

# **Behavioural Aspects of Human-AI Interaction in Theory and Application**

Der Fakultät für Wirtschaftswissenschaften der Universität Paderborn

zur Erlangung des akademischen Grades  
Doktor der Wirtschaftswissenschaften  
- Doctor rerum politicarum -

vorgelegte Dissertation von

***Anastasia Lebedeva***

geboren am 10.06.1993

in Iwanowo, Russland

Universität Paderborn  
Paderborn, Deutschland

## **Acknowledgements**

The way to my dissertation was not easy – both because I was writing it in parallel to my full-time job and because of the impacts of the corona pandemic and the war in Ukraine. Along this uneven pass I could always rely on people around me and thus would like to express my warmest thanks for their support. I thank René for enabling me to perform research I was passionately interested in, supporting my interdisciplinary approach and guiding me with his knowledge and experience. I thank Mark, Eberhard, Thomas and Björn, who, while being my superiors at work, supported me with every step of my dissertation and demonstrated a lot of understanding for my PhD activities. I also thank Kirsten and Daniel, who welcomed me in their fields of research, providing me with invaluable theoretical and methodological support, needed for my interdisciplinary studies. I thank my dear co-authors – Jaroslaw, Olesja, Jörg, Marius and Dirk for being such a great team, I could always rely on and have fun with. I thank Sabrina and Sumaya – my dearest friends, who listened to my complaints in those frustrating moments, everyone writing a PhD knows too well. And last but not least, I thank Michael – my partner, who went with me through the PhD years, with all their ups and downs, being there for me every moment of it. Finally, I thank all those, who gave me inspiration, room and opportunity to follow my passion and complete this dissertation.

# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	Scope and Context of the Studies . . . . .	3
1.2	Methodological Choices . . . . .	4
1.3	Summary of the Studies . . . . .	6
1.4	Co-authorship and Individual contributions . . . . .	10
<b>2</b>	<b>The role of response time for algorithm aversion in fast &amp; slow thinking tasks</b>	<b>11</b>
2.1	Introduction . . . . .	13
2.2	Theoretical Framework . . . . .	14
2.3	Research Question and Hypotheses . . . . .	17
2.4	Experimental Design . . . . .	18
2.5	Results . . . . .	21
2.6	Discussion . . . . .	27
2.7	Conclusion . . . . .	30
<b>3</b>	<b>Does the involvement of domain experts in the AI training affect their AI perception and AI adherence? An experiment in the context of industrial AutoML applications</b>	<b>31</b>
3.1	Introduction . . . . .	34
3.2	Method . . . . .	37
3.2.1	Experimental design . . . . .	37
3.2.2	Experimental procedure . . . . .	42
3.3	Experimental results . . . . .	44
3.3.1	Demographic and descriptive statistics . . . . .	44
3.3.2	Main analysis . . . . .	46
3.3.3	Supplementary analysis . . . . .	50

3.4	Discussion and conclusion . . . . .	52
<b>4</b>	<b>Domain knowledge acquisition for machine learning using ChatGPT</b>	<b>56</b>
4.1	Introduction . . . . .	58
4.2	Related work . . . . .	60
4.2.1	Relevant Scientific Frameworks . . . . .	60
4.2.2	Exemplary Knowledge Acquisition Methods . . . . .	62
4.2.3	Conversational Agents and ChatGPT . . . . .	63
4.3	Research Objectives and Contribution . . . . .	64
4.4	Design & Development . . . . .	66
4.5	Method & Prototype . . . . .	68
4.5.1	Method for Domain Knowledge Acquisition with ChatGPT . . .	68
4.5.2	ChatGPT Interviewer . . . . .	72
4.6	Prototype Demonstration . . . . .	74
4.7	Prototype Evaluation . . . . .	76
4.7.1	RO1: Automation . . . . .	76
4.7.2	RO2: Efficiency . . . . .	78
4.7.3	RO3: Effectiveness . . . . .	78
4.7.4	RO4: Usability . . . . .	79
4.7.5	RO5: Human-centricity . . . . .	80
4.8	Discussion . . . . .	81
4.9	Conclusion . . . . .	82
<b>5</b>	<b>Conclusion</b>	<b>85</b>
	<b>Bibliography</b>	<b>89</b>
	<b>Appendices</b>	<b>98</b>

# **Chapter 1**

## **Introduction**

## INTRODUCTION

The present dissertation has been developed cumulatively and consists of three independent interdisciplinary studies. Independence implies that each study can be considered separately – with its specific context, research question, method and empirical results. Nevertheless, all three studies shed their spotlights on a different sides of the same phenomenon – the human-AI interaction. Thus, together they are bound by the same purpose – to raise the effectiveness of human-AI interaction and to remove hurdles which prevent users and organisations from unleashing the entire potential of AI. To follow this purpose, our studies are bound to be interdisciplinary. It lies in the nature of human-AI interaction – the “human” part requires insights from behavioural and psychological sciences, while the “AI” part refers to the fields of information systems research and computer science. Followingly, in our work we do not only refer to interdisciplinary streams of literature, but also adopt methods from different research fields – the incentivised laboratory experiment from the field of behavioural economics and the Design Science Research [63] methodology from the field of information systems research. The application of either of the two methods depends on the research question we pursue. The main purpose behind the studies in Chapter 2 and Chapter 3 is to uncover influential factors behind AI aversion and AI adherence respectively. Hence, for this purpose we choose the methodology of incentivised laboratory experiment, since it allows us to reliably measure if the change of a particular influential factor leads to a change in AI aversion or AI adherence. On the other hand, in Chapter 4, we do not pursue a purely theoretical inquiry, but rather propose a new method for effective human-AI interaction in a specific context. In this case, our choice falls on the Design Science Research [63] methodology, since it allows for a scientifically grounded design of new artifacts. Additionally to applying different methods, our dissertation features various contexts. While our study in Chapter 2 is designed to be context-agnostic and seeks high results generalizability, the studies in Chapters 3 and 4 focus on the challenges of human-AI interaction in industrial applications and explicitly seek to facilitate the adoption of AI technology by industrial actors.

In the rest of this introductory Chapter, we discuss the scope of the present dissertation and our methodological choices in more detail. Moreover, we provide a short summary and an overview of co-authors’ contributions for the underlying studies.

## 1.1 Scope and Context of the Studies

All three studies of the present dissertation are bound by the same fascination for the AI technology and its potential to improve our lives in plentiful applications from health-care over policy-making to mobility and manufacturing [54]. However, the realization of the AI potential can be mitigated by various hurdles, among those a biased human behaviour towards AI [53, 46] or high complexity and resource-intensiveness of AI model development and implementation [6, 57, 73]. Besides, potential risks of AI give rise to plentiful safety and ethical concerns, resulting in a call for the Human Centered AI [72]. In our dissertation, we pursue an overarching vision of facilitating the application of the AI technology, while placing the needs and well-being of human users in the center of our considerations, as required by the idea of the Human Centered AI [72]. Thus we are interested in deeper understanding of human behaviour and needs in interactions with AI and infer how AI systems and applications can be changed to accommodate those behavioural aspects and needs.

In the first two studies we are explicitly interested in behavioural biases accompanying the human-AI interaction, in particular the irrational avoidance of algorithm advice known as algorithm aversion [53, 46]. Our scientific inquiry seeks to shed light on the influential factors behind algorithm aversion and to make suggestions on how algorithm aversion can be mitigated. Thus, our first study (see Chapter 2) provides experimental evidence that using longer response times for an AI agent can lower algorithm aversion. Besides, the type of thinking applied to a task at hand – whether a user is thinking fast or slow [16] – can influence users’ sensitivity to other factors such as task domain. In our second study (see Chapter 3) we focus on expert users and investigate, how the user involvement during the AI training process influences algorithm aversion. We demonstrate that higher user involvement is associated with a higher perceived understanding of AI functionality and higher perceived control over the resulting model. However, these perceptions are not reflected in the actual behaviour, suggesting that the users’ self-perception as experts overrules the potential positive effects of user participation. In the third study, we turn our attention to real-world problems of AI applications and how they can be mitigated through effective and user-centric human-AI interaction. In particular, we design and prototype an artifact, which allows to automate time-consuming domain knowledge acquisition tasks during the AI model building process. On the one hand, this automation substantially reduces the time needed to acquire domain knowledge, thus lowering the costs for AI model

development and the hurdles for AI adoption. On the other hand, our artifact follows the design objectives of usability and human-centricity, thus bringing additional value to the users – a notion confirmed in the prototype evaluation. The section 1.3 provides detailed summary of each study and its contribution to our stated vision.

While our motivation to facilitate the adoption of AI technology is of generic nature, the bigger part of our dissertation (Chapters 3 and 4) deals with AI applications in industrial context. This decision grounds in our personal motivation – AI can make manufacturing processes more efficient and environmentally-friendly, enhance the product quality and boost innovations. In the same time, industry faces hurdles when implementing AI solutions, among those data availability, integration of domain knowledge, limited access to data science resources, safety concerns, user acceptance and many more. By putting industrial applications in the focus of our work, we strengthen its external validity and provide context for its theoretical and practical implications.

## **1.2 Methodological Choices**

The choice of methods in our work is grounded in the research questions we address. In the first two studies (see Chapter 2 and 3) we are interested in influential factors behind advice taking behaviour in human-AI interactions. Moreover, we are primarily interested in actual behaviour and not in behavioural intentions or attitudes. Thus we refer to the methods of behavioural sciences and choose laboratory experiment as a fitting method for our inquiry. It allows us to control the experimental environment and thus ensure the *ceteris paribus* condition between the treatment groups. Further, we recruit subjects via the online recruiting system ORSEE [36] and randomly assign them to the treatments. By altering the influential factors between the treatment groups we then can make causal inferences about the relationship between those factors and the advice taking behaviour. We note that conducting a field experiment or applying MTurk or Prolific as platforms for an online experiment would increase the external validity of our results, but weaken the *ceteris paribus* condition. Since our inquiry in the first and the second study primarily seeks to empirically test the existence of a casual relationship between the variables of interest, we deem internal validity to be our priority and therefore choose the laboratory environment over other options. Moreover, by creating an experimental scenario, we can measure the actual decision making on the participants' side – if the participants actually follow the AI advice or not. We implement monetary incentives in order to



further enhance the internal validity of our study and to elicit subjects' true preferences [26]. Within the scope of our inquiry, we deem it to be an advantage over alternative methods, such as a questionnaire, which measure reported intentions and attitudes. In Chapters 2.4 and 3.2 respectively we provide further details on the experimental design, procedure and methodological limitations of the first two studies.

In our third study (see Chapter 4) we seek to both produce theoretical insight in the field of human-AI interaction and to solve a real-world problem through creation and application of an artifact. For this purpose, we refer to the design science paradigm [40]. It enables researchers to build upon the existing theory to create artifacts, which solve real-world problems. By applying and evaluating such artifacts, researchers can subsequently generate new theoretical insights and thus add to the theoretical base. The design science follows guidelines which ensure that an artifact addresses a relevant business problem and its design and evaluation are conducted rigorously [40]. To ensure the adherence to these guidelines and to increase the internal validity of our third study, we follow the Design Science Research methodology by Peffers et al. [63]. Thus, we (1) identify a relevant real-world problem; (2) derive design objectives based on the problem, its context and the existing theory; (3) design an artifact following the design objectives and instantiate it in a prototype; (4) apply the prototype in a real-world context and (5) evaluate the application results. Based on the evaluation results, we conclude about the ability of the artifact to solve the initial problem and derive insights which add to the theoretical base. Chapter 4 is structured based on these methodological steps, each step being explained in detail in its respective section.

According to Hevner et al. [40], behavioural and design sciences are like the two sides of the same coin, representing distinct but complementary paradigms which drive scientific knowledge generation. While behavioural sciences seek to develop fundamental theories to explain different phenomena, design science aims at applying those theories to real-world problems and generating knowledge out of the application [40]. Thus, we argue that the methods we apply in our work complement each other and allow us to generate both internally valid and externally relevant results – for the scientific community and practitioners alike.

## 1.3 Summary of the Studies

### **The role of response time for algorithm aversion in fast & slow thinking tasks (Chapter 2)**

Our motivation for the first study grounds in the research on algorithm aversion and appreciation. Both are forms of biased behaviour towards algorithms, when people either irrationally discount or appreciate algorithmic advice [53, 46]. We aim to explore influential factors behind algorithm aversion or appreciation. For this purpose, we bring together several streams of research. First, we build upon the proposition by Bonnefon et al. [11] to adopt the basic concepts of the Dual Process Theory [16] to the studies of human-AI interaction. In this regard we define two types of tasks – (1) fast thinking and (2) slow thinking – depending on the way how people approach the task – either with intuition or logic, respectively. Bonnefon et al. [11] argue that people seem to trust machines with tasks which require slow thinking rather than those requiring fast thinking. This proposition is supported by studies on effects of task type on human-AI interaction [13, 49]. Though the machines do not actually “think” fast or slow, users may consciously or unconsciously make inferences on whether AI is “thinking” fast or slow and if they can trust it with fast or slow tasks accordingly [11, 12, 66]. The response time of an algorithm can serve as signal for such inferences, long and short response times suggesting slow and fast “thinking” on the algorithm side respectively [11]. Thus we further refer to the literature on the effect of response time on algorithm aversion, which have been so far inconclusive – while Efendić et al. [24] find that long response times lead to higher algorithm aversion, Park et al. [61] report an opposite effect. We argue that this inconclusiveness might be attributed to the fact that the authors applied different types of tasks – while Efendić et al. [24] apply tasks we describe as slow thinking, Park et al. [61] select a fast thinking setting. Therefore we propose to study the effect of response time on algorithm aversion in light of the task at hand – fast or slow thinking – and refer to the previous notion that people may use response time to make inferences about the algorithm type of “thinking” and whether it fits the task at hand. By bringing these two streams of research together and testing our proposition empirically, we aim at gaining a more differentiated view on the influential factors behind algorithm aversion. We hypothesize that for slow thinking tasks, algorithm aversion would be higher for longer response times, while for fast thinking tasks it would be lower for longer response times. To test our hypotheses we conduct a 2x2 incentivised laboratory experiment. We manipulate the length of the AI response

time (short vs. long) and the task type (fast thinking vs. slow thinking). Additionally, we also vary the domain of the tasks. We find out that, surprisingly, long response times are associated with lower algorithm aversion both for fast and slow thinking tasks, though the effect in the fast thinking group is stronger. Moreover, we find differences in subjects' behaviour towards AI depending on their type of thinking – subjects in the fast thinking group demonstrate significant differences in advice taking depending on the task domain, while in the slow thinking group these differences disappear. We conclude that long response times can reduce algorithm aversion and that the way, how people approach the task, may make their advice taking behaviour more or less responsive to changes of the task domain. With our study, we are, to the best of our knowledge, the first to empirically test the proposition by Bonnefon et al. [11] to apply the widely known concepts of fast and slow thinking [16] to study algorithm aversion. Moreover, we are the first to provide an empirical insight on the relationship between the type of thinking and the domain of the task in context of algorithm aversion.

### **Does the involvement of domain experts in the AI training affect their AI perception and AI adherence? An experiment in the context of industrial AutoML applications (Chapter 3)**

In our second study we again consider the phenomenon of algorithm aversion and appreciation [53, 46], but this time in the context of industrial applications of Automated Machine Learning (AutoML). On the one hand, AutoML enables domain experts – users without data science skills, but with expert domain knowledge – to build machine learning models for their specific domain applications and thus lowers the financial and organisational hurdles of adopting ML technology in organisations [47, 73]. On the other hand, from the users' perspective, AutoML represents a black box process yielding a black box solution – a fact which can mitigate users' trust [82]. The trust of AutoML users might be also negatively affected by their expert nature – experts have been demonstrated to rely less on algorithms compared to lay people, even at the cost of own decision accuracy [52]. Since lack of trust can lead to inefficient decisions, we are interested in factors, which can mitigate algorithm aversion and foster trust in AutoML. Based on the notion that user participation positively affects user perceptions of a system [39] and other theoretical insights, we argue that involvement of domain experts during the model training might positively affect their perceptions and trust towards the resulting AutoML model.

We hypothesize that domain experts who have been involved in the model training are more likely to adhere to the advice of the resulting model and that their grade of adherence grows proportionally to the grade of their involvement. To test our hypothesis we conduct an incentivised laboratory experiment with three treatment groups – zero, passive and active involvement. Our experimental design mimics a real-world predictive maintenance situation in a production context. Our results demonstrate that higher grade of involvement is associated with a higher perceived influence on the AutoML model and a higher perceived understanding of the model functionality. However, this differences in perceptions between the treatment groups seem to have no effect on the actual behaviour – subjects in all treatments groups adhere to the AI advice similarly often. This result can be explained by the fact that subjects perceive themselves as experts and this self-perception might overrule the treatment effect. Last but not least, we find that subjects perceive AI training to be time well spent, feel comfortably when sharing knowledge with the AI and even enjoy the training process. With our study we are, to the best of our knowledge, the first to empirically test the effects of domain expert involvement in model training on their subsequent model advice adherence. We develop and implement a complex experimental setting which can be adopted by other researches when studying behavioural aspects of AI adherence in industrial context. Besides, we contribute to a discussion on algorithm aversion among expert users – our results show that self-perceptions about own expertise can overrule positive change in attitudes associated with higher involvement. Last but not least, our study carries valuable insights for practitioners – by involving domain experts in the model training they can enhance users’ understanding of the model functionality and even make user experience more enjoyable. To the best of our knowledge, we are the first to experimentally study the influence of user involvement in the model training on the AI advice adherence.

## **Domain knowledge acquisition for machine learning using ChatGPT (Chapter 4)**

Differently to our previous experimental studies, in our third study we are interested in solving a specific real-world problem with AI technology, while ensuring an effective human-AI interaction by design. Further, we seek to generate a theoretical insight in the field of human-AI interaction and to contribute to existing scientific frameworks by applying our solution it in a practical scenario. The Design Science Research [63] builds

the methodological foundation of our inquiry. We focus on a real-world problem of domain knowledge acquisition for machine learning. In machine learning projects, acquiring knowledge often demands significant time and effort from data scientists and other stakeholders. We propose to apply large language models in general, and ChatGPT in particular, to automate time-intensive knowledge acquisition tasks, thereby freeing a substantial share of data scientists’ capacities. Following the Design Science Research [63] methodology, we propose a novel method for automatic interview-based knowledge acquisition and align it with existing scientific frameworks. We then implement this proposed method in a prototype — the ChatGPT Interviewer. We demonstrate and evaluate the ChatGPT Interviewer in two distinct industrial applications. The evaluation confirms that ChatGPT Interviewer is able to interview domain experts in natural language, track their answers and bring their knowledge in a formalized form. Moreover, it can flexibly adopt to the application domain and to the individual user behaviour. Besides, ChatGPT Interviewer can significantly reduce knowledge acquisition effort and deliver valuable input for model building. Last but not least, the users of the ChatGPT Interviewer appreciate their interaction with it, considering it both simple and valuable, they feel in control of the interview process and its results. However, the precision and comprehensiveness of the acquired knowledge can be further enhanced. We conclude that our method can be used to automate the knowledge acquisition steps, resulting in significant efficiency gains for machine learning practitioners. Additionally, our work makes several theoretical contributions to the literature on Informed Machine Learning [77], the Cross Industry Standard Process for Data Mining [79], ChatGPT-based conversational agents and knowledge acquisition methods.

## 1.4 Co-authorship and Individual contributions

The table below summarizes contributions to the studies composing this dissertation.

Table 1.1: Summary of individual contributions

<b>Chapter 2</b>	<b>Lebedeva, A., Kornowicz, J., Lammert, O., Papenkordt, J. (2023)</b> <b>The role of response time for algorithm aversion in fast &amp; slow thinking tasks</b>
	Lebedeva, A. (25%), Kornowicz, J. (25%), Lammert, O. (25%), Papenkordt, J. (25%)
Individual contributions	<b>All authors contributed equally to the following tasks:</b> Identification of the research gap, derivation of the research question, theoretical classification, method selection, experiment design, data collection, data cleansing, implementation of the analysis, interpretation and discussion of the results, preparation and finalization of the manuscript
Conferences & Publications	This work was presented at the International Conference on Human-Computer Interaction 2023 with publication in proceedings.
<b>Chapter 3</b>	<b>Lebedeva, A., Protte, M., van Straaten, D., Fahr, R. (2024)</b> <b>Does the involvement of domain experts in the AI training affect their AI perception and AI adherence? An experiment in the context of industrial AutoML applications</b>
	Lebedeva, A. (35%), Protte, M. (35%), van Straaten, D. (15%), Fahr, R. (15%)
Individual contributions	<b>All authors contributed to the following tasks:</b> Identification of the research gap, derivation of the research question, theoretical classification, experiment design, method selection; <b>Lebedeva A. and Protte M. additionally contributed to:</b> data collection, data cleansing, implementation of the analysis, interpretation and discussion of the results, preparation and finalization of the manuscript
Conferences & Publications	This work was presented at the Future of Information and Communication Conference 2024 with publication in proceedings.
<b>Chapter 4</b>	<b>Lebedeva, A. (2024)</b> <b>Domain knowledge acquisition for machine learning using ChatGPT</b>
Individual contributions	Lebedeva, A. (100%) This paper is single-authored. All tasks were performed by the author.
Conferences & Publications	This work was submitted for publication in the Multimedia Tools and Applications journal

## **Chapter 2**

### **The role of response time for algorithm aversion in fast & slow thinking tasks**

# THE ROLE OF RESPONSE TIME FOR ALGORITHM AVERSION IN FAST & SLOW THINKING TASKS

ANASTASIA LEBEDEVA<sup>\*†</sup>, JAROSLAW KORNOWICZ<sup>\*</sup>, OLESJA LAMMERT<sup>\*</sup>,  
JÖRG PAPENKORDT<sup>\*</sup>

## Abstract

Artificial intelligence (AI) outperforms humans in plentiful domains. Despite security and ethical concerns, AI is expected to provide crucial improvements on both personal and societal levels. However, algorithm aversion is known to reduce the effectiveness of human-AI interaction and diminish the potential benefits of AI. In this paper, we build upon the Dual Process Theory [16] and investigate the effect of the AI response time on algorithm aversion for fast and slow thinking tasks. To answer our research question, we conduct a 2x2 incentivised laboratory experiment with 116 students in an advice-taking setting. We manipulate the length of the AI response time (short vs. long) and the task type (fast thinking vs. slow thinking). Additional to these treatments, we vary the domain of the task. Our results demonstrate that long response times are associated with lower algorithm aversion, both when subjects think fast and slow. Moreover, when subjects are thinking fast, we find significant differences in algorithm aversion between the task domains.

---

<sup>\*</sup>Paderborn University, Warburgerstr. 100, 33098 Paderborn, Germany

<sup>†</sup>Corresponding author, lebedeva@mail.uni-paderborn.de

This research was published in the proceedings of the International Conference on Human-Computer Interaction 2023, DOI: [https://doi.org/10.1007/978-3-031-35891-3\\_9](https://doi.org/10.1007/978-3-031-35891-3_9). The present version deviates from the published version in minor textual changes.

This research was funded by the German Federal Ministry of Education and Research (BMBF) within the “The Future of Value Creation – Research on Production, Services and Work” program (02L19C115). Olesja Lammert and Jaroslaw Kornowicz acknowledge funding by the Deutsche Forschungsgemeinschaft (TRR 318/1 2021 - 438445824). The authors thank Kirsten Thommes and René Fahr for valuable discussion and constructive comments. All errors remain our own.



## Chapter 2 Algorithm Aversion in Fast and Slow Thinking

### 2.1 Introduction

AI is designed to provide crucial improvements to healthcare, mobility, policy-making, manufacturing, and countless other domains [54]. A growing number of political as well as private decisions are being made based on algorithm recommendations [3, 64]. However, prejudice and biased behaviour toward AI often mitigate its potential as extensive research demonstrates [46, 53]. The study of biased human behaviour towards algorithms is dominated by two streams of research – algorithm aversion and algorithm appreciation [44]. Algorithm aversion describes a general rejection of algorithm advice in favor of human advice [53]. For instance, even though AI algorithms have been repeatedly proven to be more accurate in their predictions than human experts [22, 46], humans still exhibit irrational aversion towards AI [13, 80]. When the behavioural bias leans in the opposite direction, researchers speak of algorithm appreciation, which is the logical counterpart to algorithm aversion [46]. For instance, it has been demonstrated that people prefer AI recommendation over human advice in multifaceted situations, such as estimating weights, predicting music charts, or national security concerns [44, 52]. Based on the inverse definitions of the two phenomena, we consider them to represent “two sides of the same medal” [46]. Therefore, in this paper, we use only one of the two terms – algorithm aversion – to describe the entire range of human reactions to AI recommendations. Our study aims to contribute to a deeper understanding of the influential factors that may trigger algorithm aversion or appreciation. So, we build on the theoretical considerations of Bonnefon et al. [11] and are the first to experimentally investigate whether the Dual Process Theory can serve as a tool and as a perspective to study human behaviour towards AI.

The remainder of this study is structured as follows: First, we describe the theoretical approach to adopting the Dual Process Theory in the context of human-AI interaction. Subsequently, we present our methodology and then outline the results of the data analysis. In particular, we examine the effects of AI response time on the advice-taking index for fast and slow thinking tasks in three different domains. Lastly, we discuss our main findings in light of previous literature, point out possible limitations of our study and state our contribution to existing research.

### 2.2 Theoretical Framework

The search for influential factors behind human behaviour towards AI has yielded a considerable number of studies and insights [25, 35]. In their systematic literature review, Mahmud et al. [53] distinguish between individual, algorithm, task, and high-level factors. For example, individual characteristics, like personality traits [56, 71], and characteristics of the AI agent, such as its performance [22] or the explainability of the AI recommendation [1, 58, 69, 78], have been found to affect algorithm aversion. Also, contextual factors like task type or domain have been identified as factors influencing the rate of acceptance or rejection of an AI recommendation [19, 30, 38]. However, Mahmud et al. [53] emphasize that a unified theoretical framework that would comprehensively explain the nature of algorithm aversion is still lacking. One interesting approach to shed light on fundamental principles behind algorithm aversion is provided by the well-known work “Machine Behaviour” by Rahwan et al. [65]. They suggest that concepts, methods, and frameworks from social and behavioural sciences may be adapted to study machines and human-machine interactions. The idea of humans transferring human cognition to machines is not entirely new. In the field of explainable AI, for example, a large number of researchers argue that explanations for AI recommendations should be formulated verbally in a human-like manner, enabling users to construct a correct mental model of the system [18, 58]. Following Rahwan et al. [65], Bonnefon et al. [11] propose to adopt the widely known Dual Process Theory [16] and its concepts of fast and slow thinking (or System 1 and System 2, respectively) as a framework and a tool to study human-AI interaction. According to Kahneman [16], mental life can be characterized as a dynamic between two agents, System 1 and System 2, which produce fast and slow thinking, respectively. System 1 operates automatically and quickly, with little or no effort, relying on impressions, intuitions, intentions, and feelings. System 2, on the other hand, directs attention to effortful mental activities that require rules and explicit thinking, e.g., complex calculations. Besides the scientific Dual Process Theory, there exists a popular “folk theory” concerning fast and slow thinking, reflecting people’s beliefs about their own and others’ thinking processes [11]. Humans seem to consciously or unconsciously apply such “folk theory” of fast and slow thinking to explain human behaviour in everyday life. Some researchers argue that, in human-AI interactions, people might adhere to similar mechanisms to understand the behaviour of an AI agent [11, 12, 66]. In other words, humans are likely to project their beliefs about fast and slow thinking onto intelligent ma-

## Chapter 2 Algorithm Aversion in Fast and Slow Thinking

chines and try to interpret their actions accordingly. Bonnefon et al. [11] propose that humans make inferences about AI if it “thinks” fast or slow and use these inferences to assess whether AI “thinking” fits the task at hand. Alternatively stated, if people perceive an AI agent to “think” slowly, they would rather trust it with tasks that—from a human perspective—require slow thinking (e.g., logic) than with tasks that require fast thinking (e.g., intuition) [11].

This proposition is supported by several empirical studies regarding the effects of task type or task domain on human-AI interaction. However, prior to Bonnefon et al. [11], such results have not been explicitly linked to the Dual Process Theory. For instance, Lee [49] states that algorithmic and human decisions are perceived as equally trustworthy for tasks requiring “mechanical skill” (e.g., work scheduling), whereas algorithms are perceived as less trustworthy than humans for tasks requiring “human skill” (e.g., hiring). The study by Castelo et al. [13] provides similar results. The authors focus on the perceived objectivity of a task, describing an objective task as one based on measurable characteristics and requiring analytical skills (e.g., weather forecasting), and a subjective task as one that required intuition or a “gut feeling” (e.g., predicting the wittiness of jokes). Their results demonstrate once again that people prefer algorithms for objective tasks and reject them for subjective ones [13]. Generally, it seems that humans are more likely to reject an AI recommendation in tasks that, in their perception, require intuition and “human skill” even though research has revealed that even in supposedly more subjective tasks (e.g., suggesting jokes), an algorithm performs better than humans [80]. The link between these findings and the Dual Process Theory made by Bonnefon et al. [11] offers a new perspective on the question of how people perceive AI in different situations, i.e., for different tasks at hand. This perspective might offer a further understanding of how algorithm aversion can be mitigated, especially in the case of tasks perceived to require intuition. In our study, we adopt the proposition of Bonnefon et al. [11] to define different types of tasks. Specifically, we define fast thinking and slow thinking tasks through the approach people selected to solve the task — fast or slow thinking, respectively. While our definition is related to those of Castelo et al. [13] and Lee [49], it exists independently from task objectivity and task domain. For example, in the recruiting domain, if a subject makes her decision based on explicit, rule-based thinking, we define this as a slow thinking task. In the same domain, if a subject decides to rely primarily on her intuition, we define it as a fast thinking task. Additionally, we deem it irrelevant for our definition whether an objectively correct answer to the task (e.g., a calculation result) exists or

## Chapter 2 Algorithm Aversion in Fast and Slow Thinking

whether the answer is entirely subjective (e.g., a job candidate should be declined). In our definition, solely the approach chosen to solve the task determines if it is a fast thinking or a slow thinking task.

Furthermore, Bonnefon et al. [11] postulate that it is not only relevant to determine how people themselves approach a task but also how they perceive the algorithm – whether they judge it to be “thinking” fast or slow. While algorithms “think” neither fast nor slow like humans do [11, 12], they might transmit signals that enable people to make conscious or unconscious inferences about the algorithm type of “thinking.” The length of the algorithm response time is one possible signal. Consequently, its manipulation might influence human perception of an algorithm by suggesting fast or slow “thinking” [11, 61]. Efendić et al. [24] demonstrate that, for analytical tasks, people are more averse to algorithms when response times are longer. This result is contrary to inter-human interactions, where longer response times are usually associated with higher trust – answers following a longer response time are considered well thought-through. The authors attribute this contrasting effect to the fact that people perceive analytical tasks to be easy for algorithms and therefore interpret longer response times as a malfunction [24]. The study by Park et al. [61] examines the impact of response time on the acceptance of algorithm recommendations, additionally distinguishing between high- and low-accuracy algorithms. They find that, in the case of a high-accuracy algorithm, participants are more likely to follow its recommendations when response times are long. For a low-accuracy algorithm, participants are slightly more likely to follow recommendations when response times are short [61]. An apparent inconclusiveness between the results of Park et al. [61] and of Efendić et al. [24] could be explained by the fact that Park et al. [61] use a different task type than Efendić et al. [24]. While the former applies a setting we define as a fast thinking task, the latter selects a task, requiring analytical skills (a slow thinking task in our definition). Considering the results of Castelo et al. [13] and Lee [49] on the influence of task type on algorithm aversion, one might suggest that the effect of the AI response time varies for different types of tasks. Particularly, the result of Efendić et al. [24] might hold for analytical tasks (in our definition, for slow thinking tasks), while the result of [61] might be valid for fast thinking tasks. Therefore, studying the influence of AI response time on algorithm aversion for fast and slow thinking tasks might yield a more differentiated view of how algorithm aversion can be reduced, especially for tasks requiring intuition. In our work, we relate existing results on task types and response times to Dual Process Theory, as proposed by Bonnefon et al. [11], and design a behavioural

## Chapter 2 Algorithm Aversion in Fast and Slow Thinking

experiment to empirically validate possible implications.

### 2.3 Research Question and Hypotheses

So far, we have discussed two factors of interest for our investigation – the task type (fast thinking vs. slow thinking) and the AI response time (short vs. long). From the perspective of a software designer, the former is rather difficult to influence, whereas the response time is relatively simple to control. Therefore, we construct our research question and hypotheses with a primary focus on the AI response time, taking the task type as a secondary contextual factor. We ground our hypotheses in the Dual Process Theory [16] and previous work on algorithm aversion.

We hypothesize that for tasks that are approached with logic – slow thinking tasks – people expect the AI advisor to have short response times because such tasks are perceived to be easy for algorithms. Consequently, for slow thinking tasks, we expect the algorithm aversion to be higher for long response times. This result would be in accordance with Efendić et al. [24]. Castelo et al. [13] and Lee [49] showed that people perceive algorithms as being unable or less capable of solving tasks that require human intuition. Consequently, we hypothesize that people would perceive fast thinking tasks to be difficult for the AI advisor and to require additional “thinking” on the AI side. Therefore, we propose that, for fast thinking tasks, longer response times will be associated with lower algorithm aversion. This result would be in accordance with Park et al. [61].

Explicitly, we pose the following research question: Which effect does the AI response time have on algorithm aversion for slow thinking and fast thinking tasks? Our research question results in two main hypotheses:

*H1: For slow thinking tasks, the algorithm aversion is higher for a longer response time*

*H2: For fast thinking tasks, the algorithm aversion is lower for a longer response time*

Our study contributes to the research field of algorithm aversion in multiple ways. Firstly, we add to the existing empirical results on the influence of AI response time on algorithm aversion by studying its effects in two different types of tasks. Secondly, we propose and apply an experimental design to empirically test the application of the Dual Process Theory [16] to the study of algorithm aversion. Concerning possible practical applications, the results of our paper shall aid practitioners in gaining a more profound

## Chapter 2 Algorithm Aversion in Fast and Slow Thinking

understanding of the nature of algorithm aversion in the context of different task types. Moreover, it shall offer additional empirical results on how algorithm aversion may be reduced by manipulating AI response times.

### 2.4 Experimental Design

To answer our research question, we conducted a randomized controlled experiment, utilizing a 2x2 between-subjects design with student participants at the Business and Economic Research Laboratory at Paderborn University. Experimental sessions took place in attendance and in a strictly controlled environment to ensure adherence to the *ceteris paribus* condition. The experimental design was implemented as a software program using oTree [14]. The software was administered via a browser on personal computers. Each participant was seated individually on a computer and visually shielded from other participants to ensure decision privacy. Between-subject communication was prohibited.

Seven experimental sessions took place between October and November 2022. A total of 119 subjects participated in the study. Two subjects were dropped from the data set because they answered 50 % or more of comprehension questions incorrectly. Another subject was deleted because the participant was not a student. One task observation was eliminated due to a typing mistake, and another observation was canceled because the first estimation was equal to the advice, following the suggestion of Gino and Moore [33]. Consequently, our final data set comprised a total of 116 participants and 1042 observations. The gender composition is slightly skewed toward female students (58.62%) compared to males (41.4%). The average age of the subjects is 23.3 years ( $SD = 3.6$ ).

The subjects were randomly assigned to one of the four treatment conditions: (1) Thinking Slow & AI Long, (2) Thinking Slow & AI Short, (3) Thinking Fast & AI Long, and (4) Thinking Fast & AI Short. Distribution was even across the four treatment groups. The experimental setting was based on the Judge-Advisor-System (JAS) [10] framework and included 9 rounds. In each round, the participants had to solve an estimation task, designed to encourage either fast or slow thinking. During the task, they were advised by an AI with either a short or long response time. Thus, to ensure equal and controllable AI performance at all tasks, the AI advisor was simulated by a simple algorithm, with the AI advice being randomly set at either 90 % or 110 % of the true value. Subjects were unaware of the AI accuracy and its simulated nature to prevent anchoring effects.

Before the experiment started, subjects received instructions that explained the rules

## Chapter 2 Algorithm Aversion in Fast and Slow Thinking

and the setting of the experiment (see Appendix). Instructions were tailored to the treatments, and subjects' comprehension was tested with follow-up questions. During the experiment, at the beginning of each of the nine rounds, subjects were asked to provide their initial estimate of the solution for the given task. After submitting the initial estimate, subjects received a recommendation from the AI. Subjects then provided their second estimation. Both estimations were rewarded monetarily. Subjects were not informed about the accuracy of their estimates until the end of the experiment to avoid learning effects. The order of tasks was randomized to minimize any possible sequence effects. At the end of the experiment, subjects were asked to participate in a survey that included demographic factors, such as age and gender, and other variables, such as confidence in their estimates and perceived AI recommendation quality [34].

Within the fast thinking group, participants were provided with a picture of an object and asked to estimate some numeric quality of the object shown. Our object selection ensured that all participants were familiar with them. In the absence of any additional information, we, therefore, expected subjects to apply their intuition rather than analytical skills to solve the task. Within the slow thinking group, subjects received additional quantitative information about the object in a textual form. We intended this manipulation to facilitate slow thinking on the subjects' side. Being given quantitative hints, we expected subjects to apply analytical skills and logic rather than intuition to solve the task. It is also worth mentioning that, even with the additional information, the answer to the task could not be estimated with absolute accuracy and remained ambiguous.

The AI response time (the time frame between the submission of the subjects' initial estimate and the display of the AI advice) was set to two seconds in the "AI Short" treatments and ten seconds in the "AI Long" treatments. During this time frame, the task information was not visible to subjects, and a loading bar displayed the simulated progress of the AI (see Figure 2.1). After the response time had elapsed, the task became visible once again. Additionally, the AI advice and subjects' own initial estimate were displayed. Subjects were then asked to submit their second estimate (see Figure 2.2).

Additionally, the objects, whose numeric qualities subjects were asked to estimate, originated from three different domains. We chose to not introduce new domains but use settings that have been previously applied in other studies. In the "Lentils" domain, subjects were asked to estimate the number of chocolate lentils in a glass based on a photograph, following Park et al. [61]. In the slow thinking version of the task, the glass size and the number of lentils in a reference glass were additionally displayed. In the "Foot-

## Chapter 2 Algorithm Aversion in Fast and Slow Thinking

ball” domain, subjects estimated the weight of football players based on a photograph. Additional textual information in the slow thinking group included weight references for other comparable players. A similar design was used by Gino and Moore [33]. In the “Route” domain, subjects estimated the length of a car route between lesser-known German cities based on a map. In the slow thinking group, reference distances were displayed additionally to the map. This last domain was designed in accordance with Hofheinz et al. [42]. Estimation tasks were equally distributed between all three domains.

Task 3 / 9

Artificial intelligence calculates...

53%

How many chocolate lentils are in the glass pictured?



**Your previous estimate and the AI's recommendation**

Your previous estimate is **200 Lentils**.

Artificial intelligence estimates the value at **180 Lentils**.

Please make your second estimation.

Your estimation:

Lentils

**Submit estimation**

Figure 2.1: AI loading bar

Figure 2.2: AI advice

Subjects received a fixed amount of €3 for participating in the experiment. Additionally, they were able to earn a payoff for the accuracy of their estimates. Subjects had to make a total of 18 estimations during the experiment (2 estimations per round across 9 experimental rounds). For each estimation, earnings between €0.00 and €0.50 were possible, resulting in a total maximum reward of €9.00 across all estimations. The closer the subjects’ estimates were to the true value, the more they earned. In order to reward the



## Chapter 2 Algorithm Aversion in Fast and Slow Thinking

timely completion of tasks, we implemented a time pressure condition. The payoff per estimation started to gradually decrease after 45 seconds until it reached zero if participants required more than 5 minutes and 45 seconds for an estimation. Upon completion of the 9 experimental rounds, subjects were informed about their total payoff. In addition to the fixed payment of €3, subjects earned an average additional payoff of €7.39 ( $SD = 0.62$ ) based on the accuracy of their estimations.

To assess algorithm aversion, we measured the degree to which subjects followed the advice of the AI. Specifically, we used the advice-taking index [42] as the dependent variable in our experiment. The index was calculated as follows:

$$\text{Advice-taking index} = \frac{\text{Second Estimation} - \text{First Estimation}}{\text{Advice} - \text{First Estimation}}$$

The index equals 0 when the subject's first and second estimation are identical. The more subjects lean toward the AI advice, the closer the index is to 1. Consequently, at a value of 0.50, the subjects weigh the advice and their first estimation equally. This index is similar to the commonly used measurement "Weight of Advice" in the advice-taking literature [4]. However, the main difference is that the values can be negative. This can occur when a subject decreases the second estimation even though the advice recommends increasing it. Additionally, the index can be above 1 if the subject overshoots the advice, i.e., if the first estimation is 100, the advice is 200, and the second estimation is 300. To ensure accurate results, we follow the procedure of Logg et al. [52] and winsorize the values below 0 and above 1 to 0 and 1, respectively. A higher advice-taking index indicates lower algorithm aversion.

## 2.5 Results

All statistical analyses were conducted using Stata 17.0. The Mann-Whitney U-Test was applied to determine statistical significance. Table 2.1 displays the detailed descriptive statistics for the four treatment groups. Between the gender groups, no significant difference in advice-taking index was detected ( $z = -0.61, p = 0.54$ ). However, there was a significant difference ( $z = -15.08, p = 0.00$ ) in the time needed for the first estimation between the fast thinking ( $M = 17.40$  sec.,  $SD = 8.05$ ) and the slow thinking ( $M = 27.60$  sec.,  $SD = 12.42$ ) treatment groups. We can possibly attribute this to the slow thinking induced by the treatment, as well as to the amount of information that needed to be pro-

## Chapter 2 Algorithm Aversion in Fast and Slow Thinking

cessed by subjects. The time difference disappeared in the second estimation. The fast and slow thinking groups also significantly differed in the accuracy of the first estimate ( $z = -9.60, p = 0.00$ ), with subjects in the fast thinking group underestimating the true value by 18.7 % ( $SD = 27.3$ ), and in the slow thinking group by 3.7 % ( $SD = 26.1$ ) on average. A higher average estimation time and accuracy for the first estimation within the slow thinking groups compared to the fast thinking groups are identified as objective indicators, which can reasonably suggest that participants in the slow thinking treatment actually were thinking slowly, whereas subjects in the fast thinking treatment were thinking fast. Additionally, we control for domain-specific differences in estimation accuracy. In general, the participants seem to struggle particularly with the estimation of the number of lentils regardless of the treatment, whereas the deviations from the true value are lowest for the domain “Football”.

After the descriptive statistics, we now focus on the main part of our analysis. As mentioned before, we employ the advice-taking index as a proxy for algorithm aversion – a higher advice-taking index indicates lower algorithm aversion and vice versa. Further, we assume estimates of each participant in 9 rounds to be independent of each other since the participants did not receive feedback on their performance after a task and the task sequence was randomized. We start by calculating the average advice-taking index per treatment group (see Figure 2.3).

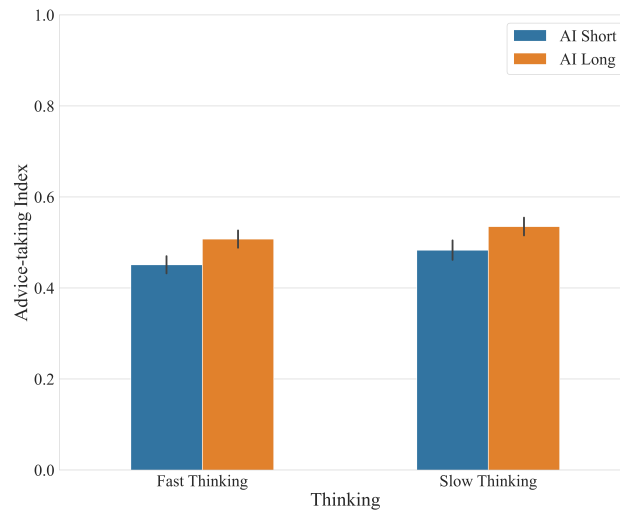


Figure 2.3: Average advice-taking index by treatment. Error bars indicate standard errors.

## Chapter 2 Algorithm Aversion in Fast and Slow Thinking

Table 2.1: Summary descriptive statistics for the N=116 observations

Variable	Estimation	Mean	SD.	Min	Max
Thinking Slow & AI Long (N=29)					
Avg. time (in sec.)	1 <sup>st</sup> estimation	28.41	10.09	5.23	58.77
	2 <sup>nd</sup> estimation	14.57	7.93	4.57	41.18
Average accuracy	1 <sup>st</sup> estimation	-0.01	0.29	-0.73	1.50
	2 <sup>nd</sup> estimation	-0.01	0.15	-0.69	1.07
Thinking Slow & AI Short (N=27)					
Avg. time (in sec.)	1 <sup>st</sup> estimation	26.73	14.47	5.49	166.58
	2 <sup>nd</sup> estimation	11.45	6.94	3.36	40.06
Average accuracy	1 <sup>st</sup> estimation	-0.07	0.23	-0.80	0.50
	2 <sup>nd</sup> estimation	-0.04	0.16	-0.70	0.33
Thinking Fast & AI Long (N=29)					
Avg. time (in sec.)	1 <sup>st</sup> estimation	16.68	6.87	5.18	36.07
	2 <sup>nd</sup> estimation	13.75	7.44	4.97	44.19
Average accuracy	1 <sup>st</sup> estimation	-0.18	0.28	-0.77	0.70
	2 <sup>nd</sup> estimation	-0.11	0.18	-0.73	0.37
Thinking Fast & AI Short (N=31)					
Avg. time (in sec.)	1 <sup>st</sup> estimation	18.07	8.97	4.76	52.14
	2 <sup>nd</sup> estimation	11.59	6.46	4.26	36.33
Average accuracy	1 <sup>st</sup> estimation	-0.19	0.28	-0.82	0.73
	2 <sup>nd</sup> estimation	-0.13	0.20	-0.75	0.38
Total (N=116)					
Avg. time (in sec.)	1 <sup>st</sup> estimation	22.31	11.57	4.76	166.58
	2 <sup>nd</sup> estimation	12.84	7.32	3.36	44.19
Average accuracy	1 <sup>st</sup> estimation	-0.11	0.28	-0.82	1.50
	2 <sup>nd</sup> estimation	-0.07	0.18	-0.75	1.07

On average, the advice-taking index was found to be higher in the groups with long AI response times than in the groups with short AI response times. The highest average advice-taking index was observed in the treatment group “Thinking Slow & AI Long” ( $M = 0.54, SD = 0.31$ ). The lowest average advice-taking index was achieved in the “Thinking Fast & AI Short” treatment group ( $M = 0.45, SD = 0.32$ ). To test our research hypotheses, we conducted non-parametric treatment group comparisons using the Mann-Whitney U-Test. Thus, to investigate our first hypothesis, we compared the advice-taking index of the groups “Thinking Slow & AI Long” and “Thinking Slow & AI Short”. The Mann-Whitney U-Test indicated that the two groups differed significantly from each other. Contrary to our H1 hypothesis, a longer response time in slow thinking

## Chapter 2 Algorithm Aversion in Fast and Slow Thinking

tasks led to a significantly higher advice-taking index ( $z = 1.79, p = 0.07$ ), i.e., to lower algorithm aversion. Similarly, to examine our second hypothesis, we compared the groups “Thinking Fast & AI Long” and “Thinking Fast & AI Short”. We found out that a longer response time led to a significantly higher advice-taking index in the fast thinking tasks ( $z = 2.16, p = 0.03$ ). Therefore, the H2 hypothesis could be confirmed.

To test for consistency in our results, we investigated whether our results concerning the two hypotheses (H1 and H2) hold within single domains. For this purpose, we additionally compared the group “Thinking Slow & AI Long” with the group “Thinking Slow & AI Short” and the group “Thinking Fast & AI Long” with the group “Thinking Fast & AI Short” within each domain separately (see Table 2.2).

Table 2.2: Treatment group comparisons of the advice-taking index within domains

Treatment		Mean	SD.	p-value
Football				
Thinking Slow	AI Long	0.51	0.30	0.57
	AI Short	0.48	0.33	
Thinking Fast	AI Long	0.41	0.28	0.63
	AI Short	0.40	0.30	
Lentils				
Thinking Slow	AI Long	0.56	0.32	0.04**
	AI Short	0.45	0.33	
Thinking Fast	AI Long	0.50	0.32	0.02**
	AI Short	0.38	0.29	
Route				
Thinking Slow	AI Long	0.54	0.32	0.71
	AI Short	0.52	0.33	
Thinking Fast	AI Long	0.62	0.29	0.45
	AI Short	0.58	0.31	

\*\*\*  $p < 0.01$ , \*\*  $p < 0.05$ , \*  $p < 0.1$

Table 2.2 presents the results of the Mann-Whitney U-Test of the different treatment comparisons within the different domains. The tests revealed significant differences only in the domain “Lentils”. Here, the results of the cross-domain analysis were confirmed. Again, longer response times led to a significantly higher advice-taking index, both for fast and slow thinking treatment groups.

Subsequently, we investigated the differences in the advice-taking index between the domains within each treatment group (see Figure 2.4). The graphic vividly illustrates

## Chapter 2 Algorithm Aversion in Fast and Slow Thinking

that the differences are larger among the domains in both fast thinking groups than in the two slow thinking groups. The Mann-Whitney U-Tests between the domains within the “Thinking Fast & AI Long” treatment group confirm that the advice-taking index differs significantly between all three domains. The same applies to the treatment group “Thinking Fast & AI Short”, except for the fact that the advice-taking index does not differ significantly between the domains “Football” and “Lentils”. In contrast, there are no significant domain-specific differences in the average advice-taking index within the two slow thinking treatment groups. Thus, we conclude that the domain-specific effects only matter in fast thinking tasks.

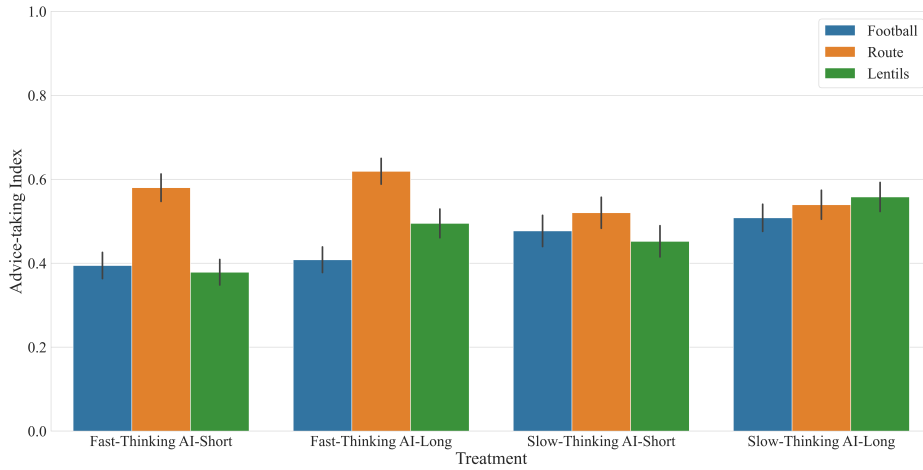


Figure 2.4: Average advice-taking index across domains within treatment groups. Error bars indicate standard errors.

To gain more insight into the underlining mechanisms behind the main treatments, we followed Gino et al. [34] and asked participants to rate the perceived quality of the AI advice and their self-confidence in their estimation. We intended to analyze those data to determine if subjectively perceived confidence in the first estimate and the perceived quality of the AI recommendation impacted the actual adaptation of the advice. Treatment group comparisons in regard to the dependent variables *perceived quality of AI advice* and *perceived confidence in their own estimation* within each domain reveal that these two dependent variables do not differ significantly between the treatments, with one exception – in the “Football” domain the *perceived quality of AI advice* differs significantly between the “Thinking Fast & AI Long” and “Thinking Fast & AI Short” groups. Moreover,

## Chapter 2 Algorithm Aversion in Fast and Slow Thinking

the quality of the AI advice in the groups with short response time is perceived to be tendentially higher, independent of the thinking type and the domains. Furthermore, it can be observed that the fast thinking groups feel more confident in their estimation than the slow thinking groups, independent of the response time and the domains. However, none of these effects becomes significant.

We further investigated whether a difference in the *confidence in their own estimation* (see Figure 2.5) and the *perceived quality of AI advice* (see Figure 2.6) between the domains within each treatment group existed. While in the “Thinking Slow & AI Short” treatment only the “Route” domain differs significantly from the other two domains in terms of perceived confidence of own estimation, in the other three treatment groups the domain “Lentils” significantly deviates from the other two. Consequently, within all treatments, the domains “Lentils” and “Route” always differ significantly from each other. Further, the perceived *confidence in their own estimation* is lowest in each treatment for the domain “Lentils”.

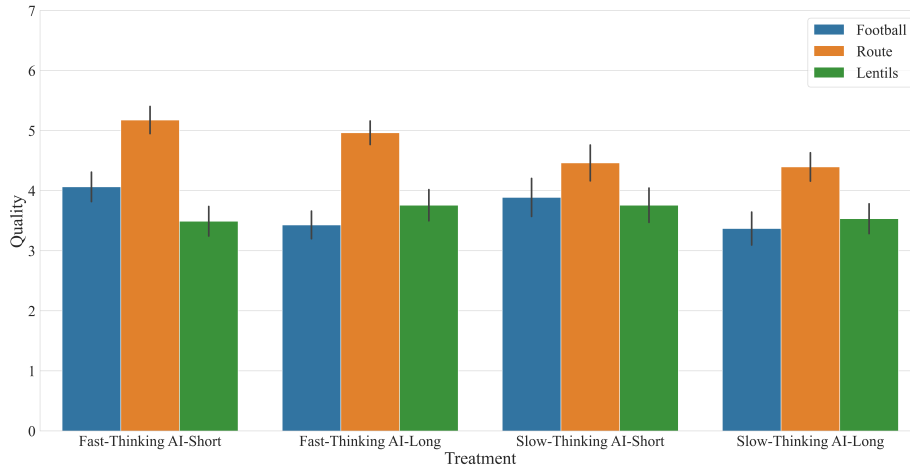


Figure 2.5: Average perceived quality of AI-advice across domains within treatment groups

If we consider the *perceived quality of AI advice* using the Mann-Whitney U-Test, it becomes apparent that, within the treatments, the domain “Route” always differs significantly from the other two domains and the perceived quality is highest in this domain (see Figure 2.6). Only in the group “Thinking Slow & AI Short” the domain “Route” is not significantly different from the domain “Football”. If we relate this to our previous re-

## Chapter 2 Algorithm Aversion in Fast and Slow Thinking

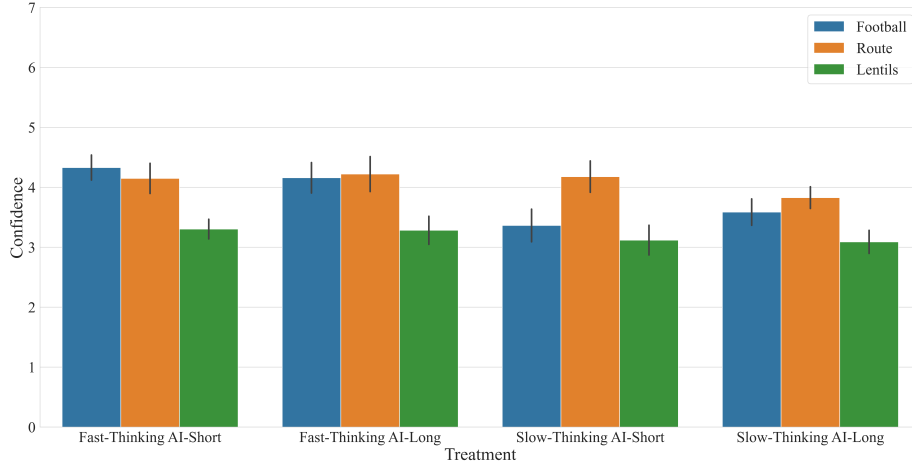


Figure 2.6: Average perceived self-confidence in own estimation across domains within treatment groups

sults, we observe that the advice-taking index for this domain is also significantly highest within the two fast thinking treatment groups (see Figure 2.4). This tendency regarding the advice-taking index can also be observed in the slow thinking treatments. We can attribute the fact that the treatment group comparisons of the *perceived quality of AI advice* and *perceived confidence in their own estimation* do not become significant (as described above) to the fact that the domains within the treatments have very similar effects.

## 2.6 Discussion

As proposed by Bonnefon et al. [11], we apply an experimental design to empirically test the application of the Dual Process Theory [16] as a framework to study algorithm aversion. In the course, we relate existing results on task types and response times to Dual Process Theory [16], derive our hypotheses and test them empirically. Our results suggest that the application of the Dual Process Theory [16] may indeed deepen the understanding of algorithm aversion.

According to our results, in slow thinking tasks, algorithm aversion is lower for longer AI response time – an effect opposite to our first hypothesis (H1). Moreover, our results for fast thinking tasks show that algorithm aversion is lower for longer response times, as we suggest in our second hypothesis (H2). The former result regarding H1 may also be

## Chapter 2 Algorithm Aversion in Fast and Slow Thinking

seen as contradicting Efendić et al. [24], who find that, in analytical tasks, long response time is associated with higher algorithm aversion. The finding concerning H2 is in line with Park et al. [61], who find that algorithm aversion is lower for longer response times in a task design that corresponds to a fast thinking task in our definition.

In our experiment, for both task types, the longer response time is associated with lower algorithm aversion, although the difference in the advice-taking index is stronger for fast thinking tasks. People associate long response times with stronger effort, both in the case of humans and algorithms [24]. Therefore, longer response times might suggest to human agents that AI is exercising a stronger effort for the task, enhancing its task capability. Interestingly, in our experiment, this effect seems to hold for both types of tasks, although previous research demonstrated that people perceive algorithms to be more capable of tasks that require analytical skill (slow thinking) and less capable of tasks that require intuition (fast thinking) [13, 49, 53]. One might attribute this result to the fact that additional quantitative information is insufficient to encourage slow thinking on the participants' side (i.e., participants apply fast thinking independently of task type), especially considering that the type of thinking applied is difficult to measure directly. However, we deem the fact that participants in the slow thinking group spent significantly more time on the tasks and performed significantly better an objective indicator of actual "slow-thinking" in the slow thinking group. Further research may investigate other ways to apply the framework of the Dual Process Theory [16] to task design in the context of algorithm aversion.

The inconclusiveness between our result concerning slow-thinking tasks and the findings of Efendić et al. [24] might be attributed to the fact that Efendić et al. [24] use verbal constructs to describe the length of algorithm response time (e.g., "after a long pause" or "after an extended period of time"). In our experiment, we apply real response times (1 sec. vs. 10 sec.). Thus, a time range could exist within which longer response times reduce algorithm aversion, as is observed in our study and by Park et al. [61]. Therefore, we advocate for more empirical research on the role of different response time ranges on algorithm aversion.

We consider the domain to be essential for the external validity of our results. Research demonstrates that people react differently to algorithmic recommendations in different domains [13, 22, 52]. We test our hypotheses in three different domains to achieve a more differentiated view of the effects of task type and AI response time on algorithm aversion. We select domains that are similar concerning complexity, moral impact, and



## Chapter 2 Algorithm Aversion in Fast and Slow Thinking

subjectivity since these factors are proven to affect algorithm aversion [13, 22, 53]. We investigate whether significant differences arise between domains within each of the main treatments. Results suggest that algorithm aversion varies significantly between the domains within each of the two fast thinking treatments. Within the slow thinking groups, the domains seem to have no effect on algorithm aversion. Notably, within each of the fast thinking groups, the highest advice-taking index is observed for the domain “Route”. In every treatment group, participants rate the AI advice quality in this domain higher than in the domains “Lentils” and “Football”. Only in the group “Thinking Slow & AI Short”, the AI advice quality in the domain “Route” is not significantly different from the domain “Football”. The fact that the AI advice quality in the domain “Route” is perceived to be higher than in other domains could be explained by the fact that people are largely familiar with algorithm recommendations in similar tasks, such as determining an optimal route by using a navigation system. In general, our finding concerning significant differences between task domains for fast thinking groups is in line with the findings of previous research [13, 22, 52].

Interestingly, these effects seem to disappear in the advice-taking index for slow thinking tasks. We propose the following interpretation for this finding. As long as participants have only a picture as a single source of information, the depicted object majorly influences participants’ behaviour. On the contrary, as soon as additional quantitative information is available to participants, the tasks become more comparable to each other (a number has to be estimated based on other numbers), and the origin of the object does not play a significant role anymore. These results are crucial when researchers want to describe different types of tasks. Our findings demonstrate clearly that the domain of a task and the way people approach it are not the same and have different effects on algorithm aversion. Additionally, the fact that we detected said difference only for fast thinking tasks but not for slow thinking tasks indicates that both groups were indeed thinking differently.

A noteworthy limitation of our results is that our observations are limited to the three domains “Lentils”, “Football”, and “Route”, which do not differ in terms of complexity, moral impact, and subjectivity. Further research might investigate whether the same result holds in other domains. Additionally, our results concerning the negative effect of long response time on algorithm aversion (i.e., its positive effect on the advice-taking index) are especially strong in the “Lentils” domain. Interestingly, in all treatment groups, participants’ confidence in their own estimation is lowest for the “Lentils” domain, suggesting that low confidence enhances the effect of longer response time on algorithm aversion.

## Chapter 2 Algorithm Aversion in Fast and Slow Thinking

Mahmud et al. [53] name self-evaluation factors, like self-efficacy, among those influencing algorithm aversion. Logg et al. [52] demonstrate that higher self-efficacy and self-confidence are associated with higher algorithm aversion, supporting our suggestion that low self-confidence could facilitate the effect of response time on algorithm aversion.

### 2.7 Conclusion

Our study yielded several insights that might prove valuable both for further research and for software developers. Firstly, we demonstrated that longer AI response times, in particular a response time of 10 seconds, are associated with lower algorithm aversion. This effect is even stronger for tasks designed to facilitate a fast thinking, intuitive approach. Secondly, among all domains, long response time had the strongest positive effect on the advice-taking index in the domain “Lentils”, in which participants displayed the lowest confidence in their own estimations. Thirdly, within the fast thinking groups, the task domain heavily impacted the advice-taking index, whereas domain differences were not significant within both slow thinking groups.

On the one hand, our results contribute to the research about the influence of response time on algorithm aversion and suggest that, at least to a certain extent, longer response times may be used to reduce algorithm aversion. On the other hand, our results indicate that advice-taking varies depending on people’s approach to the tasks and on the domain of the task. To the best of our knowledge, empirical results testing these differences between the way of thinking and the task domain broach an entirely new subject within the research of algorithm aversion. With our study, we proposed a design and delivered empirical insights following the proposal of Bonnefon et al. [11] to apply Dual Process Theory [16] to studies of human-machine interaction. Therefore, our study can aid other researchers in better understanding the nature of algorithm aversion by considering the thinking used while solving the task and not just the perceived task context.

## **Chapter 3**

**Does the involvement of domain experts in the AI training affect their AI perception and AI adherence? An experiment in the context of industrial AutoML applications**

# DOES THE INVOLVEMENT OF DOMAIN EXPERTS IN THE AI TRAINING AFFECT THEIR AI PERCEPTION AND AI ADHERENCE? AN EXPERIMENT IN THE CONTEXT OF INDUSTRIAL AUTOML APPLICATIONS

ANASTASIA LEBEDEVA<sup>\*</sup>, MARIUS PROTTE<sup>\*†</sup>, DIRK VAN STRAATEN<sup>\*</sup>,  
RENÉ FAHR<sup>\*</sup>

## Abstract

AutoML is a promising field of Machine Learning (ML) in which users can build an intelligent algorithm without any prior knowledge in data science or programming. AutoML is supposed to bring the advantages of artificial intelligence to a wide range of organizations in plentiful domains by automating the process of ML-model creation. However, AutoML often appears to end-users as a black-box model created in a black-box process – a fact which is known to negatively impact trust. Moreover, AutoML end-users are in many cases experts in their respective domains (e.g., physicians, engineers etc., being referred to as “domain experts”). Based on recent scientific insights in the field of algorithm aversion, experts have more difficulties relying on AI recommendations than lay people do – a fact that could additionally mitigate end-users’ trust in AutoML. The lack of trust in AutoML may have high-cost consequences, resulting in inefficient decisions and mitigating the overall progress in the AutoML field. Therefore, we investigate, how adherence of domain experts to AutoML recommendations can be fostered. Involvement of end-users in the product creation process was repeatedly shown to positively affect the end-user attitudes in plentiful domains. In the AutoML context, domain experts can be involved in the AutoML-model creation process by being asked to share their domain knowledge with the model. Though domain expert involvement is costly – requiring time and effort – we argue that this investment is worthy, because it may positively influence their attitudes as end-users and raise their trust in AutoML. In our experimental laboratory study we create an environment, in which subjects act as expert engineers and need to foresee malfunctions in their machine, while being advised by an AutoML-model. We apply three treatments – zero, passive and active involvement – to investigate our hypothesis. We observe that higher involvement leads to a higher perceived influence on the AutoML model and a higher perceived understanding of its functionality. However, these perceptions are not reflected in the actual behaviour – across all groups subjects demonstrate similar rates of AI adherence. We attribute this fact to subjects’ self-perception as

experts, which is equally strong among all groups and might overrule the treatment effects. Last but not least, we find that subjects perceive AI training to be time well spent, feel comfortably when sharing knowledge with the AI and even enjoy the training process.

*Keywords:* algorithm aversion, user involvement, AutoML, lab experiment

*JEL Classification:* C91, D24, L23, O14, O33

---

\*Paderborn University, Heinz-Nixdorf-Institute, Fürstenallee 11, 33102 Paderborn

†Corresponding author, [marius.protte@upb.de](mailto:marius.protte@upb.de)

This research was published in the proceedings of the Future of Information and Communication Conference 2024, DOI: [https://doi.org/10.1007/978-3-031-53960-2\\_13](https://doi.org/10.1007/978-3-031-53960-2_13). The present version deviates from the published version in minor textual changes.

This research was funded by the Deutsche Forschungsgemeinschaft within the “SFB 901: On-The-Fly (OTF) Computing – Individualised IT-Services in Dynamic Markets” program (160364472). The authors thank Dr. Carlos Paiz Gatica for valuable discussion and constructive comments. All errors remain our own.

### 3.1 Introduction

A growing number of organizations are willing to apply Machine Learning (ML) to generate value out of their data. However, building and deploying an ML-Model is generally associated with high investments and requires involvement of data scientists and other ML-professionals, who are rare on the labor market. These challenges prevent medium and small organizations from adopting ML and mitigate the overall spread of this technology. One approach to tackle this problem is known as Automated Machine Learning (AutoML) [73]. AutoML is a paradigm which seeks to automate the complex development process behind Machine Learning and to make it more accessible. The ultimate vision behind this technology is to enable domain experts – users without data-science skills, but with expertise in the application domain – to build and deploy ML-models for their specific real-world problems [47].

Though in the recent years the AutoML technology has demonstrated remarkable progress, it is still a long way till domain experts can autonomously build, apply, and interpret AutoML-models. Currently, from the user perspective, AutoML is a black-box process which leads to a black-box solution, both difficult or even impossible to interpret without data science skills. Such lack of transparency has multiply negative implications, one of them being strong mitigation of users' trust [82]. In general, a phenomenon of discounting an algorithm judgement in favor of the own or other humans' judgement, irrespective of algorithm performance, is known as algorithm aversion [53]. Research has demonstrated that people are averse towards algorithms, even when they know about algorithm superior performance [22]. The trust for AutoML can be additionally negatively impacted by the fact that most AutoML-users are experts in their respective domains – engineers, physicians etc. According to Logg et al. [52], experts rely less on algorithms in their decision making compared to lay users, even at the cost of decision accuracy. Filiz et al. [28] show that, paradoxically, people demonstrate even stronger algorithm aversion when decision stakes are higher – they strongly prefer a less accurate human advice over a more accurate algorithm advice, if a situation has critical impact on others. AutoML as a technology will eventually bring Machine Learning into a broad range of domains, high-risk domains like healthcare or industrial production among them. Consequently, algorithm aversion in AutoML-context may not only reduce users' own utility from AutoML but lead to high-stake failures. Therefore, it is a crucial question, how trust in AutoML can be fostered.

### Chapter 3 Effect of Domain Expert Involvement on AI Adherence

A comprehensive literature review by Jussupow et al. [46] summarizes that human involvement at different stages of algorithm life cycle mitigates algorithm aversion. In a broader scope of information systems research, a meta-analysis of 82 empirical studies by He and King [39] reveals that user participation has a significant positive effect on attitudinal and behavioural outcomes (user satisfaction, use intention and system use). Behavioural sciences also offer support for the notion that co-creation positively influences perception of results. For example, the widely known IKEA-effect, introduced by Norton et al. [59], describes a phenomenon when co-creation leads to higher valuation of products by consumers. Psychological ownership for the result of co-creation is shown to be one of major mechanisms underlying this behaviour [68]. In the AutoML-context, domain experts can act not only as users, but also as model contributors or even creators. Through their contribution, domain experts gain (to some extent) control over the resulting model. Dietvorst et al. [23] show that an opportunity to exhibit control over an algorithm (even slightly) reduces algorithm aversion. Besides, involvement during model training allows domain experts to insert their knowledge into the AutoML-model. Kawaguchi [48] demonstrates that sales workers become more willing to follow algorithmic advice once their expertise is integrated into the algorithm's forecasting. Additionally, by participating in the model creation process, domain experts gain better understanding about functionality and inputs of an AutoML-model. According to Yeomans et al. [80], information about algorithm functionality improves users' understanding of the algorithm and reduces algorithm aversion. Finally, Jago [45] demonstrates that highlighting human involvement in model training can have a positive effect on user perception.

Provided theoretical insights from different scientific fields demonstrate that involvement of domain experts during model creation may have a value of its own: Enable domain experts to influence the model, better understand it, develop a feeling of psychological ownership towards the model and eventually foster trust in AutoML and enhance algorithm adherence. To the best of our knowledge, current AutoML-research treats involvement of domain experts as a technical measure to insert domain-specific information into the model: The value of such involvement is determined by the value of inserted information, its carrier being simply a source. Involving domain experts has a cost of time and effort, so a rational decision maker would compare the cost with the anticipated benefit (e.g., higher accuracy of the model) to decide whether a model-training should happen with or without expert involvement. In general, it is a decision which a domain expert herself might be confronted with – should she invest time, sharing her knowledge or is it

### Chapter 3 Effect of Domain Expert Involvement on AI Adherence

sufficient to train the model based only on data? We argue that involvement of domain experts in model creation has a value beyond the technical improvement of the model and that this value must be scientifically determined in order to be considered by practitioners and decision makers. We make a first step in this direction by empirically investigating the effect of user involvement in AutoML-context.

Explicitly, we pose the following research question: How does involvement of domain experts during model training affect their adherence to model advice during its deployment, given that the model accuracy does not change through involvement? To answer our research question, we state following hypothesis:

*H1a: Expert individuals who were involved in model training are more likely to adhere to the model advice compared to individuals who were not involved.*

*H1b: Higher grade of involvement in model training is associated with higher adherence to model advice.*

Our experimental study is designed as a simplified real-world situation, modelling a predictive maintenance problem in industrial context. Within this scenario, subjects play a role of engineers, responsible for the maintenance of a manufacturing machine. They need to assess the probability of a machine malfunction and take a maintenance decision, while being advised by an AI. During the experiment, subjects gradually acquire scenario-relevant knowledge and become domain experts within the scenario. For the purpose of our inquiry, we vary the grade of subjects' involvement during model training and implement three treatment groups: zero, passive and active involvement. Following our hypothesis, we measure, how often subjects change their decision in favor of the AI advice and compare the advice adherence between the treatment groups. We implement monetary incentives for maintenance performance in order to enhance the internal validity of our study and to elicit subjects' true preferences [26].

With our paper we seek to make several contributions, valuable both for researchers and practitioners. Firstly, we empirically examine the impact of user involvement during model training on algorithm adherence. Secondly, we add to the literature on algorithm adherence among expert users. Thirdly, we add to the literature on AutoML by empirically studying domain experts as contributors and as users of an AutoML-model, holistically addressing behavioural aspects of their role. Practitioners can benefit from our results by reassessing the role of domain experts and developing measures to better



## **Chapter 3 Effect of Domain Expert Involvement on AI Adherence**

incorporate them into the AutoML-process. Additionally, they can directly influence the reliance on algorithm recommendations by varying the grade of domain expert involvement.

### **3.2 Method**

We conduct an incentivised laboratory experiment, modelling a real-world situation on a production plant, in which a decision-maker must weigh a cost of a maintenance break with a cost of a potential machine malfunction. Throughout the experiment, subjects are asked to take a role of an engineer – a domain expert for a manufacturing machine. In their decisions, subjects are confronted with an AI that predicts the probability of machine malfunctions and provides recommendations on whether to conduct a maintenance or not.

#### **3.2.1 Experimental design**

In the beginning of the experiment, subjects are given all necessary knowledge and skill to play their role as a domain expert within the experimental scenario. Further, subjects are being passively, actively, or not involved in the AI training based on the applied treatment. In the main part of the experiment, subjects are asked to assess the likelihood of a fictitious machine breaking down and to decide, whether to maintain the machine or not in multiple rounds. Conducting maintenance eliminates the possibility of a malfunction and costs a half of the payoff. If no maintenance is conducted, subjects either receive a full payoff if no malfunction occurs and a zero payoff, if malfunction takes place. The actual probability of a malfunction depends on parameters, which are not known to subjects but can be assessed based on available data. The design follows the judge-advisor principle – in every round subjects first make their own assessment and maintenance decision, then learn about the AI-assessment and decision recommendation and take a final decision, to maintain a machine or not. To prevent learning effects, subjects receive no feedback in the end of a round and learn their total payoff only after they have completed all experimental rounds. In the end of the experiment, subjects are asked to fill in a questionnaire.

As mentioned above, in advance to the main part of the experiment, subjects learn to be domain experts within the experimental scenario. For this purpose, they are provided with information on how to assess the probability of a machine malfunction. For the sake of simplicity and comprehensibility, this information is reduced to the minimum required

### Chapter 3 Effect of Domain Expert Involvement on AI Adherence

by the scenario. In our scenario, the actual probability of a malfunction is unknown to subjects, but can be estimated using three indicators: Temperature (of the motor), Speed (of the conveyor belt) and Workload (of the production line). The indicators can take unit-free values between 0 and 100. Each of the indicators has its unique optimal range, with a malfunction being less likely when an indicator lies within the optimal range. The more of the indicators lie inside their respective optimal ranges, the less the overall probability of a malfunction. For the sake of simplicity, the likelihood of malfunctions follows a four-step ordinal scale: “very unlikely”, “unlikely”, “likely” and “very likely”. These steps correspond to underlying probabilities of 5%, 35%, 65% and 95% for a malfunction occurring in a round, though these exact probabilities remain unknown to subjects. Subjects are provided with a heuristic rule: If all three indicators are located within their optimal ranges, a malfunction is considered “very unlikely”. For each indicator located outside its optimal range, a malfunction becomes one step more likely – i.e., a malfunction is “very likely”, when all indicators are outside their optimal ranges. In the example, shown in Figure 3.1, “Temperature” is located outside its optimal range and “Speed” and “Workload” are located inside their respective optimal ranges. Accordingly, a malfunction would be considered “unlikely”. To model the ambiguity of a real-world situation, the exact op-

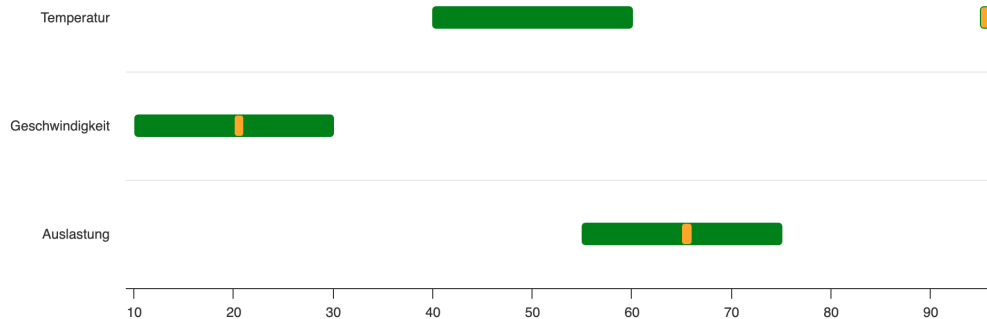


Figure 3.1: Example of the three indicators with their optimal ranges (green bars) and values (orange dots). Notation in German.

timal ranges remain unknown to subjects. Instead, they need to estimate them based on historical data. subjects’ individual assessments of optimal ranges are called “acceptable ranges”. Acceptable ranges serve subjects as an approximation of the (unknown) optimal ranges in the heuristic rule described above. To set their acceptable ranges, subjects receive – for each indicator separately – a graphic distribution of past machine states in correspondence with indicator values. Subjects are asked to choose a range of indicator values, which they consider acceptable based on the number of past malfunctions within

### Chapter 3 Effect of Domain Expert Involvement on AI Adherence

this range. As shown in the example in Figure 3.2, the distribution for each indicator contains sixty data points, thirty with a malfunction (red dots) and thirty without (green dots). Subjects are instructed to set their personal acceptable ranges to contain as many points without past malfunctions and as few points with past malfunctions as possible. This task is not trivial, as there is no obvious dominant solution.

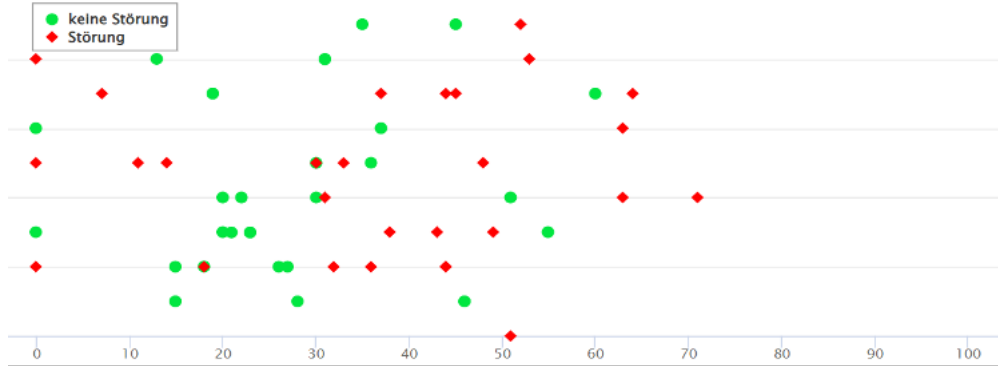


Figure 3.2: A distribution of past values for the indicator “Temperature”. The horizontal axis shows values, the vertical axis serves better visualization and has no meaningful interpretation. Notation in German.

The heuristic rule together with the individually defined acceptable ranges constitute the knowledge of subjects, necessary to take maintenance decisions. Additionally, subjects learn that they can utilize non-binding AI recommendations to aid their decisions. The accuracy of AI recommendations can be 90% or 50%, depending on the success of the AI training. The difference in the AI accuracy is introduced in order to encourage subjects to take the training seriously and to execute the necessary effort. Those who do not succeed during the training stage and receive a low-accuracy AI advisor are excluded from the data analysis afterwards. In our experiment, the AI training is a separate process which takes place after subjects have set their acceptable ranges and before they are asked to take the maintenance decisions. The design of the AI training differs between the three experimental treatments, as required by our hypothesis.

In the zero involvement treatment, subjects are informed that the AI training is fully automated. They see a brief loading bar which indicates that the training is taking place. In the passive involvement treatment, subjects are told that the AI training takes place automatically based on their acceptable ranges. Subjects are presented with 10 training situations. A training situation consists of a combination of three indicator values pre-

### Chapter 3 Effect of Domain Expert Involvement on AI Adherence

sented together with respective acceptable ranges\* (see Figure 3.3). In each situation, the probability of a machine malfunction must be assessed using the heuristic rule. Subjects are not allowed to assess the probability themselves. Instead, they observe how the heuristic rule is being automatically applied to each situation, based on their acceptable ranges. In the active involvement treatment, subjects are told that they actively train the AI with their knowledge. They are presented with the same 10 situations, but this time they can actively apply the heuristic rule and assess the malfunction probability by themselves. The design of the three treatments was chosen to mimic real-world involvement scenarios. Zero involvement serves as an equivalent to an AI training process which is carried out without domain expert participation, e.g., by data scientists or external providers. Passive involvement represents a situation in which domain experts share their knowledge about the system (in our design, in form of acceptable ranges), but do not actively apply this knowledge to the training data set. Instead, the knowledge would be processed and inserted into an AutoML system by an intermediary agent, e.g., a data scientist. Active involvement then allows domain experts to apply their knowledge directly to the training data set, without an intermediary. Based on the AutoML classification by Karmaker et al. [47], the zero involvement scenario can be seen to represent ML solutions with up to the 3rd level of automation, passive and active involvement – with the levels 4 and 5 respectively.

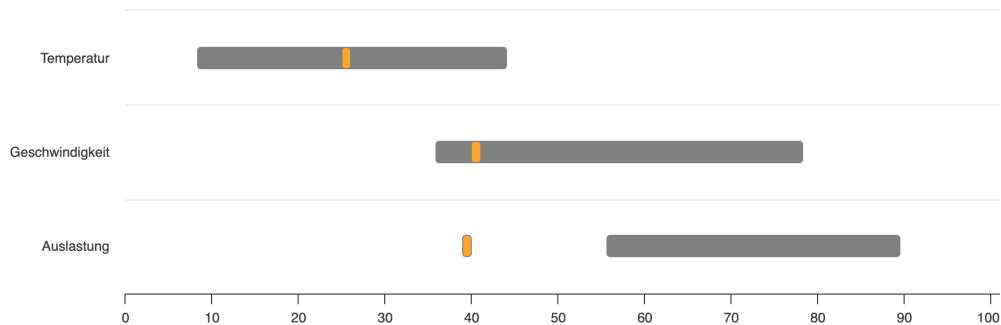


Figure 3.3: An AI training situation. Individual acceptable ranges are displayed as grey bars, the indicator values are displayed as orange dots. Notation in German.

If a malfunction has been classified as “very unlikely” or “unlikely” and no malfunction has occurred, the assessment is considered correct and otherwise incorrect. If a malfunction has been classified as “very likely” or “likely” and a malfunction has occurred,

---

\*From the training stage onward, acceptable ranges are displayed as grey bars, while the historical data points are hidden for the purpose of conciseness.

### Chapter 3 Effect of Domain Expert Involvement on AI Adherence

the assessment is considered correct and otherwise incorrect. In the zero and passive involvement groups assessments are conducted automatically, based on the subjects' acceptable ranges and the heuristic rule. In the active involvement group subjects perform assessments themselves. If at least seven training situations have been solved correctly, the AI accuracy is 90%, otherwise 50%. In all three treatment groups, in the end of the AI training subjects are informed about the resulting AI accuracy. The AI accuracy is communicated to the subjects in both probability and frequency formats to ensure proper understanding [31]. In the passive and active involvement groups, subjects additionally learn, how many situations were solved correctly.

Finally, after the AI training is completed, the main part of the experiment starts. Subjects are asked to make maintenance decisions for a fictitious machine in 25 independent rounds. In every round, they are presented with a combination of three indicator values depicted together with the individual acceptable ranges (as in Figure 3.3). Each round can be represented as a decision-tree of three binary steps resulting in a total of eight well-defined decision paths (see Appendix). In the first step, subjects are asked to assess the likelihood of a malfunction using the heuristic rule and to decide, whether a maintenance should be performed. In the second step, subjects are presented with the AI assessment of the malfunction likelihood as well as a recommendation on whether or not to perform a maintenance. In the third step, subjects are asked to take their final maintenance decision, which will determine their payoff for the round. After a round is completed, subjects receive no immediate feedback on their performance in order to minimize learning effects as well as outcome-induced decision biases [20]. Instead, subjects learn about their total performance and payoff once they have completed all 25 experimental rounds. Additionally, we purposefully do not communicate to subjects, how an average human subject would perform in a task at hand. Instead, we allow subjects to develop a feeling of the own expertise, without explicitly quantifying it. This fact aims at mirroring an ambiguity of a real word situation, where people usually do not have access to an exact statistical information about their decision accuracy and the optimal strategy (to follow or not to follow the advice) is not straight-forward.

A payoff for a round is calculated as follows. If no maintenance is performed and no malfunction occurs, the machine can produce regularly, and subjects are granted 10 Taler\*. Performing a maintenance reduces the round payoff to 5 Taler (a machine can

---

\*During the experiment, all amounts are denoted in the fictitious experimental currency "Taler", which is exchanged to Euro at €0.1 per Taler at the end of the experiment.

## Chapter 3 Effect of Domain Expert Involvement on AI Adherence

produce only half of the regular amount) and eliminates the risk of a malfunction for that round. If no maintenance has been ordered and a malfunction occurs, subjects receive 0 Taler in this round (the machine has a zero production output due to the malfunction). Consequently, subjects must weigh a higher and risky payoff (when performing no maintenance) with a secure but smaller payoff (when performing a maintenance).

Subjects have been informed that the whole scenario is fictitious. The data used during the experimental scenario has been exogenously simulated beforehand from a random normal distribution. In total, 200 situations (combinations of three indicator values together with subsequent occurrence or absence of a malfunction) were generated. Out of these, 10 have been randomly assigned to the AI training situations, while 25 have been randomly assigned to the main part of the experiment. The AI advisor has been simulated by a simple rule-based algorithm, which applied the heuristic rule while knowing the exact values of the optimal ranges. This procedure resulted in a high accuracy AI advisor with 90% accuracy and a low accuracy AI advisor with 50% accuracy. Each subject received either of the two depending on the outcome of the training stage.

### 3.2.2 Experimental procedure

The experiment has been computerized using oTree [14] and conducted in December 2022 and April 2023 at the Business and Economic Research Laboratory (BaER-Lab\*) at the University of Paderborn. In total, nine experimental sessions took place, three for each of the treatment groups. Subjects were recruited via the online recruiting system ORSEE [36] and were only allowed to participate in one session. In the beginning of every session, subjects were randomly assigned to a computer workplace. Each workplace was placed in a cubicle to ensure that every subject could see only her own screen. Subjects were informed that communication and usage of mobile devices was prohibited during the experiment. Subjects received experimental instructions in printed form and had up to 15 minutes to read them carefully. Instructions contained all the information needed in the experiment and were tailored to respective treatments. Subjects were allowed to keep the instructions till the end of the experiment and to look them up at any time. The detailed instructions can be found in Appendix.

The experimental process was logically divided into four stages (see Figure 3.4). In the first stage, subjects answered extensive comprehension questions to ensure a sufficient

---

\*[www.baer-lab.org](http://www.baer-lab.org)

## Chapter 3 Effect of Domain Expert Involvement on AI Adherence



Figure 3.4: The four stages of the experiment

understanding of the experimental scenario and rules. Subjects could only advance to the second experimental stage once all questions had been answered correctly. In the second stage, subjects were asked to set their individual acceptable ranges for the three indicators. Beforehand, subject could practice setting acceptable ranges on an example distribution. During the setting, subjects were allowed to revise their ranges as many times as they wished until ultimately confirming their desired ranges and advancing to the third stage. In the third stage, the training of the AI took place. In the zero involvement group, subjects were simply informed about the result of the training without any details on its input or process. In the passive involvement group, the training happened automatically with no decision-making authority to subjects, but a possibility to observe the training process. Additionally, subjects knew that their acceptable ranges contributed to the AI training. In the active involvement group, subjects evaluated the training data themselves, applying their acceptable ranges and the heuristic rule, thereby directly contributing to the AI training. In stage four, subjects were asked to supervise the fictitious machine over 25 rounds. In each round they had to decide, whether a maintenance should or should not take place, going through the steps described in 3.2.1. All rounds of the fourth stage were independent of each other, i.e., a decision from one round did not affect other rounds, neither did an occurrence of a malfunction. The payoffs from all 25 rounds were accumulated and exchanged to Euro at a rate of €0.1 per Taler. Additionally, subjects received a show-up fee of €2.50. The total payoff was displayed after the completion of the stage four. Noteworthy, stage four was the only stage to feature direct monetary incentives. However, at stages two and three subjects were informed that their decisions would have implications for their performance in stage four, thereby providing an indirect incentive to exert effort in the earlier stages as well.

In the end of the experiment, subjects were asked to answer a questionnaire which contained multiple standardized questionnaires to account for the subjects' affinity for technology interaction [29], ex-post confidence in their decisions [34], and general self-efficacy [8], along with treatment manipulation checks and questions on the socio-economic

## Chapter 3 Effect of Domain Expert Involvement on AI Adherence

background of the subjects, particularly age, gender, and study major. The full questionnaire can be obtained from Appendix. Upon completion of the questionnaire, subjects were called by their cabin number to receive their payment. Subject were paid their earnings in cash and thanked for their participation before being released from the lab.

### 3.3 Experimental results

The empirical analysis of our experimental observations has been conducted with Stata 17.0. It starts with reporting demographics and descriptive statistics. Then, in the main analysis, we present the results of the hypotheses testing and the treatment effectiveness evaluation. We further conduct supplementary analysis on relevant questionnaire outcomes.

#### 3.3.1 Demographic and descriptive statistics

In total, one-hundred-fifty-four ( $N=154$ ) student subjects participated in the experiment, of which 51 were assigned to the zero involvement ( $T_Z$ ), 51 to the passive involvement ( $T_P$ ) and 52 to the active involvement ( $T_A$ ) groups. Each session lasted for approximately one hour, subjects earned €17.50 on average. One subject from  $T_Z$  failed to complete the comprehension questions and has been excluded from the data set ( $N=153$ ). On average, subjects were 24 years old, their age varying between 18 and 65. Females constituted 64% of subjects (see Table 3.1). Fields of subjects' studies varied from engineering to arts, with economics (34%) and pedagogy (35%) being most common (see Appendix).

Table 3.1: Demographic statistics

	$T_Z$	$T_P$	$T_A$	Total
<b>Age</b>				
Mean	24.4	24.5	23.1	24.0
Standard deviation	6.6233	6.5065	3.8908	5.7896
Min	18	18	18	18
Max	64	65	35	65
<b>Gender</b>				
Female	0.72	0.627	0.578	0.641
<b>No of observations</b>	50	51	52	153

Depending on the success of the AI training, subjects could receive an AI advisor with a high (90%) or low (50%) accuracy. The sole purpose of this AI accuracy manipulation



### Chapter 3 Effect of Domain Expert Involvement on AI Adherence

was motivating subjects to exert effort during the AI training. For this reason, and to ensure the ceteris paribus condition between the treatments, subjects with a low AI accuracy (N=11, see Table 3.2) are excluded from the analysis. Consequently, the final data set includes N=142 independent observations. Observations are pooled at the subject level.

Table 3.2: AI accuracy

	$T_Z$	$T_P$	$T_A$	Total
<b>90% accuracy AI</b>				
Absolute	50	45	47	142
Relative	1.0	0.8824	0.9038	0.9281
<b>No of observations</b>	50	51	52	153

Table 3.3 summarizes the individual accuracy of initial assessments across the three treatments. On average, subjects' initial assessments have been accurate in 52% ( $T_Z$ ), 58% ( $T_P$ ) and 57% ( $T_A$ ) of rounds. Though it seems that subjects from  $T_Z$  have been slightly less accurate in their initial estimations, the differences in accuracy between the treatments are not statistically significant (Kruskal-Wallis H-Test with ties:  $\chi^2(2) = 3.041, p = 0.2186$ ). The standard deviation of initial accuracy is rather high at 17 p.p., 13 p.p. and 14 p.p. for  $T_Z$ ,  $T_P$  and  $T_A$  respectively. Analogously, the individual accuracy of initial assessments ranges between 0% and 88%. We note that with the AI accuracy being at 90%, subjects with high initial accuracy have had a lower probability of their initial estimation being different from the AI advice.

Table 3.3: Individual accuracy of initial assessments across treatments

	$T_Z$	$T_P$	$T_A$
<b>Accuracy of initial assessments</b>			
Mean (absolute)	13.06	14.53	14.32
Mean (relative)	0.5224	0.5813	0.5727
Standard deviation (absolute)	4.3397	3.3615	3.4332
Standard deviation (relative)	0.1736	0.1345	0.1373
Median (absolute)	13	15	15
Median (relative)	0.52	0.6	0.6
Min	0	5	5
Max	22	21	20
<b>No of observations</b>	50	45	47

After having submitted their initial assessment, subjects received advice from the AI and could revise their maintenance decision. Subjects from the  $T_Z$  group have revised

### Chapter 3 Effect of Domain Expert Involvement on AI Adherence

their initial assessment in 11% of rounds and subjects from  $T_P$  and  $T_A$  groups in 8% of rounds (see Table 3.4). These percentages incorporate total revisions – both for situations when the initial assessment has been equal or unequal to the AI advice. Subjects from  $T_Z$  appear to have revised their assessments slightly more often, however the differences between the treatment groups are not statistically significant (Kruskal-Wallis H-Test with ties:  $\chi^2(2) = 3.907, p = 0.1418$ ). Standard deviations lay at 9 p.p., 8 p.p. and 8 p.p. for  $T_Z$ ,  $T_P$  and  $T_A$  respectively. Notably, there have been subjects who never revised their decision as well as those who revised their decision in nearly 50% of rounds. Among all revisions, 70% to 82% led to a better outcome for the subjects (see Appendix).

Table 3.4: Revision decisions across treatments

	$T_Z$	$T_P$	$T_A$
<b>Revisions of initial assessments</b>			
Mean (absolute)	2.7	2.0	2.06
Mean (relative)	0.11	0.08	0.083
Standard deviation (absolute)	2.2520	1.9771	1.9605
Standard deviation (relative)	0.0895	0.0791	0.0784
Median	3	2	2
Median	0.12	0.08	0.08
Min	0	0	0
Max	8	12	9
<b>No of observations</b>	50	45	47

#### 3.3.2 Main analysis

Following our research hypotheses, we are particularly interested in situations, in which subject's initial assessment is unequal to the AI advice. For such situations, if the revised decision is equal to the AI advice, we count it as AI adherence. Per subject, we then calculate an AI adherence rate – the proportion of AI adherence decisions among situations, where one's initial assessment differs from the AI advice. Table 3.5 summarizes the corresponding empirical results across the three treatment groups. Subjects, whose initial assessment has been equal to the AI advice in all 25 rounds ( $N=10$ ), have been excluded from the calculation of the AI adherence rate. On average, subjects' initial assessment contradicted the AI advice in 5 ( $T_Z$ ) or 4 ( $T_P$ ,  $T_A$ ) out of 25 experimental rounds, at standard deviations of 4 ( $T_Z$ ) and 3 rounds ( $T_P$ ,  $T_A$ )\* respectively. While subjects in

\*Values are rounded for the purpose of readability, Table 3.5 contains precise values.

### Chapter 3 Effect of Domain Expert Involvement on AI Adherence

$T_Z$  experienced slightly more contradicting situations, the difference is not significant (Kruskal-Wallis H-Test with ties:  $\chi^2(2) = 1.89, p = 0.3887$ ). Logically, we observe that a lower initial human accuracy (see Table 3.3) corresponds to a higher number of contradicting situations for  $T_Z$  (and vice versa for  $T_P$  and  $T_A$ ). On average, the AI adherence lies at 48% for the zero involvement, 51% for the passive involvement and 47% for the active involvement group. The standard deviation is rather high at 35p.p., 37p.p. and 36p.p. respectively. Notably, in all three groups there are subjects who either always or never follow the AI advice. It is also worth mentioning that following the AI advice has been an optimal strategy, since the subjects' average initial accuracy ranges between 52% and 58%, with the best-performing subjects reaching up to 88%, while the AI accuracy is at 90%. Indeed, we observe that subjects, who follow the AI advice, benefit from it in 80% to 86% of rounds (see Appendix). However, this strategy has not been entirely obvious to the subjects since they have not been able to directly assess their own accuracy during the experiment as no feedback on previous rounds has been provided.

Table 3.5: AI adherence rate across treatments

	$T_Z$	$T_P$	$T_A$
<b>Initial assessment <math>\neq</math> AI advice</b>			
Mean (absolute)	5.14	3.84	4.1
Mean (relative)	0.2056	0.1538	0.1634
Standard deviation (absolute)	4.0255	2.5845	2.7649
Standard deviation (relative)	0.1610	0.1034	0.1106
Median	4	4	4
Min	0	0	0
Max	17	12	12
<b>No of observations</b>	50	45	47
<b>AI advice adherence rate</b>			
Mean	0.4805	0.5075	0.4660
Standard deviation	0.3480	0.3736	0.3557
Median	0.4	0.5	0.5
Min	0	0	0
Max	1	1	1
<b>No of observations</b>	46	43	43

Figure 3.5 depicts the frequency distributions of AI adherence rates for the three treatment groups. Though each group's distribution is heterogeneous within, the patterns between the groups appear to be similar. The distribution of the AI adherence rates between 0 and 1 takes a "W"-like shape, indicating that the boundary solutions – always or never

### Chapter 3 Effect of Domain Expert Involvement on AI Adherence

following the AI advice – are the prevailing strategies in every group. This meets findings from related literature of different individuals tending to be either averse or appreciative of algorithmic advice at a given task [46].

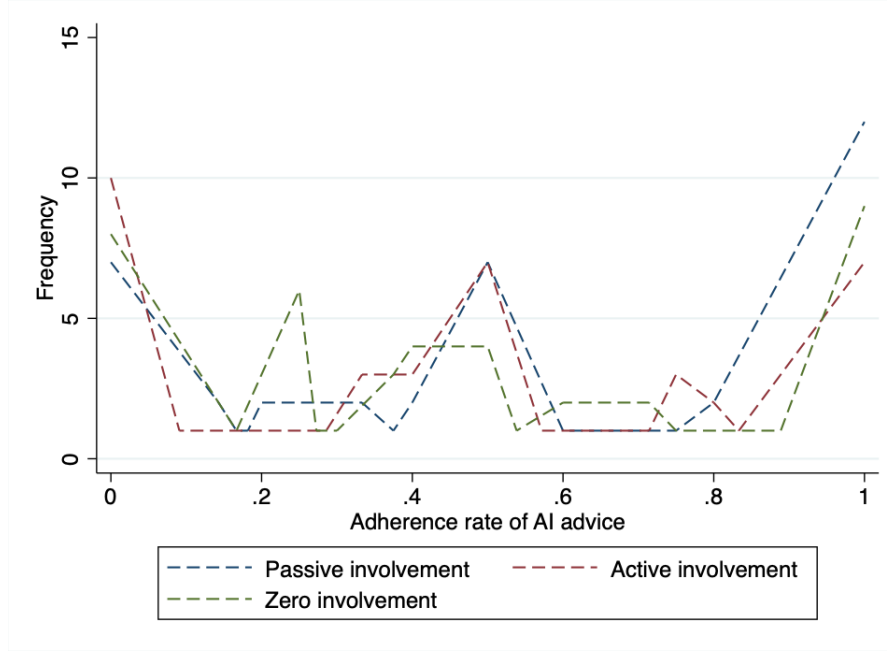


Figure 3.5: Distribution of AI adherence rate by treatment

To investigate our first research hypothesis (H1a), we conduct a Kruskal-Wallis H-Test with ties. We find no statistically significant difference in the AI adherence rate between the three treatment groups ( $\chi^2(2) = 0.299, p = 0.8610$ ). Pairwise comparisons of the groups yield the same results (Mann-Whitney U-Test, one-sided, each subject one independent observation:  $T_P$  vs.  $T_A$ :  $z = 0.529, p = 0.3000$ ;  $T_A$  vs.  $T_Z$ :  $z = 0.070, p = 0.4729$ ;  $T_P$  vs.  $T_Z$ :  $z = 0.423, p = 0.3375$ ). Additionally, we run pairwise two-sample Kolmogorov–Smirnov tests for equality of distribution functions. The tests show that distributions of AI adherence rates (between 0 and 1) do not significantly differ between groups ( $T_P$  vs.  $T_A$ :  $D = 0.1163, p = 0.938$ ;  $T_A$  vs.  $T_Z$ :  $D = 0.1067, p = 0.925$ ;  $T_P$  vs.  $T_Z$ :  $D = 0.1097, p = 0.908$ ), further supporting the results from the two previous tests as well as the impression gained from Figure 3.5.

Consequently, we conclude that the hypotheses H1a is not supported by our empirical data. Since our second hypothesis (H1b) logically requires at least one significant difference between any of the treatment groups, it is therefore also not supported. To shed light on possible causes behind the rejection of the main hypotheses, we evaluate the effective-

### Chapter 3 Effect of Domain Expert Involvement on AI Adherence

ness of our experimental design. For our experiment, it is crucial that subjects perceive themselves to be experts within the experimental scenario, independently of the treatment. Further, the treatment manipulations should lead to differences in the subjects' perception of the AI. Table 3.6 summarizes subjects' perception of the experimental scenario. The results demonstrate that subjects from all treatment groups understood the scenario and its rules well, were able to put themselves into the role of an engineer and were satisfied with their own performance. The scores for all items are remarkably high at 3.8 points and upwards on a 5-point scale. The only exception is the perceived demand of the experimental tasks, with scores between 2.6 and 2.8 points, which shows that the task was neither perceived as very easy or very difficult by the subjects.

Further, we evaluate questionnaire items which are relevant for our treatment manipulations: *perceived influence on AI training*, *perceived contribution to AI advice quality*, *perceived understanding of AI functionality*, *perceived understanding of AI advice generation* and *perceived quality of AI advice*. Table 3.7 summarizes the corresponding statistics\*. The first four items each demonstrate a pattern of subjects' consent to questionnaire statements transitively increasing from zero to active involvement treatments. In other words, subjects' perceived influence on AI and understanding of AI increases with the degree of their involvement, with the differences between  $T_Z$  and  $T_A$  being highly statistically significant for all items, while the differences between  $T_P$  and  $T_A$  are significant for half of them. Overall, we conclude that our treatment manipulations have been successful in a sense that different grades of subjects' involvement in the AI training are reflected in subjects' perception of their influence on the AI and understanding of the AI. However, these gradations in perception are not reflected in subjects' actual behaviour, as the rejection of our main hypotheses demonstrates.

Generally, the scores for subjects' perceived influence on AI and understanding of AI appear to be rather high – even in the zero-involvement group subjects perceive a certain degree of influence and understanding with scores around 3 points on a 5-point scale. Consistent with this observation, subjects assign rather low scores to the question, whether an AI advice accuracy is a gamble (see Appendix). Considering subjects' perceived quality of the AI advice, it is also high at 3.6 points and upwards for all treatment groups. However, there are no significant differences between the groups in this regard.

---

\*A full list of items as well as between-treatment comparisons can be obtained from Appendix.

### Chapter 3 Effect of Domain Expert Involvement on AI Adherence

Table 3.6: Summary statistics and between-group comparison for questionnaire items on experiment perception

	$T_Z$	$T_P$	$T_A$	Kruskal-Wallis-H
Maintenance scenario was easily understandable	4.64 (0.6312)	4.53 (0.6252)	4.66 (0.5224)	0.5203
Able to put oneself in role of production plant manager	3.9 (1.1473)	3.82 (1.1734)	4.09 (.9517)	0.6318
Sufficient information and examples provided	4.6 (0.6999)	4.51 (0.6260)	4.51 (0.6552)	0.4790
Understood experimental explanations	4.68 (0.6833)	4.58 (0.6212)	4.77 (0.4280)	0.2976
Understood experimental context	4.72 (0.4965)	4.67 (0.5222)	4.83 (0.3799)	0.2670
Certainty about experimental requirements	4.78 (0.464)	4.71 (0.5886)	4.77 (0.5197)	0.8238
Understanding of conditions for success	4.42 (0.7584)	4.38 (0.6839)	4.49 (0.6552)	0.7086
Perceived task demand	2.58 (1.0120)	2.78 (0.9508)	2.55 (0.8799)	0.4559
Satisfaction with own performance	4.3 (0.6468)	4.42 (0.5431)	4.06 (0.9185)	0.1919
Expectation of other subjects performing well	3.84 (0.8418)	3.93 (0.7508)	4.04 (0.6902)	0.4724
<b>No of observations</b>	50	45	47	

The table reports summary statistics for experiment perception items measured on a 5-point scale. Standard deviations are stated in parenthesis. Kruskal-Wallis-H reports p-values for Kruskal-Wallis H-Tests with ties.

#### 3.3.3 Supplementary analysis

To control for factors that potentially influence an individual AI adherence rates, we measured subjects' affinity for technology interaction (ATI), general self-efficacy (ASKU) and ex-post decision confidence (DC), using their standardized questionnaire metrics. Table 3.8 summarizes corresponding statistics and between-group comparisons. The average ATI scores show a moderate technological affinity among subjects, ranging between 3.5 and 4 points on a 6-point scale. The average ASKU score resides in the upper quartile of the 5-point scale for all groups, indicating that subjects generally perceive themselves

### Chapter 3 Effect of Domain Expert Involvement on AI Adherence

Table 3.7: Summary statistics and between-group comparison for questionnaire items on AI perception

	$T_Z$	$T_P$	$T_A$	Kruskal-Wallis-H	MWU ( $T_Z, T_P$ )	MWU ( $T_Z, T_A$ )	MWU ( $T_P, T_A$ )
Perceived influence on AI training	3.02 (0.9145)	3.58 (1.1178)	4.09 (0.8030)	0.0001	0.0063	0.0000	0.0270
Perceived contribution to AI advice quality	3.34 (1.0994)	3.87 (1.0574)	4.11 (0.9145)	0.0013	0.0186	0.0003	0.2908
Perceived understanding of AI functionality	3.08 (1.2094)	3.38 (1.2301)	3.91 (1.0179)	0.0024	0.1995	0.0005	0.0362
Perceived understanding of AI advice generation	2.9 (1.1473)	3.42 (1.1577)	3.72 (1.1554)	0.0020	0.0284	0.0005	0.1645
Perceived quality of AI advice	3.68 (0.7939)	3.76 (1.0478)	3.60 (0.9478)	0.4874	0.3456	0.7550	0.2726
<b>No of observations</b>	50	45	47				

The table reports summary statistics for AI perception items measured on a 5-point scale. Standard deviations are stated in parenthesis. Kruskal-Wallis-H reports p-values for Kruskal-Wallis H-Tests with ties. MWU reports pairwise two-sample between-group Mann-Whitney U-Tests.

capable of dealing with problems and accomplishing goals. Similarly, subjects appear quite confident about their decisions made during the experiment, with the average ex-post confidence for all three groups exceeding 5 points on a 7-point scale.

We do not find significant treatment differences for the general self-efficacy (Kruskal-Wallis H-Test with ties:  $\chi^2(2) = 2.442, p = 0.2950$ ) and ex-post confidence in decisions (Kruskal-Wallis H-Test with ties:  $\chi^2(2) = 0.534, p = 0.7657$ ). Notably, there appears to be a weak statistical difference in the affinity for technology interaction between the treatment groups (Kruskal-Wallis H-Test with ties:  $\chi^2(2) = 6.156, p = 0.0460$ ). This statistical significance can be attributed to the differences in means between  $T_Z$  and  $T_A$  (Mann-Whitney U-Test, two-sided:  $z = 2.514, p = 0.0115$ )\*. However, this difference must be interpreted with caution since affinity for technology interaction was measured after the experiment. As  $T_Z$  and  $T_A$  present polar opposites in terms of technology interaction quantity, this may have factored into subjects' assessment of their affinity for technology interaction.

Additionally, we have asked subjects, how they perceive the AI training in terms of

\*For the full set of pairwise Mann-Whitney U-Tests see Appendix.

### Chapter 3 Effect of Domain Expert Involvement on AI Adherence

Table 3.8: Summary statistics and between-group comparison for standardized questionnaire control items

	$T_Z$	$T_P$	$T_A$	Kruskal-Wallis-H
Affinity to Technology Index (ATI)	3.46 (0.8865)	3.55 (1.2122)	4.01 (1.1344)	0.0460
Self-efficacy (ASKU)	3.77 (0.5355)	3.78 (0.5365)	3.89 (0.5433)	0.2950
Ex-post decision confidence (DC)	5.19 (0.8460)	5.27 (0.8246)	5.08 (1.1124)	0.7657
<b>No of observations</b>	50	45	47	

The table reports summary statistics for control item scores – ATI measured on a 6-point scale, ASKU measured on a 5-point scale and DC measured on a 7-point scale. Standard deviations are stated in parenthesis. Kruskal-Wallis-H reports p-values for Kruskal-Wallis H-Tests with ties.

enjoyment and worthy time investment on a 5-point scale (see Appendix). Remarkably, subjects in all treatment groups consider the AI training a time well spent (on average 3.7-4 points) and, in parts, would have liked to invest even more time (on average 2.7-3.2 points). Similarly, subjects like sharing knowledge with AI (on average 3.9-4 points) and would have liked to share more knowledge (on average 3.3-3.5 points). Also, subjects have largely enjoyed the AI training (on average 3.3-3.7 points). The scores for opposite items (unwillingness and discomfort in sharing knowledge with AI) are consistently low and do not exceed 2 points on average.

## 3.4 Discussion and conclusion

Today's world witnesses a rapidly increasing relevance of artificial intelligence with exponentially growing number of its application domains. However, development of ML models requires programming and data science competence which is scarce on the labor market. Automated machine learning is supposed to make AI solutions more accessible to organizations without sufficient data science expertise. Domain experts play a crucial role in making this vision a reality, acting as knowledge contributors. We argue that besides providing the domain knowledge, involvement of domain experts has an additional value, since they are in many cases also the end users of AutoML solutions. By involving domain experts in the model training, one enables them to execute influence over the



### Chapter 3 Effect of Domain Expert Involvement on AI Adherence

resulting model and to enhance their understanding of its functionality. Drawing from pre-existing literature, these two factors in their turn may strengthen domain experts trust in the AutoML model and raise their adherence to its recommendations. The adherence is crucial for organizations if they want to turn the high accuracy of AI advisor systems into real efficiency gains.

Our study aims to investigate the role of domain expert involvement during the AI training on their subsequent AI advice adherence. We hypothesize that a higher grade of involvement leads to a higher advice adherence. In an incentivised laboratory experiment, we construct a predictive maintenance scenario in which subjects act as domain experts and are confronted with an AI advisor. We compare three groups: subjects who actively participate in the AI training, subjects who passively witness the AI training, and subjects who neither participate nor observe the AI training. We measure the individual adherence rate for each subject, which reflects the ratio between the number of decisions to follow an AI recommendation and the number of situations, in which the subjects' initial estimation is different from the AI advice. In our main analysis, we compare the adherence rates between the treatment groups. Further, we test whether our experimental design and treatment manipulations have an influence on subjects' perception of themselves and of the AI advisor. Additionally, we evaluate whether subjects consider the AI training to be time well spent and feel comfortable sharing their expertise with the AI.

We find no difference in AI advice adherence between groups and therefore have to reject both initial hypotheses. In other words, according to our results, the adherence behaviour of subjects does neither vary between groups nor grow with the raising grade of involvement. However remarkably, we find that treatment manipulations had an intended effect on the subjects' perception – the higher the degree of involvement in the AI training, the higher the perception of own influence on the AI model. Additionally, subjects' perception of the own understanding of the AI functionality also grows with a higher grade of involvement. According to existing literature, both higher perceived influence and higher perceived understanding should lead to a higher perceived quality of the AI advice and eventually to a higher advice adherence [22, 46, 53, 80]. In our results, we observe that a higher involvement in the AI training leads to a higher perception of influence and understanding but does not change the subjective perception of the AI advice quality, which remains similar among all treatment groups. It also does not lead to a change of actual behaviour – subject who actively participate in the AI training follow the AI advice similarly often as subjects from passive and zero involvement groups.

### Chapter 3 Effect of Domain Expert Involvement on AI Adherence

Generally, across all three involvement conditions, AI advice adherence is rather low at roughly 50%, though the subjects have been informed that AI has an accuracy of 90%. Simultaneously, subjects report to have understood the scenario well and to know exactly, what they should do in order to be successful in the tasks. These insights together with the high ex-post decision confidence among subjects lead us to conclude that subjects have felt themselves to be experts within the scenario. One possible explanation for the fact that positive effects of involvement on the AI perception do not result in behavioural change could be that these effects have been overruled by the strong feeling of expertise among the subjects, which does not depend on treatment. Subjects feel equally (and strongly) confident in their choices and might disregard the AI advice equally often because of this confidence. Experts have been repeatedly shown to behave differently in algorithm interactions than lay people and to demonstrate higher algorithm aversion [52, 53]. Our findings reinforce these results. We further observe a rather heterogeneous distribution of algorithm adherence behaviour within each group with tendencies towards the extremes, i.e., individuals either acting very averse or very appreciative towards AI-advice with only a minority adapting their decisions situation-dependent. This result matches the state of existing literature on algorithm adherence behaviour [46]. However, these within-group differences appear consistent between groups. Also, other individual factors, like affinity for technology interaction or general self-efficacy do not vary significantly between the treatment groups. Interestingly, subjects report strong positive perceptions towards the AI training – they consider it to be a well spent time, share their knowledge willingly and even enjoy the processes. These scores are equally high across all three treatment groups and go along with the “IKEA-effect”, repeatedly found in the previous literature[59].

We conclude that involvement of domain experts in the AI training enhances the feeling of control over the resulting model and the perceived understanding of the AI functionality. Both these insights are valuable in the context of the human centered AI and the AI explainability [9, 67]. Involving domain experts in the AI training helps them to influence the model, fitting it into the domain context, and to peer inside the black box. However, in our experiment, these positive effects do not lead to an increase in perceived AI advice quality nor to a higher advice adherence. We attribute this result to the fact that domain experts have a strong confidence in their expertise and this confidence could not be overruled by the treatment manipulations. This is an important result, because it highlights the challenge of changing the attitudes and behaviour of expert subjects.

Future research may investigate whether feedback on own accuracy could relativize

### **Chapter 3 Effect of Domain Expert Involvement on AI Adherence**

subjects' perception about their own expertise and make them more accessible to the treatment manipulations. However, learning effects would have to be considered in this case. Further, the number of single observations, in which the AI adherence rate could be calculated, has been rather small. By turning subjects into experts, we enable them to make accurate assessments. The high initial accuracy reduces the number of situations, in which the own initial assessment differs from the AI advice (4 to 5 out of 25 rounds, on average). This phenomenon originates in our design and mirrors a real-world situation – an expert would indeed often produce an assessment which is equal to the AI advice, however over a longer period of time the AI advisor would be on average more accurate. This may be remedied in future runs of the experiment through increasing the total number of rounds and generating more situations where the AI adherence rate can be calculated. Increasing task uncertainty would be another potential direction to go into. Also, other forms of expert involvement and knowledge sharing can be considered – an interview, a questionnaire etc. Finally, the future research can test the direct influence of better controllability and understandability on the advice adherence, for example through explicitly allowing to change the model output [23] or explaining the functionality of AutoML to domain experts [80]. As we introduced a novel experimental design through which behavioural effects of user involvement in AI-supported predictive maintenance decisions can be studied, researchers may build upon our design to investigate a variety of further research questions regarding human perception of algorithmic entities, particularly in industry- or production-centric contexts. Our results also carry two positive messages for practitioners: Through involvement, they can improve the understanding of the AutoML system by the end users. Additionally, the process of knowledge sharing can be considered well invested time because it yields both knowledge and joy.

## **Chapter 4**

# **Domain knowledge acquisition for machine learning using ChatGPT**

# DOMAIN KNOWLEDGE ACQUISITION FOR MACHINE LEARNING USING CHATGPT

ANASTASIA LEBEDEVA<sup>\*†</sup>

## Abstract

Domain knowledge can be integrated into machine learning models to account for insufficient data or specific context requirements. The challenge to acquire such knowledge is known as the *knowledge acquisition bottleneck* [41], featuring its methodological complexity. In machine learning projects, knowledge acquisition commonly requires an extensive investment of time and effort from data scientists and other stakeholders. We argue that large language models in general and ChatGPT in particular offer a technological base to automate time-consuming knowledge acquisition tasks and free the resources of data scientists for actual development. Following the methodology of Design Science Research [63], we design a novel method for automatic interview-based knowledge acquisition with ChatGPT and map it to the existing scientific frameworks. We then instantiate the proposed method in a prototype – the ChatGPT Interviewer. We demonstrate and evaluate the ChatGPT Interviewer in two separate industrial applications. Our evaluation shows that the ChatGPT Interviewer can successfully lead an interview, flexibly adopting to the individual user behaviour. However, the precision and comprehensiveness of the acquired knowledge can be further improved. We conclude that our method can be applied to automate the knowledge acquisition steps, yielding a significant efficiency gain for machine learning practitioners. Moreover, with our work, we make multiple theoretical contributions, adding to the literature on Informed Machine Learning [77], Cross Industry Standard Process for Data Mining [79], conversational agents based on ChatGPT and, last but not least, on knowledge acquisition methods.

**Keywords:** domain expert, knowledge acquisition, machine learning, ChatGPT

---

<sup>\*</sup>Paderborn University, Heinz-Nixdorf-Institute, Fürstenallee 11, 33102 Paderborn

<sup>†</sup>Corresponding author, lebedeva@mail.uni-paderborn.de

This research was submitted for publication in the Multimedia Tools and Applications journal, ISSN 1573-7721. The present version deviates from the submitted version in minor textual changes.

This research did not receive any funding or financial support. The author thanks Dr. René Fahr, Dr. Daniel Beverungen, Dr. Carlos Paiz Gatica and Tobias Liboschik for valuable discussion and constructive comments. All errors remain author's own.

### 4.1 Introduction

Domain knowledge plays an important role in many machine learning (ML) applications. When ML-practitioners are confronted with insufficient data availability or specific domain requirements, training data can be supplemented by knowledge representations, capturing the concepts, rules and relations of the domain [7, 77]. In some areas, integration of domain knowledge is especially relevant or even inevitable, for example in expert systems [41, 60], decision support systems [70], automated machine learning applications [47, 73], as well as in high stake domains or domains with security or ethical concerns [7, 77]. In order to be integrated into an ML pipeline, domain knowledge needs to be elicited and formalized, which is not a trivial issue [77]. In the information systems (IS) research, the problem of expert knowledge elicitation is known as the *knowledge acquisition bottleneck* and keeps the IS scholars and practitioners preoccupied already for decades [41, 60]. Indeed, acquiring expert knowledge can be extremely effortful and even constitute the biggest part of the development process [41].

Starting from late 1980s, methods from social and cognitive sciences, such as interviews, questionnaires, observation protocols and others, become a part of the IS toolkit to deal with the knowledge acquisition bottleneck [41, 60]. Methods from social and cognitive sciences allow IS practitioners to elicit expert knowledge in depth and to obtain information about the concepts, relations and rules of the domain [60]. However, these methods can be subject to human bias and strongly depend on human skill [41]. Moreover, methods like interview or observation are very time-consuming and often require significant ex-post effort from data scientists to formalize the output [41, 60, 70]. Time and effort are costly, besides, data scientists are a scarce resource on the labor market [6]. Many companies may even refrain from applying ML, since the investment hurdle is too high or there are not enough data scientists available [6, 73].

Data-driven approaches can eliminate or significantly reduce the effort for domain knowledge acquisition. Fueled by the advancements in machine learning and data science, such methods can automatically elicit domain features and rules from historical data, but they also have their drawbacks [70]. In most cases, the uncovered knowledge structure remains inaccessible to developers and other stakeholders due to the black box nature of ML techniques such as deep neural networks [70]. Moreover, a sufficient amount of historical data is a prerequisite for successful knowledge acquisition through ML approaches [70]. Last but not least, due to their nature, purely data-driven approaches can not account

## Chapter 4 Domain knowledge acquisition for ML using ChatGPT

for additional contextual, ethical or regulatory requirements [77].

Both social science methods and data-driven approaches to knowledge acquisition have their major advantages and drawbacks. The former are highly effortful, but output comprehensible knowledge models. The later are more efficient, but require sufficient data availability in the first place and often result in a black box. In our research, we are driven by the vision to tackle the knowledge acquisition bottleneck by proposing a method which would be both efficient and comprehensible.

The research stream on conversational agents (CAs) delivers a base for our vision. Conversational agents can be defined as artificial agents which interact with humans in natural language [21]. In the previous decades, conversational agents have evolved from rather constrained rule-based systems to flexible and context-sensitive machine learning models [5]. The application areas for conversational agents are manifold – task support, team collaboration, customer interface etc. [21]. Recent technological advancements in the field of natural language processing (NLP) and large language models (LLM) have revolutionized the design of conversational agents [5] and might offer a technological solution to the knowledge acquisition bottleneck. LLMs, in particular its flagman representative ChatGPT, are capable of leading a human-like conversation in natural language [27, 37, 51] and can be used for various information extraction tasks [50]. Leading a conversation in natural language and extracting information from relevant inputs is basically the essence of one of the most common social science methods – the interview. We therefore suggest that the tasks such as leading an interview, extracting relevant information from answers and transforming it into a formalized form could be performed by a conversational agent based on ChatGPT.

Notably, the above mentioned interview tasks are highly time consuming and require significant capacity from developers or data scientists [41, 60]. Since data science expertise is rarely available and costly [6, 57, 73], shifting their tasks over to ChatGPT would bring a significant efficiency gain. In the same time, shifting the interview tasks to ChatGPT should not affect the underlying mechanisms of the interview and preserve its methodical advantages such as natural language conversation, transparency of inputs and outputs and comprehensible results. We therefore envision a way to simplify the domain knowledge acquisition with a conversational agent based on ChatGPT, while preserving the main advantages of the interview as a social science methodology.

Following our vision, we seek to design a method to automate the knowledge acquisition interview with ChatGPT. Since the design of a novel artifact stands in the center of

## Chapter 4 Domain knowledge acquisition for ML using ChatGPT

our research, we identify the design science [40] as the underlying paradigm of our paper. Thus, we address a particular business need (the knowledge acquisition bottleneck), analyse the existing knowledge base and design a novel artifact (a method) to solve the business need and to contribute to the existing knowledge base. Specifically, we follow the Design Science Research (DSR) methodology by Peffers et al. [63]. We start in the current section by describing our motivation and defining the problem. In the following section we present the related work. Then, in Section 4.3 we explicitly define our research objectives and highlight our scientific contribution. In the Sections 4.4 and 4.5, we respectively describe the development process and the resulting artifact. Subsequently, we apply (Section 4.6) and evaluate (Section 4.7) the artifact in a medium-size German automation company. The results of the evaluation are then discussed in the light of their contribution to the initially stated problem in Section 4.8. Finally, in the Section 4.9, we wrap up with a conclusion on our contribution and an outlook for further research.

### 4.2 Related work

In this section, we present scientific work relevant for our research. We have no intention to conduct a comprehensive literature review on knowledge acquisition methods and frameworks. Rather, we seek to exemplarily demonstrate relevant scientific directions and gaps in the field. It subsequently allows us to derive our own research objectives in Section 4.3.

#### 4.2.1 Relevant Scientific Frameworks

Von Rueden et al. [77] introduce the concept of Informed Machine Learning (IML), a form of machine learning, which is based on two distinguished information sources – data and prior knowledge. On the basis of a systematic literature survey, the authors present a taxonomy of Informed Machine Learning. The taxonomy incorporates (1) different *knowledge sources*, (2) formal *knowledge representations* and (3) approaches to *knowledge integration* into ML. Analysing and comparing existing literature, von Rueden et al. [77] identify most common paths to integrate prior knowledge into ML. According to their results, the expert knowledge is commonly formalized in form of *probabilistic relations* or *human feedback*. As such, it can be then integrated into the ML pipeline via *training data*, *hypothesis set*, *learning algorithm* or *final hypothesis*. The authors high-



## Chapter 4 Domain knowledge acquisition for ML using ChatGPT

light that *human feedback* and *probabilistic relations* are most frequently integrated into the *learning algorithm*, while integration via *training data* or *hypothesis set* seems to be rather rare. This paper is informed by the work of von Rueden et al. [77] and applies the definition of Informed Machine Learning to describe machine learning based on data and expert knowledge. When designing our novel method for expert knowledge acquisition, we refer to the IML taxonomy to explicitly name, how the knowledge can be integrated into the ML pipeline. However, the main focus of our work lies on the dimension, which is not explicitly present in the IML taxonomy, namely, *how* expert knowledge can be acquired and formalized. In this regard, we contribute to the IML framework with a novel method which facilitates the formalization of expert knowledge.

Another framework we want to highlight in our work, is the Cross Industry Standard Process for Data Mining (CRISP-DM) [79], which has originated in early 2000s and has become a standard process model for data mining projects [32]. CRISP-DM is a generic framework in terms of application domain and describes a path to transform real-world problems into DM or ML tasks [32]. It consists of six successive steps – (1) *Business (Domain) Understanding*, (2) *Data Understanding*, (3) *Data Preparation*, (4) *Modeling*, (5) *Evaluation* and (6) *Deployment* – which can be repeated till a satisfactory result has been achieved [79]. With our work, we do not aim to modify the CRISP-DM process, rather we apply it to derive and systematically present the contributions of our work. Specifically, our novel method for domain knowledge acquisition seeks to automate the steps of *Domain-* and *Data Understanding* and to deliver ready-to-use formalized input for the step of *Data Preparation*. The former two steps are commonly known to require active participation of a data scientist [32, 57, 79] and can be the most resource-consuming part of a DM project [32]. Therefore, our novel approach strives to free data scientists' capacity for other steps like *Data Preparation* or *Modeling* and make a project cheaper and more efficient in general.

Though we design our envisioned method for automated knowledge acquisition with ChatGPT in most generic form – such that it can be applied both by individuals and organizations – it can be helpful to embed it in the organizational context in order to have a more structured view on methods' contributions. Thus, in our paper we refer to the research on knowledge intensive processes (KIPs) – a field on the edge between the knowledge management (KM) and the business process management (BPM) [55]. KIPs require performance of knowledge intensive tasks and are knowledge-, information- and data-centric [74]. Successful management of KIPs is especially important for organisations

## Chapter 4 Domain knowledge acquisition for ML using ChatGPT

which are active in knowledge-intensive domains like project management, emergency management, health care etc. [15]. Retrieving the knowledge required for successful KPI management is not trivial, since this knowledge often exists in undocumented or unformalized form [15]. Thus, we argue that our method enables documentation and formalization of the valuable domain knowledge, which, once documented, can be seen as a valuable knowledge source for KIPs beyond ML projects.

### 4.2.2 Exemplary Knowledge Acquisition Methods

As mentioned above, the *knowledge acquisition bottleneck* describes the difficulty and high cost of domain knowledge acquisition in IS research [41]. IS scholars strive to tackle this bottleneck by developing methods which make it easier for practitioners to acquire and integrate domain knowledge into their applications. For example, Seymoens et al. [70] highlight the lack of scientifically grounded action protocols, which would guide practitioners through the knowledge acquisition. Subsequently, the authors propose and effectively test a methodology based on a series of qualitative co-creation workshops, involving data scientists, social scientists and domain experts. Though Seymoens et al. [70] explicitly mention high efficiency and low investment among the methodology design objectives, it prescribes the involvement of several data scientists, social scientists and domain experts in multiple workshops. Such involvement requires simultaneous availability and significant time investment for all stakeholders, which could be a hurdle for companies with no or limited access to data science or social science resources. Indeed, Merkelbach et al. [57] explicitly highlight the scarcity and the expensiveness of data science expertise – a fact which makes it difficult for small and medium-sized enterprises (SMEs) to apply machine learning. The authors propose to tackle this challenge by training domain experts to implement ML applications on their own. Specifically, domain experts go through a novel version of the CRISP-DM, while being guided and trained by a data science coach. In the end, domain experts are expected to be able to implement simple data analytics applications without the support of a data scientist. As the authors themselves point out, this approach seems to be effective in terms of enabling domain experts to independently develop ML applications. However, the initial investment of time and effort for the domain expert training remains high [57]. Gill et al. [32] also work on the modification of the CRISP-DM, adding the steps of expert-centric ontology design to the framework. The authors aim to facilitate the creation of reusable ontologies, thus making follow-up projects more efficient. As concluded by the authors, reusing the

## **Chapter 4 Domain knowledge acquisition for ML using ChatGPT**

knowledge ontologies can indeed increase the efficiency, however their creation and integration into the process requires significant initial effort [32]. Besides novel workshop methods and process enhancements, scholars also seek to improve knowledge acquisition through automation. According to Hoffman et al. [41], automated knowledge acquisition tools can be more time-efficient, less dependent on human skill and able to directly deliver a formalized output. However, such tools can be very different in design, user-friendliness etc., and their effectiveness has to be evaluated individually [41]. For example, Park et al. [62] propose an interface for domain knowledge acquisition in NLP applications. The interface is designed to be more time-saving and easy-to-use for both domain experts and data scientists, compared to iterative interviews or data labeling tools. According to the authors, the interface allows domain experts to identify the key domain concepts in the textual data. Additionally, it elicits the reasoning which domain experts apply to justify a label of a data instance. As reported by the authors, the interface proves to both provide a valuable input for an NLP model and to keep the cognitive load and the stress levels of domain experts low [62]. Though the interface by Park et al. [62] seems to provide an improvement in domain knowledge acquisition for textual data, its applicability to other fields and types of data is unclear. These examples are aimed to demonstrate the variety of approaches to tackle the knowledge acquisition problem – from exact action protocols over training and empowerment strategies up to automated tools and much more. Review or comparison of all existing methodologies are out of scope of this paper. However, to the best of our knowledge, there have been no scientific publications explicitly proposing to use conversational agents in general and ChatGPT in particular to automate the knowledge acquisition process.

### **4.2.3 Conversational Agents and ChatGPT**

Diederich et al. [21] conduct a comprehensive literature review to organize and assess the research on conversational agents. The authors highlight that the majority of the studies on the application of CAs focus on customer service, marketing or sales contexts. Subsequently, they propose a new research direction – to find further types of tasks which can be effectively automated with CAs. We argue that interview automation with an explicit purpose of knowledge extraction can be a good candidate for a new task type, which was not yet explicitly discussed in the literature. Indeed, another literature review by Allouch et al. [2] provides a CA classification based on their application, featuring different tasks (e.g., question answering, entertainment, advertising etc.) and domains (e.g., business,

## Chapter 4 Domain knowledge acquisition for ML using ChatGPT

healthcare, education), while not explicitly mentioning question-asking among CA application fields. While a detailed discussion on the CA classification and frameworks is out of scope of this paper, we would like our paper to encourage CA researches to explicitly think and discuss the opportunity to apply CAs for knowledge extraction, interview automation or other applications, where asking questions might be more relevant than generating answers.

As mentioned earlier, the field of conversational agents is currently being revolutionized by the technological advancements in the field of LLM in general and ChatGPT in particular [5]. The term LLM describes a class of language models which incorporate billions of parameters and are trained on a vast amount of textual data [81]. LLMs are capable of generating human-like language and solving a broad range of NLP tasks [27, 37]. State-of-the-art LLMs base on generative pretrained transformer (GPT) architecture [81], with ChatGPT by OpenAI being the flagman example of this technology [51]. Scientific world currently sees an explosion of LLM applications in plentiful domains - medicine, education, law, finance, engineering and much more [27, 37, 51, 81]. Liu et al. [51] in their comprehensive literature survey categorize ChatGPT-related research according to the application field – e.g., question answering, text classification, text generation, code generation etc. However, among those categories and referenced papers there have been, to the best of our knowledge, no examples of adopting ChatGPT as an interviewer for the purpose of extracting and formalizing human expert knowledge. Also the comprehensive LLM literature reviews by Fan et al. [27], Hadi et al. [37], Hou et al. [43] and Zhao et al. [81] did not mention such LLM application. Li et al. [50] evaluate the performance of ChatGPT in information extraction tasks, however they focus on ChatGPT general ability to extract structured information from unstructured text, rather than on possible applications of such capabilities.

Based on the presented literature and to the best of our knowledge, we conclude that our proposal to use ChatGPT to conduct interviews with domain experts in natural language is a novel approach, which has not been yet explicitly highlighted in the literature on conversational agents, LLMs and ChatGPT.

### 4.3 Research Objectives and Contribution

Following our motivation (see Section 4.1) and based on the existing knowledge base (see Section 4.2), we aim to design a novel method for an interview-based domain knowledge

## Chapter 4 Domain knowledge acquisition for ML using ChatGPT

acquisition with ChatGPT. In the first place, this method should enable the data scientist to automate the knowledge acquisition tasks during the *Domain-* and *Data Understanding* steps of CRISP-DM [79]. By effectively doing so, it should increase the efficiency of these steps, i.e., reduce the required time and effort on the side of the data scientist. Considering the lack of data science resources on the labor market [6], this efficiency gain aims to free a part of data science capacities, which can be then invested elsewhere. Obviously, an actual efficiency improvement is only possible, when the quality of the extracted knowledge is such that it can be effectively integrated into an ML pipeline via *training data*, *hypothesis set*, *learning algorithm* or *final hypothesis*, as described by von Rueden et al. [77]. Moreover, since in our method the data scientist is replaced by a conversational agent, domain experts may be less willing to engage in a interview with the ChatGPT Interviewer than with a data scientist. The broadly established research on user technology acceptance [17, 75, 76] argues that the advantages of new technologies can only be realized when users accept and use the technology. Research has identified multiple factors which positively influence user acceptance of information technology, among those the *Perceived Ease of Use* [17]. The *Perceived Ease of Use* describes user's perception about the physical and mental effort associated with the use of a system [17]. We argue that our method should make the interview participation as easy as possible for domain experts in order to increase their acceptance for the ChatGPT Interviewer. Last but not least, an extensive research on human-centered artificial intelligence (HCAI) argues that human safety and well-being should be placed in the center of AI design and application and that meaningful human control should be ensured [72]. When designing our method, we follow the vision of HCAI and explicitly empower domain experts to steer the conversation with ChatGPT and to control the interview output in form of formalized knowledge. Based on these elaborations, we formulate the following research objectives to be fulfilled by our method (ROs):

- RO1: Automation – automate the tasks of leading an interview, extracting and formalizing relevant information with ChatGPT;
- RO2: Efficiency – ensure that the method is efficient for all stakeholders in terms of time and effort;
- RO3: Effectiveness – ensure that method outputs can be effectively used in subsequent model building steps.

## Chapter 4 Domain knowledge acquisition for ML using ChatGPT

- RO4: Usability – ensure that participating in the interview is simple and comfortable for domain experts;
- RO5: Human-centricity – ensure that domain experts can influence the course of the interview and control the quality of its results;

Following the DSR methodology, we additionally instantiate the method in a prototype for practical evaluation. The results of the evaluation should be used to assess the fulfillment of the research objectives and to derive contributions to the existing knowledge base. Specifically, with the stated research objectives, we seek to make multiple contributions to the current research. First, we respond to the problem of scarce data science resources by proposing a way which could optimize a part of data scientists' tasks and save their valuable capacity for actual development. Doing so, we also contribute to lowering a hurdle for SMEs to venture into the world of ML. Designing a new method for knowledge acquisition we also add to the research on knowledge management in general and knowledge intensive processes [55] in particular. Moreover, we contribute to the literature on conversational agents, LLMs and GhatGPT by proposing and testing a new application field – knowledge acquisition through natural language interview. Last but not least, we map our results to the existing frameworks of IML [77] and CRISP-DM [79] and therefore contribute to structured scientific discussion on knowledge acquisition in ML.

### 4.4 Design & Development

As stated by our research objectives, we aim to design a novel method for an interview-based domain knowledge acquisition with ChatGPT. Additionally, we instantiate the method in a prototype. Both the method and the prototype have been developed in several iterations – an initially envisioned method has been prototyped and tested, the test results have been used to make adjustments to the method, which has resulted in a new prototype and so on – till the research objectives have been sufficiently fulfilled. In this section, we briefly describe the development iterations, incl. the history of relevant design decisions. In all iterations, we apply the private edition of ChatGPT, the first iteration has been performed with ChatGPT 3.5 and the subsequent ones with ChatGPT 4.0.

Table 4.1 summarizes the development iterations, featuring the outcome of each iteration, the subsequent design decisions and their contribution to our research objectives. In the first iteration, we verify the underlying assumption that ChatGPT can perform an

## Chapter 4 Domain knowledge acquisition for ML using ChatGPT

Table 4.1: Design Iterations

Iteration	Participants & Roles	Outcomes & Design Decisions
(1) Test of ChatGPT ability to lead an interview	Author performs the test; Data Scientist evaluates resulting dialog	Use instructions instead of chat commands (RO1); Create a question catalogue to make the interview more deterministic (RO1)
(2) Test of ChatGPT instructions with a question catalogue	Author develops the question catalogue in consultation with Data Scientist & performs the test; Data Scientist evaluates resulting dialog	Simplify interview language (RO4); Add questions on representative data and success criteria of the ML project (RO3); Create guidelines for domain expert interaction with ChatGPT (RO4, RO5)
(3) Final prototype evaluation	Author performs the evaluation; Domain Experts participate in the interview; Data Scientist assesses results	See Section 4.7

In the column “Outcomes & Design Decisions” research objectives corresponding to specific design decisions are stated in parentheses.

interview. We start by inserting differently formulated commands into the chat interface of ChatGPT. The commands instruct ChatGPT to enquire relevant information for a particular problem and to put the answers into a formalized model. We then discuss the resulting chat dialogues with the data scientist from our demonstration case (see Section 4.6). The result of the first iteration confirms that it is generally possible to make ChatGPT lead an interview and formalize extracted knowledge. However, applying chat commands turns out to be unreliable and nondeterministic in terms of stated interview questions and their formalized output. In the second iteration, we attempt to minimize this drawback by applying custom instructions instead of chat commands and also by using ChatGPT 4.0 instead of 3.5. We precisely formulate the desired content and outcome of the interview and insert it into the custom instructions field of the ChatGPT interface. Following the adjustment, ChatGPT can repeatedly lead an interview with the desired content and outcome. While there still remains some room for nondeterministic suggestions and deviations in formulations, the results of different sessions become much more comparable

## Chapter 4 Domain knowledge acquisition for ML using ChatGPT

and ChatGPT states all required questions in all sessions. Consequently, we recognize *Instructions* for ChatGPT to be a necessary element of our method and add this element to the method overview (see Section 4.5). After the second iteration, the outcome of the ChatGPT dialog is positively evaluated by the data scientist, because it contains all the relevant information in a well formalized form. However, final adjustments need to be made, in order to make the interview easy and comfortable for domain experts. Specifically, one can not expect that domain experts are, to any extent, familiar with ChatGPT and they might have difficulties communicating with it. To address this risk, we add *Guidelines* for domain experts to our method. The *Guidelines* are supposed to explain domain experts the general purpose of the interview, inform them about ChatGPT functionality and advice them how to successfully communicate with it. After adding the *Guidelines* and making further small adjustments (see Table 4.1), we implement the final prototype and evaluate it in a German automation company (see Section 4.7).

### 4.5 Method & Prototype

In this section we present the result of the design iterations, described above. We first outline the method and then its specific instantiation in a prototype. We point out that the method has a more generic character than its instantiation. According to Hevner et al. [40], methods are processes, which provide guidance for a problem solution, while instantiations show, how a method can be applied in a working system and enable a concrete assessment of the method effectiveness. In other words, our proposed method represents a generic solution to the knowledge acquisition problem, while the prototype instantiation describes a concrete implementation example for the single elements of the method.

#### 4.5.1 Method for Domain Knowledge Acquisition with ChatGPT

We present our novel method for domain knowledge acquisition in form of a graph (see Fig. 4.1). The graph depicts the main elements of the method, structured into inputs, process and outputs. Additionally, we outline the contribution of the method to the existing frameworks of CRISP-DM [79] and IML [77].



## Chapter 4 Domain knowledge acquisition for ML using ChatGPT

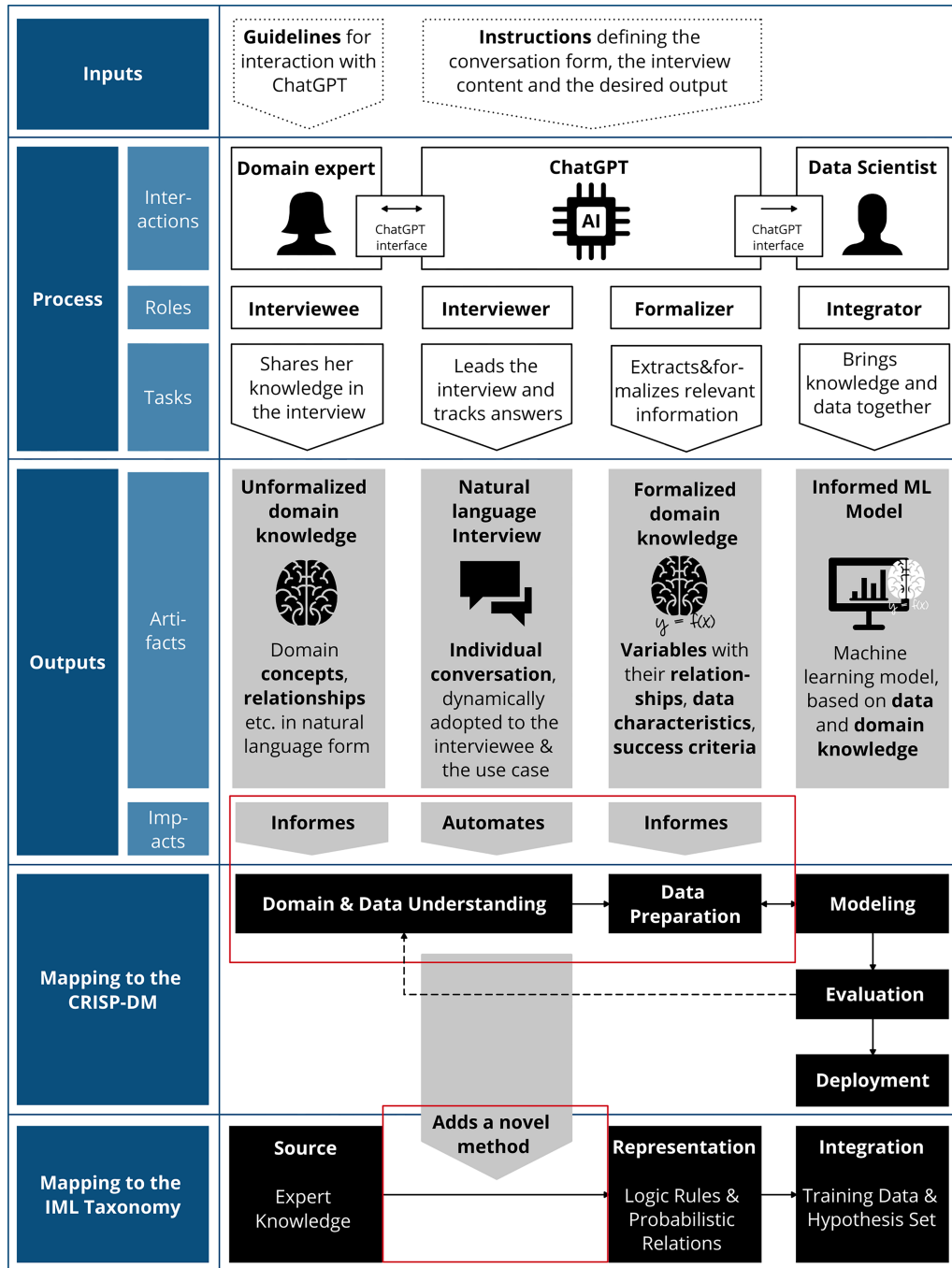


Figure 4.1: Method for Domain Knowledge Acquisition with ChatGPT. Blue boxes reflect the underlying structure of the method. White and grey boxes depict the main elements of the method with their interactions. Black boxes represent the main concepts of the frameworks CRISP-DM [79] and IML [77], while the the red frame highlights the contribution of the method to those frameworks

## Chapter 4 Domain knowledge acquisition for ML using ChatGPT

### Inputs

Our method requires two main inputs – *Guidelines* for domain experts and *Instructions* for ChatGPT. The *Guidelines* are aimed to improve the interaction between a domain expert and ChatGPT. Though ChatGPT is becoming increasingly popular among general public, one can not assume that domain experts are familiar with it. Thus, the *Guidelines* provide a basic guidance to domain experts, helping them to feel more comfortable during the interview with ChatGPT. Moreover, they inform domain experts about the possibility to ask comprehension questions, require advice or correct ChatGPT. Therefore, the *Guidelines* play a key role for the research objectives usability (RO4) and human-centricity (RO5). The *Instructions* for ChatGPT allow to steer it and make it more deterministic. ChatGPT offers an interface for custom instructions – information about how and what ChatGPT is expected to do. In our method, *Instructions* play a crucial role in transforming ChatGPT into a reliable interview agent. Through *Instructions*, we make ChatGPT to follow a defined interview protocol in a desired manner and to subsequently bring the acquired knowledge into a formalized form. Thus, they meaningfully contribute to the research objectives automation (RO1) and effectiveness (RO3). The *Instructions* can also be used to adopt the conversation to the needs of domain experts, therefore influencing usability (RO4) and human-centricity (RO5).

### Process

The process within our method incorporates interactions between the agents, their main roles and tasks. The domain expert takes over the role of an interviewee and shares her knowledge with ChatGPT, by answering its interview questions in natural language. ChatGPT plays the role of an interviewer – it poses questions, tracks answers, makes suggestions, etc. – generally guiding the domain expert through the interview based on its *Instructions*. Further, ChatGPT takes over the task to extract and formalize relevant information from domain expert answers, thus building a knowledge model. Last but not least, the data scientist takes over the formalized knowledge model and integrates it into the ML pipeline. Notably, there is no direct human-to-human interaction between the data scientist and the domain expert, this interaction being replaced by the ChatGPT interface. The replacement of data scientist through ChatGPT in the knowledge acquisition process is crucial for the research objectives automation (RO1) and efficiency (RO2).

## Chapter 4 Domain knowledge acquisition for ML using ChatGPT

### Outputs and their Contribution

Our method produces multiple artifacts as outputs, with their valuable contributions to the Cross Industry Standard Process for Data Mining [79] and the Informed Machine Learning [77] frameworks. First, the ChatGPT dialogue contains the unformalized knowledge of the domain expert, formulated in natural language. Though in the end we are primarily interested in the formalized knowledge, the ChatGPT dialogue can serve as an interview protocol, providing additional information on the domain, the problem and the whole setting. Therefore, this is a valuable output, which informs the *Domain-* and *Data Understanding* phase of the data mining process and helps the data scientist to interpret the formalized output later on. This output therefore contributes to the effectiveness research objective (RO3).

The natural language interview with ChatGPT is a valuable output artifact. It serves our main research objective to shift interview tasks over to ChatGPT (RO1). This artifact makes the the *Domain-* and *Data Understanding* phase of the CRISP-DM [79] more efficient through automation of data scientist role during the knowledge acquisition. Besides, it adds a novel way to formalize the expert knowledge, facilitating the path from knowledge source to knowledge representation in the IML framework [77]. Specifically, the interview as an output artifact encompasses the entire course of the conversation, led by ChatGPT – questions, feedback on domain expert answers, given clarifications, suggestions etc. The course of the interview is influenced by ChatGPT *Instructions*, but it is also dependent on the responses of the domain expert and on ChatGPT own processing. Therefore, though the interview follows the pattern defined in the *Instructions*, it is not entirely deterministic. Rather, it is uniquely and dynamically adopted to the use case and the individual behaviour of the domain expert. This makes this artifact different from questionnaires or other rule-based approaches.

The formalized knowledge as an output artifact contains information about the domain in a formalized form. An example can be a function, containing the goal parameter and the influence parameters, mathematical or logical rules, describing the relations between the parameters, known probabilities, limit values, time lags etc. This formalized knowledge corresponds the formal knowledge representation in the IML framework [77]. As such, it can be used by the data scientist to understand the domain and the data (see the *Domain - and Data Understanding* of CRISP-DM [79]). Subsequently, the data scientist can integrate the acquired domain knowledge into the ML model (see the *Data Preparation* and *Modeling* steps of CRISP-DM [79]). The quality of the formalized output

## Chapter 4 Domain knowledge acquisition for ML using ChatGPT

determines, how well it can be integrated into the ML pipeline, therefore playing a key role for the research objective effectiveness (RO3).

Notably, the three mentioned outputs, documented together in the the interview transcript, can be seen as a valuable resource for knowledge intensive process management (see Section 4.2).

Last but not least, the integration of the formalized knowledge into the ML pipeline results in an Informed Machine Learning Model, as defined by von Rueden et al. [77]. According to the IML taxonomy [77], the integration can take place with four different approaches – via *training data* (e.g., feature engineering), *hypothesis set* (e.g., definition of hyper-parameters), *learning algorithm* (e.g., loss function modification) or *final hypothesis* (e.g., validation of model predictions). A choice of a fitting approach can strongly vary with the use case, the data set, the form of knowledge representation etc. We note that the discussion on different integration approaches is out of scope of this paper. However, our method aims to facilitate a successful deployment of an IML model in the end of the ML project. In this regard, domain experts may also have their own, domain-specific, measurements for a successful ML implementation – for instance, an increase in manufacturing quality, a reduction of maintenance costs, etc. This particular piece of knowledge can help the data scientist to optimize the model to the needs of domain experts. Therefore, we include questions on the success criteria of the project into the ChatGPT interview (see Subsection 4.5.2).

### 4.5.2 ChatGPT Interviewer

In this subsection, we describe the final instantiation of our knowledge acquisition method in a prototype, which has been subsequently evaluated in a German automation company (see Section 4.7). We follow the structure of our method (see Fig. 4.1) to describe single elements of the prototype, specifically the inputs and the process, since the outputs are in detail described in Section 4.6. We call our prototype the ChatGPT Interviewer.

#### Inputs

The underlying elements of the prototype are the domain expert *Guidelines* and ChatGPT *Instructions*. In our evaluation use case, the *Guidelines* consist of ten simple passages, aimed at enabling a smooth and easy conversation between the domain expert and ChatGPT (see Table 4.2). While guidelines 1 to 4 mostly focus on making the domain expert

## Chapter 4 Domain knowledge acquisition for ML using ChatGPT

feel more comfortable with ChatGPT (RO4), the guidelines 5 to 9 explicitly inform the domain expert about the opportunity to state questions, ask for advice, steer the conversation and adjust ChatGPT output (RO5). Last but not least, the guideline 10 ensures that the domain expert sticks to the primary goal of the interview (RO3).

Table 4.2: Domain Expert Guidelines for Interaction with ChatGPT

No	Content
1	Do you have a problem which you would like to solve with machine learning? The ChatGPT Interviewer questions you about your problem and processes your knowledge in such a way that a data scientist can use it to train a machine learning model. (RO4)
2	To start the interview, a short greeting such as “Hello” is sufficient. (RO4)
3	The chat GPT Interviewer will guide you through the interview with questions and collect the information needed to create a mathematical model for your problem. Your expert knowledge and experience are the most important factors here. (RO4)
4	Just be yourself and answer the questions as if you were talking to one of your colleagues. (RO4)
5	If you don’t understand a question, just ask – ChatGPT will explain or rephrase it. (RO5)
6	You can control the conversation as if you were talking to a human being. For example, if there are too many questions at once, you can say something like “one at a time please”. Or if a question is too general, just say “I don’t know where to start...” – and ChatGPT will adapt. (RO5)
7	You can also ask ChatGPT to help you – for example, if you don’t know, how to measure a certain influencing factor, you can ask it what options or measurement methods are available. (RO5)
8	ChatGPT can also, at your request, improve your answers with suggestions – e.g., suggest other influencing factors that could play a role in your problem. (RO5)
9	Once the final model has been formulated, you can adjust it until you are satisfied. To do this, simply tell ChatGPT what to change and why, as if you were explaining it to your colleague. (RO5)
10	During the interview, it is advisable not to deviate much from the topic and not to <i>experiment</i> with ChatGPT. As in a human conversation, it will only distract from the main point and negatively affect the outcome. Behave professionally and remember that the better you describe your problem and your knowledge to ChatGPT, the better the ML model will be later on. (RO3)

The corresponding research objectives are stated in parentheses. Original German text has been translated into English by the author.

The ChatGPT *Instructions* are developed to fit a broad number of machine learning use cases from different domains (see Table 4.3). First, we formulate a generic task

## Chapter 4 Domain knowledge acquisition for ML using ChatGPT

for ChatGPT – eliciting a problem with questions and producing a formalized model in the end (instruction 1). Then, we specify, which questions have to be stated, while explicitly allowing ChatGPT to make own suggestions (instruction 2). The questions are chosen based on the conversations with the data scientist, as described in Section 4.4. Further, we explicitly add a question on the success criteria of the ML project at hand, as suggested by our method design (instruction 3). In the instruction 4, we make sure that ChatGPT produces a formalized knowledge model based on answers from the domain expert. Last but not least, we give ChatGPT instructions on how to communicate with domain experts (preferable conversation language and style), in order to account for domain experts’ needs (instruction 5). Table 4.3 also points out, how single instructions correspond to our research objectives. All instructions are formulated in the first person (as if domain experts would write them themselves), since they inform ChatGPT about the needs of the user, with whom it directly interacts.

### Process

In our final prototype, the process of knowledge acquisition is implemented using ChatGPT 4.0 as private edition. In our case, a ChatGPT account is first prepared using custom instructions, as described above. Then, separate sessions are created for different domain experts. Domain experts interact with ChatGPT per chat interface within their respective session. The sessions are then exported anonymously and provided to the data scientist for evaluation. We point out that our implementation of the process is, in the first place, a proof of concept for the proposed method. We acknowledge that it can be further automated and technically improved, see Sections 4.8 and 4.9 for an outlook.

## 4.6 Prototype Demonstration

The ChatGPT Interviewer has been exemplary applied in a middle-sized automation company based in Germany. Specifically, two application cases for machine learning have been chosen – (1) compressed air consumption optimization and (2) plastic parts quality control. Two domain experts – one per application case – took part in the demonstration. Additionally, a data scientist from the ML department was involved. All participants were male and older than 30 years. Demonstration took part on three separate appointments, one for each of the three participants. Appointments with the domain experts took place in person. During their respective appointments, domain experts received printed *Guidelines*

## Chapter 4 Domain knowledge acquisition for ML using ChatGPT

Table 4.3: ChatGPT Instructions

No	Content
1	I am an expert in my field of work. I want to apply machine learning for a particular problem. Elicit my field of work, my problem and information relevant for the problem with questions. Specifically, interview me and use my input to define a mathematical model or a set of logical rules, in which: (1) the depending variable is what I want to achieve, (2) independent variables are influential factors and their relations. (RO1, RO3)
2	When eliciting the relationship between the dependent and independent variables, ask: (1) how the variables can be measured or quantified? (2) are the variables continuous or categorical? (3) are there any known rules or relations between the variables? (4) what are typical example situations from my experience? (5) are there any additional context factors? (6) are there any time gaps between the change in the independent variable and the change of the dependent variable? (7) which data is available and is it representative? If you have more questions which help eliciting the problem, ask them. You may propose your own improvements to the model, additional variables or relations. (RO1, RO3)
3	Ask, how I would measure in the end of the project, if my problem has been solved by machine learning or not. (RO1, RO3)
4	In the very end, when you have all the needed information from me, provide me with a formalized mathematical model, which I can give to my data scientist to train the machine learning model. If you have final suggestions, share them too. (RO3)
5	Speak German. Use simple terms. Ask questions strictly one by one and provide information in small portions. Support questions with examples. (RO4)

The corresponding research objectives are stated in parentheses.

and a computer with their individual ready-to-use ChatGPT Interviewer, as described in Section 4.5. They had as much time to read the *Guidelines* as they needed and could ask comprehension questions to the author. Afterwards, domain experts were instructed to begin the interview with ChatGPT by typing “Hello” into the chat. During the ChatGPT interview, the author remained in the room to be able to answer questions or offer technical assistance, but did not interrupt otherwise. As soon as the interview was finished, domain experts were asked to give feedback in free form. Their feedback was noted by the author (see Section 4.7). Then the appointments were closed and domain experts were thanked for their participation.

Following the meetings with the domain experts, a remote appointment with the data scientist took place. Beforehand, the data scientist was provided with anonymized domain expert interviews. He was asked to read the interviews before the appointment and note

## Chapter 4 Domain knowledge acquisition for ML using ChatGPT

the time needed to process the content of the interviews. During the appointment, the data scientist was asked to provide feedback in free form. The feedback mainly focused on the utility of the information contained in the interviews (see Section 4.7). We note that our work has a primary focus on knowledge acquisition and not on model building. Therefore, the data scientist was asked to logically argue, if and how the knowledge acquired with ChatGPT could be used in an ML pipeline, without actually building the models. We consider such feedback to be sufficient to evaluate our prototype (see Section 4.7).

### 4.7 Prototype Evaluation

In this section, we evaluate the success of the ChatGPT Interviewer based on the content of the two interviews (see Appendix for original interview protocols) and on the feedback, collected during the prototype demonstration (see Table 4.4). We discuss, whether our research objectives are fulfilled.

#### 4.7.1 RO1: Automation

The automation research objective deals with the very fact of automating the tasks of the human interviewer, in our case a data scientist, during the knowledge acquisition process. Based on our demonstration cases, we can tell that such automation worked successfully in two application domains – compressed air consumption optimization and plastic parts quality control. We were able to shift the tasks, such as leading the interview, extracting and formalizing the knowledge, onto the ChatGPT Interviewer. The interview took place in natural language and followed the course, defined by the ChatGPT *Instructions*. Importantly, though following the same pattern, each interview has been individual. The context of questions, elaborations, suggestions and examples provided by ChatGPT varied with the application case. Moreover, ChatGPT adapted its responses to the behaviour of the domain experts – whether they gave brief or detailed answers, asked comprehension questions or required an advice. Both domain experts perceived ChatGPT to be precise, professional and competent. They reported the questions and examples to be well chosen and helpful. Though the two application domains were rather different, both domain experts felt that their interviewer knew, how to state right questions and provide good examples. Notably, the domain expert from the quality control case mentioned that the precision of the interview results depended on the details, which the domain expert was



## Chapter 4 Domain knowledge acquisition for ML using ChatGPT

Table 4.4: Feedback on the ChatGPT Interviewer

Participant	Feedback Notes
Domain Expert 1	Overall very positive experience; I had the feeling that the interview was very specialised, the questions made sense and were well chosen; I was guided well through the interview; questions were understandable, no follow-up questions were necessary; the information amount was well portioned; jumping in the interview without preparation worked well, no additional guidance or explanations are needed; I had the confidence that the result was correct and in line with my input; I had a sense of control over the situation; I appreciated very detailed questions and precise answers; the conclusion made a lot of sense.
Domain Expert 2	The result depends very much on how precisely you can answer the questions yourself – the more precise the formulation, the more precise the output (e.g. <i>pressure</i> vs. <i>injection molding pressure</i> ), also people’s willingness to communicate plays a role; I had problems getting started at the beginning, first interaction felt more difficult than with humans, but it became very easy afterwards; ChatGPT is a helpful sparring partner to sort out own thoughts, might be even better than talking to a colleague; ChatGPT has enough knowledge to be a valuable conversation partner (at least I had such a feeling); examples were helpful, the wording was understandable; I asked one question during the interview and one in the end; at the end I had the feeling that my problem was understood; an outlook would be helpful (what will happen with my knowledge?).
Data Scientist	It is great and surprising, what ChatGPT can collect; the answers are very sensible and professional, ChatGPT gives you a feeling of having understood everything; domain experts can sort things out for themselves; domain experts deal with the topic and structure their thoughts, this saves time for the initial interview; the interview also makes visible what domain experts do not know yet – perhaps they are not yet ready for a discussion or the project – it spares us wasted time due to poor preparation on the part of the domain experts; after the interview, the data scientist does not conclusively know everything to build the model and the user does not have all the information to provide the data – they need to talk to each other at least once; follow-up questions can be quickly clarified afterwards; the knowledge summary is a valuable result; the chat history helps to interpret things and acquire more detail – going over it was also very quick; An advice on how the user can acquire the data and what a sample data set might look like can be helpful; it might help to present an example file for the data; it should be also possible for the domain expert to stop the interview or to interact with the chat over longer time.

Expert 1 represents the compressed air consumption case, Expert 2 – the plastic parts quality control case. The formulations base on author’s notes. Original German text has been translated into English by the author. Small textual adjustments has been made for better readability.

able or willing to provide (e.g. *pressure* vs. *injection molding pressure*).

To sum up, the ChatGPT Interviewer was able to successfully perform two knowl-

## Chapter 4 Domain knowledge acquisition for ML using ChatGPT

edge acquisition interviews in two different application domains. Though there might be potential for further automation (see Section 4.8), our prototype implementation satisfies the requirements stated by RO1.

### 4.7.2 RO2: Efficiency

As mentioned by the second research objective, we aim our prototype to be efficient in terms of time and effort. The domain expert from the air consumption case spent as little as 15 minutes on the interview, including the reading time for the *Guidelines*. The domain expert from the quality control case spent around 30 minutes for the same activities. Therefore, our prototype demonstration showed that domain experts needed less than one hour for the ChatGPT interview. Most importantly, the ChatGPT interview did not require any data scientist's involvement. Therefore, his effort to perform the interview tasks was effectively reduced to zero. The time and effort of the data scientist, needed to process the interview results, were also reduced. The entire interview was available in digital form and did not have to be transcribed. Moreover, the content of the interview was summarized by ChatGPT into a ready-to-use domain knowledge model. In our demonstration case, the data scientist needed 15 minutes to evaluate both interviews and verify the resulting knowledge models.

Therefore, we conclude that our method for domain knowledge acquisition with ChatGPT can be very time efficient, especially in comparison with workshop-based methods (see Section 4.2).

### 4.7.3 RO3: Effectiveness

As defined by our method, the data scientist is the main consumer of the extracted domain knowledge and uses this to understand the domain and the data (see the *Domain - and Data Understanding* of CRISP-DM [79]). In our evaluation case, the domain knowledge was formalized as a set of variables with their specific types (categorical vs. continuous) and relationships (e.g. direct vs. indirect influence, influence time lag etc.). The data scientist generally described the ChatGPT Interviewer as useful, professional and able to collect valuable information. He highlighted that ChatGPT seemingly *understood* the use cases of both domain experts. Further, according to the data scientist, the formalized knowledge model was the most valuable result of the interview, while the detailed interview content helped to better understand the problem and to interpret the knowledge model.

## Chapter 4 Domain knowledge acquisition for ML using ChatGPT

The data scientist highlighted two main advantages of the ChatGPT Interviewer. Firstly, it saved data scientist's time for conducting an initial interview and simultaneously allowed domain experts to structure their thoughts. Though some questions still remained unanswered after the ChatGPT interview (e.g., the exact data structure), the effort to clarify them would be very low – so the data scientist. Secondly, the ChatGPT interview revealed whether domain experts were uncertain about their goals or did not have any representative data for their project. According to the data scientist, such evidence can save hours of unproductive discussions and enable him to focus on mature projects. To improve the prototype, the data scientist suggested to include examples of data sets and recommendations on data acquisition into the interview. He added that in some cases domain experts might want to stop the interview and return to it later (e.g., to collect some inputs or consult colleagues).

Generally, the ChatGPT Interviewer yielded meaningful output with several benefits for the data scientist. It was able to acquire and formalize the biggest part of the relevant domain knowledge and helped the data scientist to identify immature projects. However, a brief follow-up communication with domain experts was still necessary. We acknowledge that the effectiveness of the ChatGPT Interviewer can be further improved by making its output more comprehensive and the follow-up communications unnecessary. However, within the scope of the prototype, we consider the third research objective to be fulfilled.

### 4.7.4 RO4: Usability

The domain expert from the air consumption case reported to be well guided through the interview by ChatGPT. The interview questions were formulated briefly and clearly. Provided information (the *Guidelines*) together with ChatGPT explanations was sufficient, so that no additional support or guidance was necessary. The domain expert from the quality control case reported to have had difficulties answering the very first question of the interview – formulating the problem for the ML model. He felt that describing the problem to ChatGPT was more difficult than to a human interviewer. However, after overcoming this first hurdle, the rest of the interview went smoothly. The same domain expert mentioned that it was very helpful to have an agent, which acted as a sparring partner and helped him to structure his problem with relevant questions. Moreover, ChatGPT appeared to be competent both in data science and in plastic quality control, what made the conversation even more valuable. ChatGPT was able to answer his questions, deliver examples and give advice – so the domain expert from the quality control case.

## Chapter 4 Domain knowledge acquisition for ML using ChatGPT

Altogether, both domain experts reported to understand and interact with ChatGPT easily and to perceive the whole process as meaningful and helpful. Though we do not quantitatively measure the physical and mental load on the user side as recommended by Davis [17], we deem the brief time spent on interview completion (see Subsection 4.7.3) and the domain experts' feedback to be plausible indicators of actual ease of use. Thus, we conclude that the usability objective of our prototype implementation has been sufficiently fulfilled.

### 4.7.5 RO5: Human-centricity

With the human-centricity objective we aim to put domain experts in control of the interview and its results. For that purpose, both the *Guidelines* and the ChatGPT Interviewer informed domain experts about the opportunity to do adjustments, ask questions and verify results. The domain expert from the air consumption case reported that he felt to be in control of the interview result and trusted ChatGPT to formalize his knowledge correctly. He did not feel the need to state any additional questions to ChatGPT. When verifying the resulting knowledge model, he recognized it to be correct and meaningful. The domain expert from the quality control case mentioned that he used the opportunity to ask ChatGPT comprehension questions during the interview. After observing the resulting knowledge model, he felt that ChatGPT *understood* the problem and correctly summarized all relevant aspects into a knowledge model. He also used the opportunity to ask ChatGPT for a final advice. Importantly, in the end of the interview, he expected ChatGPT to mention the outlook and the next steps of the project. We note that this requirement goes beyond pure knowledge acquisition and towards stakeholder expectation management. Nonetheless, if such information is available, it could be easily integrated into the ChatGPT *Instructions*.

We conclude that with the help of *Guidelines* and ChatGPT *Instructions* we were able to place domain experts in the loop of the interview, i.e. enable them to actively influence the flow of the interview and its results. Thus we consider the human-centricity objective to be fulfilled.

### 4.8 Discussion

In our paper, we design a novel method to acquire and formalize domain expert knowledge with ChatGPT and map our contribution to the frameworks of IML [77] and CRISP-DM [79]. We instantiate the designed method in a prototype and demonstrate it in a German middle-sized automation company. Evaluating the demonstration results, we conclude that the prototype fulfills our initial research objectives and yields a valuable efficiency gain. We also uncover valuable improvement potential which can be considered by future research.

The ChatGPT Interviewer has succeeded to replace the data scientist in the initial interview with the domain expert. However, the output of ChatGPT was not sufficient to entirely automate the *Domain-* and *Data Understanding* steps of the ML project. A brief follow-up communication between the data scientist and the domain expert was still necessary to proceed with the *Data Preparation* step. Further research may look into the possibilities to make the output of the ChatGPT Interviewer more comprehensive. For instance, the ChatGPT *Instructions* can be further extended to include exemplary data samples or advice on how to acquire representative data. Moreover, domain experts can be encouraged to consult their colleagues during the interview. Such measures would make the input of domain experts more precise and might make follow-up clarifications unnecessary. In general, as one of our evaluation participants mentioned, the precision of the resulting knowledge model strongly depends on the ability and willingness of domain experts to communicate relevant information to the ChatGPT Interviewer. Therefore, additional measures to facilitate the communication between the domain expert and the ChatGPT Interviewer can be taken, such as providing additional information about ChatGPT functionality, instructing ChatGPT to ask follow-up questions or turning the ChatGPT Interviewer into a customized chat bot with a more user-friendly interface. Indeed, in our prototype we used the private edition of ChatGPT with no opportunity for custom interface designs. Such implementation has been sufficient to demonstrate and evaluate the novel method. However, the prototype can be further upgraded using enterprise edition, which allows to build customized chat bots and integrate them into customer environments. This extended functionality could be used to further automate the ChatGPT Interviewer and make it more user-friendly.

Moreover, future research might look into legal, security and ethical aspects of replacing certain activities of the data scientist with ChatGPT. In our paper, we propose a new

## Chapter 4 Domain knowledge acquisition for ML using ChatGPT

method and evaluate its reachability with a prototype. On the base of our prototype, a productive application can be developed – for example, as a custom chat bot, as described above. In this case, also non-functional requirements such as security, legal compliance etc. must be considered.

Last but not least, in the present paper, we have demonstrated and evaluated our prototype in two different industrial use cases. Though our method and the prototype have been designed in a generic way, further tests on various use cases are necessary in order to conclude if the method can be broadly applied. We therefore encourage researchers and practitioners to apply and evaluate our method in different domains. Moreover, creating a benchmark with various knowledge acquisition methods and comparing them empirically based on a set of defined criteria might deliver additional objective insight about advantages and disadvantages of single methods, proposed in the literature. Last but not least, future research might look into the ways to apply the knowledge, documented and formalized by the ChatGPT Interviewer, in various knowledge intensive processes [55] beyond ML projects.

### 4.9 Conclusion

In machine learning projects, *Domain-* and *Data Understanding* are necessary, but challenging and time-consuming steps [79]. The challenge of domain knowledge acquisition is also known in IS research as the *knowledge acquisition bottleneck* [41]. We propose to tackle this problem by automating the knowledge acquisition with a conversational agent based on ChatGPT. Guided by our research objectives – automation, efficiency, effectiveness, usability and human-centricity – we design a method for an automated knowledge acquisition with ChatGPT. We describe each element of the proposed method and highlight its contributions to the relevant theoretical frameworks – CRISP DM [79] and IML [77]. Further, we instantiate the proposed method in a prototype – the ChatGPT Interviewer. We demonstrate and evaluate the prototype in a middle-sized automation company based in Germany. The demonstration involves two separate application cases – compressed air consumption optimization and plastic parts quality control. The prototype evaluation was successful for both demonstration cases and the prototype fulfilled the research objectives. Specifically, the ChatGPT Interviewer was able to lead an interview in natural language, track the answers of domain experts and formalize their inputs in a knowledge model. Moreover, it adapted its questions to the respective application

## Chapter 4 Domain knowledge acquisition for ML using ChatGPT

case, individually reacted to interviewee's behaviour and provided context-sensitive support. Domain experts have characterized the interview with ChatGPT as meaningful and professional. They appreciated how ChatGPT *understood* their problem, stated precise questions and delivered expert-like advice. The interview with ChatGPT helped them to sort out their thoughts and to specify their problem. The data scientist also positively assessed the interview results – the final knowledge model together with the interview protocol were valuable in understanding the domain and the problem. ChatGPT was able to replace the data scientist in the initial interview. However, the acquired knowledge was not sufficient to immediately proceed with the *Data Preparation* step of the ML project. Instead, a brief follow-up communication between the two was still required. Further automation of the prototype and extension of ChatGPT *Instructions* might minimize this draw-back and make such follow-up clarifications between data scientist and domain expert unnecessary.

We conclude that our novel method for knowledge acquisition with ChatGPT can yield a valuable efficiency gain, while preserving the human-centricity of the original interview methodology. It can significantly reduce the involvement of the data scientist, while still keeping many advantages of human-to-human interaction, like natural language, context-sensitivity and individuality. Importantly, our method proved to be efficient not only for data scientists, but also for domain experts, thus reducing the overall cost of an ML project. Moreover, the method leads to the documentation and formalisation of valuable domain knowledge, which might be used by organisations in knowledge intensive processes [55] beyond ML projects.

We acknowledge that our evaluation involved only a small amount of application examples. Further demonstration cases are required in order to draw conclusions about method's general applicability. Further, if the prototype is to be developed into a productive application, security, compliance and ethical requirements have to be considered.

Our work provides multiple contributions to the existing theory. First, we show how the fast developing technology of LLM can be used to tackle the knowledge acquisition bottleneck in ML. Based on the existing literature, we design a method for automatic knowledge acquisition with ChatGPT and prove its applicability on a prototype. Evaluating the prototype, we add empirical insights to the existing knowledge base and discuss implications for further research. With our paper, we add to the theoretical work on *knowledge acquisition bottleneck* [41], Informed Machine Learning [77] and CRISP-DM [79]. We also contribute to the research on ChatGPT applications, by uncovering and

## **Chapter 4 Domain knowledge acquisition for ML using ChatGPT**

demonstrating its potential for knowledge acquisition. More generally, our results may encourage research on conversational agents and LLM to think of interview automation and knowledge acquisition as potentially new application fields, which primarily base on asking questions instead of answering them. Practitioners can benefit from our work by receiving a practical guidance on how to acquire domain knowledge with ChatGPT. The efficiency gain provided by our method could make ML more accessible for small and medium-sized organizations and help them to exploit the potential of ML.



## **Chapter 5**

## **Conclusion**

## CONCLUSION

In our dissertation we aim to facilitate the application of AI technology through effective human-AI interaction. For this purpose, we conduct three independent studies, applying distinct methods and highlighting different aspects of human-AI interaction in various contexts. While the first two studies apply the experimental methods of behavioural sciences to investigate influential factors behind algorithm aversion and AI advice taking, our third study follows the design science paradigm and generates theoretical insight by designing an artifact and applying it to a real-world problem. Further, while our first study is designed for generic context, the second and the third studies explicitly focus on industrial applications and how effective human-AI interaction may foster AI usage in the industry. Though the results of each study have been discussed in the respective chapters in detail, in this chapter we would like to summarize the most relevant results and draw an overall conclusion about the contributions of our dissertation.

In our first study (see Chapter 2) we are the first to empirically test the proposition by Bonnefon et al. [11] to apply the widely known concepts of fast and slow thinking [16] to study algorithm aversion. In an incentivised laboratory experiment we manipulate the way, how subjects approach the task (fast vs. slow thinking) and the type of AI “thinking” (suggested by short and long AI response times, respectively). We find out that longer response times are associated with lower algorithm aversion both for fast and slow thinking tasks. We conclude that, at least to some extent, prolonging AI response times may increase AI advice taking and reduce algorithm aversion. Moreover, we report differences in subjects’ advice taking behaviour depending on their type of thinking – advice taking index in the fast thinking groups varies with the domain of the task, while in the slow thinking groups it remains constant across domains. We conclude that the way, how people approach the task – either fast or slow – seem to influence the advice taking behaviour indirectly, affecting subjects’ responsiveness to the changes of task domain. To the best of our knowledge, we are the first to provide an empirical insight on the relationship between the type of thinking and the domain of the task in context of algorithm aversion.

In our second study (see Chapter 3) we are, to the best of our knowledge, the first to experimentally study the influence of user involvement in the model training on the AI advice adherence. We design a comprehensive experimental scenario which mimics real-world predictive maintenance applications. We find out that higher grade of user involvement in the model training is associated with higher perceived control over the

resulting model and higher perceived understanding of the model functionality. Besides, subjects consider the time invested in model training to be well-spent and joyful. However, these positive changes in perceptions do not lead to behavioural changes, i.e. the advice adherence remains constant across treatments. We attribute this result to subjects' self perceptions as experts, which seem to overrule the treatment effects. We conclude that user involvement can be used to positively influence the user perceptions about the resulting model, however more research on algorithm aversion among experts is necessary to facilitate actual changes in user behaviour.

In our third study (see Chapter 4) we follow the Design Science Methodology [63] to develop a novel method for domain knowledge acquisition in ML projects. With this method, we tackle a real world problem of a *knowledge acquisition bottleneck* [41] in ML. We ground our design in existing theory on knowledge acquisition methods and AI-based conversational agents. Moreover, we map it to the existing theoretical frameworks of Informed Machine Learning [77] and the Cross Industry Standard Process for Data Mining [79]. Benefiting from the technology of large language models, the method automates the data scientists' tasks of leading the interview, tracking the answers and formalizing the acquired knowledge. This automation enables data scientists to save their time and capacity for actual development. Moreover, the method preserves the nature of a human-led interview, such as natural language or flexible adaptation to interviewee needs. By following usability and human-centricity as our research objectives, we ensure that an interviewee – a domain expert – feels comfortable during the interview and remains in control of its results. We instantiate the proposed method in a prototype, using ChatGPT – a ChatGPT Interviewer. We evaluate the prototype in two industrial applications, demonstrating its effectiveness in interviewing domain experts and formalizing their knowledge. We are able to substantially reduce the time and effort for domain knowledge acquisition. Domain experts consider an interview with the ChatGPT to be simple, enjoyable and valuable. However, the quality and detail of the extracted knowledge can be further improved. With our research, we contribute to the literature on knowledge acquisition by demonstrating how large language models (LLM) in general and ChatGPT in particular can be used to tackle the *knowledge acquisition bottleneck* [41]. Further, we add to the theoretical work on Informed Machine Learning [77] and CRISP-DM [79] by proposing a new way to acquire and formalize domain knowledge for ML projects. Our study opens up a new application field for LLM-based conversational agents in general and ChatGPT in particular. For practitioners, we offer a guidance how to automate knowledge acquisition tasks with

ChatGPT, enabling them to reduce costs of an ML project and thus lower hurdles for ML adoption.

We conclude that our work delivers valuable contributions to multiple strings of research – from empirical insights adding to the theory on algorithm aversion and AI advice taking to actual AI applications in real-world settings. Our results are also relevant for researchers and practitioners interested in design of AI agents and organisation of ML projects. For example, they can consider prolonging AI response times to facilitate user advice taking, involving domain experts during the model training to enhance users’ understanding of the model or implementing the ChatGPT-Interviewer to increase the efficiency of an ML project. In addition, our work is interdisciplinary in theoretical scope, context and applied methods. This approach enables us to benefit from a broad theoretical foundation and to consider research questions on the edge between behavioural sciences, information systems research and computer science. Thus, with our interdisciplinary contributions we deliver both theoretical and practical insights, fostering effective and beneficial human-AI interaction. Doing so, we add to the vision behind the present dissertation – to facilitate AI application and to support organisations, especially in the industry, to unfold the entire potential of AI.

# Bibliography

- [1] Ashraf Abdul, Jo Vermeulen, Danding Wang, Brian Y. Lim, and Mohan Kankanhalli. Trends and trajectories for explainable, accountable and intelligible systems: An HCI research agenda. In *Proceedings of the 2018 CHI conference on human factors in computing systems*, pages 1–18, 2018.
- [2] Merav Allouch, Amos Azaria, and Rina Azoulay. Conversational agents: Goals, technologies, vision and challenges. *Sensors*, 21(24), 2021.
- [3] Theo Araujo, Natali Helberger, Sanne Kruikemeier, and Claes de Vreese. In AI we trust? Perceptions about automated decision-making by artificial intelligence. *AI and SOCIETY*, 35, 09 2020.
- [4] Phoebe E. Bailey, Tarren Leon, Natalie C. Ebner, Ahmed A. Moustafa, and Gabrielle Weidemann. A meta-analysis of the weight of advice in decision-making. *Current Psychology*, pages 1–26, 2022.
- [5] Gaurang Bansal, Vinay Chamola, Amir Hussain, Mohsen Guizani, and Dusit Niyato. Transforming conversations with AI – a comprehensive study of ChatGPT. *Cognitive Computation*, pages 1–24, 2024.
- [6] Nadja Bauer, Lukas Stankiewicz, Malte Jastrow, Daniel Horn, Jens Teubner, Kristian Kersting, Jochen Deuse, and Claus Weihs. Industrial data science: Developing a qualification concept for machine learning in industrial production. *Archives of Data Science, Series A*, 5(1), 2018.
- [7] Katharina Beckh, Sebastian Müller, Matthias Jakobs, Vanessa Toborek, Hanxiao Tan, Raphael Fischer, Pascal Welke, Sebastian Houben, and Laura von Rueden. Explainable machine learning with prior knowledge: An overview, 2021.

- [8] Constanze Beierlein, Christoph J. Kemper, Anastassiya Kovaleva, and Beatrice Rammstedt. Short scale for measuring general self-efficacy beliefs (ASKU). *methods, data, analyses*, 7(2):251–278, 2013.
- [9] William J. Bingley, Caitlin Curtis, Steven Lockey, Alina Bialkowski, Nicole Gillespie, S. Alexander Haslam, Ryan K.L. Ko, Niklas Steffens, Janet Wiles, and Peter Worthy. Where is the human in human-centered AI? insights from developer priorities and user experiences. *Computers in Human Behavior*, 141, 2023.
- [10] Silvia Bonaccio and Reeshad S. Dalal. 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, 2006.
- [11] Jean-Francois Bonnefon and Iyad Rahwan. Machine thinking, fast and slow. *Trends in Cognitive Sciences*, 24(12):1019–1027, 2020.
- [12] Grady Booch, Francesco Fabiano, Lior Hoshen, Kiran Kate, Jonathan Lenchner, Nick Linck, Andrea Loreggia, Keerthiram Murugesan, Nicholas Mattei, Francesca Rossi, and Biplav Srivastava. Thinking fast and slow in AI, 10 2020.
- [13] Noah Castelo, Maarten W. Bos, and Donald R. Lehmann. Task-dependent algorithm aversion. *Journal of Marketing Research*, 56(5):809–825, 2019.
- [14] Daniel L. Chen, Martin Schonger, and Chris Wickens. oTree – an open-source platform for laboratory, online, and field experiments. *Journal of Behavioral and Experimental Finance*, 9:88–97, 2016.
- [15] Claudio Di Ciccio, Andrea Marrella, and Alessandro Russo. Knowledge-intensive processes: Characteristics, requirements and analysis of contemporary approaches. *Journal on Data Semantics*, 4:29–57, 2015.
- [16] Kahneman Daniel. *Thinking, fast and slow*. Farrar, Straus and Giroux, 2011.
- [17] Fred Davis. *A Technology Acceptance Model for Empirically Testing New End-User Information Systems*. PhD thesis, Massachusetts Institute of Technology, 1985.
- [18] Maartje M.A. De Graaf and Bertram F. Malle. How people explain action (and autonomous intelligent systems should too). In *2017 AAAI Fall Symposium Series*, 2017.

- [19] Joost C.F. De Winter and Dimitra Dodou. Why the fitts list has persisted throughout the history of function allocation. *Cognition, Technology & Work*, 16(1):1–11, 2014.
- [20] Jerker Denrell and James G. March. Adaptation as information restriction: The hot stove effect. *Organization Science*, 12(5):523–659, 2001.
- [21] Stephan Diederich, Alfred Brendel, Stefan Morana, and Lutz Kolbe. On the design of and interaction with conversational agents: An organizing and assessing review of human-computer interaction research. *Journal of the Association for Information Systems*, 23(1):96–138, 2022.
- [22] Berkeley J. Dietvorst, Joseph P. Simmons, and Cade Massey. Algorithm aversion: people erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, 144(1):114–126, 2015.
- [23] Berkeley J. Dietvorst, Joseph P. Simmons, and Cade Massey. Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. *Management Science*, 64(3):1155–1170, 2018.
- [24] Emir Efendić, Philippe P.F.M. Van de Calseyde, and Anthony M. Evans. Slow response times undermine trust in algorithmic (but not human) predictions. *Organizational Behavior and Human Decision Processes*, 157:103–114, 2020.
- [25] Ida Merete Enholm, Emmanouil Papagiannidis, Patrick Mikalef, and John Krogstie. Artificial intelligence and business value: A literature review. *Information Systems Frontiers*, 24(5):1709–1734, 2022.
- [26] Armin Falk and James J. Heckman. Lab experiments are a major source of knowledge in the social sciences. *Science*, 326(5952):535–538, 2009.
- [27] Lizhou Fan, Lingyao Li, Zihui Ma, Sanggyu Lee, Huizi Yu, and Libby Hemphill. A bibliometric review of large language models research from 2017 to 2023, 2023.
- [28] Ibrahim Filiz, Jan Judek, Marco Lorenz, and Markus Spiwoks. The tragedy of algorithm aversion. *Wolfsburg Working Papers No. 21-02*, 02 2021.
- [29] Thomas Franke, Christiane Attig, and Daniel Wessel. A personal resource for technology interaction: Development and validation of the affinity for technology interaction (ATI) scale. *International Journal of Human–Computer Interaction*, 35(6):456–467, 2018.

- [30] Ilaria Gaudiello, Elisabetta Zibetti, Sébastien Lefort, Mohamed Chetouani, and Serena Ivaldi. Trust as indicator of robot functional and social acceptance. an experimental study on user conformation to iCub answers. *Computers in Human Behavior*, 61:633–655, 2016.
- [31] Gerd Gigerenzer and Ulrich Hoffrage. How to improve bayesian reasoning without instruction: Frequency formats. *Psychological Review*, 102(4):684–704, 1995.
- [32] Milapji Singh Gill, Tom Westermann, Marvin Schieseck, and Alexander Fay. Integration of domain expert-centric ontology design into the CRISP-DM for cyber-physical production systems, 2023.
- [33] Francesca Gino and Don A. Moore. Effects of task difficulty on use of advice. *Journal of Behavioral Decision Making*, 20(1):21–35, 2007.
- [34] Franceso Gino, Alison. W. Brooks, and Maurice. E. Schweitzer. Anxiety, advice, and the ability to discern: feeling anxious motivates individuals to seek and use advice. *Journal of Personality and Social Psychology*, 102(3):497–512, 2012.
- [35] Ella Glikson and Anita Williams Woolley. Human trust in artificial intelligence: Review of empirical research. *Academy of Management Annals*, 14(2):627–660, 2020.
- [36] Ben Greiner. Subject pool recruitment procedures: organizing experiments with ORSEE. *Journal of the Economic Science Association*, 1:114–125, 2015.
- [37] Muhammad Usman Hadi, Qasem Al-Tashi, Rizwan Qureshi, Abbas Shah, Amgad Muneer, Muhammad Irfan, Anas Zafar, Muhammad Shaikh, Naveed Akhtar, Jia Wu, and Seyedali Mirjalili. Large language models: A comprehensive survey of its applications, challenges, limitations, and future prospects, 2023.
- [38] Peter A. Hancock, Deborah R. Billings, Kristin E. Schaefer, Jessie Y.C. Chen, Ewart J. De Visser, and Raja Parasuraman. A meta-analysis of factors affecting trust in human-robot interaction. *Human factors*, 53(5):517–527, 2011.
- [39] Jun He and William R. King. The role of user participation in information systems development: Implications from a meta-analysis. *Journal of Management Information Systems*, 25(1):301–331, 2008.



- [40] Alan R. Hevner, Salvatore T. March, Jinsoo Park, and Sudha Ram. Design science in information systems research. *MIS Quarterly*, 28(1):75–105, 2004.
- [41] Robert R. Hoffman, Nigel R. Shadbolt, A.Mike Burton, and Gary Klein. Eliciting knowledge from experts: A methodological analysis. *Organizational Behavior and Human Decision Processes*, 62(2):129–158, 1995.
- [42] Christine Hofheinz, Markus Germar, Thomas Schultze, Johannes Michalak, and Andreas Mojzisch. Are depressed people more or less susceptible to informational social influence? *Cognitive Therapy and Research*, 41, 10 2017.
- [43] Xinyi Hou, Yanjie Zhao, Yue Liu, Zhou Yang, Kailong Wang, Li Li, Xiapu Luo, David Lo, John Grundy, and Haoyu Wang. Large language models for software engineering: A systematic literature review, 2023.
- [44] Yoyo Tsung-Yu Hou and Malte F. Jung. Who is the expert? Reconciling algorithm aversion and algorithm appreciation in AI-supported decision making. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2):1–25, 2021.
- [45] Arthur S. Jago. Algorithms and authenticity. *Academy of Management Discoveries*, 5(1):38–56, 2019.
- [46] Ekaterina Jussupow, Izak Benbasat, and Armin Heinzl. Why are we averse towards algorithms? A comprehensive literature review on algorithm aversion. *Proceedings of the 28th European Conference on Information Systems (ECIS)*, pages 1–16, 2020.
- [47] Shubhra Kanti Karmaker, Md. Mahadi Hassan, Micah J. Smith, Lei Xu, Chengxiang Zhai, and Kalyan Veeramachaneni. AutoML to date and beyond: Challenges and opportunities. *ACM Computing Surveys*, 54(8):1–36, 2020.
- [48] Kohei Kawaguchi. When will workers follow an algorithm? A field experiment with a retail business. *Management Science*, 67(3):1670–1695, 2021.
- [49] Min Kyung Lee. Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. *Big Data & Society*, 5(1):1–16, 2018.
- [50] Bo Li, Gexiang Fang, Yang Yang, Quansen Wang, Wei Ye, Wen Zhao, and Shikun Zhang. Evaluating ChatGPT’s information extraction capabilities: An assessment of performance, explainability, calibration, and faithfulness, 2023.

- [51] Yiheng Liu, Tianle Han, Siyuan Ma, Jiayue Zhang, Yuanyuan Yang, Jiaming Tian, Hao He, Antong Li, Mengshen He, Zhengliang Liu, Zihao Wu, Lin Zhao, Dajiang Zhu, Xiang Li, Ning Qiang, Dingang Shen, Tianming Liu, and Bao Ge. Summary of ChatGPT-related research and perspective towards the future of large language models. *Meta-Radiology*, 1(2):100017, 2023.
- [52] Jennifer M. Logg, Julia A. Minson, and Don A. Moore. Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes*, 151:90–103, 2019.
- [53] Hasan Mahmud, A.K.M. Najmul Islam, Syed Ishtiaque Ahmed, and Kari Smolander. What influences algorithmic decision-making? A systematic literature review on algorithm aversion. *Technological Forecasting and Social Change*, 175, 2022.
- [54] Spyros Makridakis. The forthcoming artificial intelligence (AI) revolution: Its impact on society and firms. *Futures*, 90:46–60, 2017.
- [55] Olivera Marjanovic and Ronald Freeze. Knowledge intensive business processes: Theoretical foundations and research challenges. In *Proceedings of the Annual Hawaii International Conference on System Sciences*, pages 1–10, 2011.
- [56] Maranda McBride, Lemuria Carter, and Celestine Ntuen. The impact of personality on nurses’ bias towards automated decision aid acceptance. *International Journal of Information Systems and Change Management*, 6(2):132–146, 2012.
- [57] Silke Merkelbach, Sebastian Von Enzberg, Arno Kühn, and Roman Dumitrescu. Towards a process model to enable domain experts to become citizen data scientists for industrial applications. In *2022 IEEE 5th International Conference on Industrial Cyber-Physical Systems (ICPS)*, pages 1–6, 2022.
- [58] Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*, 267:1–38, 2019.
- [59] Michael I. Norton, Daniel Mochon, and Dan Ariely. The ikea effect: When labor leads to love. *Journal of Consumer Psychology*, 22(3):453–460, 2012.
- [60] Judith Reitman Olson and Henry H. Rueter. Extracting expertise from experts: Methods for knowledge acquisition. *Expert Systems*, 4(3):152–168, 1987.

- [61] Joon Sung Park, Rick Barber, Alex Kirlik, and Karrie Karahalios. A slow algorithm improves users' assessments of the algorithm's accuracy. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–15, 2019.
- [62] Soya Park, April Yi Wang, Ban Kawas, Q. Vera Liao, David Piorkowski, and Marina Danilevsky. Facilitating knowledge sharing from domain experts to data scientists for building NLP models. In *26th International Conference on Intelligent User Interfaces*. ACM, 2021.
- [63] Ken Peffers, Tuure Tuunanen, Marcus A. Rothenberger, and Samir Chatterjee. A design science research methodology for information systems research. *Journal of Management Information Systems*, 24(3):45–77, 2007.
- [64] Andrew Prael and Lyn Van Swol. Understanding algorithm aversion: When is advice from automation discounted? *Journal of Forecasting*, 36(6):691–702, 2017.
- [65] Iyad Rahwan, Manuel Cebrian, Nick Obradovich, Josh Bongard, Jean-François Bonnefon, Cynthia Breazeal, Jacob Crandall, Nicholas Christakis, Iain Couzin, Matthew Jackson, Nicholas Jennings, Ece Kamar, Isabel Kloumann, Hugo Larochelle, David Lazer, Richard McElreath, Alan Mislove, David Parkes, Alex Pentland, and Michael Wellman. Machine behaviour. *Nature*, 568:477–486, 04 2019.
- [66] Francesca Rossi and Andrea Loreggia. Preferences and ethical priorities: Thinking fast and slow in AI. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*, AAMAS '19, pages 3–4, Richland, SC, 2019. International Foundation for Autonomous Agents and Multiagent Systems.
- [67] A. Saranya and R. Subhashini. A systematic review of explainable artificial intelligence models and applications: Recent developments and future trends. *Decision Analytics Journal*, 7, 2023.
- [68] Marko Sarstedt, Doreen Neubert, and Kati Barth. The IKEA effect. a conceptual replication. *Journal of Marketing Behavior*, 2:307–312, 11 2016.
- [69] Tjeerd A.J. Schoonderwoerd, Wiard Jorritsma, Mark A. Neerincx, and Karel Van Den Bosch. Human-centered XAI: Developing design patterns for explanations of

clinical decision support systems. *International Journal of Human-Computer Studies*, 154:102684, 2021.

- [70] Tom Seymoens, Femke Ongenaes, An Jacobs, Stijn Verstichel, and Ann Ackaert. A methodology to involve domain experts and machine learning techniques in the design of human-centered algorithms. In Barbara Rita Barricelli, Virpi Roto, Torkil Clemmensen, Pedro Campos, Arminda Lopes, Frederica Gonçalves, and José Abdelnour-Nocera, editors, *Human Work Interaction Design. Designing Engaging Automation*, pages 200–214, Cham, 2019. Springer International Publishing.
- [71] Navya Nishith Sharan and Daniela Maria Romano. The effects of personality and locus of control on trust in humans versus artificial intelligence. *Heliyon*, 6(8), 2020.
- [72] Ben Shneiderman. Bridging the gap between ethics and practice: Guidelines for reliable, safe, and trustworthy human-centered AI systems. *ACM Transactions on Interactive Intelligent Systems*, 10(4):1–31, 2020.
- [73] Vivek Singh and Kailash Joshi. Automated machine learning (AutoML): An overview of opportunities for application and research. *Journal of Information Technology Case and Application Research*, 24:1–11, 05 2022.
- [74] Roman Vaculin, Richard Hull, Fenno Heath, Craig Cochran, Anil Nigam, and Noi Sukaviriya. Declarative business artifact centric modeling of decision and knowledge intensive business processes. In *Proceedings - IEEE International Enterprise Distributed Object Computing Workshop, EDOC*, pages 151–160, 2011.
- [75] Viswanath Venkatesh, Michael G. Morris, Gordon B. Davis, and Fred D. Davis. User acceptance of information technology: Toward a unified view. *MIS Quarterly*, 27(3):425–478, 2003.
- [76] Viswanath Venkatesh, James Thong, and Xin Xu. Unified theory of acceptance and use of technology: A synthesis and the road ahead. *Journal of the Association for Information Systems*, 17(5):328–376, 2016.
- [77] Laura von Rueden, Sebastian Mayer, Katharina Beckh, Bogdan Georgiev, Sven Giesselbach, Raoul Heese, Birgit Kirsch, Michal Walczak, Julius Pfrommer, Annika Pick, Rajkumar Ramamurthy, Jochen Garcke, Christian Bauckhage, and Jannis Schuecker. Informed machine learning - a taxonomy and survey of integrating prior

knowledge into learning systems. *IEEE Transactions on Knowledge and Data Engineering*, 2021.

- [78] Xinru Wang and Ming Yin. Are explanations helpful? A comparative study of the effects of explanations in AI-assisted decision-making. In *26th International Conference on Intelligent User Interfaces*, pages 318–328, 2021.
- [79] Rüdiger Wirth and Jochen Hipp. CRISP-DM: Towards a standard process model for data mining. *Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining*, 01 2000.
- [80] Michael Yeomans, Anuj Shah, Sendhil Mullainathan, and Jon Kleinberg. Making sense of recommendations. *Journal of Behavioral Decision Making*, 32(4):403–414, 2019.
- [81] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. A survey of large language models, 2023.
- [82] Marc-André Zöller, Waldemar Titov, Thomas Schlegel, and Marco F. Huber. Xautoml: A visual analytics tool for understanding and validating automated machine learning. *ACM Transactions on Interactive Intelligent Systems*, 13(4):1–39, 2023.

# Appendices

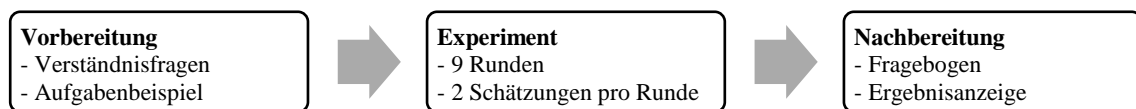
## **Supplementary Materials to Chapter 2**

# Instruktionen

## Allgemeine Hinweise

- Die Gesamtdauer des Experimentes beträgt ca. 30 Minuten.
- Während der Durchführung ist keine Kommunikation gestattet. Mobiltelefone müssen während der kompletten Experimentdauer ausgeschaltet sein.
- Sämtliche Aktionen, die Sie im Rahmen dieses Experiments tätigen, erfolgen **anonym**.
- Bitte **nutzen Sie die Knöpfe “Weiter” oder “Antwort einreichen”**, um zur nächsten Seite im Experiment zu gelangen. Es ist technisch nicht möglich, auf eine bereits verlassene Seite zurück zu gelangen. Bitte **klicken Sie in Ihrem Browserfenster niemals auf “zurück” oder auf “neu laden”**, sonst wird der Experimentverlauf gestört und Sie verlieren Ihre Auszahlung.

## Ablauf



## Vorbereitung

- **Verständnisfragen:** Sie erhalten Fragen zum Inhalt der Instruktionen. Sie kommen erst weiter, wenn Sie alle Fragen richtig beantwortet haben. Die Antworten haben keinen Einfluss auf Ihre Auszahlung.
- **Aufgabenbeispiel:** Sie erhalten ein imaginäres Beispiel für eine Runde des nachfolgenden Experimentes. Sie haben die Möglichkeit, sich mit dem Aufbau einer Runde und deren einzelnen Bausteinen vertraut zu machen. Das Beispiel hat keinen Einfluss auf die Auszahlung.

## Experiment

- Das Experiment geht über **neun Runden**.



- Ihre **Aufgabe** ist es, anhand von Abbildungen die Eigenschaften verschiedener Objekte zu schätzen. In jeder Runde muss ein neues Objekt eingeschätzt werden. [Außerdem erhalten Sie zusätzliche Informationen zu den abgebildeten Objekten.]
- Eine Runde erhält folgende Aktionen:
  - Sie machen Ihre **1. Schätzung** des abgebildeten Objektes,
  - Die **Künstliche Intelligenz (KI)** macht eine Schätzung,
  - Das Ergebnis der **KI-Schätzung** wird Ihnen als **eine unverbindliche Empfehlung** angezeigt,
  - Sie machen Ihre **2. Schätzung** des abgebildeten Objektes.
- Ihre **Auszahlung** setzt sich wie folgt zusammen:
  - Sie erhalten eine Show-up Fee in Höhe von **3 Euro**.
  - Außerdem haben Sie die Möglichkeit, bis zu **9 Euro** über die neun Runden des Experimentes (maximal **1 Euro** pro Runde) zu verdienen:
    - In jeder Runde machen Sie 2 Schätzungen. Für jede Schätzung können Sie bis zu **0,50 Euro** verdienen.
    - Die **Auszahlung pro Schätzung** setzt sich wie folgt zusammen:
 
$$[ \text{Auszahlung für eine Schätzung} ] = [ \text{erspielter Betrag für die Genauigkeit einer Schätzung} ] - [ \text{Kosten für die zusätzliche Zeit für eine Schätzung} ]$$
  - Die Show-up Fee und die Auszahlung aus den neun Runden werden aufaddiert und das ergibt Ihre **Gesamtauszahlung**. Die maximale Gesamtauszahlung für dieses Experiment beträgt **12 Euro**.
- **Genauigkeit einer Schätzung:**
  - Für jede Schätzungsaufgabe existiert eine einzige richtige Antwort, ein **wahrer Wert**, der ausschließlich der Experimentleitung bekannt ist.

- Die **Genauigkeit einer Schätzung** wird an ihrer **Abweichung zum wahren Wert** gemessen. Die Abweichung wird in Prozent dargestellt. Die Richtung der Abweichung (ob die Schätzung unter oder über dem wahren Wert liegt) spielt keine Rolle.
- Grundsätzlich gilt: **Je geringer die Abweichung zum wahren Wert ist, desto höher ist Ihre Auszahlung für diese Schätzung. Oder anders gesagt, je näher Sie mit Ihrer Schätzung am wahren Wert liegen, desto höher Ihre Auszahlung.**
- **Zeit pro Schätzung:**
  - Sie haben für Ihre beiden Schätzungen jeweils **45 Sekunden kostenfrei** Zeit. Innerhalb dieser Zeit wird Ihre Auszahlung pro Schätzung lediglich durch ihre Genauigkeit bestimmt.
  - Danach kosten Sie **jede zusätzlichen 30 Sekunden 1/10 (ein Zehntel)** ihrer erspielten Auszahlung für die Genauigkeit der jeweiligen Schätzung.
  - Das heißt, wenn Sie 345 Sekunden (entspricht 5 Minuten und 45 Sekunden) oder länger für eine Schätzung brauchen, wird Ihre Auszahlung für diese Schätzung Null sein.
  - Sie können kein Geld verlieren, d.h., **Ihre Auszahlung pro Schätzung kann nicht unter null gehen**, egal, wie lange Sie brauchen.

### Nachbereitung

- **Fragebogen:** Nach dem Experiment erhalten Sie einen Fragebogen. Die **vollständige und ehrliche Beantwortung der Fragen ist sehr wichtig** für die anschließende Auswertung des Experiments. Die Auswertung wird ausschließlich für wissenschaftliche Zwecke verwendet. Ihre Antworten in diesem Fragebogen haben keinen Einfluss auf die Auszahlung.
- **Ergebnisanzeige:** Als letztes wird Ihnen aus Ihren sämtlichen Schätzungen resultierende Gesamtauszahlung angezeigt.

## **Supplementary Materials to Chapter 3**

# Experimental Instructions

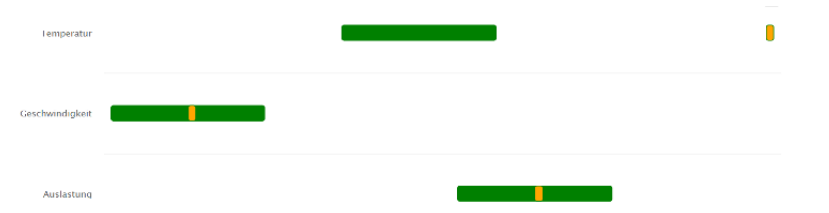
## Scenario

- Over the course of the experiment, you assume the role of a skilled worker in an industrial company. You will be responsible for the operation of a production facility.
- There is a certain probability that a malfunction may occur in the production facility. To avoid potential malfunctions, maintenance can be performed.
- Your task is to evaluate the probability of malfunctions in multiple rounds and then decide whether the production plant should be maintained in a given round. You will be supported in your task by an artificial intelligence (AI).

## Malfunction probability

- The probability of a malfunction is unknown but can be estimated using three indicators.
- These indicators are: Temperature, Speed and Voltage. Each of the indicators can take values between 0 and 100.
- Each of the indicators has its own optimal range. If an indicator is in its optimal range, this is particularly good for the production plant corresponds to a malfunction being less likely.
- The more indicators' values are located outside their respective optimum ranges, the more likely a malfunction becomes.
  - If three indicators are within the optimal ranges and none are outside, a malfunction is very UNlikely.
  - If two indicators are within the optimal ranges and one is outside, a malfunction is UNlikely.
  - If one indicator is within the optimal range and two are outside, a malfunction is likely.
  - If all three indicators are outside the optimal ranges and none inside, a malfunction is very likely.

- Important: In the experiment, you do not know the optimal ranges. Instead, you must estimate them as accurately as possible, based on the data points of past malfunctions. This estimation is called "acceptable ranges" (see "Procedure").



**Example:** The graphic above displays an example for the optimal ranges (green bars) for the three indicators (orange dots). In this example, the "Temperature" indicator is located outside its optimal range and the "Speed" and "Voltage" indicators are located inside their respective optimal ranges. Accordingly, a malfunction would be considered **unlikely** in this case.

## Support by an AI

- In each round, the AI predicts the probability of a malfunction and, based on this prediction, gives you a non-binding recommendation as to whether maintenance should be performed.
- The accuracy of the AI predictions can vary. It depends on how the AI has been trained. Training the AI is part of the experiment (see below). You will be informed about the achieved accuracy of the AI (in percent) in the experiment at the end of Stage 3.

## Procedure

The experiment consists of four stages that build upon each other.

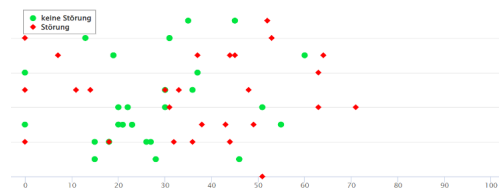


## Stage 1: Comprehension checks

- In this stage, comprehension checks are conducted about the instructions. Only once you have answered all the control questions correctly the experiment can begin. You have an unlimited number of attempts to answer the questions correctly.

## Stage 2: Selection of acceptable ranges

- As mentioned, the optimal ranges of the individual indicators are unknown to you. Instead, you must define an acceptable range for each indicator.
- An acceptable range is an approximation of the actual (unknown) optimal range. The closer the acceptable ranges you set are to the optimal range, the better the AI's advice will be.
- Data about past malfunctions is available for you to set your acceptable ranges:
  - For each indicator individually, you can see at which values there were malfunctions in the past (red dots) and at which there were not (green dots).
  - You are now asked to define a lower limit (minimum) and an upper limit (maximum) of your acceptable range (green dashes). In general, an acceptable range should contain as many points without malfunctions (green) and as few points with malfunctions (red) as possible.



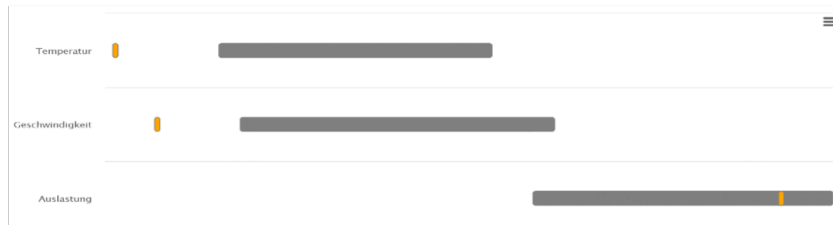
- At the beginning, you will be given an example that you can use to practice setting the limits (technical note: the limit that is closer to your mouse pointer moves in each case).
- After you have set and confirmed the acceptable ranges, they will be displayed as gray bars in the further course of the experiment for the sake of conciseness (see figure). The individual data points are hidden.



- You define a total of three acceptable ranges (one for each indicator), which you will necessarily need in the further course of the experiment.
- Your acceptable ranges will be displayed for all further decisions, so you do not have to memorize or note them.

### Stage 3: Training the AI

- In this stage, the AI is trained based on your acceptable ranges defined in Stage 2. The AI thus learns how to evaluate the probability of malfunctions for different indicator combinations.
- The training of the AI happens through ten training situations as follows:
  - Each training situation represents a combination of the three indicators' values. These values are shown together with their acceptable ranges.
  - The following figure provides an example of a training situation. Orange bars represent the indicators' values. Gray bars represent the acceptable ranges.



- For each training situation, you can see which indicators are within and which are outside your defined acceptable ranges.
- Your task is to tell the AI how each training situation is to be evaluated regarding the likelihood of a malfunction. In doing so, you help the AI learn.
- Use your acceptable ranges and your knowledge about the probability of malfunctions for the evaluation:
  - \* If three indicators are within your acceptable ranges and zero are outside, a malfunction is very UNlikely.
  - \* If two indicators are within your acceptable ranges and one is outside, a malfunction is UNlikely.

- \* If one indicator is within your acceptable ranges and two are outside, a malfunction is likely.
  - \* If there are zero indicators inside your acceptable ranges and three outside, a malfunction is very likely.
- Each of your ten malfunction probability assessments is then checked for correctness:
  - \* If a malfunction has been classified as "very unlikely" or "unlikely" and no malfunction has actually occurred, the assessment is considered correct and otherwise incorrect.
  - \* If a malfunction was classified as "very likely" or "likely" and a malfunction actually occurred, the evaluation is considered correct and otherwise incorrect.
  - \* Whether a malfunction actually occurs or not depends on the actual optimum ranges, which remain unknown.
- The result of the training, and thus the quality of the AI, depends on how many training situations have been correctly assessed. You will be informed about the result at the end of the training stage. Two results are possible:
  - If at least seven training situations were evaluated correctly, you will receive an AI with the accuracy of 90% (on average it is correct in 9 out of 10 cases and wrong in one out of 10 cases).
  - If less than seven training situations were evaluated correctly, you will receive an AI with the accuracy of 50% (it is correct on average in 5 out of 10 cases and wrong in 5 out of 10 cases).
- After completing this stage, the AI has learned to evaluate malfunction probabilities in comparable situations through the training situations.
- In stage 4, you can use the AI for decision support.

## **Stage 4: Production plant surveillance**

- This stage consists of 25 rounds.



- In each round, you have to make the decision whether to maintain the production facility.
- All rounds are independent of each other, i.e., the decision in one round does not affect other rounds.
- In each round, you will receive a graphic showing the values of the three indicators and your self-defined acceptable ranges (see Stage 3).
- In In each round you make your decision in two steps:
  - In the first step, you evaluate the given situation in terms of the probability of failure and decide whether maintenance should be performed.
  - In the second step, the AI's recommendation is displayed to you. Afterwards, you are asked again whether you want to perform maintenance.
- Only the decision in the second step is relevant for your payoff in the respective round.
- Whether a malfunction actually occurs or not depends on the optimal ranges, which remain unknown. You will only find out at the end of the experiment how often you were correct and how high your payoff will be.

## Payoffs

- During the experiment, all amounts are denoted in the fictitious currency "Taler".
- Per round you receive following payoffs depending on your maintenance decision and occurrence/non-occurrence of a malfunction:
  - You decide that maintenance should be performed.
    - \* Maintenance limits your production capacities. Therefore, your payoff this round is 5 Taler.
  - You decide that no maintenance should be performed.
    - \* If no malfunction occurs and you can therefore produce fully, your payoff from this round is 10 Taler.

\* If a malfunction occurs and therefore you cannot produce, your payoff from this round is 0 Taler.

- The payoffs from all rounds are cumulated.
- At the end of the experiment, you will receive your payoffs at an exchange rate of € 1 per 10 Taler. In addition, you will receive a show-up fee of € 2.50.

## **Additional remarks**

- All communication is prohibited for the duration of the experiment except for communication explicitly permitted by the instructions.
- Mobile phones must be turned off for the duration of the experiment.
- All decisions within the scope of the experiment will remain completely anonymous.
- After completing the main part of the experiment, we kindly ask you to answer some additional questions. Answering the questions honestly and in full is very important for the subsequent analysis of the experiment. The answers to the questions remain anonymous and will only be evaluated for scientific purposes. Your answers in this questionnaire have no impact on your payoff achieved in the experiment.

# Questionnaire

**Please answer the following questions.**

What is your age?

What is your gender?

- Male
- Female
- Non-Binary

What is your highest level of education?

- Highschool / GED
- Undergraduate degree
- Graduate degree
- Else / Prefer not to say

What is your current study major?

- Arts & Design
- Business & Economics
- Computer Science
- Cultural & Social Studies
- Engineering
- Natural Sciences
- Pedagogics
- Non-student

### Please answer the following questions.

Please indicate your consent with the following statement on a scale from 1 (= completely disagree) to 7 (= completely agree).

	1	2	3	4	5	6	7
I believe that my estimates of malfunction probabilities were correct.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I believe that my estimates of malfunction probabilities are close to the true value.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I was very confident about the accuracy of my estimates of malfunction probabilities.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I was confident I would do well on the task.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I have no doubt that my estimates of malfunction probabilities are close to the true values.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

### Please answer the following questions.

Please indicate to what extent the following statements apply to you personally.

	Does not apply at all	Applies little	Somewhat applies	Pretty much applies	Fully applies
In difficult situations I can rely on my abilities.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I can cope well with most of the problems on my own power.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Even strenuous and complicated tasks I can usually solve well.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

### Please answer the following questions.

Please indicate your consent with the following statement on a scale from 1 (= Do not consent at all) to 5 (= Fully consent).

	1	2	3	4	5
I could easily understand the maintenance scenario presented in the experiment.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I was able to put myself in the role of a production plant manager.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
The information and examples provided to me prepared me well for the decisions in the experiment.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I could understand most of the explanations in the experiment.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I understood the contexts presented in the experiment.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I knew exactly what was required of me in the experiment.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I knew what I had to do to be as successful as possible in the experiment.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
The tasks of the experiment were demanding.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I am satisfied with my performance in the experiment.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I think most of the experiment participants did well on the tasks.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

**Please answer the following questions.**

Please indicate your consent with the following statement on a scale from 1 (= Do not consent at all) to 5 (= Fully consent).

	1	2	3	4	5
The AI-training stage was time well spent.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I would have liked to invest more time in training the AI.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I was instrumental in how good the quality of the AI recommendations was.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I felt like it was a gamble whether you got a high-quality AI or a low-quality AI at the end of the training stage.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Training the AI has helped me better understand how the AI makes its recommendations.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Training the AI helped me better understand how it works.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I liked that my expert knowledge was shared with the AI in the form of my acceptable intervals.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I found it uncomfortable that the AI had access to my expert knowledge in the form of my acceptable intervals.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
If I had the chance, I would have preferred to share even more knowledge with the AI.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
If I had a choice, I would rather not have shared any knowledge with the AI at all.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
My influence on training the AI was great.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I enjoyed training the AI.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I found the AI recommendations to be of high quality.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
When deciding whether to perform maintenance, I weighed things logically.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I made decisions about whether to perform maintenance intuitively rather than strategically.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

**Please answer the following questions.**

In the following questionnaire, we will ask you about your interaction with technical systems. The term “technical systems” refers to apps and other software applications, as well as entire digital devices (e.g., mobile phone, computer, TV, car navigation).

Please indicate to what extent you agree to the following statements.

[illegible]

## Additional tables and graphics

Table 1: Distribution of study majors among participants

Major	Frequency	Percentage
Arts & Design	13	8.50
Business & Economics	52	33.99
Computer Science	11	7.19
Cultural & Social Studies	11	7.19
Engineering	10	6.54
Natural Sciences	1	0.65
Pedagogics	53	34.64
Non-student	2	1.31
Total	153	100.00

Table 2: Pairwise between-group comparisons of standardized questionnaire scores

	$T_Z$ vs. $T_P$	$T_Z$ vs. $T_A$	$T_P$ vs. $T_A$
Affinity to Technology Interaction (ATI)	0.477 (0.6359)	2.514 (0.0115)	-1.618 (0.1063)
Self-efficacy (ASKU)	-0.160 (0.8749)	1.167 (0.2452)	-1.532 (0.1264)
Ex-post decision confidence (DC)	0.606 (0.5476)	-0.167 (0.8698)	0.646 (0.5211)

This table reports results for pairwise two-sample between-group Mann-Whitney U-Tests between experimental groups. z-values reported with p-values in parenthesis.

Table 3: Summary statistics and between-group comparison of questionnaire items on treatment perception

	$T_Z$	$T_P$	$T_A$	Kruskal-Wallis-H
<b>Items on treatment perception</b>				
AI-training was time well spent	4.02 (.9366)	4.02 (.9412)	3.72 (1.2283)	.5221
Would have liked to spend more time for training	3.22 (1.2664)	2.73 (1.3382)	2.77 (1.5211)	.1388
Perceived contribution to AI-advice quality	3.34 (1.0994)	3.87 (1.0574)	4.11 (.9145)	.0013
AI-quality outcome perceived as gamble	2.52 (1.3130)	2.13 (1.1200)	1.96 (.9079)	.1163
Perceived understanding of AI's functionality	3.08 (1.2095)	3.38 (1.2301)	3.91 (1.0179)	.0024
Perceived understanding of how AI generates advice	2.9 (1.1473)	3.42 (1.1578)	3.72 (1.1554)	.0020
Liked sharing knowledge with AI	4.00 (.7825)	3.89 (.9101)	3.96 (.8836)	.8598
Felt uncomfortable sharing knowledge with AI	1.54 (.9941)	1.51 (.8427)	1.40 (.8251)	.8251
Would have liked to share more knowledge with AI	3.52 (1.3589)	3.42 (1.1578)	3.34 (1.3717)	.7511
Would have preferred not to share any knowledge with AI	1.54 (.7616)	1.64 (.9806)	1.64 (1.0092)	.9923
Perceived influence on AI-training	3.02 (.9145)	3.58 (1.1178)	4.09 (.8030)	.0001
Enjoyed training the AI	3.26 (1.1920)	3.42 (1.0333)	3.70 (1.0818)	.1646
Perceived quality of AI-advice	3.68 (.7939)	3.76 (1.0478)	3.60 (.9478)	.4874
Maintenance decisions made rather strategically	4.28 (.7570)	4.44 (.6236)	4.26 (.6416)	.3548
Maintenance decisions made rather intuitively	2.28 (1.0887)	2.09 (1.0406)	1.98 (.9205)	.3967
Number of observations	50	45	47	

This table reports summary statistics of questionnaire items on treatment perception measured on a 5-point scale by treatment. Standard deviations reported in parenthesis. Kruskal-Wallis-H reports the p-values for Kruskal-Wallis H-Tests with ties between experimental groups. Includes only subjects with high-accuracy AI.

Table 4: Illustration of decision process for each round of maintenance stage

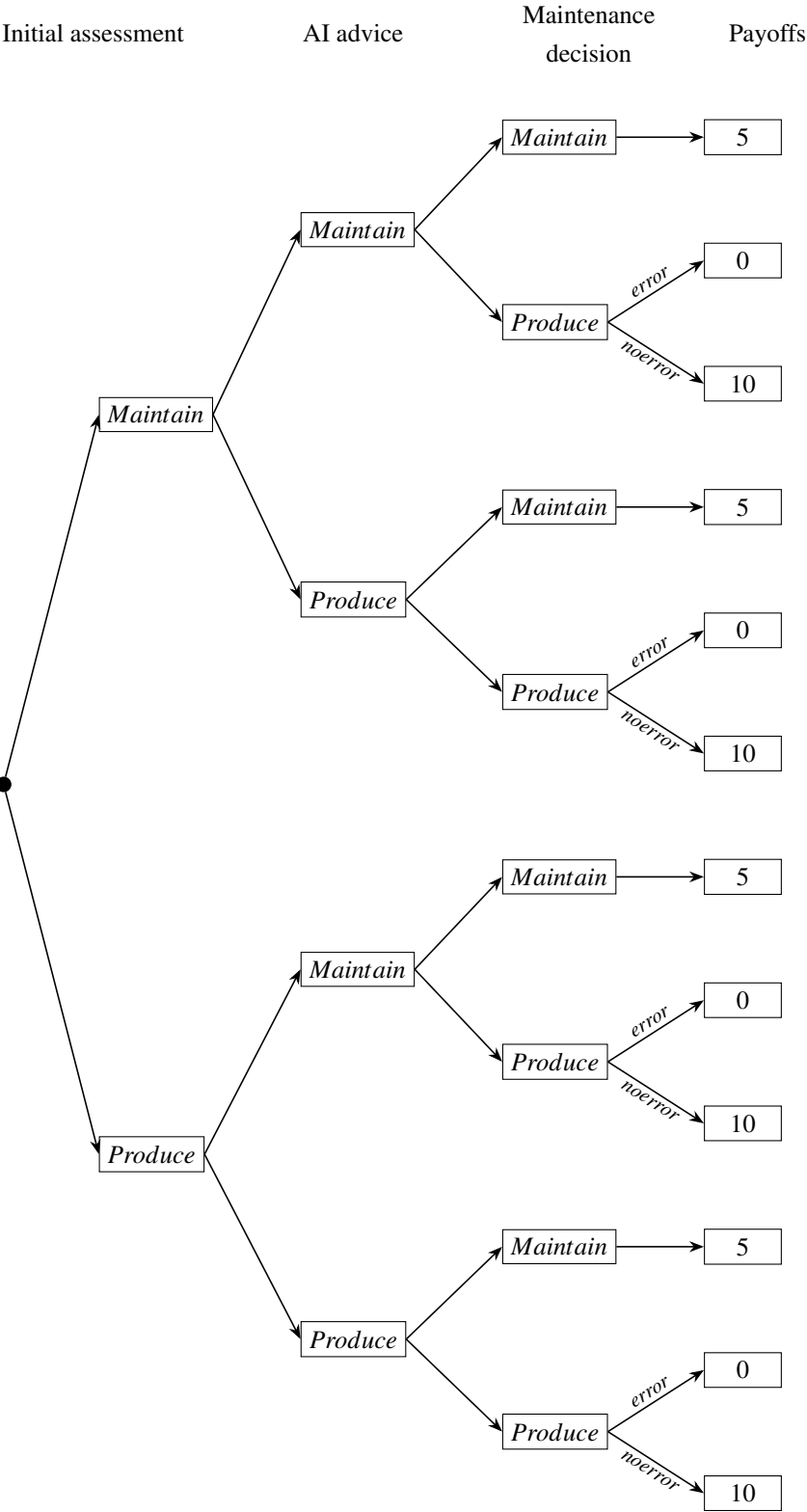




Table 5: Maintenance recommendation by the AI in each round of the maintenance stage (high-accuracy AI)

Round	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
AI maintenance recommendation	no	yes	yes	yes	yes	no	no	no	yes	no	yes	no	no	no	no	no	no	no	no	no	yes	no	yes	no	no
Resulting payoff	10	5	5	5	5	0	10	10	5	0	5	10	10	10	0	10	10	10	0	10	5	10	5	10	10

This table reports the (high-accuracy) AI's maintenance recommendations for each round of the maintenance stage, as well as the corresponding payoffs. As the high-accuracy AI is 90% accurate, errors occur in a total of four rounds, i.e., no maintenance is recommended despite being necessary, therefore the respective round's payoff is zero.

Table 6: Efficient revision rate and AI net benefit

	$T_Z$	$T_P$	$T_A$	Kruskal-Wallis-H
<b>Efficient revision rate</b>				
Mean	.7338	.8207	.6960	.2306
Standard deviation	.2558	.2352	.3363	
Min	0	.33	0	
Median	.73	1	.77	
Max	1	1	1	
Number of observations	40	37	35	
<b>AI net benefit</b>				
Mean	.8582	.8662	.7842	.4867
Standard deviation	.1905	.2067	.2892	
Min	.5	.33	0	
Median	1	1	1	
Max	1	1	1	
Number of observations	37	36	33	

In total, 30 subjects did not revise their initial decision intent at least once. Therefore, they were excluded from this calculation, as a rate of efficient revisions could not be computed for them. Analogously, a total of 36 subjects either did not revise their initial decision intent at least once or follow the AI's advice at least once. Therefore, they were excluded from this calculation, as an AI net benefit could not be computed for them. Kruskal-Wallis-H reports the p-value two-sample Kruskal-Wallis H-Tests with ties between experimental groups.

## **Supplementary Materials to Chapter 4**

# Originale Interviewprotokolle

*Die Interviewprotokolle liegen in der originalen Form vor. Die Rechtschreibung wurde vollständig übernommen.*

## Interview 1: Druckluftverbrauch (28. November 2023)

**Anonym**

Hallo

**Expert Interviewer**

Guten Tag! Wie kann ich Ihnen bei der Anwendung von maschinellem Lernen in Ihrem Fachgebiet helfen? Bitte beschreiben Sie zunächst Ihr Fachgebiet und das spezifische Problem, das Sie lösen möchten.

**Anonym**

Druckluftsteuerung

**Expert Interviewer**

Verstanden, Sie arbeiten also im Bereich der Druckluftsteuerung. Können Sie mir bitte mehr über das spezifische Problem erzählen, das Sie mit maschinellem Lernen lösen möchten? Was genau möchten Sie erreichen?

**Anonym**

angepasste drücke zu verschiedenen zeiten

**Expert Interviewer**

Danke für die Präzisierung. Sie möchten also mithilfe von maschinellem Lernen Drücke in einem Druckluftsystem zu verschiedenen Zeiten anpassen. Lassen Sie uns nun die abhängige Variable definieren, also das, was Sie erreichen wollen. Ist die abhängige Variable in diesem Fall der Druck in Ihrem System zu einem bestimmten Zeitpunkt?

**Anonym**

ja

**Expert Interviewer**

Perfekt, die abhängige Variable ist also der Druck im System zu einem bestimmten Zeitpunkt. Jetzt würde ich gerne mehr über die unabhängigen Variablen erfahren, also die Faktoren, die den Druck beeinflussen können. Können Sie einige dieser Faktoren nennen?

**Anonym**

maschinen laufen zu verschiedenen zeiten mit unterschiedlichem druckverbrauch, bzw luftverbrauch

**Expert Interviewer**

Vielen Dank für diese Information. Die unabhängigen Variablen scheinen also die Betriebszeiten der Maschinen und deren Druck- bzw. Luftverbrauch zu sein.

Um weiterzumachen, möchte ich fragen: Wie können diese Variablen gemessen oder quantifiziert werden? Gibt es zum Beispiel Sensoren, die den Luftverbrauch der Maschinen messen, oder Aufzeichnungen über deren Betriebszeiten?

**Anonym**

leider nicvht

**Expert Interviewer**

Verstehe, Sie haben keine direkten Messungen für den Luftverbrauch oder die Betriebszeiten der Maschinen. In diesem Fall könnten wir alternative Methoden in Betracht ziehen, um diese Variablen zu schätzen oder indirekt zu messen. Gibt es andere Indikatoren oder Daten, die auf den Luftverbrauch oder die Betriebszeiten der Maschinen hinweisen könnten? Beispielsweise Produktionsdaten, Schichtpläne oder Wartungsprotokolle?

**Anonym**

Schichtpläne

**Expert Interviewer**

Gut, Schichtpläne können als indirekter Indikator für Betriebszeiten der Maschinen dienen. Um weiter zu gehen: Sind die Schichtpläne und damit verbundenen Maschinenbetriebszeiten kontinuierliche oder kategoriale Variablen? Mit kontinuierlich meine ich, ob sie in einem stetigen Zeitrahmen (zum Beispiel jede Minute oder Stunde) erfasst werden, oder kategorial, das heißt in festen Gruppen oder Klassen (zum Beispiel Frühschicht, Spätschicht, Nachtschicht)?

**Anonym**

zu verschiedenen Schichtzeitpunkten

**Expert Interviewer**

Verstanden, die Schichtzeitpunkte sind also kategoriale Variablen, da sie in festgelegten Gruppen (wie Frühschicht, Spätschicht, Nachtschicht) auftreten.

Nun zur nächsten Frage: Gibt es bekannte Regeln oder Beziehungen zwischen den Schichtzeitpunkten und dem Luftverbrauch bzw. dem Druck im System? Zum Beispiel, dass zu bestimmten Schichten mehr oder weniger Luft verbraucht wird?

**Anonym**

ja

**Expert Interviewer**

Ausgezeichnet, es gibt also eine bekannte Beziehung zwischen den Schichtzeitpunkten und dem Luftverbrauch bzw. dem Druck im System. Könnten Sie mir ein paar typische Beispielsituationen aus Ihrer Erfahrung nennen, in denen diese Beziehung deutlich wird? Zum Beispiel, in welchen Schichten der Luftverbrauch besonders hoch oder niedrig ist und wie sich das auf den Druck auswirkt?

**Anonym**

in der Frühschicht besonders hoch, bei Pausenzeiten niedrig

**Expert Interviewer**

Danke für das Beispiel. Es zeigt, dass der Luftverbrauch in der Frühschicht besonders hoch ist und während der Pausenzeiten niedrig. Das hilft uns, die Beziehung zwischen den Schichtzeiten und dem Luftverbrauch besser zu verstehen.

Nächste Frage: Gibt es zusätzliche Kontextfaktoren, die den Luftverbrauch oder den Druck im System beeinflussen könnten? Denken Sie beispielsweise an Umgebungstemperaturen, die Anzahl der aktiven Maschinen oder Wartungszyklen.

**Anonym**

nur die Anzahl der verschiedenen Maschinen

**Expert Interviewer**

Verstanden, die Anzahl der verschiedenen Maschinen ist also ein weiterer relevanter Kontextfaktor.

Nun zur nächsten Frage: Gibt es eine zeitliche Verzögerung zwischen der Änderung in den unabhängigen Variablen (wie Schichtzeiten und Anzahl der Maschinen) und der Änderung des Drucks, der abhängigen Variable? Zum Beispiel, dauert es eine gewisse Zeit, bis sich eine Änderung in den Schichtzeiten oder der Anzahl der Maschinen auf den Druck im System auswirkt?

**Anonym**

ja

**Expert Interviewer**

Verstanden, es gibt also eine zeitliche Verzögerung zwischen den Änderungen in den unabhängigen Variablen (Schichtzeiten und Anzahl der Maschinen) und der Auswirkung auf den Druck im System.

Nun zur Datenverfügbarkeit: Welche Daten stehen Ihnen aktuell zur Verfügung, die für das Training des maschinellen Lernmodells genutzt werden könnten? Sind diese Daten repräsentativ für die normalen Betriebsbedingungen? Zum Beispiel Daten über den aktuellen Druck im System, Schichtpläne, Anzahl der Maschinen, etc.

**Anonym**

ja

**Expert Interviewer**

Großartig, Sie haben also Zugang zu Daten, die repräsentativ für die normalen Betriebsbedingungen sind. Diese umfassen den aktuellen Druck im System, Schichtpläne und die Anzahl der Maschinen.

Zum Abschluss, wie würden Sie am Ende des Projekts messen, ob Ihr Problem durch das maschinelle Lernen gelöst wurde? Welche Kriterien oder Metriken würden Sie verwenden, um den Erfolg des Projekts zu bewerten?

**Anonym**

eine übergeordnete steuerung, die ausgelesen werden kann

**Expert Interviewer**

Verstanden, der Erfolg des Projekts kann durch eine übergeordnete Steuerung bewertet werden, die auslesbar ist. Dies könnte bedeuten, dass Sie nach der Implementierung des maschinellen Lernmodells prüfen, ob die Steuerung des Druckluftsystems effizienter und den variierenden Anforderungen entsprechend angepasst wird.

Basierend auf unseren Diskussionen, hier ist ein vereinfachtes Modell für Ihr maschinelles Lernprojekt:

1. Abhängige Variable (Ziel):

- Druck im System zu einem bestimmten Zeitpunkt.

2. Unabhängige Variablen (Einflussfaktoren):



- Schichtzeitpunkte (kategorial): Frühschicht, Spätschicht, Nachtschicht, etc.
- Anzahl der aktiven Maschinen (kontinuierlich oder kategorial, je nach Datenerfassung).

### 3. Beziehungen:

- Höherer Luftverbrauch und damit höherer Druckbedarf während der Frühschicht und niedrigerer Verbrauch während der Pausenzeiten.
- Zeitliche Verzögerung zwischen der Änderung in Schichtzeiten bzw. Maschinenanzahl und der Auswirkung auf den Druck.

### 4. Zusätzliche Kontextfaktoren:

- Möglicherweise Umgebungseinflüsse, Wartungszustände oder andere Betriebsbedingungen.

### 5. Daten:

- Aktuelle Druckdaten, Schichtpläne, Anzahl der Maschinen.

### 6. Erfolgsmessung:

- Effizienz und Anpassungsfähigkeit der übergeordneten Druckluftsteuerung.

Für die Umsetzung dieses Modells in ein maschinelles Lernprojekt empfehle ich, die Daten sorgfältig zu sammeln und aufzubereiten, insbesondere die Zeitverzögerungen zwischen den Änderungen der unabhängigen Variablen und den Auswirkungen auf den Druck zu erfassen. Ein Data Scientist kann dann ein geeignetes maschinelles Lernmodell auswählen und trainieren, zum Beispiel ein Zeitreihen-Vorhersagemodell oder ein regelbasiertes System, je nach Datenstruktur und -qualität.

Haben Sie noch weitere Fragen oder benötigen Sie weitere Unterstützung bei der Modellentwicklung?

**Anonym**

nein

### **Expert Interviewer**

In Ordnung, falls Sie in der Zukunft weitere Unterstützung benötigen oder Fragen haben, stehe ich Ihnen gerne zur Verfügung. Viel Erfolg bei Ihrem Projekt zur Druckluftsteuerung! Auf Wiedersehen!

### **Anonym**

danke

### **Expert Interviewer**

Gern geschehen! Wenn Sie in Zukunft weitere Fragen haben, zögern Sie nicht, mich zu kontaktieren. Alles Gute für Ihr Projekt! Auf Wiedersehen!

## **Interview 2: Qualität der Kunststoffteile (28. November 2023)**

### **Anonym**

hi

### **Expert Interviewer**

Hallo! Wie kann ich Ihnen helfen, maschinelles Lernen auf Ihr spezifisches Problem anzuwenden? Können Sie mir zuerst Ihren Arbeitsbereich und das Problem, das Sie lösen möchten, beschreiben?

### **Anonym**

ich möchte gerne anhand von datenanalyse nicht voll ausgespritzte teile im kunststoffprozess entdecken oder vorhersagen

### **Expert Interviewer**

Verstehe, Sie möchten also maschinelles Lernen nutzen, um im Kunststoffverarbeitungsprozess nicht vollständig ausgespritzte Teile zu identifizieren oder deren Auftreten vorherzusagen. Um Ihnen bei der Erstellung eines mathematischen Modells zu helfen, benötige ich einige weitere Informationen.

Lassen Sie uns mit den abhängigen Variablen beginnen, also dem, was Sie erreichen möchten. In Ihrem Fall scheint das die Erkennung oder Vorhersage von nicht vollständig ausgespritzten Teilen zu sein. Wie würden Sie dieses Ergebnis quantifizieren oder messen? Gibt es bestimmte Merkmale oder Indikatoren, die anzeigen, dass ein Teil nicht vollständig ausgespritzt ist?

**Anonym**

ja, es ist ein optischer Fehler, der immer an 1 oder 2 vorhersehbaren Stellen am Einzelteil auftritt

**Expert Interviewer**

Verstanden. Der optische Fehler an ein oder zwei vorhersehbaren Stellen am Einzelteil ist also die abhängige Variable, die wir messen möchten.

Nun zu den unabhängigen Variablen, also den Faktoren, die möglicherweise Einfluss auf das Auftreten dieser Fehler haben. Können Sie einige dieser Faktoren benennen? Welche Aspekte des Kunststoffverarbeitungsprozesses könnten Ihrer Meinung nach dazu beitragen, dass diese Fehler auftreten?

**Anonym**

meinst du mit Faktoren die Ursache?

**Expert Interviewer**

Ja, genau. Mit Faktoren meine ich die möglichen Ursachen oder Einflüsse, die dazu führen könnten, dass der optische Fehler auftritt. Das können Aspekte des Herstellungsprozesses, der verwendeten Materialien, der Maschineneinstellungen oder ähnliche Dinge sein. Können Sie einige dieser potenziellen Ursachen nennen?

**Anonym**

ok, die Ursache ist meistens durch Wechselwirkung zwischen Material und Ist-Parameter. Das bedeutet also, dass beispielsweise Schwankungen im Material die Stabilität der Prozessparameter negativ beeinflussen

### **Expert Interviewer**

Vielen Dank für die Klarstellung. Die Wechselwirkung zwischen dem Material und den Ist-Parametern des Prozesses ist also eine wesentliche Ursache für das Problem.

Um dies weiter zu spezifizieren:

a) Wie können diese Wechselwirkungen gemessen oder quantifiziert werden? Gibt es spezifische Parameter des Materials oder des Prozesses, die Sie bereits überwachen?

### **Anonym**

ja, wir überwachen einspritzzeit, restmassepolster, einspritzdruck und dosierzeit und belege sie mit einem toleranzfenster

### **Expert Interviewer**

Ausgezeichnet, diese überwachten Parameter – Einspritzzeit, Restmassepolster, Einspritzdruck und Dosierzeit – sind wichtige unabhängige Variablen. Sie haben auch erwähnt, dass Sie diese mit einem Toleranzfenster belegen.

Nun zur nächsten Frage:

b) Sind diese Variablen (Einspritzzeit, Restmassepolster, Einspritzdruck, Dosierzeit) und die Wechselwirkungen zwischen Material und diesen Parametern kontinuierlich oder kategorial? Zum Beispiel, werden Einspritzzeit und Druck als kontinuierliche Werte (wie 2,5 Sekunden, 150 bar) oder in Kategorien (wie hoch, mittel, niedrig) erfasst?

### **Anonym**

es sind kontinuierliche werte

### **Expert Interviewer**

Verstanden, diese Variablen sind also kontinuierlich. Das ist wichtig für die Art des mathematischen Modells, das wir entwickeln.

c) Gibt es bekannte Regeln oder Beziehungen zwischen diesen Variablen (Einspritzzeit, Restmassepolster, Einspritzdruck, Dosierzeit) und dem Auftreten des optischen Fehlers?

Zum Beispiel, wenn der Einspritzdruck zu hoch oder zu niedrig ist, tritt der Fehler häufiger auf? Oder wenn die Einspritzzeit außerhalb eines bestimmten Bereichs liegt, erhöht sich das Risiko eines Fehlers?

### **Anonym**

tendenziell kann man sagen, dass sich das Risiko auf nicht voll ausgespritzte Teile erhöht indem moment wo der Einspritzdruck sinkt. Sinken bedeutet vom Referenzpunkt abweichen. Da der Einspritzdruck eigentlich konstant ist, kann man Schwankungen im Material erkennen, wenn er sinkt.

### **Expert Interviewer**

Das ist eine wichtige Information. Also besteht eine Beziehung zwischen dem Sinken des Einspritzdrucks (vom Referenzpunkt abweichen) und dem erhöhten Risiko für nicht vollständig ausgespritzte Teile. Dies deutet darauf hin, dass der Einspritzdruck ein Schlüsselfaktor ist, der eng überwacht werden sollte.

d) Könnten Sie typische Beispielsituationen aus Ihrer Erfahrung nennen, in denen diese Art von Fehlern auftrat? Dies könnte uns helfen, das Problem besser zu verstehen und das Modell präziser zu gestalten.

### **Anonym**

Ich weiß noch nicht was eine typische Beispielsituation sein könnte.

### **Expert Interviewer**

Kein Problem, wir können diesen Punkt überspringen und direkt zur nächsten Frage übergehen.

e) Gibt es zusätzliche Kontextfaktoren, die berücksichtigt werden sollten? Zum Beispiel Umgebungsbedingungen wie Temperatur oder Luftfeuchtigkeit, die Einfluss auf den Prozess haben könnten? Oder andere Faktoren, die nicht direkt mit den bisher genannten Parametern zusammenhängen, aber dennoch relevant sein könnten?

**Anonym**

ja, umgebungs-klima hat einfluss auf das granulat. vorwiegend luftfeuchtigkeit

**Expert Interviewer**

Vielen Dank für diese Information. Die Luftfeuchtigkeit als Teil des Umgebungsklimas kann also das Granulat beeinflussen und ist somit ein weiterer wichtiger Faktor in unserem Modell.

f) Gibt es Zeitverzögerungen zwischen der Änderung der unabhängigen Variablen (wie Einspritzdruck, Luftfeuchtigkeit) und der Änderung des abhängigen Variablen (dem Auftreten des optischen Fehlers)? Anders gefragt: Wenn sich zum Beispiel der Einspritzdruck oder die Luftfeuchtigkeit ändert, wie schnell wirkt sich das auf die Qualität der Teile aus?

**Anonym**

wenn sich der einspritzdruck ändert, also sinkt, dann wirkt es sich inline auf das teil aus. wenn sich luftfeuchtigkeit ändert, dann wirkt sich das verzögert aus. es ist also so, dass sich erst die luftfeuchtigkeit (warscheinlich) ändert, bevor dann irgendwann der Einspritzdruck sinkt oder steigt

**Expert Interviewer**

Das ist eine wichtige Unterscheidung. Direkte Auswirkungen des Einspritzdrucks auf das Teil und verzögerte Effekte der Luftfeuchtigkeit auf das Granulat und letztendlich auf den Einspritzdruck.

g) Welche Daten stehen Ihnen zur Verfügung, und sind diese Daten repräsentativ? Haben Sie beispielsweise historische Daten über Einspritzdruck, Luftfeuchtigkeit, und aufgetretene Fehler? Sind diese Daten umfassend und decken sie verschiedene Situationen und Bedingungen ab?

**Anonym**

nein, es sind noch keine daten vorhanden

## **Expert Interviewer**

Okay, das Fehlen von historischen Daten ist eine Herausforderung, aber nicht unüberwindbar. Für die Entwicklung eines präzisen maschinellen Lernmodells wäre es ideal, Daten über einen gewissen Zeitraum zu sammeln, einschließlich der Einspritzdruckwerte, Luftfeuchtigkeitswerte und der Auftretshäufigkeit des optischen Fehlers. Diese Daten würden es ermöglichen, Muster und Beziehungen zwischen den Variablen zu identifizieren.

Zum Abschluss, wie würden Sie am Ende des Projekts messen, ob Ihr Problem durch maschinelles Lernen gelöst wurde oder nicht? Welche Kriterien oder Metriken würden Sie verwenden, um den Erfolg zu bewerten?

## **Anonym**

da gibt es 2 möglichkeiten. low cost variante wäre: Stichproben. best practice wäre: Kameraprüfung, also 100% prüfung

## **Expert Interviewer**

Verstanden. Die Erfolgsmessung könnte entweder durch Stichproben oder, in einer umfassenderen Variante, durch eine 100%ige Prüfung mittels Kamerakontrolle erfolgen. Beide Methoden haben ihre Vor- und Nachteile in Bezug auf Kosten und Genauigkeit.

Basierend auf den Informationen, die Sie mir gegeben haben, könnte das mathematische Modell wie folgt aussehen:

### **1. Abhängige Variable (Zielvariable):**

- Auftreten des optischen Fehlers (Ja/Nein oder Wahrscheinlichkeit eines Fehlers)

### **2. Unabhängige Variablen (Einflussfaktoren):**

- Einspritzdruck (kontinuierlich)
- Luftfeuchtigkeit (kontinuierlich)
- Einspritzzeit (kontinuierlich)
- Restmassepolster (kontinuierlich)

- Dosierzeit (kontinuierlich)

### **3. Beziehungen:**

- Direkter Einfluss des Einspritzdrucks auf die Fehlerwahrscheinlichkeit
- Indirekter und verzögerter Einfluss der Luftfeuchtigkeit auf den Einspritzdruck und somit auf die Fehlerwahrscheinlichkeit
- Mögliche Wechselwirkungen zwischen den anderen Parametern und der Fehlerwahrscheinlichkeit

### **4. Datenerfassung und Modelltraining:**

- Da aktuell keine Daten vorhanden sind, empfiehlt es sich, Daten über einen Zeitraum zu sammeln, um das Modell zu trainieren und zu validieren.

### **5. Erfolgsmessung:**

- Über Stichproben oder eine 100%ige Kameraprüfung.

Ein Vorschlag wäre, ein maschinelles Lernmodell zu entwickeln, das auf Regressionsanalyse oder Klassifikation basiert, je nachdem, ob Sie die Wahrscheinlichkeit eines Fehlers vorhersagen oder eine Ja/Nein-Entscheidung treffen wollen. Es könnte sinnvoll sein, mit einem einfacheren Modell zu beginnen und es zu verfeinern, wenn mehr Daten verfügbar sind.

Haben Sie noch weitere Fragen oder benötigen Sie zusätzliche Informationen?

### **Anonym**

gibt es vorschläge deinerseits zum stichproben iintervall? Also wie häufig sollten stichproben durchgeführt werden um das modell angemessen zu trainieren, wenn ich keine 100% implementieren kann



## Expert Interviewer

Die Festlegung des Stichprobenintervalls für die Datenerfassung ist eine wichtige Entscheidung, die die Qualität und Aussagekraft Ihres maschinellen Lernmodells beeinflussen kann. Hier einige Überlegungen und Vorschläge:

1. **Variabilität des Prozesses:** Betrachten Sie, wie oft sich die Bedingungen in Ihrem Prozess ändern. Wenn Sie beispielsweise häufige Schwankungen in Faktoren wie Einspritzdruck oder Luftfeuchtigkeit haben, sollten Sie häufiger Stichproben nehmen, um diese Variabilität zu erfassen.
2. **Auftretenshäufigkeit des Fehlers:** Wenn der Fehler relativ selten auftritt, benötigen Sie mehr Stichproben über einen längeren Zeitraum, um genügend Instanzen des Fehlers für das Training des Modells zu sammeln.
3. **Ressourcen und Machbarkeit:** Berücksichtigen Sie Ihre verfügbaren Ressourcen. Häufigere Stichproben bieten mehr Daten, können aber auch teurer und aufwendiger in der Handhabung sein.
4. **Statistische Signifikanz:** Je mehr Daten Sie haben, desto genauer kann das Modell werden. Allerdings gibt es einen Punkt, an dem mehr Daten nur noch marginale Verbesserungen bringen.

Als allgemeine Empfehlung könnte ein gutes Ausgangsintervall für Stichproben einmal pro Schicht oder einmal täglich sein, je nachdem, wie schnell sich Ihre Prozessbedingungen ändern. Wenn möglich, versuchen Sie, Stichproben in verschiedenen Betriebsbedingungen (z.B. zu verschiedenen Tageszeiten, unter verschiedenen Wetterbedingungen) zu nehmen, um eine breite Datenbasis zu gewährleisten.

Es wäre auch hilfreich, eine Pilotphase durchzuführen, in der Sie das Stichprobenintervall testen und anpassen, um zu sehen, wie gut es funktioniert und ob die gesammelten Daten repräsentativ und nützlich für das Training Ihres Modells sind.

Haben Sie weitere Fragen oder gibt es noch etwas, bei dem ich helfen kann?