveness of these explanations through a behavioral experiment. Our study is the first to empirically examine the impact of anomaly detection explanations on anomaly investigation. The validation of this potential will inspire new use cases of anomaly detection that may substantially change the approach to anomaly detection in general.

In our second study, which addresses RQ3, we critically assess the potential drawbacks of AI-assisted decision-making (cf. Chapter 9). A major challenge is human deskilling, which is induced by the low cognitive effort required to validate AI advice as opposed to doing it alone (Buçinca et al., 2021; Fuegener et al., 2022). To address this challenge, we discuss different combinations of levels of automation and explainability and finally follow the idea of informative guidance that refrains from providing explicit recommendations (Silver, 1991). In line with this notion, we propose to withhold the decision from the AI agents and let the human interact with

the AI agent by providing explanations of the AI agent (Adadi & Berrada, 2018), such as examples, counterfactuals, or feature importance. We call this new form of collaboration intelligent decision assistance (IDA). After conceptualizing IDA and deriving hypotheses about its effects, we provide initial evidence for its validity through a systematic review of empirical studies in the literature. With our work, we contribute to research by conceptualizing a new form of human-AI collaboration that addresses current drawbacks of AI-assisted decision-making—human deskilling. We thereby continue the research started by Fügener et al. (2021) and provide potential means to counteract deskilling and thereby preserve complementarity potential.

To conclude, this thesis contributes to research by providing a comprehensive analysis of the current state of empirical studies, introducing new theoretical concepts, developing research models, and addressing critical challenges. In doing so, we are extending the research on human-AI collaboration in information systems, human-computer interaction, and computer science research.

10.2 Managerial Implications

After having outlined the theoretical contribution of our work, in the following, we discuss the managerial implications. The economic potential of AI continues to grow, now fueled by the emergence of new applications such as generative AI (Jo, 2023).³ According to a 2023 study, the implementation of generative AI in 63 use cases studied could generate annual economic benefits ranging from \$2.6 trillion to \$4.4 trillion. (McKinsey, 2023). This would increase the predicted monetary impact of AI by 15 to 40 percent. AI budgets are continuously increasing as well and are expected to hit \$154 billion in 2023—up 27% over 2022 (Needham, 2023). However, a recent study shows that while AI adoption is steadily increasing, the return on investment (ROI) that can be realized from AI projects is still lacking (Ashoori et al., 2023). Specifically, the average ROI for enterprise-wide initiatives is only 5.9%, well below the typical 10% cost of capital, and even best-in-class companies are only scratching 13%. This gap highlights a critical need for developing more effective strategies to enhance the ROI of AI deployments. We argue that properly designed human-AI collaboration addresses this lack of ROI. For example, in real estate market transactions, one of the most important questions is whether a house is appropriately valued, i.e., the house is offered within a price range that reflects its assets and is comparable to similar houses in terms of the characteristics of the property and

³Generative AI usually refers to the process of extracting intent information from human instructions and generating content according to the extracted intentions (Cao et al., 2023).

its surroundings. It is difficult for humans to correctly quantify a property's value because of the many factors that affect it (Kucklick et al., 2021). For this reason, more and more companies are using AI agents in real estate appraisals (Madhuri et al., 2019). As we have shown in Chapter 5, the availability of additional human information, e.g., by visiting the house, could lead to a complementarity potential between human and AI agents in appraising houses. We argue that with the right design, human-AI collaboration could lead to CTP, i.e., prediction accuracy that exceeds the individual performance of both AI and humans. This would lead to a better prediction of house prices and thus improve the ROI of the AI investment.⁴ Our work is concerned with deriving foundations and insights for such effective human-AI collaboration, i.e., how to achieve CTP. Thus, the results presented have direct practical relevance, as they can guide managers to pursue their AI initiatives, leading to improved ROI. In the following, we highlight the specific managerial implications of our thesis.

First of all, we present means to facilitate systematic discussions in organizations about AI initiatives and their strategic positioning. Discussions about human-AI collaboration in practical contexts are typically characterized by inconsistent terminology and vague statements. For example, it is often said humans need to monitor AI, but no reason is provided. In this work, we provide a formalization of human-AI complementarity and precise terminology. Furthermore, we provide a classification of sources of complementarity potential and different integration mechanisms. We believe that by providing clarity on terminology, these foundations can contribute to a more informed and systematic strategic exchange and, thus, to the development of AI in organizations.

The foundation of effective human-AI collaboration is complementarity potential. We provide managers with guidance on where to identify complementarity potential by providing an overview of possible sources. We highlight the potential of information asymmetry as one practical, relevant source of complementarity potential. Especially on the human side, access to further information not available to the AI agent often exists in practical applications. Not all data might be available in a digital format due to technical or economic reasons. Also, information is not always digitizable, especially if it exists only implicitly (Ibrahim et al., 2021). While some of these obstacles can be overcome, others will likely persist. We empirically validate that information asymmetry can increase the complementarity potential and, beyond that, has further beneficial implications in the collaboration of human and AI agents. In general, managers could invest in training that enhances the complementary skills

⁴This example does not account for the costs related to human-AI collaboration. We will discuss this in the next section.

of their employees, such as searching and interpreting unique human information, to improve the inherent complementarity potential in human-AI teams.

Complementarity potential needs to be harnessed in practical applications. Our research provides guidance on how to leverage existing complementarity potential in the most widely used practical form of human-AI collaboration—AI-assisted decision-making. In practice, managers typically focus on the adoption of AI and compliance with AI advice maximizing the reliance (Alnowami et al., 2022; Jussupow et al., 2021; Kerasidou et al., 2021; Shin, 2021; Siau & Wang, 2018). However, to realize the potential benefits of human-AI collaboration, employees need to rely appropriately on AI advice (Bansal et al., 2021; Schemmer et al., 2023d). We show that the appropriateness of reliance (AoR) can be positively influenced by proper design, e.g., by using adequate explanations of the AI’s decision-making.

Besides deriving guidance on the proper design for effective human-AI collaboration, our research shows empirically that with the right design of human-AI collaboration, practitioners can not only achieve CTP by using AI as a “teacher” but also increase the knowledge level of their workforce. We show the potential to learn from each other in a human-AI collaboration. Thus, these insights can be used to guide not only designers of AI systems but also knowledge managers within organizations to enhance the knowledge of practitioners.

Lastly, we introduce and inspire the use of novel concepts and methods. We develop a new method for harnessing complementarity in anomaly investigation, which has implications for any use case with large amounts of data and rare classes of interest. To cope with advantages induced through AI-assisted decision-making, we derive Intelligent Decision Assistance (IDA), which addresses deskilling in human-AI collaboration.

In summary, our research contributes to practice by providing terminology, structure, and insights into the design of effective human-AI collaboration and deriving novel, inspiring concepts. In doing so, we provide support for human-AI collaboration in general. Managers often see only the short-term benefits of cost reduction through automation while neglecting the potential value of human-AI collaboration. Even when they do support human involvement, it is usually for flimsy reasons based on external pressures, such as complying with regulations. Our research shows that when faced with external pressure, managers should not view it as a burden but rather as an opportunity to leverage complementarity potential, realize CTP, and improve ROI.

10.3 Limitations and Future Research

The research presented in this thesis certainly has limitations that motivate future research. Since Chapter 3 through Chapter 9 discuss limitations and future research for each study individually, in this section, we reflect the most important ones that should be taken into consideration regarding the generalizability of the results.

First and foremost, this thesis is limited by the focus on *effectiveness* of human-AI collaboration, i.e., achieving CTP. In this thesis, we focused on effectiveness because we observed that CTP was not consistently achieved in empirical work. For this reason, we wanted to demonstrate the possibility of achieving CTP by unraveling its impact factors. Now that we have investigated systematically impact factors, future research should analyze the overall economic impact of human-AI collaboration. While a marginal performance improvement may be theoretically interesting to demonstrate the existence of CTP, it may not be practically relevant if it does not improve the ROI. Including the cost of human-AI collaboration implementations could strengthen the practical implications of this thesis.

An additional limitation arises from the methodological focus on empirical research in this thesis. To answer RQ2-RQ4, we conducted behavioral experiments. The generalizability of our research may be limited to the domains and tasks we selected. The focus of our research was primarily concentrated on singular tasks, which inherently imposes constraints on the broad applicability of the results across diverse tasks. Therefore, while our findings provide valuable insights into human-AI collaboration in this specific task, extrapolating these conclusions to encompass all potential tasks within different domains or industries might not be appropriate. However, each experimental setting was consciously and carefully chosen for the individual studies. Additionally, we employ a variety of tasks over all the studies, ranging from image classification to text analysis to sensor data, which should increase generalizability. Nevertheless, future work must validate our concepts in other tasks and domains.

The backbone of this thesis is the assumption that complementarity potential is available. In this thesis, we provide guidance on sources of complementarity potential. Providing these sources in a structured way may inspire researchers to look for complementarity beyond arbitrary statements about complementarity capabilities, such as human creativity and AI computational power. Furthermore, we introduce a promising and practical source, information asymmetry (Ibrahim et al., 2019). We conduct a behavioral experiment to show the impact of unique human contextual information on complementarity potential. Future work needs to validate additional

sources empirically. First ideas for, e.g., training AI agents to be more complementary can be found in Hemmer et al. (2022a) and Hemmer et al. (2023).

Another limitation is the focus on one particular instantiation of AI assistance elements throughout the thesis—explainable AI. We focus on explainable AI for our analysis of the state of the art of empirical human-AI collaboration research as well as the main assistance element to harness the complementarity potential in AI-assisted as well as beyond AI-assisted decision-making. Since explainable AI is certainly the most used form to support human-AI collaboration (Lai et al., 2023), we believe this is a good starting point for research. However, future work should evaluate other elements of AI assistance, such as providing in-depth information on training data.

Lastly, with two exceptions, this thesis focuses on AI-assisted decision-making, encompassing classification and regression tasks. However, AI agents can address a magnitude of additional tasks (Carbonell et al., 1983), such as clustering, forecasting, etc. Especially the generation of text or images, commonly referred to as generative AI (Jo, 2023), is growing. The scope of this thesis has implications for all kinds of tasks conducted by AI agents. Future work needs to investigate the applicability of our foundations and insights.

Overall, we hope that this work will contribute to making the most of current improvements in AI. We believe that our research will prove useful in designing human-AI collaboration that allows human and AI agents to complement each other, thus moving from competition to complementarity.

A.1 Appendix Human-AI Complementarity Conceptualization

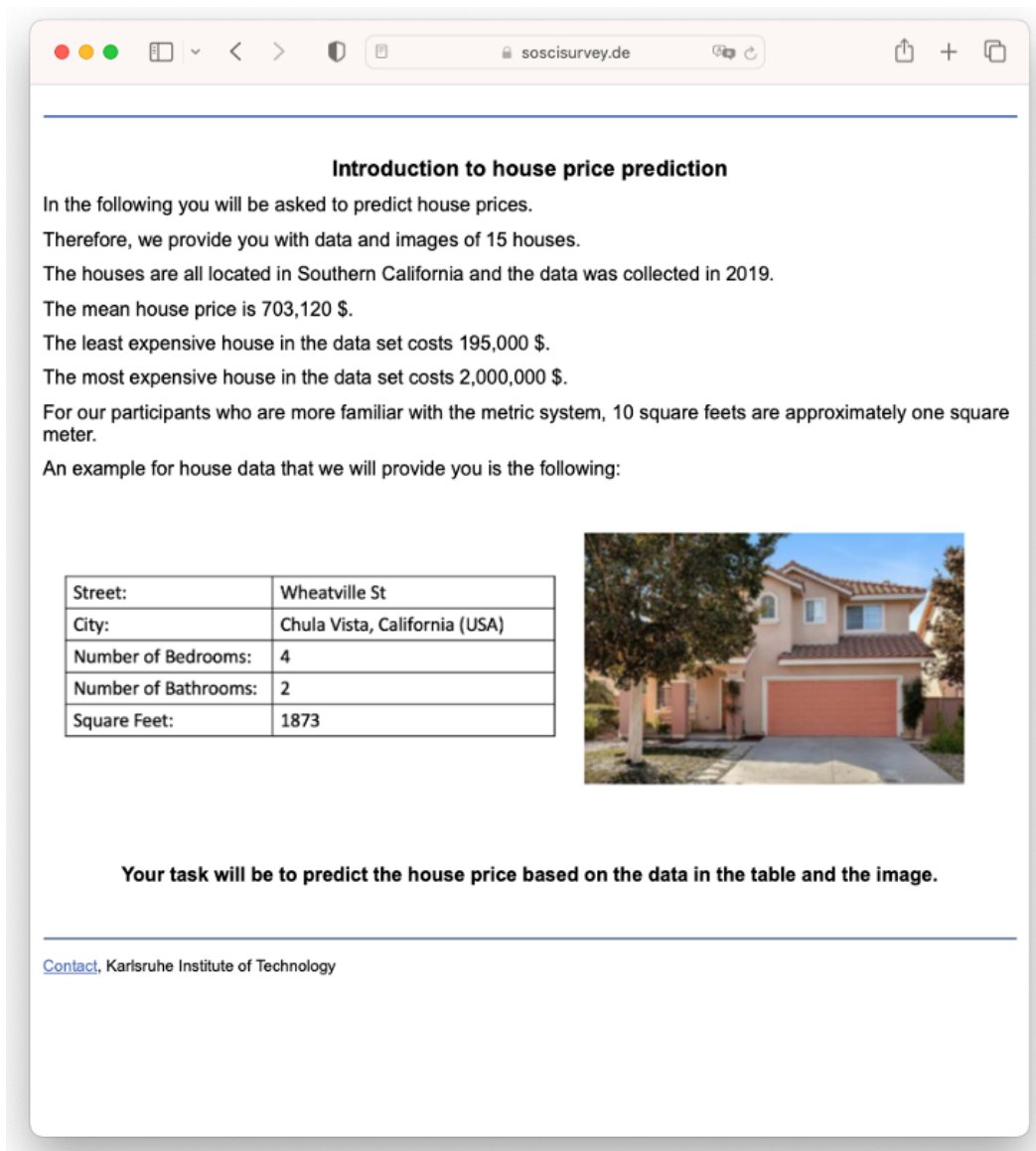
A.1.1 In-Person Pilot Study

For our in-person pilot study, we conducted two expert workshops. Each one lasted 90 minutes and aimed to elaborate on the usefulness of the house images for the participants and generate insights for the experimental design. The participants were 13 interviewees and two researchers. The first workshop was conducted with a smaller group to facilitate more extensive exchange. The second workshop focused on collecting broad ideas. We discussed three prediction cases in each session. After each house, we collected feedback in a structured way by asking the participants to make notes about how the AI and the information supported their decision-making.

In general, both workshops confirmed the usefulness of unique human contextual information (UHCI). We also received helpful comments for further refinement of our study. In the first workshop, a participant mentioned that the “picture was the first indication for me” and another one stated “I was already 70% sure what I was going to type when I saw the picture.” Both comments indicate the importance that participants see in the UHCI.

In the second workshop, participants highlighted the need for more information about the underlying data by stating “show the user the summary statistics.” They also mentioned the importance of a training section, for example by saying “I need examples of wrong (and right) predictions”.

A.1.2 Task and AI Tutorial




The screenshot shows a web browser window with the URL `soscisurvey.de`. The page content is as follows:

Introduction to house price prediction

In the following you will be asked to predict house prices.
Therefore, we provide you with data and images of 15 houses.
The houses are all located in Southern California and the data was collected in 2019.
The mean house price is 703,120 \$.
The least expensive house in the data set costs 195,000 \$.
The most expensive house in the data set costs 2,000,000 \$.
For our participants who are more familiar with the metric system, 10 square feet are approximately one square meter.
An example for house data that we will provide you is the following:

Street:	Wheatville St
City:	Chula Vista, California (USA)
Number of Bedrooms:	4
Number of Bathrooms:	2
Square Feet:	1873



Your task will be to predict the house price based on the data in the table and the image.

[Contact](#), Karlsruhe Institute of Technology


Figure A.1.: Online experiment graphical user interface for the task tutorial.

Introduction to Artificial Intelligence

After conducting the prediction on your own, you will be asked to adjust the AI's prediction of the same task. The AI produces its prediction based on the tabular data and has not access to the image.

To better understand how recommendations of artificial intelligence look like, we want to show you an example. In the following, we show you an AI prediction.

Street:	Wheatville St
City:	Chula Vista, California (USA)
Number of Bedrooms:	4
Number of Bathrooms:	2
Square Feet:	1873



Predicted Price:
589,464 \$

5% Quantile: 433,538 \$ 95% Quantile: 705,727 \$

0 \$ 195,000 \$ (Least expensive house price) 2,000,000 \$ (Most expensive house price)

From the plot you can see that the average predicted price of the AI for the particular house is 589,464 \$. Additionally, we provide you with information of the certainty of the AI in the form of quantiles. The 5% quantile tells you that with a probability of 95% the prediction of the AI is greater than the respective quantile value (433,538 \$). Similarly, the 95% quantile tells you that the AI's prediction is with a probability of 95% smaller than the quantile's value (705,727 \$).

In our example, with a probability of 95% the AI's prediction is greater than 433,538 \$ and also smaller than 705,727 \$.

The AI can be seen as a strong and reasonable base estimator. Therefore, your task will be to adjust the AI's base estimate in a best possible way.

Figure A.2.: Online experiment graphical user interface for the AI tutorial.

A.1.3 Participant Statistics

Table A.1.: Summary of participants' characteristics (unique human contextual information: UHCI).

Number per condition	Without UHCI = 53 With UHCI = 48
Age	Mean = 24.44 Standard deviation = 6.4
Gender	Female = 56% Male = 43% Non-binary = 1%
Education	High school = 41% Bachelor's = 44% Master's = 10% Other = 5%

Bibliography

- Abasolo, J. M., & Gomez, M. (2000). MELISA: An Ontology-Based Agent for Information Retrieval in Medicine. *Proceedings of the First International Workshop on the Semantic Web*, 73–82.
- Abdel-Karim, B. M., Pfeuffer, N., & Hinz, O. (2021). Machine Learning in Information Systems - a Bibliographic Review and Open Research Issues. *Electronic Markets*, 31(3), 643–670.
- Adadi, A., & Berrada, M. (2018). Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access*, 6, 52138–52160.
- Adam, M., Wessel, M., & Benlian, A. (2021). AI-based Chatbots in Customer Service and Their Effects on User Compliance. *Electronic Markets*, 31(2), 427–445.
- Adebayo, J., Muelly, M., Liccardi, I., & Kim, B. (2020). Debugging Tests for Model Explanations. *Proceedings of the 34th International Conference on Neural Information Processing Systems*, 700–712.
- Adelstein, J. (2007). Disconnecting Knowledge From the Knower: The Knowledge Worker as Icarus. *Equal Opportunities International*, 26(8), 853–871.
- Adhikari, A., Tax, D. M., Satta, R., & Faeth, M. (2019). LEAFAGE: Example-Based and Feature Importance-Based Explanations for Black-Box ML Models. *2019 IEEE International Conference on Fuzzy Systems*, 1–7.
- Ågerfalk, P. J., Conboy, K., & Myers, M. D. (2020). Information Systems in the Age of Pandemics: COVID-19 and Beyond. *European Journal of Information Systems*, 29(3), 203–207.
- Albert, E. T. (2019). AI in Talent Acquisition: A Review of AI-applications Used in Recruitment and Selection. *Strategic HR Review*, 18(5), 215–221.
- Alfeo, A. L., Cimino, M. G., Manco, G., Ritacco, E., & Vaglini, G. (2020). Using an Autoencoder in the Design of an Anomaly Detector for Smart Manufacturing. *Pattern Recognition Letters*, 136, 272–278.
- Alipour, K., Ray, A., Lin, X., Cogswell, M., Schulze, J. P., Yao, Y., & Burachas, G. T. (2021). Improving Users' Mental Model With Attention-Directed Counterfactual Edits. *Applied AI Letters*, 2(4), 47–59.
- Alnowami, M., Abolaban, F., Hijazi, H., & Nisbet, A. (2022). Regression Analysis of Rectal Cancer and Possible Application of Artificial Intelligence (AI) Utilization in Radiotherapy. *Applied Sciences*, 12(2), 725–736.
- Alt, R. (2018). Electronic Markets and Current General Research. *Electronic Markets*, 28, 123–128.

- Alt, R. (2021). Electronic Markets on the Next Convergence. *Electronic Markets*, 31, 1–9.
- Alufaisan, Y., Marusich, L. R., Bakdash, J. Z., Zhou, Y., & Kantarcioglu, M. (2021). Does Explainable Artificial Intelligence Improve Human Decision-Making? *Proceedings of the AAAI Conference on Artificial Intelligence*, 6618–6626.
- Alvarez Melis, D., & Jaakkola, T. (2018). Towards Robust Interpretability With Self-Explaining Neural Networks. *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, 7786–7795.
- Amruthnath, N., & Gupta, T. (2018). A Research Study on Unsupervised Machine Learning Algorithms for Early Fault Detection in Predictive Maintenance. *2018 5th International Conference on Industrial Engineering and Applications*, 355–361.
- Arjun, C., Viraj, P., Deshraj, Y., Prithvijit, C., & Devi, P. (2018). Do Explanations Make VQA Models More Predictable to a Human? *arXiv preprint arXiv:1810.12366*.
- Arnott, D., & Pervan, G. (2015). A Critical Analysis of Decision Support Systems Research. *Formulating Research Methods for Information Systems*, 2, 127–168.
- Arnott, D., & Pervan, G. (2016). A Critical Analysis of Decision Support Systems Research Revisited: The Rise of Design Science. *Formulating Research Methods for Information Systems*, 3, 43–103.
- Aronsson, L., & Bengtsson, A. (2021). *Security Log Analysis With Explainable Machine Learning* (Master's thesis). University of Gothenburg.
- Asatiani, A., Penttinen, E., Rinta-Kahila, T., & Salovaara, A. (2019). Implementation of Automation as Distributed Cognition in Knowledge Work Organizations: Six Recommendations for Managers. *40th International Conference on Information Systems*, 1–16.
- Ashoori, M., Goehring, B., Humphrey, T., Naghshineh, M., & Reese, C. R. (2023). *Generating ROI With AI* [IBM Institute for Business Value]. Retrieved July 16, 2023, from <https://www.ibm.com/thought-leadership/institute-business-value/en-us/report/ai-capabilities>
- Ates, E., Aksar, B., Leung, V. J., & Coskun, A. K. (2020). *Explainable Machine Learning Frameworks for Managing HPC Systems*. (tech. rep.). Sandia National Lab.(SNL-NM), Albuquerque, NM (United States).
- Ates, E., Aksar, B., Leung, V. J., & Coskun, A. K. (2021). Counterfactual Explanations for Multivariate Time Series. *2021 International Conference on Applied Artificial Intelligence*, 1–8.
- Audibert, J., Michiardi, P., Guyard, F., Marti, S., & Zuluaga, M. A. (2022). Do Deep Neural Networks Contribute to Multivariate Time Series Anomaly Detection? *Pattern Recognition*, 132, 108945–108966.
- Baier, L., Hofmann, M., Kühn, N., Mohr, M., & Satzger, G. (2020). Handling Concept Drifts in Regression Problems—the Error Intersection Approach. *Proceedings of 15th International Conference on Wirtschaftsinformatik, 2020, Potsdam, Germany*, 1–15.

- Baier, L., Köhl, N., & Satzger, G. (2019). How to Cope With Change?-Preserving Validity of Predictive Services Over Time. *Proceedings of the 52nd Hawaii International Conference on System Sciences*, 1085–1094.
- Bailey, N. R., & Scerbo, M. W. (2007). Automation-Induced Complacency for Monitoring Highly Reliable Systems: The Role of Task Complexity, System Experience, and Operator Trust. *Theoretical Issues in Ergonomics Science*, 8(4), 321–348.
- Bainbridge, L. (1983). Ironies of Automation. In *Analysis, Design and Evaluation of Man-Machine Systems* (pp. 129–135).
- Bakos, J. Y., & Treacy, M. E. (1986). Information Technology and Corporate Strategy: A Research Perspective. *MIS Quarterly*, 10(2), 107–119.
- Baldi, P. (2012). Autoencoders, Unsupervised Learning, and Deep Architectures. *Proceedings of ICML Workshop on Unsupervised and Transfer Learning*, 37–49.
- Bank, D., Koenigstein, N., & Giryas, R. (2020). Autoencoders. *arXiv Preprint arXiv:2003.05991*, 1–22.
- Bansal, G., Nushi, B., Kamar, E., Lasecki, W. S., Weld, D. S., & Horvitz, E. (2019a). Beyond Accuracy: The Role of Mental Models in Human-AI Team Performance. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, 7(1), 2–11.
- Bansal, G., Nushi, B., Kamar, E., Weld, D. S., Lasecki, W. S., & Horvitz, E. (2019b). Updates in Human-AI Teams: Understanding and Addressing the Performance/Compatibility Tradeoff. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01), 2429–2437.
- Bansal, G., Smith-Renner, A. M., Buçinca, Z., Wu, T., Holstein, K., Hullman, J., & Stumpf, S. (2022). Workshop on Trust and Reliance in AI-Human Teams. *CHI Conference on Human Factors in Computing Systems Extended Abstracts*, 1–6.
- Bansal, G., Wu, T., Zhou, J., Fok, R., Nushi, B., Kamar, E., Ribeiro, M. T., & Weld, D. (2021). Does the Whole Exceed Its Parts? The Effect of AI Explanations on Complementary Team Performance. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–16.
- Barrick, M. R., Stewart, G. L., Neubert, M. J., & Mount, M. K. (1998). Relating Member Ability and Personality to Work-Team Processes and Team Effectiveness. *Journal of Applied Psychology*, 83(3), 377.
- Basu, S., & Christensen, J. (2013). Teaching Classification Boundaries to Humans. *Proceedings of the AAAI Conference on Artificial Intelligence*, 27(1), 109–115.
- Bauer, K., Hinz, O., van der Aalst, W., & Weinhardt, C. (2021). Expl(AI)n It to Me—Explainable AI and Information Systems Research. *Business & Information Systems Engineering*, 63(2), 79–82.
- Bederson, B. B., & Shneiderman, B. (2003). *The Craft of Information Visualization: Readings and Reflections*. Morgan Kaufmann.
- Bellman, R. E. (1978). Artificial Intelligence: Can Computers Think?, 1–146.

- Bengio, Y., et al. (2009). Learning Deep Architectures for AI. *Foundations and Trends extreg-istered in Machine Learning*, 2(1), 1–127.
- Bentler, P. M. (1990). Comparative Fit Indexes in Structural Models. *Psychological Bulletin*, 107(2), 238.
- Berns, K., & Hirth, J. (2006). Control of Facial Expressions of the Humanoid Robot Head ROMAN. *2006 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 3119–3124.
- Bhattacharya, I., & Lindgreen, E. R. (2020). A Semi-Supervised Machine Learning Approach to Detect Anomalies in Big Accounting Data. *Proceedings of the 28th European Conference on Information Systems*, 100–115.
- Bhuyan, M. H., Bhattacharyya, D. K., & Kalita, J. K. (2013). Network Anomaly Detection: Methods, Systems and Tools. *IEEE Communications Surveys & Tutorials*, 16(1), 303–336.
- Billings, D., Davidson, A., Schaeffer, J., & Szafron, D. (2002). The Challenge of Poker. *Artificial Intelligence*, 134(1-2), 201–240.
- Blattberg, R., & Hoch, S. (2010). Database Models and Managerial Intuition: 50% MODEL+ 50% Manager. *Management Science*, 36(8), 887–899.
- Blazquez-Garcia, A., Conde, A., Mori, U., & Lozano, J. A. (2021). A Review on Outlier/Anomaly Detection in Time Series Data. *ACM Computing Surveys*, 54(3), 1–33.
- Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., et al. (2021). On the Opportunities and Risks of Foundation Models. *arXiv Preprint arXiv:2108.07258*.
- Bonaccio, S., & Dalal, R. S. (2006). Advice Taking and Decision-Making: An Integrative Literature Review, and Implications for the Organizational Sciences. *Organizational Behavior and Human Decision Processes*, 101(2), 127–151.
- Bondi, E., Koster, R., Sheahan, H., Chadwick, M., Bachrach, Y., Cemgil, T., Paquet, U., & Dvijotham, K. (2022). Role of Human-AI Interaction in Selective Prediction. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(5), 5286–5294.
- Borenstein, M., Hedges, L. V., Higgins, J. P., & Rothstein, H. R. (2010). A Basic Introduction to Fixed-Effect and Random-Effects Models for Meta-Analysis. *Research Synthesis Methods*, 1(2), 97–111.
- Borenstein, M., Hedges, L. V., Higgins, J. P., & Rothstein, H. R. (2021). *Introduction to Meta-Analysis*. John Wiley & Sons.
- Brady, M. (1985). Artificial Intelligence and Robotics. *Artificial Intelligence*, 26(1), 79–121.
- Braei, M., & Wagner, S. (2020). Anomaly Detection in Univariate Time-Series: A Survey on the State-of-the-Art. *arXiv Preprint arXiv:2004.00433*.
- Braga, A., & Logan, R. K. (2017). The Emperor of Strong AI Has No Clothes: Limits to Artificial Intelligence. *Information*, 8(4), 156.

- Breitenbacher, D., Homoliak, I., Aung, Y. L., Tippenhauer, N. O., & Elovici, Y. (2019). HADES-IoT: A Practical Host-Based Anomaly Detection System for IoT Devices. *Proceedings of the 2019 ACM Asia Conference on Computer and Communications Security*, 479–484.
- Brink, J. A. (2017). Big Data Management, Access, and Protection. *Journal of the American College of Radiology*, 14(5), 579–580.
- Briscoe, E., & Feldman, J. (2011). Conceptual Complexity and the Bias/Variance Tradeoff. *Cognition*, 118(1), 2–16.
- Brown, N., & Sandholm, T. (2019). Superhuman AI for Multiplayer Poker. *Science*, 365, 885–890.
- Brynjolfsson, E., & McAfee, A. (2017). Artificial Intelligence, for Real. *Harvard Business Review*, 1, 1–31.
- Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., et al. (2023). Sparks of Artificial General Intelligence: Early Experiments With GPT-4. *arXiv Preprint arXiv:2303.12712*.
- Buçinca, Z., Lin, P., Gajos, K. Z., & Glassman, E. L. (2020). Proxy Tasks and Subjective Measures Can Be Misleading in Evaluating Explainable AI Systems. *Proceedings of the 25th International Conference on Intelligent User Interfaces*, 454–464.
- Buçinca, Z., Malaya, M. B., & Gajos, K. Z. (2021). To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-assisted Decision-Making. *Proceedings of the ACM on Human-Computer Interaction*, 5(1), 1–21.
- Burkart, N., & Huber, M. F. (2021). A Survey on the Explainability of Supervised Machine Learning. *Journal of Artificial Intelligence Research*, 70, 245–317.
- Bussone, A., Stumpf, S., & O’Sullivan, D. (2015). The Role of Explanations on Trust and Reliance in Clinical Decision Support Systems. *2015 International Conference on Healthcare Informatics*, 160–169.
- Cai, C. J., Jongejan, J., & Holbrook, J. (2019). The Effects of Example-Based Explanations in a Machine Learning Interface. *Proceedings of the 24th International Conference on Intelligent User Interfaces*, 258–262.
- Cakmak, M., & Lopes, M. (2012). Algorithmic and Human Teaching of Sequential Decision Tasks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 26(1), 1536–1542.
- Camerer, C. F. (2018). Artificial Intelligence and Behavioral Economics. In *The Economics of Artificial Intelligence: An Agenda* (pp. 587–608). University of Chicago Press.
- Camposato, O. (2020). *Artificial Intelligence, Machine Learning, and Deep Learning*. Mercury Learning; Information.
- Campos, G., Zimek, A., Sander, J., Campello, R., Micenkova, B., Schubert, E., Assent, I., & Houle, M. (2016). On the Evaluation of Unsupervised Outlier Detection: Measures, Datasets, and an Empirical Study. *Data Mining and Knowledge Discovery*, 30, 891–927.

- Cao, Y., Li, S., Liu, Y., Yan, Z., Dai, Y., Yu, P. S., & Sun, L. (2023). A Comprehensive Survey of AI-Generated Content (AIGC): A History of Generative AI From Gan to Chatgpt. *arXiv Preprint arXiv:2303.04226*.
- Carbonell, J. G., Michalski, R. S., & Mitchell, T. M. (1983). An Overview of Machine Learning. *Machine Learning*, 1, 3–23.
- Carnap, R. (1955). Meaning and Synonymy in Natural Languages. *Philosophical Studies*, 6, 33–47.
- Carton, S., Mei, Q., & Resnick, P. (2018). Extractive Adversarial Networks: High-Recall Explanations for Identifying Personal Attacks in Social Media Posts. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 3497–3507.
- Carton, S., Mei, Q., & Resnick, P. (2020). Feature-Based Explanations Don't Help People Detect Misclassifications of Online Toxicity. *Proceedings of the International AAAI Conference on Web and Social Media*, 14, 95–106.
- Casolla, G., Cuomo, S., Di Cola, V. S., & Piccialli, F. (2019). Exploring Unsupervised Learning Techniques for the Internet of Things. *IEEE Transactions on Industrial Informatics*, 16(4), 2621–2628.
- Chalapathy, R., & Chawla, S. (2019). Deep Learning for Anomaly Detection: A Survey. *arXiv preprint arXiv:1901.03407*.
- Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly Detection: A Survey. *ACM Computing Surveys (CSUR)*, 41(3), 1–58.
- Chandrasekaran, A., Yadav, D., Chattopadhyay, P., Prabhu, V., & Parikh, D. (2017). It Takes Two to Tango: Towards Theory of AI's Mind. *arXiv Preprint arXiv:1704.00717*.
- Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., Wirth, R., et al. (2000). CRISP-DM 1.0: Step-by-Step Data Mining Guide. *SPSS Inc*, 9(13), 1–73.
- Charniak, E., & McDermott, D. (1985). Introduction to Artificial Intelligence.
- Chemweno, P., Pintelon, L., Muchiri, P., et al. (2016). I-Rcam: Intelligent Expert System for Root Cause Analysis in Maintenance Decision Making. *2016 IEEE International Conference on Prognostics and Health Management*, 1–7.
- Chen, Y. (2020). Exploring the Impact of Similarity Model to Identify the Most Similar Image From a Large Image Database. *Journal of Physics: Conference Series*, 1693(1).
- Cheng, Z., Wang, S., Zhang, P., Wang, S., Liu, X., & Zhu, E. (2021). Improved Autoencoder for Unsupervised Anomaly Detection. *International Journal of Intelligent Systems*, 36(12), 7103–7125.
- Chiang, C.-W., & Yin, M. (2021). You'd Better Stop! Understanding Human Reliance on Machine Learning Models Under Covariate Shift. *13th ACM Web Science Conference 2021*, 120–129.

- Choi, H., Kim, D., Kim, J., Kim, J., & Kang, P. (2022). Explainable Anomaly Detection Framework for Predictive Maintenance in Manufacturing Systems. *Applied Soft Computing*, 125.
- Choi, K., Yi, J., Park, C., & Yoon, S. (2021). Deep Learning for Anomaly Detection in Time-Series Data: Review, Analysis, and Guidelines. *IEEE Access*, 9, 120043–120065.
- Chong, L., Zhang, G., Goucher-Lambert, K., Kotovsky, K., & Cagan, J. (2022). Human Confidence in Artificial Intelligence and in Themselves: The Evolution and Impact of Confidence on Adoption of AI Advice. *Computers in Human Behavior*, 127, 107018.
- Chu, E., Roy, D., & Andreas, J. (2020). Are Visual Explanations Useful? A Case Study in Model-in-the-Loop Prediction. *arXiv Preprint arXiv:2007.12248*.
- Chuang, T.-T., & Yadav, S. B. (1997). An Agent-Based Architecture of an Adaptive Decision Support System. *Proceedings of the 41st Americas Conference on Information Systems*, 41–46.
- Clark, A., Fox, C., & Lappin, S. (2012). *The Handbook of Computational Linguistics and Natural Language Processing* (Vol. 118). John Wiley & Sons.
- Coombs, C., Hislop, D., Taneva, S. K., & Barnard, S. (2020). The Strategic Impacts of Intelligent Automation for Knowledge and Service Work: An Interdisciplinary Review. *The Journal of Strategic Information Systems*, 29(4), 101600–101630.
- Copeland, M. K. (2016). The Impact of Authentic, Ethical, Transformational Leadership on Leader Effectiveness. *Journal of Leadership, Accountability and Ethics*, 13(3), 79–100.
- Crosby, L. A., Evans, K. R., & Cowles, D. (1990). Relationship Quality in Services Selling: An Interpersonal Influence Perspective. *Journal of Marketing*, 54(3), 68–81.
- D'Amour, A., Heller, K., Moldovan, D., Adlam, B., Alipanahi, B., Beutel, A., Chen, C., Deaton, J., Eisenstein, J., Hoffman, M. D., et al. (2022). Underspecification Presents Challenges for Credibility in Modern Machine Learning. *The Journal of Machine Learning Research*, 23(1), 10237–10297.
- Das, A., & Rad, P. (2020). Opportunities and Challenges in Explainable Artificial Intelligence (XAI): A Survey. *arXiv Preprint arXiv:2006.11371*.
- D'Atri, A., De Marco, M., & Casalino, N. (2008). *Interdisciplinary Aspects of Information Systems Studies: The Italian Association for Information Systems*. Springer Science & Business Media.
- Davenport, T. H., & Kirby, J. (2016). *Only Humans Need Apply: Winners and Losers in the Age of Smart Machines*. Harper Business New York.
- Davey, B., & Cope, C. (2008). Requirements Elicitation—What's Missing? *Issues in Informing Science & Information Technology*, 5, 543–551.
- Davis, B., Glenski, M., Sealy, W., & Arendt, D. (2020). Measure Utility, Gain Trust: Practical Advice for XAI Researchers. *2020 IEEE Workshop on TRust and EXPertise in Visual Analytics*, 1–8.

- Day, M.-Y., Cheng, T.-K., & Li, J.-G. (2018). AI Robo-Advisor With Big Data Analytics for Financial Services. *International Conference on Advances in Social Networks Analysis and Mining*, 1027–1031.
- De Benedetti, M., Leonardi, F., Messina, F., Santoro, C., & Vasilakos, A. (2018). Anomaly Detection and Predictive Maintenance for Photovoltaic Systems. *Neurocomputing*, 310, 59–68.
- De-Arteaga, M., Romanov, A., Wallach, H., Chayes, J., Borgs, C., Chouldechova, A., Geyik, S., Kenthapadi, K., & Kalai, A. T. (2019). Bias in Bios: A Case Study of Semantic Representation Bias in a High-Stakes Setting. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 120–128.
- Debener, J., Heinke, V., & Kriebel, J. (2021). Insurance Fraud and Isolation Forests. *Proceedings of the 42nd International Conference on Information Systems*, 15–25.
- Dellermann, D., Ebel, P., Söllner, M., & Leimeister, J. M. (2019a). Hybrid Intelligence. *Business & Information Systems Engineering*, 61(5), 637–643.
- Dellermann, D., Lipusch, N., Ebel, P., & Leimeister, J. M. (2019b). Design Principles for a Hybrid Intelligence Decision Support System for Business Model Validation. *Electronic Markets*, 29, 423–441.
- Demajo, L. M., Vella, V., & Dingli, A. (2020). Explainable Ai for Interpretable Credit Scoring. *arXiv Preprint arXiv:2012.03749*.
- DerSimonian, R., & Laird, N. (1986). Meta-Analysis in Clinical Trials. *Controlled Clinical Trials*, 7(3), 177–188.
- Devine, D. J., & Phillips, J. L. (2001). Do Smarter Teams Do Better: A Meta-Analysis of Cognitive Ability and Team Performance. *Small Group Research*, 32(5), 507–532.
- Dietterich, T. G. (2009). Machine Learning in Ecosystem Informatics and Sustainability. *21st International Joint Conference on Artificial Intelligence*, 8–13.
- Dillon, B., Plehn, T., Sauer, C., & Sorrenson, P. (2021). Better Latent Spaces for Better Autoencoders. *SciPost Physics*, 11(3), 061.
- Dix, M. (2021). A Three-Step Machine Learning Pipeline for Detecting and Explaining Anomalies in the Time Series of Industrial Process Plants. *The International Conference on Deep Learning, Big Data and Blockchain*, 15–26.
- Donahue, K., Chouldechova, A., & Kenthapadi, K. (2022). Human-Algorithm Collaboration: Achieving Complementarity and Avoiding Unfairness. *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 1639–1656.
- Doney, P. M., & Cannon, J. P. (1997). An Examination of the Nature of Trust in Buyer–Seller Relationships. *Journal of Marketing*, 61(2), 35–51.
- Dong, S., Wang, P., & Abbas, K. (2021). A Survey on Deep Learning and Its Applications. *Computer Science Review*, 40, 100379.
- Doshi-Velez, F., & Kim, B. (2017). Towards a Rigorous Science of Interpretable Machine Learning. *arXiv Preprint arXiv:1702.08608*.

- Du, M., Li, F., Zheng, G., & Srikumar, V. (2017). Deeplog: Anomaly Detection and Diagnosis From System Logs Through Deep Learning. *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, 1285–1298.
- Dua, D., & Graff, C. (2019). *UCI Machine Learning Repository* [University of California, Irvine, School of Information and Computer Science]. Retrieved July 16, 2023, from <http://archive.ics.uci.edu/ml>
- Dunin-Barkowski, W. (2020). Toward and Beyond Human-Level AI. *Frontiers in Neurorobotics*, 14, 617446–617448.
- Dzindolet, M. T., Peterson, S. A., Pomranky, R. A., Pierce, L. G., & Beck, H. P. (2003). The Role of Trust in Automation Reliance. *International Journal of Human-Computer Studies*, 58(6), 697–718.
- Eaden, J., Abrams, K., & Mayberry, J. (2001). The Risk of Colorectal Cancer in Ulcerative Colitis: A Meta-Analysis. *Gut*, 48(4), 526–535.
- Edwards, C., Edwards, A., Spence, P. R., & Lin, X. (2018). I, Teacher: Using Artificial Intelligence (AI) and Social Robots in Communication and Instruction. *Communication Education*, 67(4), 473–480.
- Ehsan, U., Liao, Q. V., Muller, M., Riedl, M. O., & Weisz, J. D. (2021). Expanding Explainability: Towards Social Transparency in Ai Systems. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–19.
- Endsley, M. R., et al. (1997). The Role of Situation Awareness in Naturalistic Decision Making. *Naturalistic Decision Making*, 269, 284.
- Endsley, M. R., & Kaber, D. B. (1999). Level of Automation Effects on Performance, Situation Awareness and Workload in a Dynamic Control Task. *Ergonomics*, 42(3), 462–492.
- Engbom, N. (2020). *Firm and Worker Dynamics in an Aging Labor Market* (tech. rep.). Federal Reserve Bank of Minneapolis Minneapolis.
- Engel, C., Ebel, P., & Leimeister, J. M. (2022). Cognitive Automation. *Electronic Markets*, 32(1), 339–350.
- European Union. (2018). *Art. 22 GDPR - Automated Individual Decision-Making, Including Profiling*. Retrieved July 16, 2023, from <https://gdpr-info.eu/art-22-gdpr/>
- Evans, G. E., & Riha, J. R. (1989). Assessing DSS Effectiveness Using Evaluation Research Methods. *Information & Management*, 16(4), 197–206.
- Fahse, T. B., Blohm, I., & van Giffen, B. (2022). Effectiveness of Example-Based Explanations to Improve Human Decision Quality in Machine Learning Forecasting Systems. *Proceedings of the 43rd International Conference on Information Systems*, 1–10.
- Fazlollahi, B., Parikh, M. A., & Verma, S. (1995). Evaluation of Decisional Guidance in Decision Support Systems: An Empirical Study. *Proceedings of the 3rd International Conference of the Decision Science Institute*, 2, 2012–2030.

- Ferreira, N., Poco, J., Vo, H. T., Freire, J., & Silva, C. T. (2013). Visual Exploration of Big Spatio-Temporal Urban Data: A Study of New York City Taxi Trips. *IEEE Transactions on Visualization and Computer Graphics*, 19(12), 2149–2158.
- Filali Boubrahimi, S., & Hamdi, S. M. (2022). On the Mining of Time Series Data Counterfactual Explanations Using Barycenters. *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, 3943–3947.
- Frank, U. (1998). Increasing the Level of Automation in Organisations: Some Remarks on Formalisation, Contingency and the Social Construction of Reality. *The Systemist*, 20, 98–113.
- Frey, C. B., & Osborne, M. A. (2017). The Future of Employment: How Susceptible Are Jobs to Computerisation? *Technological Forecasting and Social Change*, 114, 254–280.
- Frye, M., Mohren, J., & Schmitt, R. H. (2021). Benchmarking of Data Preprocessing Methods for Machine Learning-Applications in Production [2 citations (Semantic Scholar/-DOI) [2022-11-04]]. *Procedia CIRP*, 104, 50–55.
- Fuegener, A., Grahl, J., Gupta, A., & Ketter, W. (2022). Cognitive Challenges in Human-Artificial Intelligence Collaboration: Investigating the Path Toward Productive Delegation. *Information Systems Research*, 33(2), 678–696.
- Fügener, A., Grahl, J., Gupta, A., & Ketter, W. (2021). Will Humans-in-the-Loop Become Borgs? Merits and Pitfalls of Working With AI. *Management Information Systems Quarterly*, 45, 1–30.
- Fukuda, T., Michelini, R., Potkonjak, V., Tzafestas, S., Valavanis, K., & Vukobratovic, M. (2001). How Far Away Is "Artificial Man". *IEEE Robotics & Automation Magazine*, 8(1), 66–73.
- Gama, J., Žliobaitė, I., Bifet, A., Pechenizkiy, M., & Bouchachia, A. (2014). A Survey on Concept Drift Adaptation. *ACM Computing Surveys*, 46(4), 1–37.
- Gamboa, J. C. B. (2017). Deep Learning for Time-Series Analysis. *arXiv Preprint arXiv:1701.01887*.
- Ganesan, S. (1994). Determinants of Long-Term Orientation in Buyer-Seller Relationships. *Journal of Marketing*, 58(2), 1–19.
- Gao, S., & Xu, D. (2009). Conceptual Modeling and Development of an Intelligent Agent-Assisted Decision Support System for Anti-Money Laundering. *Expert Systems With Applications*, 36(2), 1493–1504.
- Garg, A., Zhang, W., Samaran, J., Savitha, R., & Foo, C.-S. (2022). An Evaluation of Anomaly Detection and Diagnosis in Multivariate Time Series. *IEEE Transactions on Neural Networks and Learning Systems*, 33(6–16), 2508–2517.
- Gefen, D., Karahanna, E., & Straub, D. W. (2003). Trust and TAM in Online Shopping: An Integrated Model. *MIS Quarterly*, 27(1), 51–90.
- Geirhos, R., Narayanappa, K., Mitzkus, B., Thieringer, T., Bethge, M., Wichmann, F. A., & Brendel, W. (2021). Partial Success in Closing the Gap Between Human and Machine Vision. *Advances in Neural Information Processing Systems*, 34, 23885–23899.

- Geller, T. (2014). How Do You Feel? Your Computer Knows. *Communications of the ACM*, 57(8), 24–26.
- Gerber, A., Derckx, P., Döppner, D. A., & Schoder, D. (2020). Conceptualization of the Human-Machine Symbiosis—a Literature Review. *Proceedings of the 53rd Hawaii International Conference on System Sciences*.
- Ghalehtaki, R. A., Ebrahimzadeh, A., Wuhib, F., & Glitho, R. H. (2022). An Unsupervised Machine Learning-Based Method for Detection and Explanation of Anomalies in Cloud Environments. *2022 25th Conference on Innovation in Clouds, Internet and Networks*, 24–31.
- Ghavamipoor, H., & Hashemi Golpayegani, S. A. (2020). A Reinforcement Learning Based Model for Adaptive Service Quality Management in E-Commerce Websites. *Business & Information Systems Engineering*, 62, 159–177.
- Ghorbani, A., Abid, A., & Zou, J. (2019). Interpretation of Neural Networks Is Fragile. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(1), 3681–3688.
- Gilpin, L. H., Testart, C., Fruchter, N., & Adebayo, J. (2019). Explaining Explanations to Society. *arXiv Preprint arXiv:1901.06560*.
- Glaser, R. (1986). *Intelligence as Acquired Proficiency*. Psychology Press.
- Goddard, K., Roudsari, A., & Wyatt, J. C. (2012). Automation Bias: A Systematic Review of Frequency, Effect Mediators, and Mitigators. *Journal of the American Medical Informatics Association*, 19(1), 121–127.
- Goddard, K., Roudsari, A., & Wyatt, J. C. (2014). Automation Bias: Empirical Results Assessing Influencing Factors. *International Journal of Medical Informatics*, 83(5), 368–375.
- Goldenberg, S. L., Nir, G., & Salcudean, S. E. (2019). A New Era: Artificial Intelligence and Machine Learning in Prostate Cancer. *Nature Reviews Urology*, 16(7), 391–403.
- Gonzalez, A. V., Bansal, G., Fan, A., Jia, R., Mehdad, Y., & Iyer, S. (2020). Human Evaluation of Spoken vs. Visual Explanations for Open-Domain Qa. *arXiv Preprint arXiv:2012.15075*.
- Goodhue, D. L., & Thompson, R. L. (1995). Task-Technology Fit and Individual Performance. *MIS Quarterly*, 19(2), 213–236.
- Görnitz, N., Kloft, M., Rieck, K., & Brefeld, U. (2013). Toward Supervised Anomaly Detection. *Journal of Artificial Intelligence Research*, 46, 235–262.
- Gorry, G. A., & Scott Morton, M. S. (1971). *A Framework for Management Information Systems*. [Cambridge, MIT].
- Goyal, Y., Wu, Z., Ernst, J., Batra, D., Parikh, D., & Lee, S. (2019). Counterfactual Visual Explanations. *International Conference on Machine Learning*, 97, 2376–2384.
- Green, B., & Chen, Y. (2019). The Principles and Limits of Algorithm-in-the-Loop Decision Making. *Proceedings of the ACM on Human-Computer Interaction*, 3(1), 1–24.

- Gregor, S., & Benbasat, I. (1999). Explanations From Intelligent Systems: Theoretical Foundations and Implications for Practice. *MIS Quarterly*, 23(4), 497–530.
- Grootswagers, T. (2020). A Primer on Running Human Behavioural Experiments Online. *Behavior Research Methods*, 52, 2283–2286.
- Grosu, R. (2022). Can Artificial Intelligence Improve Our Health? In *Strategies for Sustainability of the Earth System* (pp. 273–281).
- Guidotti, R. (2022). Counterfactual Explanations and How to Find Them: Literature Review and Benchmarking. *Data Mining and Knowledge Discovery*, 36, 1–55.
- Guizzo, E. (2014). How Aldebaran Robotics Built Its Friendly Humanoid Robot, Pepper. *IEEE Spectrum*.
- Gulshan, V., Rajan, R. P., Widner, K., Wu, D., Wubbels, P., Rhodes, T., Whitehouse, K., Coram, M., Corrado, G., Ramasamy, K., et al. (2019). Performance of a Deep-Learning Algorithm vs Manual Grading for Detecting Diabetic Retinopathy in India. *JAMA Ophthalmology*, 137(9), 987–993.
- Ha, D. T., Hoang, N. X., Hoang, N. V., Du, N. H., Huong, T. T., & Tran, K. P. (2022). Explainable Anomaly Detection for Industrial Control System Cybersecurity. *IFAC-PapersOnLine*, 55(10), 1183–1188.
- Halford, G., Baker, R., McCredden, J., & Bain, J. (2005). How Many Variables Can Humans Process? *Psychological Science*, 16, 70–6.
- Haq, A. U., Li, J. P., Saboor, A., Khan, J., Wali, S., Ahmad, S., Ali, A., Khan, G. A., & Zhou, W. (2021). Detection of Breast Cancer Through Clinical Data Using Supervised and Unsupervised Feature Selection Techniques. *IEEE Access*, 9, 22090–22105.
- Harris, J. G., & Davenport, T. H. (2005). Automated Decision Making Comes of Age. *MIT Sloan Management Review*, 46(4), 2–10.
- Harter, J. K., Schmidt, F. L., & Hayes, T. L. (2002). Business-Unit-Level Relationship Between Employee Satisfaction, Employee Engagement, and Business Outcomes: A Meta-Analysis. *Journal of Applied Psychology*, 87(2), 268.
- Harvey, N., & Fischer, I. (1997). Taking Advice: Accepting Help, Improving Judgment, and Sharing Responsibility. *Organizational Behavior and Human Decision Processes*, 70(2), 117–133.
- Hase, P., & Bansal, M. (2020). Evaluating Explainable AI: Which Algorithmic Explanations Help Users Predict Model Behavior? *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5540–5552.
- Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (Vol. 2). Springer.
- Hatzilygeroudis, I., & Prentzas, J. (2004). Using a Hybrid Rule-Based Approach in Developing an Intelligent Tutoring System With Knowledge Acquisition and Update Capabilities. *Expert Systems With Applications*, 26(4), 477–492.
- Haugeland, J. (1989). *Artificial Intelligence: The Very Idea*. MIT press.

- Hawkins, D. M. (1980). *Identification of Outliers* (Vol. 11). Springer.
- He, G., Kuiper, L., & Gadiraju, U. (2023). Knowing About Knowing: An Illusion of Human Competence Can Hinder Appropriate Reliance on AI Systems. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 1–18.
- He, K., Zhang, X., Ren, S., & Sun, J. (2015). Delving Deep Into Rectifiers: Surpassing Human-Level Performance on Imagenet Classification. *Proceedings of the IEEE International Conference on Computer Vision*, 1026–1034.
- He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep Residual Learning for Image Recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770–778.
- He, R., & McAuley, J. (2016). Ups and Downs: Modeling the Visual Evolution of Fashion Trends With One-Class Collaborative Filtering. *Proceedings of the 25th International Conference on World Wide Web*, 507–517.
- He, S., Rui, H., & Whinston, A. B. (2018). Social Media Strategies in Product-Harm Crises. *Information Systems Research*, 29, 362–380.
- Hedges, L. V., & Olkin, I. (1985). *Statistical Methods for Meta-Analysis*. Academic Press.
- Heer, J. (2019). Agency Plus Automation: Designing Artificial Intelligence Into Interactive Systems. *Proceedings of the National Academy of Sciences*, 116(6), 1844–1850.
- Hegazy, I. M., Faheem, H. M., Al-Arif, T., & Ahmed, T. (2005). Performance Evaluation of Agent-Based IDS. *Proceedings of the 2nd International Conference on Intelligent Computing and Information Systems (ICICIS 2005)*, 314–319.
- Hein, A., Weking, J., Schreieck, M., Wiesche, M., Böhm, M., & Krcmar, H. (2019). Value Co-Creation Practices in Business-to-Business Platform Ecosystems. *Electronic Markets*, 29, 503–518.
- Hemmer, P., Kühl, N., & Schöffner, J. (2020). DEAL: Deep Evidential Active Learning for Image Classification. *19th IEEE International Conference on Machine Learning and Applications*, 865–870.
- Hemmer, P., Schellhammer, S., Vössing, M., Jakubik, J., & Satzger, G. (2022a). Forming Effective Human-AI Teams: Building Machine Learning Models That Complement the Capabilities of Multiple Experts. *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*, 2478–2484.
- Hemmer, P., Schemmer, M., Kühl, N., Vössing, M., & Satzger, G. (2022b). On the Effect of Information Asymmetry in Human-AI Teams. *ACM CHI 2022 Workshop on Human-Centered Explainable AI*.
- Hemmer, P., Schemmer, M., Vössing, M., & Kühl, N. (2021). Human-AI Complementarity in Hybrid Intelligence Systems: A Structured Literature Review. *Proceedings of the 17th Pacific Asia Conference on Information Systems*, 7–39.

- Hemmer, P., Westphal, M., Schemmer, M., Vetter, S., Vössing, M., & Satzger, G. (2023). Human-AI Collaboration: The Effect of AI Delegation on Human Task Performance and Task Satisfaction. *Proceedings of the 28th International Conference on Intelligent User Interfaces*, 453–463.
- Hidemichi, F., & Shunsuke, M. (2017). *Trends and Priority Shifts in Artificial Intelligence Technology Invention: A Global Patent Analysis* (tech. rep.). Research Institute of Economy, Trade and Industry (RIETI).
- Higgins, J. P., Thomas, J., Chandler, J., Cumpston, M., Li, T., Page, M. J., & Welch, V. A. (2019). *Cochrane Handbook for Systematic Reviews of Interventions*. John Wiley & Sons.
- Hillman, N. L. (2019). The Use of Artificial Intelligence in Gauging the Risk of Recidivism. *The Judges' Journal*, 58, 36–39.
- Hirt, R., Kühl, N., & Satzger, G. (2019). Cognitive Computing for Customer Profiling: Meta Classification for Gender Prediction. *Electronic Markets*, 29(1), 93–106.
- Hitomi, K. (1994). Automation—its Concept and a Short History. *Technovation*, 14(2), 121–128.
- Hogan, J., & Holland, B. (2003). Using Theory to Evaluate Personality and Job-Performance Relations: A Socioanalytic Perspective. *Journal of Applied Psychology*, 88(1), 100.
- Hu, L.-t., & Bentler, P. M. (1999). Cutoff Criteria for Fit Indexes in Covariance Structure Analysis: Conventional Criteria Versus New Alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, 6(1), 1–55.
- Huang, J., Kurniawan, E., & Sun, S. (2022). Cellular KPI Anomaly Detection With GAN and Time Series Decomposition. *ICC 2022 - IEEE International Conference on Communications*, 4074–4079.
- Hunke, F., Heinz, D., & Satzger, G. (2022). Creating Customer Value From Data: Foundations and Archetypes of Analytics-Based Services. *Electronic Markets*, 1–19.
- Hunter, J. E., & Hunter, R. F. (1984). Validity and Utility of Alternative Predictors of Job Performance. *Psychological Bulletin*, 96(1), 72.
- Hurtz, G. M., & Donovan, J. J. (2000). Personality and Job Performance: The Big Five Revisited. *Journal of Applied Psychology*, 85(6), 869.
- Hussain, M. T., & Perera, C. (2022). Explainable Sensor Data-Driven Anomaly Detection in Internet of Things Systems. *2022 IEEE/ACM Seventh International Conference on Internet-of-Things Design and Implementation (IoTDI)*, 80–81.
- Ibrahim, M., Louie, M., Modarres, C., & Paisley, J. (2019). Global Explanations of Neural Networks: Mapping the Landscape of Predictions. *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 279–287.
- Ibrahim, R., Kim, S.-H., & Tong, J. (2021). Eliciting Human Judgment for Prediction Algorithms. *Management Science*, 67(4), 2314–2325.

- Infosys. (2019). *How FinTechs Can Enable Better Support to FIs' Credit Decisioning?* Retrieved July 16, 2023, from <https://www.infosys.com/industries/financial-services/insights/documents/fintechs-fi-partners-credit-decision.pdf>
- Insurance Europe. (2019). *Insurance Fraud: Not a Victimless Crime*. Retrieved July 16, 2023, from <https://insuranceeurope.eu/publications/703/Insurance%5C%20fraud%5C%20-%5C%20not%5C%20a%5C%20victimless%5C%20crime.pdf>
- IntHout, J., Ioannidis, J. P., Rovers, M. M., & Goeman, J. J. (2016). Plea for Routinely Presenting Prediction Intervals in Meta-Analysis. *BMJ Open*, 6(7).
- Jackson, D., & Bowden, J. (2016). Confidence Intervals for the Between-Study Variance in Random-Effects Meta-Analysis Using Generalised Heterogeneity Statistics: Should We Use Unequal Tails? *BMC Medical Research Methodology*, 16(1), 1–15.
- Jacovi, A., Marasović, A., Miller, T., & Goldberg, Y. (2021). Formalizing Trust in Artificial Intelligence: Prerequisites, Causes and Goals of Human Trust in AI. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 624–635.
- Jakubik, J., Schöffner, J., Hoge, V., Vössing, M., & Kühl, N. (2023). An Empirical Evaluation of Predicted Outcomes as Explanations in Human-Ai Decision-Making. *Machine Learning and Principles and Practice of Knowledge Discovery in Databases: International Workshops of ECML PKDD*, 353–368.
- Jakubowski, J., Stanisiz, P., Bobek, S., & Nalepa, G. (2021). Explainable Anomaly Detection for Hot-Rolling Industrial Process. *2021 IEEE 8th International Conference on Data Science and Advanced Analytics (DSAA)*.
- Janiesch, C., Zschech, P., & Heinrich, K. (2021). Machine Learning and Deep Learning. *Electronic Markets*, 31(3), 685–695.
- Jankauskas, M., Serackis, A., Šapurov, M., Pomarnacki, R., Baskys, A., Hyunh, V. K., Vaimann, T., & Zakis, J. (2023). Exploring the Limits of Early Predictive Maintenance in Wind Turbines Applying an Anomaly Detection Technique. *Sensors*, 23(12), 5695–5705.
- Janus, P., Ganzha, M., Bicki, A., & Paprzycki, M. (2021). Applying Machine Learning to Study Infrastructure Anomalies in a Mid-Size Data Center—Preliminary Considerations. *Proceedings of the 54th Hawaii International Conference on System Sciences*, 218–228.
- Jarrah, M. H. (2018). Artificial Intelligence and the Future of Work: Human-AI Symbiosis in Organizational Decision Making. *Business Horizons*, 61(4), 577–586.
- Jo, A. (2023). The Promise and Peril of Generative AI. *Nature*, 614(1), 214–216.
- John M., D. (1990). Personality Structure: Emergence of the Five-Factor Model. *Annual Reviews Psychology*, 50, 116–123.
- Jones, D., & Gregor, S. (2007). The Anatomy of a Design Theory. *Journal of the Association for Information Systems*, 8(5), 1–27.
- Jordan, M. I., & Mitchell, T. M. (2015). Machine Learning: Trends, Perspectives, and Prospects. *Science*, 349(6245), 255–260.

- Jorge, A. M., Leal, J. P., Anand, S. S., & Dias, H. (2014). A Study of Machine Learning Methods for Detecting User Interest During Web Sessions. *Proceedings of the 18th International Database Engineering & Applications Symposium*, 149–157.
- Jung, D., Dorner, V., Glaser, F., & Morana, S. (2018). Robo-Advisory. *Business & Information Systems Engineering*, 60(1), 81–86.
- Jussupow, E., Benbasat, I., & Heinzl, A. (2020). Why Are We Averse Towards Algorithms? A Comprehensive Literature Review on Algorithm Aversion. *Proceedings of the 28th European Conference on Information Systems*, 168–185.
- Jussupow, E., Spohrer, K., Heinzl, A., & Gawlitzka, J. (2021). Augmenting Medical Diagnosis Decisions? An Investigation Into Physicians' Decision-Making Process With Artificial Intelligence. *Information Systems Research*, 32(3), 713–735.
- Kaelbling, L. P., Littman, M. L., & Moore, A. W. (1996). Reinforcement Learning: A Survey. *Journal of Artificial Intelligence Research*, 4, 237–285.
- Kahneman, D. (2011). *Thinking, Fast and Slow*. Macmillan.
- Karimipour, H., Dehghantanha, A., Parizi, R. M., Choo, K.-K. R., & Leung, H. (2019). A Deep and Scalable Unsupervised Machine Learning System for Cyber-Attack Detection in Large-Scale Smart Grids. *IEEE Access*, 7, 80778–80788.
- Kaur, H., Nori, H., Jenkins, S., Caruana, R., Wallach, H., & Wortman Vaughan, J. (2020). Interpreting Interpretability: Understanding Data Scientists' Use of Interpretability Tools for Machine Learning. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–14.
- Kenny, E. M., & Keane, M. T. (2021). On Generating Plausible Counterfactual and Semi-Factual Explanations for Deep Learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(13), 11575–11585.
- Kerasidou, C. X., Kerasidou, A., Buscher, M., & Wilkinson, S. (2021). Before and Beyond Trust: Reliance in Medical AI. *Journal of Medical Ethics*, 852–856.
- Keynes, J. M. (1923). *A Tract on Monetary Reform*. Cosimo Classics.
- Kieu, T., Yang, B., Guo, C., & Jensen, C. S. (2019). Outlier Detection for Time Series With Recurrent Autoencoder Ensembles. *International Joint Conference on Artificial Intelligence*, 2725–2732.
- Kitts, B., & Leblanc, B. (2004). Optimal Bidding on Keyword Auctions. *Electronic Markets*, 14(3), 186–201.
- Klapp, O. E. (1986). *Overload and Boredom: Essays on the Quality of Life in the Information Society*. Greenwood Publishing Group Inc.
- Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., & Mullainathan, S. (2018). Human Decisions and Machine Predictions. *The Quarterly Journal of Economics*, 133(1), 237–293.

- Kloker, A., Fleiß, J., Koeth, C., Kloiber, T., Ratheiser, P., & Thalmann, S. (2022). Caution or Trust in AI? How to Design XAI in Sensitive Use Cases? *Americas Conference on Information Systems*, 16–28.
- Klör, B., Monhof, M., Beverungen, D., & Bräuer, S. (2018). Design and Evaluation of a Model-Driven Decision Support System for Repurposing Electric Vehicle Batteries. *European Journal of Information Systems*, 27(2), 171–188.
- Koehler, D. J. (1991). Explanation, Imagination, and Confidence in Judgment. *Psychological Bulletin*, 110(3), 499.
- Koza, J. R., Bennett, F. H., Andre, D., & Keane, M. A. (1996). Automated Design of Both the Topology and Sizing of Analog Electrical Circuits Using Genetic Programming. Springer.
- Krawczyk, B. (2016). Learning From Imbalanced Data: Open Challenges and Future Directions. *Progress in Artificial Intelligence*, 5(4), 221–232.
- Kucklick, J.-P., Müller, J., Beverungen, D., & Müller, O. (2021). Quantifying the Impact of Location Data for Real Estate Appraisal - a GIS-based Deep Learning Approach. *Proceedings of the 29th European Conference on Information Systems*, 23–36.
- Kühl, N., Hirt, R., Baier, L., Schmitz, B., & Satzger, G. (2021). How to Conduct Rigorous Supervised Machine Learning in Information Systems Research: The Supervised Machine Learning Report Card. *Communications of the Association for Information Systems*, 48–76(1), 46.
- Kühl, N., Lobana, J., & Meske, C. (2019). Do You Comply With AI?—Personalized Explanations of Learning Algorithms and Their Impact on Employees' Compliance Behavior. *Proceedings of the 40th International Conference on Information Systems: Paper-A-Thon*.
- Kühl, N., Mühlthaler, M., & Goutier, M. (2020). Supporting customer-oriented marketing with artificial intelligence: automatically quantifying customer needs from social media. *Electronic Markets*, 30, 351–367.
- Kühl, N., Schemmer, M., Goutier, M., & Satzger, G. (2022). Artificial Intelligence and Machine Learning. *Electronic Markets*, 32(4), 2235–2244.
- 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. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–12.
- Kvaløy, O., Nieken, P., & Schöttner, A. (2015). Hidden Benefits of Reward: A Field Experiment on Motivation and Monetary Incentives. *European Economic Review*, 76, 188–199.
- Lacity, M. C., & Willcocks, L. P. (2016). A New Approach to Automating Services. *MIT Sloan Management Review*, 58(1), 41–49.
- Lai, K. (2018). Estimating Standardized SEM Parameters Given Nonnormal Data and Incorrect Model: Methods and Comparison. *Structural Equation Modeling: A Multidisciplinary Journal*, 25(4), 600–620.

- Lai, V., Carton, S., Bhatnagar, R., Liao, Q. V., Zhang, Y., & Tan, C. (2022). Human-AI Collaboration via Conditional Delegation: A Case Study of Content Moderation. *CHI Conference on Human Factors in Computing Systems*, 1–18.
- Lai, V., Chen, C., Smith-Renner, A., Liao, Q. V., & Tan, C. (2023). Towards a Science of Human-AI Decision Making: An Overview of Design Space in Empirical Human-Subject Studies. *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 1369–1385.
- Lai, V., Liu, H., & Tan, C. (2020). "Why Is' Chicago'deceptive?" Towards Building Model-Driven Tutorials for Humans. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–13.
- Lai, V., & Tan, C. (2019). On Human Predictions With Explanations and Predictions of Machine Learning Models: A Case Study on Deception Detection. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 29–38.
- Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). Building Machines That Learn and Think Like People. *Behavioral and Brain Sciences*, 40.
- LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep Learning. *Nature*, 521(7553), 436–444.
- Lee, J., & Moray, N. (1992). Trust, Control Strategies and Allocation of Function in Human-Machine Systems. *Ergonomics*, 35(10), 1243–1270.
- Lee, J. D., & See, K. A. (2004). Trust in Automation: Designing for Appropriate Reliance. *Human Factors*, 46(1), 50–80.
- Lee, W. K., & Sohn, S. Y. (2020). A Large-Scale Data-Based Investigation on the Relationship Between Bad Weather and Taxi Tipping. *Journal of Environmental Psychology*, 70, 101458.
- Legg, S., & Hutter, M. (2007). Universal Intelligence: A Definition of Machine Intelligence. *Minds and Machines*, 17, 391–444.
- Leibig, C., Brehmer, M., Bunk, S., Byng, D., Pinker, K., & Umutlu, L. (2022). Combining the Strengths of Radiologists and AI for Breast Cancer Screening: A Retrospective Analysis. *The Lancet Digital Health*, 4(7), 507–519.
- Léon, E., & Dejoux, C. (2018). *Métamorphose Des Managers...: À L'ère Du Numérique Et De l'Intelligence Artificielle*. Pearson.
- LePine, J. A., Hollenbeck, J. R., Ilgen, D. R., & Hedlund, J. (1997). Effects of Individual Differences on the Performance of Hierarchical Decision-Making Teams: Much More Than G. *Journal of Applied Psychology*, 82(5), 803.
- Levitt, B., & March, J. G. (1988). Organizational Learning. *Annual Review of Sociology*, 14(1), 319–338.
- Leys, C., Ley, C., Klein, O., Bernard, P., & Licata, L. (2013). Detecting Outliers: Do Not Use Standard Deviation Around the Mean, Use Absolute Deviation Around the Median. *Journal of Experimental Social Psychology*, 49(4), 764–766.

- Li, D., Kulasegaram, K., & Hodges, B. D. (2019a). Why We Needn't Fear the Machines: Opportunities for Medicine in a Machine Learning World. *Academic Medicine*, 94(5), 623–625.
- Li, X., Feng, J., Meng, Y., Han, Q., Wu, F., & Li, J. (2019b). A Unified MRC Framework for Named Entity Recognition. *arXiv Preprint arXiv:1910.11476*.
- Licklider, J. C. (1960). Man-Computer Symbiosis. *IRE Transactions on Human Factors in Electronics*, (1), 4–11.
- Liebman, E., Saar-Tsechansky, M., & Stone, P. (2014). Dj-Mc: A Reinforcement-Learning Agent for Music Playlist Recommendation. *arXiv Preprint arXiv:1401.1880*.
- Lieto, A., Bhatt, M., Oltramari, A., & Vernon, D. (2018). The Role of Cognitive Architectures in General Artificial Intelligence. *Cognitive Systems Research*, 48, 1–3.
- Liu, D., Alnegheimish, S., Zyttek, A., & Veeramachaneni, K. (2022). MTV: Visual Analytics for Detecting, Investigating, and Annotating Anomalies in Multivariate Time Series. *Proceedings of the ACM on Human-Computer Interaction*, 6(1), 1–30.
- Liu, H., Lai, V., & Tan, C. (2021). Understanding the Effect of Out-of-Distribution Examples and Interactive Explanations on Human-AI Decision Making. *Proceedings of the ACM on Human-Computer Interaction*, 5(2), 1–45.
- Longoni, C., Bonezzi, A., & Morewedge, C. K. (2019). Resistance to Medical Artificial Intelligence. *Journal of Consumer Research*, 46(4), 629–650.
- Lundberg, S. M., Erion, G. G., & Lee, S.-I. (2018). Consistent Individualized Feature Attribution for Tree Ensembles. *arXiv Preprint arXiv:1802.03888*.
- Lundberg, S. M., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. *Advances in Neural Information Processing Systems*, 30, 4768–4777.
- Madhuri, C. R., Anuradha, G., & Pujitha, M. V. (2019). House Price Prediction Using Regression Techniques: A Comparative Study. *2019 International Conference on Smart Structures and Systems*, 1–5.
- Maedche, A., Legner, C., Benlian, A., Berger, B., Gimpel, H., Hess, T., Hinz, O., Morana, S., & Söllner, M. (2019). AI-based Digital Assistants: Opportunities, Threats, and Research Perspectives. *Business & Information Systems Engineering*, 61, 535–544.
- Malhotra, P., Ramakrishnan, A., Anand, G., Vig, L., Agarwal, P., & Shroff, G. (2016). LSTM-based Encoder-Decoder for Multi-Sensor Anomaly Detection. *arXiv Preprint arXiv:1607.00148*.
- Mallari, K., Inkpen, K., Johns, P., Tan, S., Ramesh, D., & Kamar, E. (2020). Do I Look Like a Criminal? Examining How Race Presentation Impacts Human Judgement of Recidivism. *Proceedings of the 2020 Chi Conference on Human Factors in Computing Systems*, 1–13.
- Malone, T., Vaccaro, M., Campero, A., Song, J., Wen, H., & Almaatouq, A. (2023). A Test for Evaluating Performance in Human-AI Systems.

- Manzey, D., Reichenbach, J., & Onnasch, L. (2012). Human Performance Consequences of Automated Decision Aids: The Impact of Degree of Automation and System Experience. *Journal of Cognitive Engineering and Decision Making*, 6(1), 57–87.
- Martin-de-Castro, G., López-Sáez, P., & Navas-López, J. E. (2008). Processes of Knowledge Creation in Knowledge-Intensive Firms: Empirical Evidence From Boston's Route 128 and Spain. *Technovation*, 28(4), 222–230.
- Matschak, T., Prinz, C., Masuch, K., & Trang, S. (2021). Healthcare in Fraudster's Crosshairs: Designing, Implementing and Evaluating a Machine Learning Approach for Anomaly Detection on Medical Prescription Claim Data. *Pacis*, 89.
- McAuley, J., Leskovec, J., & Jurafsky, D. (2012). Learning Attitudes and Attributes From Multi-Aspect Reviews. *12th International Conference on Data Mining*, 1020–1025.
- McDonald, K., Fisher, S., & Connelly, C. E. (2017). E-HRM Systems in Support of "Smart" Workforce Management: An Exploratory Case Study of System Success. *Electronic HRM in the Smart Era*, 87–108.
- McGrath, C., Palmgren, P. J., & Liljedahl, M. (2019). Twelve Tips for Conducting Qualitative Research Interviews. *Medical Teacher*, 41(9), 1002–1006.
- McKinney, S. M., Sieniek, M., Godbole, V., Godwin, J., Antropova, N., Ashrafian, H., Back, T., Chesus, M., Corrado, G. S., Darzi, A., et al. (2020). International Evaluation of an AI System for Breast Cancer Screening. *Nature*, 577(7788), 89–94.
- McKinsey. (2023). *The Economic Potential of Generative AI: The Next Productivity Frontier*. Retrieved July 16, 2023, from <https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/the-economic-potential-of-generative-ai-the-next-productivity-frontier%5C#key-insights>
- McKnight, D. H., Choudhury, V., & Kacmar, C. (2002). The Impact of Initial Consumer Trust on Intentions to Transact With a Web Site: A Trust Building Model. *The Journal of Strategic Information Systems*, 11(3-4), 297–323.
- Meehl, P. E. (1957). When Shall We Use Our Heads Instead of the Formula? *Journal of Counseling Psychology*, 4(4), 268–273.
- Meske, C., Bunde, E., Schneider, J., & Gersch, M. (2022). Explainable Artificial Intelligence: Objectives, Stakeholders, and Future Research Opportunities. *Information Systems Management*, 39(1), 53–63.
- Minarini, F., & Decker, L. (2020). Time-Series Anomaly Detection Applied to Log-Based Diagnostic System Using Unsupervised Machine Learning Approach. *Conference of Open Innovations Association, FRUCT*, (27), 343–348.
- Misra, P., & Yadav, A. S. (2019). Impact of Preprocessing Methods on Healthcare Predictions. *Proceedings of 2nd International Conference on Advanced Computing and Software Engineering*.
- Mitchell, T. M. (1997). *Machine Learning*.
- Montazemi, A. R., Wang, F., Nainar, S. K., & Bart, C. K. (1996). On the Effectiveness of Decisional Guidance. *Decision Support Systems*, 18(2), 181–198.

- Morana, S., Schacht, S., Scherp, A., & Maedche, A. (2017). A Review of the Nature and Effects of Guidance Design Features. *Decision Support Systems*, 97, 31–42.
- Morgeson, F. P., Delaney-Klinger, K., & Hemingway, M. A. (2005). The Importance of Job Autonomy, Cognitive Ability, and Job-Related Skill for Predicting Role Breadth and Job Performance. *Journal of Applied Psychology*, 90(2), 399–407.
- Mosier, K. L., & Skitka, L. J. (1999). Automation Use and Automation Bias. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 43(3), 344–348.
- Moustafa, N., Hu, J., & Slay, J. (2019). A Holistic Review of Network Anomaly Detection Systems: A Comprehensive Survey. *Journal of Network and Computer Applications*, 128, 33–55.
- Mozannar, H., Bansal, G., Fourney, A., & Horvitz, E. (2022). Reading Between the Lines: Modeling User Behavior and Costs in AI-Assisted Programming. *arXiv Preprint arXiv:2210.14306*.
- Mozannar, H., & Sontag, D. (2020). Consistent Estimators for Learning to Defer to an Expert. *International Conference on Machine Learning*, 7076–7087.
- Müller, O., Junglas, I., Brocke, J. v., & Debortoli, S. (2016). Utilizing Big Data Analytics for Information Systems Research: Challenges, Promises and Guidelines. *European Journal of Information Systems*, 25, 289–302.
- Muruti, G., Rahim, F. A., & bin Ibrahim, Z.-A. (2018). A Survey on Anomalies Detection Techniques and Measurement Methods. *2018 IEEE Conference on Application, Information and Network Security*, 81–86.
- Nassif, A. B., Talib, M. A., Nasir, Q., & Dakalbab, F. M. (2021). Machine Learning for Anomaly Detection: A Systematic Review. *IEEE Access*, 9, 78658–78700.
- Nawrocki, T., Maldjian, P. D., Slasky, S. E., & Contractor, S. G. (2018). Artificial Intelligence and Radiology: Have Rumors of the Radiologist’s Demise Been Greatly Exaggerated? *25(8)*, 967–972.
- Needham, M. (2023). *Worldwide Spending on AI-Centric Systems Forecast to Reach \$154 Billion in 2023* [IDC: The premier global market intelligence company]. Retrieved July 15, 2023, from <https://www.idc.com/getdoc.jsp?containerId=prUS50454123>
- Neisser, U. (2014). *Cognitive Psychology: Classic Edition*. Psychology press.
- Neuhofer, B., Buhalis, D., & Ladkin, A. (2015). Smart Technologies for Personalized Experiences: A Case Study in the Hospitality Domain. *Electronic Markets*, 25, 243–254.
- Neuman, G. A., & Wright, J. (1999). Team Effectiveness: Beyond Skills and Cognitive Ability. *Journal of Applied Psychology*, 84(3), 376.
- Newell, A., & Simon, H. A. (1961). GPS, a Program That Simulates Human Thought.
- Nguyen, G., Kim, D., & Nguyen, A. (2021). The Effectiveness of Feature Attribution Methods and Its Correlation With Automatic Evaluation Scores. *Advances in Neural Information Processing Systems*, 34, 26422–26436.

- Nguyen, G., Taesiri, M. R., & Nguyen, A. (2022). Visual Correspondence-Based Explanations Improve AI Robustness and Human-Ai Team Accuracy. *Neural Information Processing Systems*.
- Niese, B., & Adya, M. (2022). Who Can I Trust? Use and Disuse of AI-Based Systems Under Conditions of High and Low Stakes, 7–9.
- Nourani, M., Honeycutt, D. R., Block, J. E., Roy, C., Rahman, T., Ragan, E. D., & Gogate, V. (2020a). Investigating the Importance of First Impressions and Explainable AI With Interactive Video Analysis. *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–8.
- Nourani, M., King, J., & Ragan, E. (2020b). The Role of Domain Expertise in User Trust and the Impact of First Impressions With Intelligent Systems. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, 8(1), 112–121.
- Nunes, I., & Jannach, D. (2017). A Systematic Review and Taxonomy of Explanations in Decision Support and Recommender Systems. *User Modeling and User-Adapted Interaction*, 27, 393–444.
- Oliveira, D. F. N., Vismari, L. F., Nascimento, A. M., Almeida Jr, J. R. d., Cugnasca, P. S., Camargo Jr, J. B., Almeida, L., Gripp, R., & Neves, M. (2022). A New Interpretable Unsupervised Anomaly Detection Method Based on Residual Explanation. *IEEE Access*, 10, 1401–1409.
- Ongsulee, P. (2017). Artificial Intelligence, Machine Learning and Deep Learning. *2017 15th International Conference on ICT and Knowledge Engineering (ICT&KE)*, 1–6.
- Oroszi, F., & Ruhland, J. (2010). An Early Warning System for Hospital Acquired Pneumonia, 93–107.
- Ott, M., Cardie, C., & Hancock, J. T. (2013). Negative Deceptive Opinion Spam. *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 497–501.
- Ott, M., Choi, Y., Cardie, C., & Hancock, J. T. (2011). Finding Deceptive Opinion Spam by Any Stretch of the Imagination. *arXiv Preprint arXiv:1107.4557*.
- Pang, G., Shen, C., Cao, L., & Hengel, A. V. D. (2021). Deep Learning for Anomaly Detection: A Review. *ACM Computing Surveys (CSUR)*, 54(2), 1–38.
- Pang, G., Shen, C., & van den Hengel, A. (2019). Deep Anomaly Detection With Deviation Networks. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 353–362.
- Papenmeier, A., Englebienne, G., & Seifert, C. (2019). How Model Accuracy and Explanation Fidelity Influence User Trust. *arXiv Preprint arXiv:1907.12652*.
- Parasuraman, R., Molloy, R., & Singh, I. L. (1993). Performance Consequences of Automation-Induced 'complacency'. *The International Journal of Aviation Psychology*, 3(1), 1–23.

- Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2000). A Model for Types and Levels of Human Interaction With Automation. *IEEE Transactions on Systems, Man, and Cybernetics Part A: Systems and Humans.*, 30(3).
- Parkes, A. (2012). Persuasive Decision Support: Improving Reliance on Decision Aids. *Pacis*, 4(3), 2.
- Peng, A., Nushi, B., Kiciman, E., Inkpen, K., & Kamar, E. (2022). Investigations of Performance and Bias in Human-AI Teamwork in Hiring. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(11), 12089–12097.
- Pennisi, M., Kavasidis, I., Spampinato, C., Schinina, V., Palazzo, S., Salanitri, F. P., Bellitto, G., Rundo, F., Aldinucci, M., Cristofaro, M., et al. (2021). An Explainable AI System for Automated COVID-19 Assessment and Lesion Categorization From CT-scans. *Artificial Intelligence in Medicine*, 118, 102114–102127.
- Peterson, D. K., & Pitz, G. F. (1988). Confidence, Uncertainty, and the Use of Information. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 14(1), 85–92.
- Petropoulos, F., Fildes, R., & Goodwin, P. (2016). Do ‘Big Losses’ in Judgmental Adjustments to Statistical Forecasts Affect Experts’ Behaviour? *European Journal of Operational Research*, 249(3), 842–852.
- Phillips-Wren, G., Power, D. J., & Mora, M. (2019). Cognitive Bias, Decision Styles, and Risk Attitudes in Decision Making and DSS. *Journal of Decision Systems*, 28(2), 63–66.
- Pisano, E. D. (2020). *AI Shows Promise for Breast Cancer Screening*. Retrieved July 16, 2023, from <https://www.nature.com/articles/d41586-019-03822-8>
- Plumb, G., Molitor, D., & Talwalkar, A. S. (2018). Model Agnostic Supervised Local Explanations. *Advances in Neural Information Processing Systems*, 31, 2520–2529.
- Poole, D. I., Goebel, R. G., & Mackworth, A. K. (1998). *Computational Intelligence* (Vol. 1). Oxford University Press Oxford.
- Poursabzi-Sangdeh, F., Goldstein, D. G., Hofman, J. M., Wortman Vaughan, J. W., & Wallach, H. (2021). Manipulating and Measuring Model Interpretability. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–52.
- Power, D. J. (2007). A Brief History of Decision Support Systems. *DSS Resources*, 3.
- Power, D. J., Cyphert, D., & Roth, R. M. (2019). Analytics, Bias, and Evidence: The Quest for Rational Decision Making. *Journal of Decision Systems*, 28(2), 120–137.
- ProPublica. (2016). Machine Bias: There’s Software Used Across the Country to Predict Future Criminals. And It’s Biased Against Blacks. <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>
- Qian, M., & Qian, D. (2020). Defining a Human-Machine Teaming Model for AI-Powered Human-Centered Machine Translation Agent by Learning From Human-Human Group Discussion: Dialog Categories and Dialog Moves. *International Conference on Human-Computer Interaction*, 70–81.

- Qian, Y., Ying, S., & Wang, B. (2020). Anomaly Detection in Distributed Systems via Variational Autoencoders. *2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 2822–2829.
- Qing, C., Parfenov, S., & Kim, L.-J. (2015). Identifying Travel Patterns During Extreme Weather Using Taxi GPS Data. *Transportation Research Board 94th Annual Meeting*.
- Rai, A., Constantinides, P., & Sarker, S. (2019). Next Generation Digital Platforms: Toward Human-AI Hybrids. *Mis Quarterly*, 43(1), iii–ix.
- Rastogi, C., Leqi, L., Holstein, K., & Heidari, H. (2022). A Unifying Framework for Combining Complementary Strengths of Humans and ML Toward Better Predictive Decision-Making. *arXiv Preprint arXiv:2204.10806*.
- Ren, H., Hou, Z., Wang, H., Zarzhitsky, D., & Etingov, P. (2018). Pattern Mining and Anomaly Detection Based on Power System Synchrophasor Measurements. *Proceedings of the Annual Hawaii International Conference on System Sciences*.
- Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster R-Cnn: Towards Real-Time Object Detection With Region Proposal Networks. *Advances in Neural Information Processing Systems*, 28, 91–99.
- Reverberi, C., Rigon, T., Solari, A., Hassan, C., Cherubini, P., & Cherubini, A. (2022). Experimental Evidence of Effective Human-AI Collaboration in Medical Decision-Making. *Scientific Reports*, 12(1), 14952.
- Ribeiro, M. T., Singh, S., & Guestrin, C. (2016a). "Why Should I Trust You?" Explaining the Predictions of Any Classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144.
- Ribeiro, M. T., Singh, S., & Guestrin, C. (2016b). Model-Agnostic Interpretability of Machine Learning. *arXiv Preprint arXiv:1606.05386*.
- Ribeiro, M. T., Singh, S., & Guestrin, C. (2018). Anchors: High-Precision Model-Agnostic Explanations. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1).
- Rich, E., & Knight, K. (1991). *Artificial Intelligence*. McGraw-Hill, New York.
- Riefle, L., & Benz, C. (2021). User-Specific Determinants of Conversational Agent Usage: A Review and Potential for Future Research. *Innovation Through Information Systems: Volume II: A Collection of Latest Research on Technology Issues*, 115–129.
- Riley, V. (2018). Operator Reliance on Automation: Theory and Data. In *Automation and Human Performance: Theory and Applications* (pp. 19–35).
- Risdal, M., Prasanth, RumiGhosh, Soundar, Stefanie, & Cukierski, W. (2016). *Bosch Production Line Performance* [Kaggle]. Retrieved July 16, 2023, from <https://kaggle.com/competitions/bosch-production-line-performance>
- Ritchie, S. G. (1990). A Knowledge-Based Decision Support Architecture for Advanced Traffic Management. *Transportation Research Part A: General*, 24(1), 27–37.
- Roelofs, C. M., Lutz, M.-A., Faulstich, S., & Vogt, S. (2021). Autoencoder-Based Anomaly Root Cause Analysis for Wind Turbines. *Energy and AI*, 4, 100065–100074.

- Roohi, A., Faust, K., Djuric, U., & Diamandis, P. (2020). Unsupervised Machine Learning in Pathology: The Next Frontier. *Surgical Pathology Clinics*, 13(2), 349–358.
- Ross, S. I., Martinez, F., Houde, S., Muller, M., & Weisz, J. D. (2023). The Programmer's Assistant: Conversational Interaction With a Large Language Model for Software Development. *Proceedings of the 28th International Conference on Intelligent User Interfaces*, 491–514.
- Rosseel, Y. (2012). Lavaan: An R Package for Structural Equation Modeling. *Journal of Statistical Software*, 48(2), 1–36.
- Rostami, M., Kolouri, S., Kim, K., & Eaton, E. (2017). Multi-Agent Distributed Lifelong Learning for Collective Knowledge Acquisition. *arXiv Preprint arXiv:1709.05412*.
- Rousseeuw, P. J., & Croux, C. (1993). Alternatives to the Median Absolute Deviation. *Journal of the American Statistical Association*, 88(424), 1273–1283.
- Ruelens, F., Iacovella, S., Claessens, B. J., & Belmans, R. (2015). Learning Agent for a Heat-Pump Thermostat With a Set-Back Strategy Using Model-Free Reinforcement Learning. *Energies*, 8(8), 8300–8318.
- Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1985). *Learning Internal Representations by Error Propagation* (tech. rep.). California Univ San Diego La Jolla Inst for Cognitive Science.
- Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al. (2015a). ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, 115(3), 211–252.
- Russakovsky, O., Li, L.-J., & Fei-Fei, L. (2015b). Best of Both Worlds: Human-Machine Collaboration for Object Annotation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2121–2131.
- Russell, S. J. (2010). *Artificial Intelligence a Modern Approach*. Pearson Education, Inc.
- Šabić, E., Keeley, D., Henderson, B., & Nannemann, S. (2021). Healthcare and Anomaly Detection: Using Machine Learning to Predict Anomalies in Heart Rate Data. *Ai & Society*, 36(1), 149–158.
- Saeed, W., & Omlin, C. (2023). Explainable AI (XAI): A Systematic Meta-Survey of Current Challenges and Future Opportunities. *Knowledge-Based Systems*, 263, 110273.
- Samtani, S., Chinn, R., Chen, H., & Nunamaker Jr, J. F. (2017). Exploring Emerging Hacker Assets and Key Hackers for Proactive Cyber Threat Intelligence. *Journal of Management Information Systems*, 34(4), 1023–1053.
- Sanchez, J. C., & Principe, J. C. (2009). Prerequisites for Symbiotic Brain-Machine Interfaces. *2009 IEEE International Conference on Systems, Man and Cybernetics*, 1736–1741.
- Sanders, N. R., & Ritzman, L. P. (1991). On Knowing When to Switch From Quantitative to Judgemental Forecasts. *International Journal of Operations & Production Management*, 11(6), 27–37.

- Sanders, N. R., & Ritzman, L. P. (1995). Bringing Judgment Into Combination Forecasts. *Journal of Operations Management*, 13(4), 311–321.
- Sanders, N. R., & Ritzman, L. P. (2001). Judgmental Adjustment of Statistical Forecasts. *Principles of Forecasting: A Handbook for Researchers and Practitioners*, 405–416.
- Sandini, G., Mohan, V., Sciutti, A., & Morasso, P. (2018). Social Cognition for Human-Robot Symbiosis—challenges and Building Blocks. *Frontiers in Neurorobotics*, 12, 34–53.
- Sarker, I. H. (2021). Machine Learning: Algorithms, Real-World Applications and Research Directions. *SN Computer Science*, 2(3), 1–21.
- Sarter, N. B., & Schroeder, B. (2001). Supporting Decision Making and Action Selection Under Time Pressure and Uncertainty: The Case of in-Flight Icing. *Human Factors*, 43(4), 573–583.
- Scharowski, N., Perrig, S. A., von Felten, N., & Brühlmann, F. (2022). Trust and Reliance in XAI—Distinguishing Between Attitudinal and Behavioral Measures. *arXiv Preprint arXiv:2203.12318*.
- Schemmer, M., Bartos, A., Spitzer, P., Hemmer, P., Kühl, N., Liebschner, J., & Satzger, G. (2023a). Towards Effective Human-AI Decision-Making: The Role of Human Learning in Appropriate Reliance on AI Advice [Working paper].
- Schemmer, M., Hemmer, P., Kühl, N., Benz, C., & Satzger, G. (2022a). Should I Follow AI-based Advice? Measuring Appropriate Reliance in Human-AI Decision-Making. *Workshop on Trust and Reliance in AI-Human Teams at CHI 2022*.
- Schemmer, M., Hemmer, P., Kühl, N., Vössing, M., & Satzger, G. (2023b). Human-AI Complementarity: Conceptualization and the Effect of Information Asymmetry [Working paper].
- Schemmer, M., Hemmer, P., Nitsche, M., Kühl, N., & Vössing, M. (2022b). A Meta-Analysis on the Utility of Explainable Artificial Intelligence in Human-AI Decision-Making. *Proceedings of the 2022 AAI/ACM Conference on AI, Ethics, and Society*, 617–626.
- Schemmer, M., Holstein, J., Kühl, N., Vössing, M., & Satzger, G. (2023c). From Anomaly Detection to Anomaly Investigation: Support by Explainable AI [Working paper].
- Schemmer, M., Kuehl, N., Benz, C., Bartos, A., & Satzger, G. (2023d). Appropriate Reliance on AI Advice: Conceptualization and the Effect of Explanations. *Proceedings of the 28th International Conference on Intelligent User Interfaces*, 410–422.
- Schemmer, M., Kühl, N., Benz, C., & Satzger, G. (2022c). On the Influence of Explainable AI on Automation Bias. *Proceedings of the 30th European Conference on Information Systems*, 51–64.
- Schemmer, M., Kühl, N., & Satzger, G. (2022d). Intelligent Decision Assistance Versus Automated Decision-Making: Enhancing Knowledge Work Through Explainable Artificial Intelligence. *Proceedings of the 55th Hawaii International Conference on System Sciences*, 617–626.
- Schleiffer, R. (2005). An Intelligent Agent Model. *European Journal of Operational Research*, 166(3), 666–693.

- Schmidt, P., & Biessmann, F. (2019). Quantifying Interpretability and Trust in Machine Learning Systems. *arXiv Preprint arXiv:1901.08558*.
- Schoeffer, J., De-Arteaga, M., & Kuehl, N. (2022a). On Explanations, Fairness, and Appropriate Reliance in Human-AI Decision-Making. *arXiv Preprint arXiv:2209.11812*.
- Schoeffer, J., De-Arteaga, M., & Kuehl, N. (2022b). On the Relationship Between Explanations, Fairness Perceptions, and Decisions. *ACM CHI 2022 Workshop on Human-Centered Explainable AI*.
- Schoeffer, J., Machowski, Y., & Kuehl, N. (2021). A Study on Fairness and Trust Perceptions in Automated Decision Making. *arXiv Preprint arXiv:2103.04757*.
- Schotten, M., Meester, W. J., Steinginga, S., Ross, C. A., et al. (2017). A Brief History of Scopus: The World's Largest Abstract and Citation Database of Scientific Literature. In *Research Analytics* (pp. 31–58). Auerbach Publications.
- Schreiber, C., Schiefer, G., Alpers, S., Take, M., & Oberweis, A. (2020). Prozessorientiertes Reinforcement Learning: Grafische Modellierung Zur Unterstützung Der Erklärbarkeit. *Modellierung-C 2020: Modellierung 2020 Short, Workshop and Tools & Demo Papers*. Edited by Judith Michael and Dominik Bork, 172–178.
- Schuetz, S., & Venkatesh, V. (2020). The Rise of Human Machines: How Cognitive Computing Systems Challenge Assumptions of User-System Interaction. *Journal of the Association for Information Systems*, 21(2), 460–482.
- Schultz, M., Gnos, N., & Tropmann-Frick, M. (2022). XAI in the Audit Domain—Explaining an Autoencoder Model for Anomaly Detection. *Wirtschaftsinformatik 2022 Proceedings*, 1–14.
- Seeber, I., Bittner, E., Briggs, R. O., De Vreede, T., De Vreede, G.-J., Elkins, A., Maier, R., Merz, A. B., Oeste-Reiß, S., Randrup, N., et al. (2020). Machines as Teammates: A Research Agenda on AI in Team Collaboration. *Information & Management*, 57(2).
- Senoner, J., Netland, T., & Feuerriegel, S. (2022). Using Explainable Artificial Intelligence to Improve Process Quality: Evidence From Semiconductor Manufacturing. *Management Science*, 68(8), 5704–5723.
- Shen, H., & Huang, T.-H. (2020). How Useful Are the Machine-Generated Interpretations to General Users? A Human Evaluation on Guessing the Incorrectly Predicted Labels. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, 8, 168–172.
- Shin, D. (2021). The Effects of Explainability and Causability on Perception, Trust, and Acceptance: Implications for Explainable AI. *International Journal of Human-Computer Studies*, 146, 102551–102561.
- Shirazi, P., Ico, G., Anderson, C. S., Ma, M. C., Kim, B. S., Nam, J., & Myung, N. V. (2017). Size-Dependent Piezoelectric Properties of Electrospun BaTiO₃ for Enhanced Energy Harvesting. *Advanced Sustainable Systems*, 1(11), 1700091–1700099.
- Shneiderman, B. (2020). Human-Centered Artificial Intelligence: Reliable, Safe & Trustworthy. *International Journal of Human-Computer Interaction*, 36(6), 495–504.

- Siau, K., & Wang, W. (2018). Building Trust in Artificial Intelligence, Machine Learning, and Robotics. *Cutter Business Technology Journal*, 31(2), 47–53.
- Siemon, D., Li, R., & Robra-Bissantz, S. (2020). Towards a Model of Team Roles in Human-Machine Collaboration. *Proceedings of the 41st International Conference on Information Systems*, 7–17.
- Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., Lanctot, M., Sifre, L., Kumaran, D., Graepel, T., et al. (2018). A General Reinforcement Learning Algorithm That Masters Chess, Shogi, and Go Through Self-Play. *Science*, 362(6419), 1140–1144.
- Silver, M. S. (1991). Decisional Guidance for Computer-Based Decision Support. *MIS Quarterly*, 15, 105–122.
- Šimić, I., Sabol, V., & Veas, E. (2021). XAI Methods for Neural Time Series Classification: A Brief Review. *arXiv Preprint arXiv:2108.08009*.
- Simon, H. A. (1960). *The New Science of Management Decision*. Harper & Brothers.
- Singla, A., Bogunovic, I., Bartók, G., Karbasi, A., & Krause, A. (2014). Near-Optimally Teaching the Crowd to Classify. *International Conference on Machine Learning*, 154–162.
- Siu, H. C., Peña, J., Chen, E., Zhou, Y., Lopez, V., Palko, K., Chang, K., & Allen, R. (2021). Evaluation of Human-AI Teams for Learned and Rule-Based Agents in Hanabi. *Advances in Neural Information Processing Systems*, 34, 16183–16195.
- Skitka, L. J., Mosier, K., & Burdick, M. D. (2000). Accountability and Automation Bias. *International Journal of Human-Computer Studies*, 52(4), 701–717.
- Smith, K. (2013). *Environmental Hazards: Assessing Risk and Reducing Disaster*. Routledge.
- Sniezek, J. A., & Van Swol, L. M. (2001). Trust, Confidence, and Expertise in a Judge-Advisor System. *Organizational Behavior and Human Decision Processes*, 84(2), 288–307.
- Soldani, J., & Brogi, A. (2022). Anomaly Detection and Failure Root Cause Analysis in (Micro) Service-Based Cloud Applications: A Survey. *ACM Computing Surveys (CSUR)*, 55(3), 1–39.
- Song, F., Diao, Y., Read, J., Stiegler, A., & Bifet, A. (2018). EXAD: A System for Explainable Anomaly Detection on Big Data Traces. *2018 IEEE International Conference on Data Mining Workshops*, 1435–1440.
- Song, X., Wu, M., Jermaine, C., & Ranka, S. (2007). Conditional Anomaly Detection. *IEEE Transactions on Knowledge and Data Engineering*, 19(5), 631–645.
- Sparrow, B., Liu, J., & Wegner, D. M. (2011). Google Effects on Memory: Cognitive Consequences of Having Information at Our Fingertips. *Science*, 333(6043), 776–778.
- Spitzer, P., Kühl, N., & Goutier, M. (2022). Training Novices: The Role of Human-Ai Collaboration and Knowledge Transfer. *arXiv Preprint arXiv:2207.00497*.

- Stark, R., Gruber, H., Hinkofer, L., & Mandl, H. (2004). Overcoming Problems of Knowledge Application and Transfer. *Professional Learning: Gaps and Transitions on the Way From Novice to Expert*, 49–70.
- Stauder, M., & Kühn, N. (2022). AI for in-Line Vehicle Sequence Controlling: Development and Evaluation of an Adaptive Machine Learning Artifact to Predict Sequence Deviations in a Mixed-Model Production Line. *Flexible Services and Manufacturing Journal*, 1–39.
- Steenwinckel, B., De Paepe, D., Hautte, S. V., Heyvaert, P., Bentefrit, M., Moens, P., Dimou, A., Van Den Bossche, B., De Turck, F., Van Hoecke, S., et al. (2021). FLAGS: A Methodology for Adaptive Anomaly Detection and Root Cause Analysis on Sensor Data Streams by Fusing Expert Knowledge With Machine Learning. *Future Generation Computer Systems*, 116, 30–48.
- Steyvers, M., Tejada, H., Kerrigan, G., & Smyth, P. (2022). Bayesian Modeling of Human-AI Complementarity. *Proceedings of the National Academy of Sciences*, 119(11), 1–7.
- Sulem, D., Donini, M., Zafar, M. B., Aubet, F.-X., Gasthaus, J., Januschowski, T., Das, S., Kenthapadi, K., & Archambeau, C. (2022). Diverse Counterfactual Explanations for Anomaly Detection in Time Series. *arXiv Preprint arXiv:2203.11103*.
- Suresh, H., Gomez, S. R., Nam, K. K., & Satyanarayan, A. (2021). Beyond Expertise and Roles: A Framework to Characterize the Stakeholders of Interpretable Machine Learning and Their Needs. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–16.
- Susto, G. A., Terzi, M., & Beghi, A. (2017). Anomaly Detection Approaches for Semiconductor Manufacturing. *Procedia Manufacturing*, 11, 2018–2024.
- Sutton, S. G., Arnold, V., & Holt, M. (2018). How Much Automation Is Too Much? Keeping the Human Relevant in Knowledge Work. *Journal of Emerging Technologies in Accounting*, 15(2), 15–25.
- Swartout, W. R., & Moore, J. D. (1993). Explanation in Second Generation Expert Systems. *Second Generation Expert Systems*, 543–585.
- Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., & Rabinovich, A. (2015). Going Deeper With Convolutions. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1–9.
- Takeda, S., Kishimoto, A., Hamada, L., Nakano, D., & Smith, J. R. (2023). Foundation Model for Material Science. *Proceedings of the AAAI Conference on Artificial Intelligence*, 15376–15383.
- Talone, A. (2019). *The Effect of Reliability Information and Risk on Appropriate Reliance in an Autonomous Robot Teammate* (Doctoral dissertation). University of Central Florida.
- Taudien, A., Fuegener, A., Gupta, A., & Ketter, W. (2022). Calibrating Users' Mental Models for Delegation to AI. *Proceedings of the 43rd International Conference on Information Systems*, 16–26.

- Team, L. P. B. (2017). *LSAT Prep Book Study Guide: Quick Study & Practice Test Questions for the Law School Admissions Council's (LSAC) Law School Admission Test*. Mometrix Test Preparation, Beaumont, TX.
- Tejeda, H., Kumar, A., Smyth, P., & Steyvers, M. (2022). AI-assisted Decision-Making: A Cognitive Modeling Approach to Infer Latent Reliance Strategies. *Computational Brain & Behavior*, 5, 491–508.
- Tereschenko, O., Raff, S., Rose, S., & Wentzel, D. (2022). Are You Trying to Be Funny? The Impact of Affiliative Humor of Smart Home Technologies on Human-Like Trust. *Proceedings of the 43rd International Conference on Information Systems*, 6–24.
- Terveen, L. G. (1995). Overview of human-computer collaboration. *Knowledge-Based Systems*, 8(2-3), 67–81.
- Thiebes, S., Lins, S., & Sunyaev, A. (2021). Trustworthy Artificial Intelligence. *Electronic Markets*, 31, 447–464.
- TLC. (2022). *TLC Trip Record Data* [City of New York]. Retrieved July 16, 2023, from <https://www.nyc.gov/site/tlc/about/tlc-trip-record-data.page>
- Tomsett, R., Harborne, D., Chakraborty, S., Gurram, P., & Preece, A. (2020). Sanity Checks for Saliency Metrics. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(4), 6021–6029.
- Treiss, A., Walk, J., & Kühn, N. (2020). An Uncertainty-Based Human-in-the-Loop System for Industrial Tool Wear Analysis. *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 85–100.
- Turban, E., Sharda, R., & Delen, D. (2010). *Decision Support and Business Intelligence Systems*. Prentice Hall Learning Outcomes.
- Turiel, J., & Aste, T. (2020). Peer-to-Peer Loan Acceptance and Default Prediction With Artificial Intelligence. *Royal Society Open Science*, 7(6), 191649–191668.
- Turing, A. M. (2012). Computing Machinery and Intelligence. *Mind*, 433–464.
- Ullman, S. (2019). Using Neuroscience to Develop Artificial Intelligence. *Science*, 363(6428), 692–693.
- United States Department of Justice. (2014). *State Court Processing Statistics, 1990-2009: Felony Defendants in Large Urban Counties*.
- Van der Waa, J., Nieuwburg, E., Cremers, A., & Neerinx, M. (2021). Evaluating XAI: A Comparison of Rule-Based and Example-Based Explanations. *Artificial Intelligence*, 291, 103404–103423.
- Van Lent, M., Fisher, W., & Mancuso, M. (2004). An Explainable Artificial Intelligence System for Small-Unit Tactical Behavior. *Proceedings of the National Conference on Artificial Intelligence*, 900–907.
- Verkerken, M., D'hooge, L., Wauters, T., Volckaert, B., & De Turck, F. (2022). Towards Model Generalization for Intrusion Detection: Unsupervised Machine Learning Techniques. *Journal of Network and Systems Management*, 30, 1–25.

- Vickers, N. J. (2017). Animal Communication: When I'm Calling You, Will You Answer Too? *Current Biology*, 27(14), 713–715.
- Viriato, J. C. (2019). AI and Machine Learning in Real Estate Investment. *The Journal of Portfolio Management*, 45(7), 43–54.
- vom Brocke, J., Simons, A., Niehaves, B., Riemer, K., Plattfaut, R., & Cleven, A. (2009). Reconstructing the Giant: On the Importance of Rigour in Documenting the Literature Search Process. *Proceedings of the 17th European Conference on Information Systems*, 161–175.
- Vössing, M. (2020). *Designing Human-Computer Collaboration: Transparency and Automation for Intelligence Augmentation* (Doctoral dissertation). Karlsruher Institut für Technologie (KIT).
- Vössing, M., Kühn, N., Lind, M., & Satzger, G. (2022). Designing Transparency for Effective Human-AI Collaboration. *Information Systems Frontiers*, 24(3), 877–895.
- Wachter, S., Mittelstadt, B., & Russell, C. (2017). Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR. *Harv. JL & Tech.*, 31, 841–893.
- Wagner, R. K. (1997). Intelligence, Training, and Employment. *American Psychologist*, 52(10), 1059–1069.
- Wah, C., Branson, S., Welinder, P., Perona, P., & Belongie, S. (2011). *The Caltech-Ucsd Birds-200-2011 Dataset*. Retrieved July 16, 2023, from <https://authors.library.caltech.edu/27452/>
- Wang, J., Cai, L., Yu, A., & Meng, D. (2019). Embedding Learning With Heterogeneous Event Sequence for Insider Threat Detection. *2019 IEEE 31st International Conference on Tools With Artificial Intelligence (ICTAI)*, 947–954.
- Wang, K., Wang, B., & Peng, L. (2009). CVAP: Validation for Cluster Analyses. *Data Science Journal*, 8, 88–93.
- Wang, L., Jamieson, G. A., & Hollands, J. G. (2008). Selecting Methods for the Analysis of Reliance on Automation. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 52(4), 287–291.
- Wang, P., & Vasconcelos, N. (2020). Scout: Self-Aware Discriminant Counterfactual Explanations. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8981–8990.
- Wang, X., & Yin, M. (2021). Are Explanations Helpful? A Comparative Study of the Effects of Explanations in AI-Assisted Decision-Making. *26th International Conference on Intelligent User Interfaces*, 318–328.
- Wang, X., & Du, X. (2018). Why Does Advice Discounting Occur? The Combined Roles of Confidence and Trust. *Frontiers in Psychology*, 9, 2381.

- Wanner, J., Herm, L.-V., Heinrich, K., Janiesch, C., & Zschech, P. (2020). White, Grey, Black: Effects of XAI Augmentation on the Confidence in AI-based Decision Support Systems. *Proceedings of the 41st International Conference on Information Systems*, 14–23.
- Waseem, Z., & Hovy, D. (2016). Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter. *Proceedings of the NAACL Student Research Workshop*, 88–93.
- Watson, H. J. (2014). Tutorial: Big Data Analytics: Concepts, Technologies, and Applications. *Communications of the Association for Information Systems*, 34(1), 65.
- Wilder, B., Horvitz, E., & Kamar, E. (2020). Learning to Complement Humans. *arXiv Preprint arXiv:2005.00582*.
- Wilson, H. J., & Daugherty, P. R. (2018). Collaborative Intelligence: Humans and AI Are Joining Forces. *Harvard Business Review*, 96(4), 114–123.
- Witten, I. H., Frank, E., & Hall, M. A. (2011). *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann.
- Wolpert, D. H. (1992). Stacked Generalization. *Neural Networks*, 5(2), 241–259.
- Wu, N., Phang, J., Park, J., Shen, Y., Huang, Z., Zorin, M., Jastrzebski, S., Févry, T., Katsnelson, J., Kim, E., Wolfson, S., Parikh, U., Gaddam, S., Lin, L., Ho, K., Weinstein, J. D., Reig, B., Gao, Y., Toth, H., . . . Geras, K. (2020). Deep Neural Networks Improve Radiologists' Performance in Breast Cancer Screening. *IEEE Transactions on Medical Imaging*, 39(4), 1184–1194.
- Wulczyn, E., Thain, N., & Dixon, L. (2017). Ex Machina: Personal Attacks Seen at Scale. *Proceedings of the 26th International Conference on World Wide Web*, 1391–1399.
- Xu, X., Wickens, C. D., & Rantanen, E. M. (2007). Effects of Conflict Alerting System Reliability and Task Difficulty on Pilots' Conflict Detection With Cockpit Display of Traffic Information. *Ergonomics*, 50(1), 112–130.
- Xue, F., & Yan, W. (2022). Multivariate Time Series Anomaly Detection With Few Positive Samples. *2022 International Joint Conference on Neural Networks (IJCNN)*, 1–7.
- Yang, F., Huang, Z., Scholtz, J., & Arendt, D. L. (2020a). How Do Visual Explanations Foster End Users' Appropriate Trust in Machine Learning? *Proceedings of the 25th International Conference on Intelligent User Interfaces*, 189–201.
- Yang, L., Kenny, E. M., Ng, T. L. J., Yang, Y., Smyth, B., & Dong, R. (2020b). Generating Plausible Counterfactual Explanations for Deep Transformers in Financial Text Classification. *arXiv Preprint arXiv:2010.12512*.
- Yaniv, I. (2004). Receiving Other People's Advice: Influence and Benefit. *Organizational Behavior and Human Decision Processes*, 93(1), 1–13.
- Yeung, A., Joshi, S., Williams, J. J., & Rudzicz, F. (2020). Sequential Explanations With Mental Model-Based Policies. *arXiv Preprint arXiv:2007.09028*.

- Young, L. F. (1983). Right-Brained Decision Support Systems. *ACM SIGMIS Database: The Database for Advances in Information Systems*, 14(4), 28–36.
- Yu, K., Berkovsky, S., Taib, R., Zhou, J., & Chen, F. (2019). Do I Trust My Machine Teammate? An Investigation From Perception to Decision. *Proceedings of the 24th International Conference on Intelligent User Interfaces*, 460–468.
- Yu, Y., Eshghi, A., & Lemon, O. (2017). VOILA: An Optimised Dialogue System for Interactively Learning Visually-Grounded Word Meanings (Demonstration System). *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, 197–200.
- Zarnoth, P., & Sniezek, J. A. (1997). The Social Influence of Confidence in Group Decision Making. *Journal of Experimental Social Psychology*, 33(4), 345–366.
- Zhang, Q., Lee, M. L., & Carter, S. (2022). You Complete Me: Human-AI Teams and Complementary Expertise. *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, 1–28.
- Zhang, Y., Liao, Q. V., & Bellamy, R. K. (2020). Effect of Confidence and Explanation on Accuracy and Trust Calibration in AI-assisted Decision Making. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 295–305.
- Zhao, Y., & Feng, Y. (2019). A Study on the Impact of Team Heterogeneity on Organizational Performance in Start-Ups. *Proceedings of the 5th International Conference on Economics and Management*, 215–218.
- Zheng, N.-n., Liu, Z.-y., Ren, P.-j., Ma, Y.-q., Chen, S.-t., Yu, S.-y., Xue, J.-r., Chen, B.-d., & Wang, F.-y. (2017). Hybrid-Augmented Intelligence: Collaboration and Cognition. *Frontiers of Information Technology & Electronic Engineering*, 18(2), 153–179.
- Zheng, Q., Wu, Z., Cheng, X., Jiang, L., & Liu, J. (2013). Learning to Crawl Deep Web. *Information Systems*, 38(6), 801–819.
- Zheng, Z., Zheng, L., & Yang, Y. (2018). Pedestrian Alignment Network for Large-Scale Person Re-Identification. *IEEE Transactions on Circuits and Systems for Video Technology*, 29(10), 3037–3045.
- Zhou, L., Paul, S., Demirkan, H., Yuan, L., Spohrer, J., Zhou, M., & Basu, J. (2021). Intelligence Augmentation: Towards Building Human-Machine Symbiotic Relationship. *AIS Transactions on Human-Computer Interaction*, 13(2), 243–264.
- Zhou, Z.-J., Hu, C.-H., Yang, J.-B., Xu, D.-L., & Zhou, D.-H. (2009). Online Updating Belief Rule Based System for Pipeline Leak Detection Under Expert Intervention. *Expert Systems With Applications*, 36(4), 7700–7709.
- Zhu, X., Singla, A., Zilles, S., & Rafferty, A. N. (2018). An Overview of Machine Teaching. *arXiv Preprint arXiv:1801.05927*.
- Zhu, X. J. (2005). *Semi-Supervised Learning Literature Survey* (tech. rep.). University of Wisconsin.

- Zschech, P., Walk, J., Heinrich, K., Vössing, M., & Kühl, N. (2021). A Picture Is Worth a Collaboration: Accumulating Design Knowledge for Computer-Vision-Based Hybrid Intelligence Systems. *Proceedings of the 29th European Conference on Information Systems*, 127–145.
- Zuboff, S. (1985). Automate/Informate: The Two Faces of Intelligent Technology. *Organizational Dynamics*, 14(2), 5–18.

Declarations

Eidesstattliche Versicherung

gemäß §13 Absatz 2 Ziffer 3 der Promotionsordnung des Karlsruher Instituts für
Technologie für die KIT-Fakultät für Wirtschaftswissenschaften

1. Bei der eingereichten Dissertation zu dem Thema "*From Competition to Complementarity: Foundations and Evidence for Effective Human-AI Collaboration*" handelt es sich um meine eigenständig erbrachte Leistung.
2. Ich habe nur die angegebenen Quellen und Hilfsmittel benutzt und mich keiner unzulässigen Hilfe Dritter bedient. Insbesondere habe ich wörtlich oder sinngemäß aus anderen Werken übernommene Inhalte als solche kenntlich gemacht.
3. Die Arbeit oder Teile davon habe ich bislang nicht an einer Hochschule des In- oder Auslands als Bestandteil einer Prüfungs- oder Qualifikationsleistung vorgelegt.
4. Die Richtigkeit der vorstehenden Erklärungen bestätige ich.
5. Die Bedeutung der eidesstattlichen Versicherung und die strafrechtlichen Folgen einer unrichtigen oder unvollständigen eidesstattlichen Versicherung sind mir bekannt.

Ich versichere an Eides statt, dass ich nach bestem Wissen die reine Wahrheit erklärt und nichts verschwiegen habe.

Karlsruhe, den 16.07.2023

Max Richard Schemmer

