

---

# From Assistance to Empowerment: Human-AI Collaboration in High-Risk Decision Making

---



TECHNISCHE  
UNIVERSITÄT  
DARMSTADT

Department of Law and Economics  
at the Technical University of Darmstadt

## **Dissertation**

by

Sara Jourdan (formerly Ellenrieder)

submitted in fulfilment of the requirements for the degree of  
Doctor rerum politicarum (Dr. rer. pol.)

First Assessor: Prof. Dr. Peter Buxmann  
Second Assessor: Prof. Dr. Alexander Benlian

Darmstadt 2024

---

Sara Jourdan (formerly Ellenrieder)

From Assistance to Empowerment: Human-AI Collaboration in High-Risk Decision Making

Darmstadt, Technical University of Darmstadt

Year thesis published in TUprints: 2024

Date of the viva voce: 11.11.2024

TUprints under CC BY-SA 4.0 International

<https://creativecommons.org/licenses/>

**Declaration of Authorship**

I hereby declare that the submitted thesis is my own work. All quotes, whether word by word or in my own words, have been marked as such.

The thesis has not been published anywhere else nor presented to any other examination board.

Ich erkläre hiermit ehrenwörtlich, dass ich die vorliegende Arbeit selbstständig angefertigt habe. Sämtliche aus fremden Quellen direkt oder indirekt übernommenen Gedanken sind als solche kenntlich gemacht.

Die Arbeit wurde bisher weder einer anderen Prüfungsbehörde vorgelegt noch veröffentlicht.

Sara Jourdan (formerly Ellenrieder)

Darmstadt, November 16, 2024

## Abstract

The increasing availability of large amounts of valuable data and the development of ever more powerful machine learning (ML) algorithms enable ML systems to quickly and independently identify complex relationships in data. As a result, ML systems not only generate new knowledge, but also offer significant potential to augment human capabilities and assist decision makers in challenging tasks.

In high-risk areas such as aviation or healthcare, humans retain final decision-making responsibility, but will increasingly collaborate with ML systems to improve decision-making processes. However, since ML systems rely on statistical approaches, they are susceptible to error, and the complexity of modern algorithms often renders the output of ML systems opaque to humans. While initial approaches from the field of explainable artificial intelligence (XAI) aim to make the output of ML systems more understandable and comprehensible to humans, current research investigating the impact of ML systems on human decision makers is limited and lacks approaches on how humans can improve their capabilities through collaboration to make better decisions in the long run. To fully exploit the potential of ML systems in high-risk areas, both humans and ML systems should be able to learn from each other to enhance their performance in the context of collaboration. Furthermore, it is essential to design effective collaboration that considers the unique characteristics of ML systems and enables humans to critically assess system decisions. This dissertation comprises five published papers that use a mixed-methods study, two quantitative experiments and two qualitative design science research (DSR) studies to explore the collaboration and bilateral influences between humans and ML systems in decision-making contexts within high-risk areas from three perspectives: (1) the *human perspective*, (2) the *ML system perspective*, and (3) the *collaborative perspective*.

From a *human perspective*, this dissertation examines how humans can learn from ML systems in collaboration to enhance their own capabilities and avoid the risk of false learning due to erroneous ML output. In a mixed-methods study, radiologists segmented 690 brain tumors in MRI scans supported by either high-performing or low-performing ML systems, which provided explainable or non-explainable output design. The study shows that human decision makers can learn from ML systems to improve their decision performance and confidence. However, incorrect system outputs also lead to false learning and pose risks for decision makers. Explanations from the XAI field can significantly improve the learning success of radiologists and prevent false

learning in the case of incorrect ML system output. In fact, some radiologists were even able to learn from mistakes made by low-performing ML systems when local explanations were provided with the system output. This study provides first empirical insights into the human learning potential in the context of collaborating with ML systems. The finding that explainable design of ML systems enables radiologists to identify erroneous output may facilitate earlier adoption of explainable ML systems that can improve their performance over time.

The *ML system perspective*, on the other hand, examines how ML systems must be designed to respond flexibly to changes in human problem perception and their dynamic deployment environment. This allows the systems to also learn from humans and ensures reliable system performance in dynamic collaborative environments. Through 15 qualitative interviews with data science and ML experts in the context of a DSR study, challenges for the long-term deployment of ML systems are identified. The results show that the requirements for flexible adaptation of systems in long-term use must be established in the early phases of the ML development process. Tangible design requirements and principles for ML systems that can learn from their environment and humans are derived for all phases of the CRISP-ML(Q) process model for the development and deployment of ML models. Implementing these principles allows ML systems to maintain or even improve their performance in the long run despite occurring changes, thus creating the prerequisites for a sustainable lifecycle of ML systems.

Finally, the *collaborative perspective* examines how the collaboration between humans and ML systems should be designed to account for the unique characteristics of ML systems, such as error proneness and opacity, as well as the cognitive biases that are inherent to human decision making. In this context, pilots were provided with different ML systems for the visual detection of other aircraft in the airspace during 222 recorded flight simulations. The experiment examines the influence of different ML error types and XAI approaches in collaboration, and shows that an explainable output design can significantly reduce ML error-induced pilot trust and performance degradation for individual error types. However, processing explanations from the XAI field increases the pilot's mental workload. While ML errors erode the trust of human decision makers, a DSR study is conducted to derive design principles for acceptance-promoting artifacts for collaboration between humans and ML systems. Finally, the last part of the analysis shows how cognitive biases such as the IKEA effect cause humans to overvalue the results of collaboration with ML systems when a high level of personal effort is invested in the collaboration. The findings provide a broad foundation for designing effective human-AI collaboration in organizations, especially in high-risk areas where humans will be involved in decision making for the long term.

Overall, the papers show how by designing effective collaboration, both humans and ML systems can benefit from each other in the long run and enhance their own capabilities. The explainable design of ML system outputs can serve as a catalyst for the adoption of ML systems, especially in

high-risk areas. This dissertation defines novel requirements for the collaboration between humans and ML systems and provides guidance for ML developers, scientists, and organizations that aspire to involve both human decision makers and ML systems in decision-making processes and ensure high and robust performance in the long term.

**Abstract (German version)**

Die zunehmende Verfügbarkeit großer Mengen an wertvollen Daten und die Entwicklung immer leistungsfähiger Algorithmen im Bereich des maschinellen Lernens (ML) erlauben es ML-Systemen auch komplexe Zusammenhänge in Daten schnell und eigenständig zu erkennen. Hierdurch können ML-Systeme nicht nur neues Wissen generieren, sondern bieten insbesondere großes Potential, menschliche Fähigkeiten zu erweitern und Entscheidungsträger auch bei anspruchsvollen Tätigkeiten zu unterstützen.<sup>1</sup>

In Hochrisikobereichen wie der Luftfahrt oder dem Gesundheitswesen trägt der Mensch die finale Entscheidungsverantwortung, wird allerdings zunehmend mit ML-Systemen kollaborieren, um Entscheidungsprozesse zu verbessern. Da ML-Systeme jedoch auf statistischen Ansätzen beruhen, sind sie fehleranfällig und die Komplexität moderner Algorithmen lässt ML-Systemausgaben für den Menschen oft undurchsichtig erscheinen. Während erste Ansätze aus dem Forschungsfeld der erklärbaren künstlichen Intelligenz (XAI) bereits darauf abzielen ML-Systemausgaben für den Menschen verständlicher und nachvollziehbarer zu gestalten, greift die aktuelle Forschung, die den Einfluss von ML-Systemen auf den menschlichen Entscheidungsträger untersucht, zu kurz. Es fehlt an Ansätzen, die es dem Menschen ermöglichen, seine Fähigkeiten durch die Kollaboration zu erweitern, um langfristig bessere Entscheidungen zu treffen. Um das Potential von ML-Systemen in Hochrisikobereichen ausschöpfen zu können, ist es erforderlich, dass sowohl der Mensch als auch das System voneinander lernen können. Auf diese Weise kann die Leistungsfähigkeit beider Parteien im Rahmen der Kollaboration verbessert werden. Darüber hinaus muss eine effektive Zusammenarbeit gestaltet werden, die die besonderen Eigenschaften von ML-Systemen berücksichtigt und dem Menschen erlaubt Systementscheidungen kritisch zu hinterfragen. Die vorliegende Dissertation umfasst fünf veröffentlichte Beiträge, die mittels einer Mixed-Methods-Studie, zwei quantitativen Experimenten und zwei qualitativen Design Science Research (DSR) Studien die Kollaboration und bilateralen Einflüsse zwischen Mensch und ML-System im Kontext von Entscheidungen in

---

<sup>1</sup> Im Sinne einer verbesserten Lesbarkeit wird in diesem Text das generische Maskulinum verwendet. Dies schließt explizit Personen aller Geschlechteridentitäten ein.

Hochrisikobereichen aus drei Perspektiven untersuchen: der (1) *Mensch-*, (2) *ML-System-* und (3) *Kollaborationsperspektive*.

Im Rahmen der *Mensch-Perspektive* wird untersucht, wie der Mensch von ML-Systemen in der Kollaboration lernen kann, um seine eigenen Fähigkeiten auszubauen und Risiken für falsches Lernen durch fehlerhafte ML-Systemausgaben vermieden werden. In einer Mixed-Methods Studie segmentieren Radiologen 690 Hirntumore in MRT-Bildern unter Einbezug leistungsstarker oder leistungsschwacher ML-Systeme, die ein erklärbares oder nicht-erklärbares Design für Systemausgaben bereitstellen. Die Studie offenbart, dass menschliche Entscheidungsträger von ML-Systemen lernen können, um ihre Entscheidungsperformance und -sicherheit zu verbessern. Im Falle von fehlerhaften Systemausgaben führt dies jedoch auch zu falschem Lernen und birgt ein Risiko für den Entscheidungsträger. Erklärungen aus dem XAI-Bereich können den Lernerfolg der Radiologen signifikant verbessern und verhindern falsches Lernen im Falle inkorrekturer ML-Systemausgaben. Tatsächlich können einige Radiologen sogar von Fehlern leistungsschwacher ML-Systeme lernen, wenn mit den Systemausgaben lokale Erklärungen bereitgestellt werden. Die Studie liefert erste empirische Erkenntnisse zum menschlichen Lernpotential im Rahmen der Kollaboration mit ML-Systemen. Die Erkenntnis, dass erklärbares Design von ML-Systemen Radiologen befähigt, fehlerhafte Ausgaben zu identifizieren, kann eine frühere Adoption von erklärbaren ML-Systemen, die ihre Leistungsfähigkeit über die Zeit ausbauen können, ermöglichen.

Die *ML-Systemperspektive* untersucht hingegen, wie ML-Systeme entwickelt werden müssen, sodass diese flexibel auf Änderungen in der Problemwahrnehmung des Menschen und ihrer dynamischen Umgebung reagieren können. Dies erlaubt den Systemen auch vom Mensch zu lernen und stellt eine zuverlässige Leistungsfähigkeit der Systeme in dynamischen Kollaborationsumgebungen sicher. Mittels 15 qualitativer Interviews mit Data Science und ML-Experten im Rahmen einer DSR-Studie werden Herausforderungen für den langfristigen Einsatz von ML-Systemen identifiziert. Die Ergebnisse verdeutlichen, dass bereits in frühen Phasen im ML-Entwicklungsprozess Voraussetzungen für die flexible Anpassung der Systeme im realen Langzeiteinsatz geschaffen werden müssen. Es werden konkrete Designanforderungen und Designprinzipien für ML-Systeme, die von ihrer Umgebung und dem Menschen lernen können, für alle Phasen des CRISP-ML(Q) Prozessmodells für die Entwicklung und den Einsatz von ML-Systemen abgeleitet. Die Umsetzung dieser Prinzipien erlaubt ML-Systemen ihre Performance auch langfristig trotz auftretender Veränderungen zu erhalten oder sogar zu verbessern und schafft damit die Voraussetzungen für einen nachhaltigen Lebenszyklus von ML-Systemen.

Abschließend untersucht die (3) *Kollaborationsperspektive*, wie die Zusammenarbeit von Mensch und ML-System gestaltet sein sollte, um die speziellen Eigenschaften wie Fehleranfälligkeit und Undurchsichtigkeit von ML-Systemen, aber auch kognitive Verzerrungen, die beim menschlichen



Entscheidungsträger auftreten zu berücksichtigen und vorteilhaft in die Kollaboration einzubeziehen. Hierbei werden Piloten für 222 aufgezeichnete Flugsimulationen verschiedene ML-Systeme zur visuellen Detektion anderer Flugzeuge im Luftraum bereitgestellt. Das Experiment untersucht den Einfluss verschiedener ML-Fehlertypen und XAI-Ansätze in der Kollaboration und zeigt, dass erklärbares Design ML-fehlerbedingte Vertrauens- und Leistungseinbrüche für einzelne Fehlertypen signifikant reduzieren kann. Allerdings erhöht die Verarbeitung der Erklärungen die mentale Arbeitslast der Piloten. Während ML-Fehler das Vertrauen von menschlichen Entscheidungsträgern schädigen, werden in einer DSR-Studie zudem Designprinzipien für akzeptanzfördernde Artefakte für die Zusammenarbeit von Mensch und ML-System abgeleitet. Abschließend wird im letzten Teil der Analyse gezeigt wie kognitive Verzerrungen wie der IKEA-Effekt verursachen, dass Menschen die Ergebnisse der Kollaboration mit ML-Systemen mehr wertschätzen, wenn ein hohes Maß an eigenem Aufwand in die Kollaboration eingebracht wurde. Die Erkenntnisse bieten ein breites Fundament für die Gestaltung wirkungsvoller Kollaboration in Organisationen und insbesondere Hochrisikobereichen, wo Menschen auch langfristig in die Entscheidungsfindung eingebunden sein werden.

Übergreifend zeigen die Studien, wie über die Gestaltung effektiver Kollaboration sowohl Menschen als auch ML-Systeme langfristig voneinander profitieren und ihre eigenen Fähigkeiten verbessern können. Dabei kann erklärbares Design von ML-Systemausgaben als Katalysator für die Adoption von ML-Systemen insbesondere in Hochrisikobereichen dienen. Diese Dissertation definiert neue Ansprüche für die Kollaboration von Mensch und ML-System und bietet Orientierung für ML-Entwickler, Wissenschaftler und Organisationen, die sowohl menschliche Entscheidungsträger als auch ML-Systeme in Entscheidungsprozesse einbeziehen und eine langfristig hohe Leistungsfähigkeit sicherstellen möchten.

**Table of Contents**

<b>List of Figures</b> .....	<b>XIII</b>
<b>List of Tables</b> .....	<b>XIV</b>
<b>List of Abbreviations</b> .....	<b>XV</b>
<b>1. Introduction</b> .....	<b>1</b>
1.1 Overarching Motivation.....	1
1.2 Overarching Research Questions.....	3
1.3 Structure of the Dissertation .....	5
<b>2. Research Context</b> .....	<b>11</b>
2.1 Artificial Intelligence and Machine Learning.....	11
2.2 Explainable AI .....	13
2.3 Human-AI Collaboration.....	15
2.4 Theoretical Foundations .....	16
2.4.1 Learning Theories.....	17
2.4.2 Error Management Theory .....	18
2.4.3 Unified Theory of Acceptance and Use of Technology .....	19
2.4.4 Cognitive Biases.....	20
<b>3. Paper A: Promoting Learning Through Explainable Artificial Intelligence: An Experimental Study in Radiology</b> .....	<b>22</b>
3.1 Introduction.....	23
3.2 Theoretical Background .....	25
3.2.1 Machine Learning .....	25
3.2.2 ML-based Decision Support Systems for Medical Diagnostics.....	26
3.2.3 Explainable Design of ML Systems .....	26
3.2.4 Promoting Human Learning Through Machine Learning Systems .....	27
3.2.5 Human Learning.....	28
3.2.6 Hypothesizing Human Learning Through Interaction With ML-based DSSs.....	30
3.3 Methodology.....	30
3.3.1 Empirical Context .....	31
3.3.2 Research Design .....	32
3.4 Results.....	37
3.5 Discussion.....	40

---

3.6	Conclusion.....	43
<b>4.</b>	<b>Paper B: Toward the Sustainable Development of Machine Learning Applications in Industry 4.0 .....</b>	<b>44</b>
4.1	Introduction.....	45
4.2	Related Work.....	47
4.2.1	Definition of ML and Process Models for Structuring ML Projects .....	47
4.2.2	Sustainable Long-term Deployment of ML Systems.....	49
4.3	Research Methodology.....	50
4.3.1	Design Science Research Approach.....	50
4.3.2	Semi-structured Expert Interviews.....	51
4.4	Results.....	52
4.4.1	Challenges for Long-term Operation of ML Systems in Industry 4.0 .....	53
4.4.2	Deriving Design Requirements and Principles for a Solution Instantiation .....	57
4.5	Discussion and Conclusion .....	61
4.6	Acknowledgements .....	62
<b>5.</b>	<b>Paper C: Pilots and Pixels: A Comparative Analysis of Machine Learning Error Effects on Aviation Decision Making.....</b>	<b>63</b>
5.1	Introduction.....	64
5.2	Theoretical Background .....	66
5.2.1	Machine Learning .....	66
5.2.2	ML Error Types and Error Management Theory .....	66
5.2.3	Explainable Design of ML Systems .....	68
5.2.4	Collaborative Decision Making: Performance, Trust, and Mental Workload.....	69
5.2.5	Hypothesizing Human Interaction With Erroneous ML-based DSSs .....	70
5.3	Methodology.....	71
5.3.1	Empirical Context .....	71
5.3.2	Study Procedure and Data Collection.....	72
5.3.3	Experimental Setup.....	74
5.3.4	Data Analysis and Pilot Statistics .....	76
5.4	Results.....	76
5.5	Discussion.....	79
5.6	Limitations and Future Research Directions .....	81
<b>6.</b>	<b>Paper D: Design for Acceptance and Intuitive Interaction: Teaming Autonomous Aerial Systems with Non-experts.....</b>	<b>82</b>
6.1	Introduction.....	83
6.2	Theoretical Background .....	84
6.2.1	Interaction for Human–Autonomy Teaming .....	85
6.2.2	Designing Human–Drone Collaboration .....	85

---

6.2.3 Unified Theory of Acceptance and Use of Technology .....	86
6.3 Design Science Research Approach .....	88
6.4 Results.....	91
6.4.1 Awareness of the Problem .....	91
6.4.2 Suggestion.....	92
6.4.3 Deriving Design Principles.....	95
6.4.4 Development.....	96
6.4.5 Evaluation and Discussion of Results.....	97
6.5 Conclusion, Limitations, and Directions for Future Research .....	102
<b>7. Paper E: The Influence of Effort on the Perceived Value of Generative AI: A Study of the IKEA Effect .....</b>	<b>105</b>
7.1 Introduction.....	106
7.2 Theoretical Background .....	107
7.2.1 Artificial Intelligence .....	107
7.2.2 Cognitive Bias .....	108
7.2.3 Understanding the IKEA Effect.....	109
7.2.4 Hypothesizing the IKEA Effect in Human-AI Collaboration .....	111
7.3 Methodology.....	113
7.3.1 Research Design and Measures.....	114
7.3.2 Data Collection and Sample .....	117
7.4 Data Analysis and Results .....	117
7.5 Discussion.....	121
7.5.1 Contributions.....	122
7.5.2 Limitations and Future Research.....	123
7.6 Acknowledgement .....	124
<b>8. Discussion of Contributions and Conclusion .....</b>	<b>125</b>
8.1 Theoretical Contributions.....	125
8.2 Practical Contributions .....	128
8.3 Concluding Remarks .....	130
<b>References .....</b>	<b>133</b>

**List of Figures**

Figure 1: Outline of the dissertation .....	8
Figure 2: The UTAUT model with its four direct determinants of user acceptance and usage behavior (Venkatesh et al., 2003).....	20
Figure 3: Exemplary MRI scan of a high-grade glioma with edema in FLAIR sequence (Bakas et al., 2017; Kaggle, 2020; Menze et al., 2015) .....	31
Figure 4: Procedure of the study, which includes pre- and post-experimental interviews (gray), and the segmentation tasks within the experiment (blue).....	32
Figure 5: Between-subject experimental study design for segmentation of brain tumors and different designs of ML-based DSSs.....	34
Figure 6: Measured deltas of pre- and post-experiment IoU-based performance and decision confidence .....	37
Figure 7: Qualitative comparison of the CRISP-DM (Wirth and Hipp, 2000) model phases to the more recent CRISP-ML (Studer et al., 2021), which provides the structure for the categorization of challenges and derived design requirements.....	48
Figure 8: Design principles for ML applications suitable for long-term deployment in dynamic environments .....	60
Figure 9: Design and procedure of the between-subject online experiment for detecting aircraft in vicinity with different variants of erroneous ML-based DSSs.....	73
Figure 10: Exemplary MSFS2020 scene with detected aircraft in magenta bounding box (NEEF).....	75
Figure 11: The UTAUT model with its four direct determinants of user acceptance and usage behavior (Venkatesh et al., 2003).....	87
Figure 12: DSR project structure (cf. Kuechler & Vaishnavi, 2008) including expert, non-expert and mixed design cycles. This study focuses on the cycles highlighted in grey.....	89
Figure 13: Design requirements derived from expert interviews.....	93
Figure 14: Instantiated solution for teaming non-experts with autonomous drone delivery systems and respective design features (icons made by Freepik, Pixel Perfect and Satawatdesign from Flaticon, 2022).....	97
Figure 15: a) Experimental setup for the image task and b) example task with AI collaboration. Participants followed the same procedure for the text generation task. ....	116

**List of Tables**

Table 1: List of publications included in this dissertation .....	6
Table 2: Overview of interviewee symbols and corresponding industry roles .....	52
Table 3: Challenges and corresponding solution approaches identified as design requirements.....	58
Table 4: Changes in FPs and FNs for increasing the confidence thresholds of the YOLOv8 model trained on the AOT dataset from 20% to 40%, 40% to 60%, and 60% to 80%.....	72
Table 5: Performance, mental workload and trust assessment following interaction with ML-based DSSs and results of the conducted Welch-ANOVA .....	77
Table 6: Hypotheses testing and results from the online experiment.....	78
Table 7: Roles of expert interview participants .....	90
Table 8: Non-expert focus group study sample.....	99
Table 9: Results of focus group discussions by non-experts on derived design principles .....	101
Table 10: Assessment of reliability and convergent validity (values are 1.000 for one-dimensional constructs).....	118
Table 11: Discriminant validity.....	119
Table 12: Hypotheses testing.....	120

**List of Abbreviations**

ACIS	Australasian Conference on Information Systems
AE	Aviation expert
AED	Automated external defibrillator
AI	Artificial intelligence
ANOVA	Analysis of variance
AOT	Amazon Prime Air airborne object tracking
AVE	Average variance extracted
AWS	Amazon Web Services
BMBF	German Federal Ministry of Education and Research
BVLOS	Beyond visual line of sight
CE	European conformity
CMB	Common method bias
CPS	Cyber-physical systems
CRISP-DM	Cross industry standard process for data mining
CRISP-ML(Q)	Cross industry standard process model for the development of machine learning applications with quality assurance methodology
DAA	Detect and avoid
DE	Drone expert
DF	Design feature
DP	Design principle
DR	Design requirement
DSR	Design science research
DSS	Decision support system
E	Expert
EASA	European Union Aviation Safety Agency

---

ECIS	European Conference on Information Systems
EEF	Explainable error-free
EFN	Explainable false negative
EFP	Explainable false positive
EHP	Explainable high-performing
ELP	Explainable low-performing
EMT	Error management theory
ESF	European social fund
EU	European Union
FAA	Federal Aviation Agency
FDA	Food and Drug Administration
FLAIR	Fluid attenuated inversion recovery
FN	False negative
FP	False positive
G	Group
GenAI	Generative artificial intelligence
GD	Gadolinium
HICSS	Hawaii International Conference on System Sciences
H	Hypothesis
ICIS	International Conference on Information Systems
IoT	Internet of things
IoU	Intersection over union
IP	Interview partner
IS	Information systems
M	Mean
MC	Mass customization
ML	Machine learning
MRI	Magnetic resonance imaging
MSFS	Microsoft flight simulator
NASA	National Aeronautics and Space Administration



---

NEEF	Non-explainable error free
NEFP	Non-explainable false positive
NEFN	Non-explainable false negative
NEHP	Non-explainable high-performing
NELP	Non-explainable low-performing
P	Participant
PACIS	Pacific Asia Conference on Information Systems
RQ	Research question
SD	Standard deviation
SDG	Sustainable development goal
T1CE	Contrast-enhanced T1 sequences
TLX	Task load index
UAS	Unmanned aerial systems
UAV	Unmanned aerial vehicles
UPC-UA	Open platform communications - unified architecture
UTAUT	Unified theory of acceptance and use of technology
VFR	Visual flight rules
VTOL	Vertical take-off and landing
WTP	Willingness to pay
XAI	Explainable artificial intelligence

# 1. Introduction

## 1.1 Overarching Motivation

As the application areas for the use of artificial intelligence (AI) are becoming ever more diverse and the technology is increasingly demonstrating its disruptive potential (e.g., Benbya et al., 2021, 2024; Berente et al., 2021; Jordan & Mitchell, 2015), it is essential to expand our understanding of the mutual impact between humans and AI systems in collaboration (e.g., Abdel-Karim et al., 2023; Alavi et al., 2024; Benbya et al., 2021; Gaube et al., 2023; Pumplun et al., 2023). Today, AI has even found its way into high-risk environments such as healthcare or aviation, supporting tasks previously performed by human experts (e.g., Lebovitz et al., 2022; Reyes et al., 2020; Rudin, 2019; Shen et al., 2019; Sutton et al., 2020). However, due to the great risks associated with decisions in high-risk areas, AI systems are primarily used to augment human capabilities as support systems, which requires humans and AI systems to collaborate for decision making (e.g., Jha & Topol, 2016; Lebovitz et al., 2022; Meskó & Görög, 2020). In this context, the European Aviation Safety Agency (EASA) recently announced that it aims to approve AI systems for human cognitive assistance in decision making for civil aviation by 2025 (European Union Aviation Safety Agency, 2023) and there are already multiple certified AI solutions on the market to support medical diagnostics in radiology (e.g., Benjamins et al., 2020; Radboud University Medical Center, 2023). For instance, AI systems have been developed to assist physicians by analyzing medical imaging data to diagnose cancer, strokes, and other abnormalities (e.g., Calisto et al., 2021; Cheng et al., 2016; Gaube et al., 2023; Jussupow et al., 2021; Lebovitz et al., 2021; Pumplun et al., 2023), achieving diagnostic accuracy which exceeds that of human experts in certain scenarios (e.g., McKinney et al., 2020; Shen et al., 2019).

The capabilities of modern applications in the field of AI are based on machine learning (ML), which is also known as a “general-purpose technology” (Brynjolfsson & Mitchell, 2017). ML-based systems can independently recognize patterns in large datasets and make predictions for new data based on the correlations learned (Brynjolfsson & Mitchell, 2017; Russell & Norvig, 2021). In the case of generative AI (GenAI), advanced ML approaches are used to generate new data based on learned patterns (Benbya et al., 2024; Dwivedi et al., 2023; Teubner et al., 2023). Contrary to traditional information systems, ML systems possess the ability to learn autonomously from examples, eliminating the need for solution instructions created by humans (Samuel, 1959). This

ability empowers ML systems to discover new solutions that humans may not find and thus provide complementary knowledge (Fügenger et al., 2022; Mitchell, 1997). However, ML systems rely on statistical patterns, are highly inscrutable, and remain prone to errors (Berente et al., 2021; Russell & Norvig, 2021). In the context of human-AI collaboration, the probabilistic nature and complexity of modern ML systems pose significant challenges, as end users struggle to comprehend the underlying mechanisms through which these systems arrive at their output (Diakopoulos, 2016; Lebovitz et al., 2022; Reyes et al., 2020; Rudin, 2019). Due to these unique characteristics of ML systems, they are often viewed as “black boxes” (Adadi & Berrada, 2018; Brasse et al., 2023; Guidotti et al., 2018; Rudin, 2019). To address this issue, the research stream of explainable AI (XAI) explores methods and develops explanations that aim to render ML output more human-understandable (Arrieta et al., 2020; Guidotti et al., 2018; Miller, 2019). In addition, the recently agreed EU AI Act also requires that end users of ML systems must be able to understand and interpret the systems’ characteristics and outputs (European Commission, 2023; Panigutti et al., 2023; Sovrano et al., 2022).

Especially in high-risk environments, the use of ML systems requires that decision makers not only critically question the ML output, but also do not lose their own skills and experience performance degradation over time (e.g., Fügenger et al., 2021; Goddard et al., 2012; Jussupow et al., 2021). Several studies have already shown the significant impact of ML systems on end users’ decision-making processes and capabilities (e.g., Abdel-Karim et al., 2023; Fügenger et al., 2021; Gaube et al., 2023; Jussupow et al., 2021; Lebovitz et al., 2022). In this context, recent research emphasizes the importance of exploring how ML systems can not only support human decision makers but also enhance their ability to learn from these systems and improve their performance (e.g., Abdel-Karim et al., 2023; Gaube et al., 2023; Pumplun et al., 2023; Sturm, Koppe, et al., 2021). This also applies vice versa. In the context of collaborative decision making, it is crucial to acknowledge that both humans and ML systems should not be perceived as static entities. Thus, the system design should not only promote human learning but also allow ML systems to learn from and adapt to their dynamic environment to maintain or potentially improve performance over time (e.g., Grønsund & Aanestad, 2020; Jourdan et al., 2021; Sculley et al., 2015; Studer et al., 2021; Stumpf et al., 2009; Sturm, Gerlach, et al., 2021).

This dissertation seeks to explore the multifaceted challenges inherent in collaboration between humans and ML systems. Specifically, it aims to uncover strategies to ensure that both humans and ML systems can continuously improve their capabilities and performance through collaboration, with a focus on further promoting effective teamwork.

## 1.2 Overarching Research Questions

The safe and sustainable deployment of ML systems within high-risk environments requires a symbiotic relationship between responsible human decision makers and ML systems, one where collaboration not only yields successful decision outcomes but also promotes long-term mutual growth and enhancement. To ensure that both parties benefit from collaboration, both humans and ML systems must continuously evolve in order to prevent a decline in performance, while also working effectively as a team. I thus consider the (1) *human*, (2) *ML system*, and (3) *collaborative perspective* of human-AI collaboration in this dissertation.

In recent years, research in the field of human-AI collaboration has focused on the risks that arise for the human decision maker (e.g., Fügener et al., 2021; Jussupow et al., 2021). A notable concern is, for example, the tendency for humans to overly rely on ML recommendations, accepting them without questioning them critically—a phenomenon known as automation bias (Goddard et al., 2012). This poses a particular risk due to the systems' susceptibility to errors (Berente et al., 2021; Russell & Norvig, 2021). However, ML systems also hold substantial promise for professionals operating in high-risk environments: By independently identifying patterns in data (Brynjolfsson & Mitchell, 2017; Russell & Norvig, 2021), ML systems can discover knowledge that augments human capabilities (Fügener et al., 2022; Sturm, Koppe, et al., 2021). This offers the potential for humans to learn from ML systems in order to improve their own capabilities. However, the complexity of ML systems leaves their outputs opaque (Berente et al., 2021; Lebovitz et al., 2022; Russell & Norvig, 2021). While research has shown that explanations from the XAI field enhance human capabilities in assessing ML system performance (Jussupow et al., 2021; Lebovitz et al., 2021) and promote trust (Benbya et al., 2021; Glikson & Woolley, 2020), the potential of ML systems and XAI approaches to facilitate human learning remains unexplored. Instead of risking performance degradation due to human-AI collaboration (e.g., Fügener et al., 2021; Goddard et al., 2012), current research calls for studies to explore how ML systems can promote human learning (Abdel-Karim et al., 2023; Gaube et al., 2023; Pumplun et al., 2023).

Within the *human perspective* of the analysis, I aim to combine findings from the field of human-AI collaboration, XAI and educational research to enhance our understanding of whether and how humans can actually learn from ML systems and how explanations from the field of XAI affect learning outcomes. This novel approach could enable human decision makers to benefit from collaboration with ML systems in high-risk environments and I thus pose the following first, overarching research question (RQ):

**RQ1:** *In order to enhance human performance, how can the human decision maker learn from the ML system?*

However, performance-related challenges that arise during human-AI collaboration are not exclusive to human decision makers. Long-term deployment can also negatively affect the performance of the ML systems (Jourdan et al., 2021; Rudin & Wagstaff, 2014; Salama et al., 2021; Sculley et al., 2015). ML system performance is highly dependent on the quantity and quality of data describing the problem at hand and used for training (Grover et al., 2018; Smith, 2020). While ML systems are trained on static datasets, the human perception of the problem may change during deployment (Russell & Norvig, 2021; Sculley et al., 2015), or other shifts in the dynamic deployment environment may lead to a change in data distribution, risking a degradation in system performance (Jourdan et al., 2021; Sculley et al., 2015; Sturm, Gerlach, et al., 2021). It is important that ML systems learn through feedback from humans, for example, if they have incorrectly assessed the quality of a component, but also if humans want to share new findings about component defects.

While research increasingly addresses the need for continuous auditing and altering activities to adjust systems to dynamic changes in human problem perceptions or the deployment environment, it focuses on approaches such as *human-in-the-loop* (e.g., Amershi et al., 2019; Grønsund & Aanestad, 2020; Stumpf et al., 2009; Sturm, Koppe, et al., 2021). However, existing human-in-the-loop approaches to provide human feedback do not cover the various dynamics in the environment that can weaken the performance of ML systems in the long term. Moreover, these approaches take effect only after deployment and do not explain how ML systems need to be designed and developed to flexibly respond to shifts in data and problem perception, thus making them easier to maintain. Information Systems (IS) research recognizes this challenge and calls for methods to continuously maintain or even improve the performance of ML systems over time (e.g., Audzeyeva & Hudson, 2016; Grønsund & Aanestad, 2020). In addition, it is crucial to prevent a decay in accuracy and to maintain or even enhance performance over time for the adoption of ML systems in high-risk environments (Huyen, 2022). To further advance the development of ML systems that can flexibly respond to dynamic changes and learn from their environment and the human decision maker, I pose the following research question (RQ):

**RQ2:** *In order to enhance ML system performance, how can the ML system learn from the human decision maker and the dynamic environment in which it is deployed?*

In addition to ensuring that both parties—humans and ML systems—benefit from collaboration, it is critical that they can work together effectively as a team to ensure high performance in collaborative decision making. While research initially viewed the use of ML systems primarily as a tool to augment human capabilities, the increasing capacity for autonomy in ML systems has led to a shift toward speaking of collaboration (e.g., Berente et al., 2021; Schuetz & Venkatesh, 2020), in which both humans and ML systems contribute in independent roles to pursue a common goal or perform a common task (e.g., McNeese et al., 2018; Siemon, 2022; Zercher et al., 2023). The

deployment of ML systems as team members presents new challenges for collaborative decision making, which requires further investigation (Seeber et al., 2020). For example, human-AI teaming raises new questions about the trust of human decision makers, especially in the face of conflicting or erroneous ML advice (e.g., McNeese et al., 2021; Seeber et al., 2020; Zercher et al., 2023). Machine teammates will further influence human cognitive biases in decision making (e.g., Balakrishnan et al., 2021; Seeber et al., 2020), and the adoption of autonomous ML systems poses challenges to user acceptance (e.g., Berente et al., 2021; Pumplun et al., 2021). To facilitate effective teaming of decision makers and ML systems, it is essential to enhance our understanding of emerging biases and develop systems that leverage them while fostering trust and acceptance. Therefore, I pose the following RQ:

**RQ3:** *In order to enhance collaborative performance, how can a human decision maker and an ML system effectively work as a team?*

### 1.3 Structure of the Dissertation

Addressing the three RQs outlined, this dissertation comprises five research papers that have been published in various peer-reviewed outlets as presented in Table 1. The following section summarizes research methodologies and findings of the five papers and describes how they contribute to the objectives of this dissertation.

Paper A addresses the first RQ by exploring whether and how humans can learn from collaborating with ML-based decision support systems (DSSs) to enhance their performance and decision-making confidence over time. In this mixed-methods study, radiologists segmented 690 brain tumors in collaboration with explainable vs non-explainable ML-based decision support systems (DSSs) that were either high- vs low-performing. The quantitative results of the experiment show that decision makers can learn from ML systems, but they can also acquire incorrect knowledge. Explainable ML output design improves learning outcomes and can prevent false learning. In some cases, radiologists were even able to learn from the errors of ML systems when the output was made explainable. Addressing several calls for research (e.g., Asatiani et al., 2021; Grønsund & Aanestad, 2020; Sturm, Gerlach, et al., 2021), paper A empirically demonstrates that collaboration with ML systems can promote human learning in the context of a high-risk application area and that explainable output design improves potential learning outcomes. The findings pave the way for early adoption strategies of explainable ML systems and link the value of explanations from the XAI domain to explanations used in education.

Paper B examines the system perspective (RQ2) of human-AI collaboration. Following a design science research (DSR) approach, challenges for the performance of ML systems in long-term deployment are identified based on qualitative interviews with data science and ML experts. The

paper further derives design requirements (DRs) and principles (DPs), structured along the cross industry standard process model for the development of machine learning applications with quality assurance methodology (CRISP-ML(Q)), that can guide the development and deployment of ML systems which are able to learn from human decision makers and adapt to changes in their dynamic deployment environment. This allows ML systems to be maintained and flexibly adapted during deployment to ensure robust model performance over the long term. In particular, this holistic approach extends existing *human-in-the-loop* approaches, which are applied only after deployment, although important development steps in earlier phases already influence the ability of ML systems to learn from, e.g., the dynamic problem perception of human decision makers.

**Table 1: List of publications included in this dissertation**

RQ1 Human perspective	Paper A	Ellenrieder, S., Kallina, E. M., Pumplun, L., Gawlitza, J. F., Ziegelmayr, S., & Buxmann, P. (2023). <b>Promoting Learning Through Explainable Artificial Intelligence: An Experimental Study in Radiology.</b> <sup>2</sup> In <i>Proceedings of the 44th International Conference on Information Systems (ICIS)</i> , Hyderabad, India. VHB-Rating <sup>3</sup> : A.
RQ2 ML system perspective	Paper B	Ellenrieder, S., Jourdan, N., Biegel, T., Cassoli, B. B., Metternich, J., & Buxmann, P. (2023). <b>Toward the Sustainable Development of Machine Learning Applications in Industry 4.0.</b> In <i>Proceedings of the 31st European Conference on Information Systems (ECIS)</i> , Kristiansand, Norway. VHB-Rating: A.
RQ3 Collaborative perspective	Paper C	Ellenrieder, S., Ellenrieder, N., Hendriks, P., & Mehler, M. F. (2024). <b>Pilots and Pixels: A Comparative Analysis of Machine Learning Error Effects on Aviation Decision Making.</b> <sup>4</sup> In <i>Proceedings of the 32nd European Conference on Information Systems (ECIS)</i> , Paphos, Cyprus. VHB-Rating: A.
	Paper D	Ellenrieder, S., Mehler, M. F., & Turan Akdag, M. (2023). <b>Design for Acceptance and Intuitive Interaction: Teaming Autonomous Aerial Systems with Non-experts.</b> In <i>Proceedings of the 27th Pacific Asia Conference on Information Systems (PACIS)</i> , Nanchang, China. VHB-Rating: C.
	Paper E	Mehler, M. F.*, Ellenrieder, S.*, & Buxmann, P. (2024). <b>The Influence of Effort on the Perceived Value of Generative AI: A Study of the IKEA Effect.</b> In <i>Proceedings of the 32nd European Conference on Information Systems (ECIS)</i> , Paphos, Cyprus. VHB-Rating: A.  *shared first authorship

<sup>2</sup> Awarded with the **ICIS 2023 Best Paper Award** in Honor of TP Liang.

<sup>3</sup> The latest VHB Publication Media Rating 2024 is selected as the preferred source for assessing the quality of peer-reviewed papers and articles for my doctoral study program by the Technical University of Darmstadt.

<sup>4</sup> Awarded with the **Claudio Ciborra First Runner Up Award**.

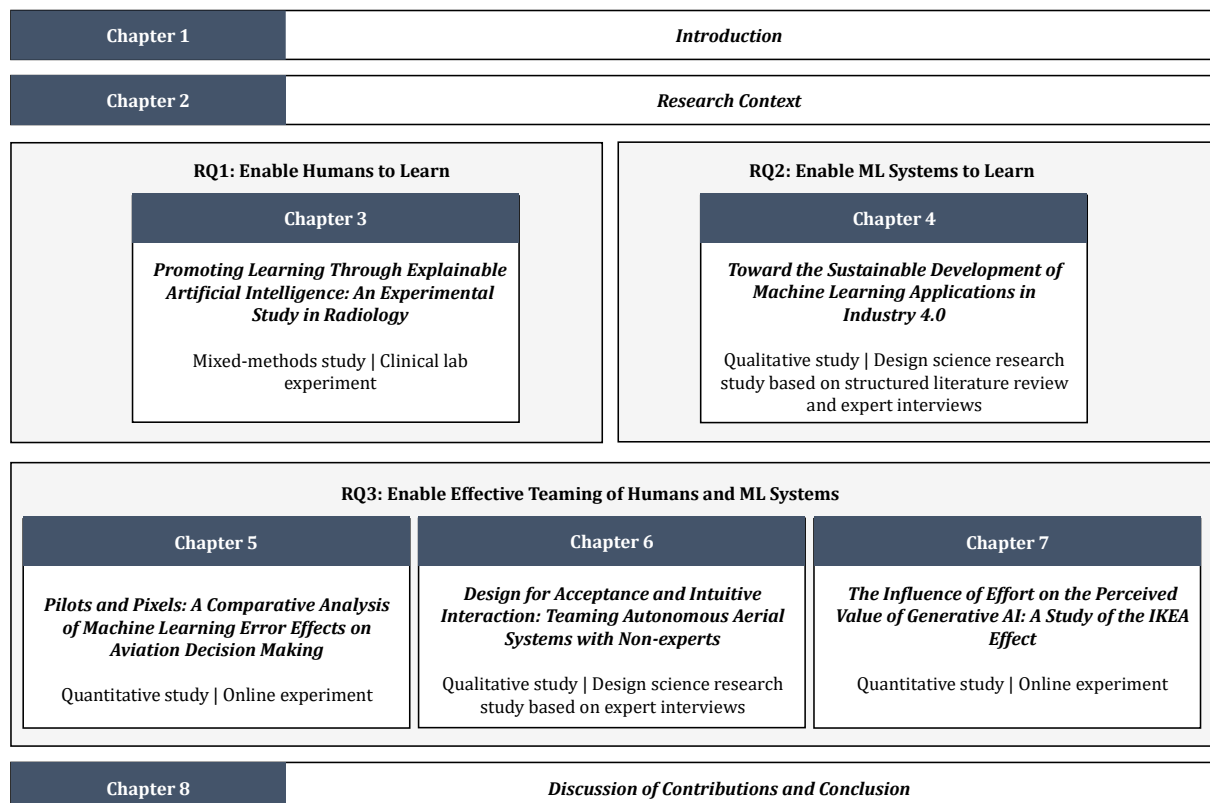
From a collaborative perspective (RQ3), the increasing autonomy of ML systems, their remaining susceptibility to error, as well as human cognitive biases in decision making, present several challenges and opportunities to design effective collaboration, which are investigated by papers C, D, and E. However, the existing research that examines the impact of ML systems on human decision makers does not distinguish between the different types of ML errors. Paper C explores how different types of errors—false positives (FPs) and false negatives (FNs)—of ML systems affect the performance, trust, and mental workload of decision makers, aiming to design systems aligned with error management theory (EMT). In an online experiment, pilots are provided with 222 recordings of flight simulations and different ML system variants that output FPs and FNs for detecting other aircraft in the airspace, with either an explainable or non-explainable output design. The study shows how both types of errors negatively affect the performance and trust of decision makers in high-risk areas. Explainable output design can mitigate these negative effects in terms of performance and trust, especially for FPs, but increases the mental workload of pilots in the case of FPs. The results can be used as a guide for the optimization of the sensitivity and explainability of the ML system in relation to the application area and the needs of the end user. In addition to the risk of erroneous behavior, the increasing degree of autonomy poses new challenges for human-AI collaboration and barriers to the acceptance of ML systems. At the same time, however, it increasingly enables non-experts to work with this complex technology by reducing the need for human supervision. Again, in the context of aviation, Paper D presents a DSR study that builds on qualitative interviews with drone pilots and aviation experts to derive DRs and DPs for artefacts that enable teaming of autonomous systems and non-experts and further promote technology acceptance. The study also shows that established design guidelines for improving explainability in human-AI collaboration need to be reconsidered with increasing autonomy and use by non-experts. The increasing level of autonomy not only changes the roles in human-AI collaboration, but also risks affecting how humans perceive the value of the solutions created. Paper E examines the bias in collaboration, where people overvalue collaboratively created solutions when a lot of their own effort went into them. The results of an online experiment provide a basis for rethinking current deployment strategies for the use of AI, as in some cases the effort that humans put into collaboration is essential to appreciate the results. The study suggests that meaningful collaboration to accomplish joint tasks may be advantageous over task splitting and automation due to cognitive biases.

To explore how both humans and ML systems can learn from collaboration and work effectively as a team, different methodologies and theoretical backgrounds are applied. Paper A is a mixed-methods study that includes both quantitative results from a clinical lab experiment and qualitative insights from interviews and think-aloud protocols with radiologists. Papers B and D are DSR studies based on qualitative expert interviews. Papers C and E employ a quantitative



approach in the context of two online experiments. To investigate human learning potential in collaboration, study A builds on Hunt's (2003) learning theory. Paper B is based on the extended CRISP-ML(Q) process model (Studer et al., 2021) to derive requirements for ML systems that can learn from human decision makers for system maintenance after deployment. Paper C draws on the error management theory (Haselton & Nettle, 2006) to examine the impact of different ML error types on collaboration and align systems accordingly to minimize the costs incurred by the decision maker. Paper D uses the unified theory of acceptance and use of technology (UTAUT) as the kernel theory (Venkatesh et al., 2003) in the DSR approach to derive DRs and DPs for system design that promote user acceptance. Finally, paper E builds upon the cognitive bias known as the IKEA effect (Norton et al., 2012) and examines the influence of human effort on the perceived value of collaboratively developed solutions.

The research papers under discussion are detailed in Chapters 3 through 7. Preceding this, Chapter 2 provides a comprehensive description of the research context and summarizes the relevant theoretical background. To conclude, Chapter 8 reflects on the theoretical and practical contributions of the research papers, and outlines avenues for future research. Figure 1 offers a schematic overview of the structure of the dissertation.



**Figure 1: Outline of the dissertation**

In addition to the research papers presented in this dissertation, I also contributed to the following peer-reviewed publications and journal submissions during my time as a Ph.D. candidate. However, they are not included in this dissertation.

- Wagner, L.\*, Ellenrieder, S.\*, Mayer, L., Müller, C., Bernhard, L., Kolb, S., Harb, F., Jell, A., Berlet, M., Feussner, H., Buxmann, P., Knoll, A., & Wilhelm, D. (2024). **Robotic Scrub Nurse as Mind Reader: Anticipating Required Surgical Instruments Based on Real-Time Laparoscopic Video Analysis.** *Communications Medicine* 4(1), 156, <https://doi.org/10.1038/s43856-024-00581-0>.  
*\*shared first authorship*
- Buxmann, P., & Ellenrieder, S. (2024). **Unlocking AI's Potential: Human Collaboration as the Catalyst.** *Weizenbaum Journal of the Digital Society* 4(1), <https://doi.org/10.34669/WI.WJDS/4.1.7>.
- Gräf, M., Mehler, M., & Jourdan, S. (2024) **Crisis Management in the Metaverse: Designing Virtual Worlds for Real-World Resilience.** In *Proceedings of the 45th International Conference on Information Systems (ICIS)*, Bangkok, Thailand. VHB-Ranking: A.
- Unzicker, D., Mehler, M., Kammholz, L., Sturm, T., Jourdan, S., & Buxmann, P. (2024). **All Eyes on the Reviewer: Understanding the Impact of GenAI on Mental Workload and Performance in Code Reviews.** In *Proceedings of the 45th International Conference on Information Systems (ICIS)*, Bangkok, Thailand. VHB-Ranking: A.
- Gräf, M., Mehler, M., & Ellenrieder, S. (2024). **AI Strategy in Action: A Case Study on Make-or-Buy for AI-based Services.** In *Proceedings of the 28th Pacific Asia Conference on Information Systems (PACIS)*, Ho Chi Minh City, Vietnam. VHB-Ranking: C.
- Mehler, M., Ellenrieder, S., Turan Akdag, M., Wagner, A., Benbasat, I. (2023). **How to Survey: A Framework for Developing Cross-Sectional Surveys.** In *Proceedings of the 44th International Conference on Information Systems (ICIS)*, Hyderabad, India. VHB-Ranking: A.
- Ellenrieder, S., Jourdan, N., & Reuter-Oppermann, M. (2023). **Delivery Drones - Just a Hype? Towards Autonomous Air Mobility Services at Scale.** In *Proceedings of the 56th Hawaii International Conference on System Sciences (HICSS)*, Maui, USA. VHB-Ranking: B.
- Gräf, M., Zöll, A., Wahl, N., Ellenrieder, S., Hager, F., Sturm, T., & Vetter, O. A. (2023). **Designing the Organizational Metaverse for Effective Socialization.** In *Proceedings of the 27th Pacific Asia Conference on Information Systems (PACIS)*, Nanchang, China. VHB-Ranking: C.
- Joglekar, S.; Ellenrieder, S.; and Reuter-Oppermann, M., **Unlocking Solver Potential: A Framework for Analysis and Inter-Comparison of Optimisation Solvers** (2023). In

---

*Proceedings of the 34<sup>th</sup> Australasian Conference on Information Systems (ACIS), Wellington, New Zealand.*

## 2. Research Context

This chapter lays the foundation for this dissertation by outlining the research context and relevant theories for the included papers. First, an overview of AI and ML, XAI and human-AI collaboration is given and the current state of research is described. Theories on human learning, technology acceptance and biases in human-AI collaboration are then presented.

### 2.1 Artificial Intelligence and Machine Learning

As early as the 1950s, AI research set itself the goal of developing machines that are capable of performing tasks that require human intelligence (McCarthy et al., 2006; Rai et al., 2019). While a large number of new definitions have emerged over the years, the concept of an intelligent agent that can perceive and act upon their environment has become established in the IS field (Benbya et al., 2021; Berente et al., 2021; Russell & Norvig, 2021; Schuetz & Venkatesh, 2020). Intelligent agents enable AI systems to perform cognitive functions “that we associate with human minds, such as perceiving, reasoning, learning, interacting with the environment, problem solving, decision-making, and even demonstrating creativity” (Rai et al., 2019, p. 3). Initially, the field pursued rule-based approaches (Russell & Norvig, 2021), but today’s modern applications are primarily based on statistical machine learning approaches (Brynjolfsson & Mitchell, 2017). This shift has overcome the challenge that humans often cannot describe their own decision-making rules (Brynjolfsson & Mitchell, 2017; Fügener et al., 2021).

ML, as the data-driven subfield of AI, utilizes learning algorithms that independently recognize patterns in large datasets. Subsequently, trained ML models can apply learned patterns to new data to make predictions, offer decision recommendations, classify data, or trigger further actions (Mitchell, 1997; Russell & Norvig, 2021). In the context of GenAI, sophisticated ML techniques are employed to create new data by utilizing identified patterns (Benbya et al., 2024; Dwivedi et al., 2023; Teubner et al., 2023). Since ML models independently find solutions to problems, there is no longer a need for human solution instructions in the form of code (Samuel, 1959). Moreover, this approach allows for the discovery of new solutions and thus the generation of knowledge that is complementary to human knowledge (Fügener et al., 2022; Sturm, Gerlach, et al., 2021). Driven by rapid developments in the ML field, applications have transitioned into real-world use in

various organizations (e.g., Benbya et al., 2021; Lebovitz et al., 2021; McKinsey & Company, 2023), and ML systems can sometimes match or even surpass the performance of human experts in task execution (e.g., McKinney et al., 2020; Shen et al., 2019). This makes the use of ML increasingly attractive for high-risk environments such as healthcare or aviation (Lebovitz et al., 2021; Pumplun et al., 2023; Rudin, 2019; Sambasivan et al., 2021).

However, specific characteristics of ML systems pose challenges for real-world deployment and, in particular, collaboration with human decision makers: Since ML learning algorithms are based on statistical approaches, ML-based systems are inherently probabilistic and remain susceptible to errors (Berente et al., 2021; Russell & Norvig, 2021). Furthermore, these approaches can lead to inconsistent system behavior (Amershi et al., 2019; Schuetz & Venkatesh, 2020), which is a particular problem due to the inscrutability of the systems (Asatiani et al., 2021; Berente et al., 2021; Rudin, 2019). The complexity of modern ML models makes the inner workings of the models and their outcomes incomprehensible to human decision makers and, in addition, models cannot provide reasoning for their decisions (Diakopoulos, 2016; Rudin, 2019). This is often referred to as a “black-box” problem (Adadi & Berrada, 2018; Castelveccchi, 2016; Guidotti et al., 2018).

Despite the inherent risks, ML systems have been successfully explored and deployed in high-risk areas of application, including radiology. A well-known example of the use of ML in radiology is the detection, classification and segmentation of tumors in medical images (e.g., Calisto et al., 2021; Lebovitz et al., 2021; McKinney et al., 2020; Pumplun et al., 2023; Silva & Ribeiro, 2011). In the case of a detection model, two types of errors can occur: The ML system falsely detects a tumor in a healthy patient—a false positive—or the system fails to detect an existing tumor—a false negative error (Goutte & Gaussier, 2005; Padilla et al., 2020; Silva & Ribeiro, 2011). The medical example clearly shows how significant the implications of ML errors are in high-risk environments. However, ML developers can influence the rates of both error types by setting the confidence threshold of the system, choosing to reduce the rate of one error type at the expense of the other (Padilla et al., 2020). The confidence threshold determines the minimum level of confidence necessary for a model’s detections to be recognized as valid (Asatiani et al., 2021; Sculley et al., 2015; Sturm, Gerlach, et al., 2021). Reducing the confidence threshold increases the ML system’s sensitivity and reduces FNs, but can increase the rate of FPs. Conversely, increasing the confidence threshold decreases sensitivity, reducing FPs but increasing the rate of FNs (Padilla et al., 2020; Wenkel et al., 2021).

Setting the confidence threshold is just one decision in the iterative development process of ML systems, which encompasses several phases (Sambasivan et al., 2021; Wirth & Hipp, 2000). Initially, the problem to be solved must be identified, and system requirements defined. Based on these requirements, data and metadata that best describe the real-world problem are collected

and analyzed. This is followed by data cleaning, preprocessing, feature extraction, and selection to build meaningful datasets. The parameters of selected learning algorithms can then be trained based on a partial data set (Wirth & Hipp, 2000). Using various performance metrics and held-out test data, an offline evaluation of the trained models typically follows (Jourdan et al., 2021). If the predefined requirements are met, the model can then be deployed to production (Wirth & Hipp, 2000). Although well-known process models like the popular CRISP-DM model end with the deployment phase (Wirth & Hipp, 2000), new process models like CRISP-ML(Q) (Studer et al., 2021) and various research studies indicate the need for continuous monitoring and maintenance of ML systems in real-world deployment to maintain performance over the long term (Asatiani et al., 2021; Sculley et al., 2015; Sturm, Gerlach, et al., 2021). ML systems are trained and often evaluated on static datasets but are deployed in dynamic environmental conditions. For example, human perception of a problem may change (Russell & Norvig, 2021), or other causes may lead to deviations that negatively affect the performance of the ML system (Jourdan et al., 2021; Sculley et al., 2015; Sturm, Gerlach, et al., 2021). IS research has also recognized the need for continuous auditing and altering activities to maintain the performance and value of ML applications over the long term (e.g., Asatiani et al., 2021; Grønsund & Aanestad, 2020; Sturm, Gerlach, et al., 2021). While the focus is on integrating human feedback after deployment through, for example, *human-in-the-loop* approaches (e.g., Grønsund & Aanestad, 2020; Stumpf et al., 2009), there is still a lack of comprehensive guidance on how ML systems must be developed throughout all phases to be able to learn from humans and the dynamic environment over time and respond to changes to maintain or even improve their performance—a requirement that is paramount for the use of ML in high-risk environments and for promoting sustainable ML lifecycles.

## 2.2 Explainable AI

As described above, the complexity of modern ML systems renders their outcomes opaque, creating what is commonly referred to as the “black-box” problem (Asatiani et al., 2021; Berente et al., 2021; Rudin, 2019). For human decision makers, collaboration with modern ML systems poses a major challenge due to this lack of explainability (Arrieta et al., 2020; Lebovitz et al., 2022; Meske et al., 2022; Reyes et al., 2020). To address this issue, the research field of XAI has emerged at the intersection of human-computer interaction, computer science, and social science (Arrieta et al., 2020). XAI approaches primarily aim to make the decision-making processes and results of ML systems more understandable to humans (Adadi & Berrada, 2018; Arrieta et al., 2020; Meske et al., 2022). This is achieved by providing explanations that offer reasoning for system decisions, thereby improving the explainability of the outputs (Meske et al., 2022; Miller, 2019). The terms *explainability*, *interpretability*, and *transparency* are often used interchangeably (Meske et al., 2022; Pumplun et al., 2023). However, in this dissertation, I adhere to the distinction that

transparent ML systems are inherently understandable to humans due to their underlying models, such as decision trees, which do not require additional explanations to be comprehensible (Panigutti et al., 2023; Pumplun et al., 2023; Rosenfeld & Richardson, 2019). Interpretability is closely aligned with the concept of explainability. Yet, while interpretability focuses on how easily users can predict an ML system's output as inputs and parameters change, explainability enables humans to understand the underlying mechanisms of ML systems through explanations (Arrieta et al., 2020; Meske et al., 2022; Panigutti et al., 2023; Pumplun et al., 2023; Reyes et al., 2020; Rosenfeld & Richardson, 2019). Thus, explainable systems allow humans to understand *why* ML systems make certain decisions (Arrieta et al., 2020; Guidotti et al., 2018).

Three categories of explanations can be distinguished in the XAI field: Firstly, model explanations, which provide meta-information about the underlying ML model, the system's development, and its performance (Cai et al., 2019; Diakopoulos, 2016; Meske et al., 2022; Pumplun et al., 2023). Secondly, global explanations, which can enhance the overall understanding of the system by clarifying, for example, the influence of individual features on the ML system's decision making process (e.g., Ghorbani et al., 2019). Lastly, local explanations are designed to improve the comprehension of local outputs from an ML system, for instance, by detailing pixel contributions, providing uncertainty estimates, or conducting sensitivity analyses (Guo et al., 2017; Pumplun et al., 2023).

Explanations in the XAI field are developed for various purposes: Initially, ML developers were the primary target group, as an understanding of the inner workings of ML systems is crucial for identifying flaws and thus for the improvement of the systems. Furthermore, explanations may be necessary for AI regulators to test and certify systems (Bhatt et al., 2020; Meske et al., 2022). On December 9, 2023, a political consensus on the AI Act was achieved between the European Parliament and the Council which imposes strict obligations on the transparency and explainability of *high-risk AI systems* (European Commission, 2023). Although the AI Act does not directly enforce the use of XAI methods, it demands that users who are responsible for overseeing the systems must be able to understand the system's characteristics and correctly interpret its output (Panigutti et al., 2023; Sovrano et al., 2022). It is noteworthy that the AI Act places particular emphasis on end users as the intended recipients of the information embedded in, for instance, explanations from the XAI field. Similar requirements have been proposed in practice: The European Aviation Safety Agency (EASA) published a concept paper in March 2024 that requires safety-related ML applications in aviation to meet high explainability requirements. EASA emphasizes that not only software developers, but also operational users of ML applications, such as flight crews, have a need for explainability. In this context, explainability should not only improve understanding of the decisions made by ML systems for the operational users, but also contribute to building trust (European Union Aviation Safety Agency, 2024). IS research has also

increasingly dealt with the impact of explanations on the human decision maker as the end user in recent years (e.g., Asatiani et al., 2021; Gaube et al., 2023; Pumplun et al., 2023). In this context, it is important to consider the often limited prior knowledge of end users regarding ML and to avoid providing overly complex explanations (Adadi & Berrada, 2018; Bhatt et al., 2020; Meske et al., 2022), which can be challenging as explanations are usually statistical in nature (Bhatt et al., 2020).

The positive impact of explanations from the XAI field on human decision makers has been demonstrated in recent studies (e.g., Gaube et al., 2023; Jussupow et al., 2021; Lebovitz et al., 2021). For example, explainable design allows for better system evaluation by humans (e.g., Jussupow et al., 2021; Lebovitz et al., 2021), improved trust building (Benbya et al., 2021; Schuetz & Venkatesh, 2020), more effective communication with the system (Kane et al., 2021), and recent studies even suggest the potential of explanations to promote human learning (Abdel-Karim et al., 2023; Gaube et al., 2023; Jussupow et al., 2021; Lebovitz et al., 2021; Meske et al., 2022). However, the provision of explanations introduces additional information into the decision-making process, which—despite all the mentioned advantages—can also have a negative impact, as it requires additional cognitive resources (e.g., Lebovitz et al., 2022; Pumplun et al., 2023). Moreover, explanations can cause adjustments to mental models, potentially leading to biased decisions (Bauer et al., 2023). Nonetheless, the use of explainable ML systems holds great potential for human decision makers. Particularly, the potential of explanations to promote human learning is of great relevance for the current research on the use of ML systems for decision support in high-risk areas, as humans will continue to bear responsibility in the long term—a research area that remains under-explored and will be examined in this dissertation.

### **2.3 Human-AI Collaboration**

While AI systems were initially developed with the goal of assisting humans in performing tasks and augmenting human capabilities (e.g., Brynjolfsson & McAfee, 2014; Maedche et al., 2019), the increasing autonomy and performance of modern algorithms (e.g., Berente et al., 2021; Schuetz & Venkatesh, 2020) has led to research that now speaks of human-AI collaboration (e.g., Fügener et al., 2021; Mirbabaie et al., 2022; Vössing et al., 2022). In the context of human-AI collaboration, both humans and AI pursue a common goal and perform joint tasks in independent roles (e.g., McNeese et al., 2018; Siemon, 2022; Vössing et al., 2022; Zercher et al., 2023). For this definition, it is no longer sufficient for AI systems to simply act as digital assistants, for example, scheduling meetings based on email requests to improve the performance of individuals or human teams in the workplace (e.g., Maedche et al., 2019; Seeber et al., 2020). When humans and AI systems collaborate, often referred to as human-AI teaming (e.g., Gurney et al., 2022; McNeese et al., 2018; Seeber et al., 2020; Zercher et al., 2023), AI systems have the ability to engage in complex problem-



solving and decision-making processes as team partners (Maedche et al., 2019; Seeber et al., 2020), taking on tasks that previously required human intelligence, such as diagnosing diseases like cancer (e.g., Lebovitz et al., 2022; McKinney et al., 2020; Vössing et al., 2022).

This view has already been adopted in practice for the classification of AI systems in high-risk areas such as aviation. In its recently published AI Roadmap 2.0, the EASA divides AI applications in aviation into three levels. Level 1 AI systems are used to cognitively support and augment humans, while Level 2 AI systems collaborate with humans, also referred to here as human-AI teaming. Level 2 AI systems can already select and execute actions automatically but are supervised by humans who retain final decision-making authority and responsibility. At Level 3, systems are given full authority to make and execute decisions without supervision, also referred to as advanced automation (European Union Aviation Safety Agency, 2023, 2024).

Due to technological advancements and changing role distributions, existing knowledge about human-computer interaction in the context of AI needs to be reconsidered and expanded, creating new avenues for future research (e.g., Maedche et al., 2019; Seeber et al., 2020). For example, Seeber et al. (2019) provide a comprehensive overview of the challenges and opportunities for the design of socio-technical systems that emerge in the context of human-AI collaboration. Characteristics such as the error-proneness and lack of transparency and explainability of AI systems (e.g., Adadi & Berrada, 2018; Berente et al., 2021) affect, for instance, how humans develop trust in AI (Glikson & Woolley, 2020; Gurney et al., 2022; Vössing et al., 2022) and their acceptance of the technology (e.g., Berente et al., 2021; Pumplun et al., 2021). In addition, Fügener et al. (2021) point out that humans have difficulty assessing their own capabilities, which makes them poor delegators when working with AI systems. Overall, IS research is divided with respect to the impact of AI systems on the human decision maker. Studies have shown that collaboration with AI systems can indeed lead to better performance and that humans can improve their decision making as a result (e.g., Abdel-Karim et al., 2023; Gaube et al., 2023), while others point to risks to reasoning and the potential loss of unique knowledge in decision makers (Fügener et al., 2021; Jussupow et al., 2021). In this dissertation, I aim to expand our understanding of the impact of AI systems on human decision makers, with a focus on high-risk domains such as healthcare and aviation. In particular, the potential for humans to learn and improve through collaboration with AI systems, as well as the effects of characteristics such as error-proneness and the explainability of system output on trust, cognitive biases, technology acceptance, and mental workload, will be explored.

## **2.4 Theoretical Foundations**

In what follows, the theoretical foundations that are applied to explore the human, system, and collaborative perspectives on human-AI collaboration are presented. In order to analyze whether

and how humans can learn from ML systems, I rely on Hunt's (2003) theory of human learning. Error management theory, on the other hand, is applied to analyze how the error rates of ML systems should be biased in order to minimize the costs incurred by the decision maker. In addition, I describe the unified theory of acceptance and use of technology, which serves as a kernel theory to derive DRs and DPs that promote technology acceptance for highly autonomous ML systems. Finally, the IKEA effect is used to investigate the extent to which the human effort invested in collaboration leads to an overvaluation of the solutions developed.

#### 2.4.1 Learning Theories

Aiming to describe and understand the complex process of human learning, various learning theories have emerged over the years (e.g., Chase & Simon, 1973; Hunt, 2003; Miller, 1956). Human learning encompasses the absorption, processing, and storage of information into short-term and long-term memory (Gagné, 1970). According to Miller's (1956) learning theory, humans cluster related *bits of information* into so-called *chunks of information* during the learning process. Only when an individual manages to recognize and establish connections between new and existing chunks from long-term memory, can they acquire new skills or expand existing ones, thereby changing their behavioral potential (Chase & Simon, 1973; Miller, 1976). More modern learning theories, like that of Hunt (2003), define learning in general as the acquisition and retention of new knowledge in human memory. To measure learning, therefore, a change in knowledge must be observed. Unfortunately, knowledge is an intangible asset for which there is no standardized measurement method (Sveiby, 1997). Many studies in both educational research and IS research fields that aim to measure learning, therefore, investigate changes in behavior (e.g., Abdel-Karim et al., 2020; Calisto et al., 2021; Regueras et al., 2009). However, these studies do not distinguish between changes in knowledge—also referred to as behavioral potential—and behavior, although not every change in knowledge results in a change in behavior (Hunt, 2003). This is particularly true in high-risk application areas, where the consequences of incorrect decisions are severe. Here, it is important for decision makers to consider how confident they are that their decision is correct (Hunt, 2003). The question of how certain a person must be about a belief for it to guide behavior is referred to in this context as the *boundary problem* (Quine, 1987). Building on these insights, Hunt (2003) proposes an epistemetric method to measure changes in knowledge and thus learning. In addition to measuring behavior (i.e., performance measurement), this method also captures the certainty with which a person possesses knowledge. For the effective acquisition and storage of knowledge, explanations that include necessary contextual information and facilitate the integration of new information into existing knowledge are considered essential by educational research (Crowley & Siegler, 1999; Fender & Crowley, 2007). Yet, unanswered in this context is the question of whether explanations from the XAI field can also

serve to support the integration of information provided by ML systems, to promote human learning in human-AI collaboration. Paper A investigates the impact of ML systems and XAI approaches on human learning, applying the epistemic method proposed by Hunt (2003).

#### 2.4.2 Error Management Theory

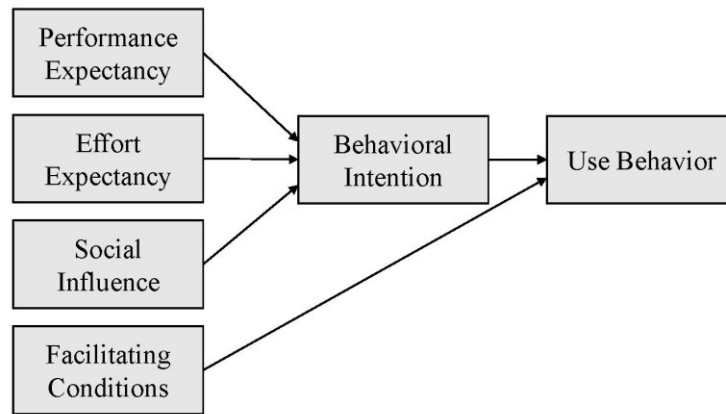
While human decision making is inherently biased, there are also impulses from research to incorporate bias into “*humanly engineered systems*” (Haselton & Nettle, 2006). In this context, Haselton and Nettle (2006) introduce the error management theory, which posits that in decisions made under uncertainty where the costs of different types of errors—FNs and FPs—are asymmetric over evolutionary time, the system should be biased to prefer the less costly error type. Taking this bias into account during system development can reduce the overall costs by minimizing the more costly error type. In order to avoid more costly errors, it is also accepted that the overall frequency of errors may increase (Green & Swets, 1966; Johnson et al., 2013). Natural selection has caused human decision making to adapt over evolutionary time to make predictable errors that avoid particularly costly mistakes (Haselton & Nettle, 2006; Johnson et al., 2013). Introducing such a bias thus aligns the system with human psychology (Haselton & Nettle, 2006). However, in order to bias systems according to EMT and decide on feasible system decision strategies under uncertainty, it is necessary to understand the different costs of errors and why decision makers react to each type of error differently (Swets et al., 2000). Especially in high-risk areas, where the consequences of wrong decisions are severe, many studies investigate which type of error is preferable. For example, Arkes & Mellers (2002) conducted a survey among students to determine whether they were less inclined to accept FPs (false convictions) or FNs (false acquittals) of people who had been charged with serious crimes. The authors found that there was a clear bias, and the majority of participants accepted a higher rate of FNs, thereby acquitting the guilty (Arkes & Mellers, 2002). Swets et al. (2000) demonstrated that introducing a system bias regarding different error types can improve the accuracy of medical diagnoses in the collaboration between physicians and DSSs. While it is usually easy to consider FNs as more severe compared to FPs (e.g., overlooking a tumor compared to a false over-detection) in high-risk application areas, it is also relevant to assess the costs of FPs on the human decision maker (Luce & Kahn, 1999). This allows for biasing systems according to EMT and understanding the impact of both error types on the decision maker and collaborative decision performance (Haselton & Nettle, 2006; Wardle & Pope, 1992). A system that is overly sensitive and produces a large number of FPs can also result in costs if decision makers stop using the system. While EMT can generally be applied to decisions under uncertainty (Haselton & Nettle, 2006), it becomes particularly relevant for ML systems used in collaboration with human decision makers. Research often aims to improve the accuracy of ML systems (Roy et al., 2022). However, it is important to

acknowledge that even highly accurate ML systems are susceptible to errors (Russell & Norvig, 2021). Therefore, it is crucial to also consider strategies for error management. Although the technical implementation of favoring one type of error, as described in Section 2.1, is straightforward for ML systems, IS research investigating the impact of erroneous ML systems on the human decision maker has not yet distinguished between the different types of errors that can occur (e.g., Abdel-Karim et al., 2020; Jussupow et al., 2021; Pumplun et al., 2023), and we know little about the cost that both types of errors incur for the human decision maker. In this context, it is also important to consider the integration of approaches from the XAI area and their impact on the costs of different ML error types. Paper C conducted an experimental study with pilots to investigate the costs related to performance, trust, and mental workload caused by different ML error types which provides as broad foundation to design ML systems in line with EMT for collaboration with human decision makers.

#### 2.4.3 *Unified Theory of Acceptance and Use of Technology*

Although the deployment of ML systems in high-risk areas such as medical diagnostics offers great potential to improve decisions, the adoption of these systems in practice remains low. One factor hindering adoption is the low ML acceptance by decision makers such as physicians (e.g., Pumplun et al., 2021). For successful human-AI collaboration, it is accordingly essential that advice given by ML systems is considered and that there is acceptance for the technology among end users. The UTAUT model is a theoretical approach that aids in comprehending the factors that influence technology acceptance and subsequent usage behavior (Venkatesh et al., 2003). Furthermore, the UTAUT model can be used to target potential users who are less inclined to adopt a technology (Venkatesh et al., 2003, 2016). Based on eight prominent theories explaining individual technology acceptance, Venkatesh et al. (2003) developed the UTAUT model, which includes four key constructs (see Figure 2) that serve as direct determinants of usage intention and behavior.

*Performance expectancy* generally refers to the belief of individuals that utilizing a particular system will enhance their job performance. *Effort expectancy* concerns the perceived ease associated with the use of the technology or system. This expectancy plays a crucial role at the initial stages of adoption, although its impact tends to diminish as users become more accustomed to the technology. Moreover, the intention to adopt and use a new technology is influenced by *social influence*, which is the perception of an individual regarding the expectations of important others that they should adopt the new technology. This determinant is particularly relevant at the beginning of technology usage when the user's experience is limited. Lastly, *facilitating conditions* relate to the individual's belief in the availability of necessary organizational and technical support systems to facilitate the use of the technology (Brown & Venkatesh, 2005; Venkatesh et al., 2003, 2012).



**Figure 2: The UTAUT model with its four direct determinants of user acceptance and usage behavior (Venkatesh et al., 2003)**

In addition to these core constructs, the UTAUT model also identifies four key moderators of the effects of these constructs on usage intention and behavior: gender, age, experience, and voluntariness of use. These moderators help to explain the variability in technology adoption and use across different user groups and contexts (Venkatesh et al., 2003, 2016). While the UTAUT model was derived based on data from employees in organizations and serves to predict behavioral intentions to use a technology by professionals (Venkatesh et al., 2003), the extended UTAUT2 is applicable in a consumer context (Venkatesh et al., 2012). Since this work investigates the technology acceptance of ML systems by decision makers in organizations, such as pilots and radiologists, UTAUT is applied accordingly. Paper D uses the UTAUT model as a kernel theory to derive DRs and DPs that promote technology acceptance for collaboration with highly autonomous ML systems.

#### 2.4.4 Cognitive Biases

When exploring human-AI collaboration, it is critical to acknowledge that humans do not always think and act rationally (Das & Teng, 1999; Kahnemann, 2012). Cognitive biases lead to systematic deviations from rational decisions, resulting in individuals not processing all information in decision-making processes, being influenced by emotions and prejudices in their interpretations, or applying heuristics to make quicker decisions (Arnott, 2006; Das & Teng, 1999; Kahnemann, 2012). To date, research has identified numerous cognitive biases and developed different taxonomies (e.g., Arnott, 2006; French et al., 2023). For example, the anchoring bias causes people to give more weight to information that is provided first than to information that is provided later (Ahsen et al., 2019; French et al., 2023; Kahnemann, 2012). The overconfidence bias, on the other hand, causes people to overestimate their knowledge and ability to solve difficult or novel problems (Arnott, 2006; Odean, 1998). Moreover, people tend to attribute positive outcomes to themselves and negative ones to others, known as the self-serving bias (Mezulis et al., 2004; Miller

& Ross, 1975), and the status quo bias can lead to the rejection of new technologies to avoid change (Kim & Kankanhalli, 2009). Research has shown that cognitive biases can also occur when collaborating with AI systems. Balakrishnan et al. (2021) demonstrated this for the example of the status quo bias, which impairs the acceptance of AI-based assistance systems, and Ahsen et al. (2024) for the example of weighted information provision by classification algorithms in the assessment of mammograms by radiologists.

One cognitive bias that has not yet been studied in the context of human-AI collaboration and that can significantly affect the perceived value of collaboratively created solutions is the so-called IKEA effect (Norton et al., 2012). The potential for AI systems to automate tasks and reduce human effort in accomplishing common tasks and goals (Berente et al., 2021; Brynjolfsson & Mitchell, 2017; Russell & Norvig, 2021) carries the risk that task outcomes and generated solutions will not be fully valued, as findings from psychological and behavioral research have shown that “labour leads to love” (Norton et al., 2012). The IKEA effect suggests that people overvalue objects into whose assembly or creation they have invested a lot of their own effort (Mochon et al., 2012; Norton et al., 2012). Named after the Swedish manufacturer whose products often require significant customer effort in final assembly, this effect has been demonstrated for a variety of physical products, such as preparing food, assembling cardboard boxes, and creating clothing and pictures (e.g., Dohle et al., 2014; Ling et al., 2020; Mochon et al., 2012; Norton et al., 2012; Radtke et al., 2019). In fact, objects in which a great deal of personal effort has been invested are valued even more highly than identical objects created by experts (Norton et al., 2012). Although the IKEA effect has been established for the creation of many physical objects (e.g., Dohle et al., 2014; Norton et al., 2012), it is unknown how human effort in collaboration with AI systems will influence the perceived value and appreciation of the created content and solutions. In addition, characteristics of AI systems such as lack of transparency and explainability (Arrieta et al., 2020; Berente et al., 2021) make it difficult for human decision makers to assess the impact of their own efforts in collaboration. Understanding the impact of human effort invested in collaborating with AI systems on the valuation of the content and solutions produced is essential for designing future collaboration and effective teamwork between humans and AI systems, and this cognitive bias is therefore explored in more detail in paper E.

### **3. Paper A: Promoting Learning Through Explainable Artificial Intelligence: An Experimental Study in Radiology**

#### **Title**

Promoting Learning Through Explainable Artificial Intelligence: An Experimental Study in Radiology

#### **Authors**

Ellenrieder, Sara; Kallina, Emma M.; Pumplun, Luisa; Gawlitza, Joshua F.; Ziegelmayr, Sebastian; and Buxmann, Peter

#### **Publication Outlet**

Proceedings of the 44<sup>th</sup> International Conference of Information Systems (ICIS), Hyderabad, India, 2023

Awarded with the Best Paper Award in Honor of TP Liang

#### **Abstract**

The deployment of machine learning (ML)-based decision support systems (DSSs) in high-risk environments such as radiology is increasing. Despite having achieved high decision accuracy, they are prone to errors. Thus, they are primarily used to assist radiologists in their decision making. However, collaborative decision making poses risks to the decision maker, e.g. automation bias and long-term performance degradation. To address these issues, we propose combining findings of the research streams of explainable artificial intelligence and education to promote human learning through interaction with ML-based DSSs. We provided radiologists with explainable vs non-explainable decision support that was high- vs low-performing in a between-subject experimental study to support manual segmentation of 690 brain tumor scans. Our results show that explainable ML-based DSSs improved human learning outcomes and prevented false learning triggered by incorrect decision support. In fact, radiologists were able to learn from errors made by the low-performing explainable ML-based DSS.

#### **Keywords**

Explainable artificial intelligence, human learning, machine learning, experimental study, radiology, human-AI collaboration

### 3.1 Introduction

The increasing availability of large amounts of data and improved computational power have led to rapid improvements in the field of artificial intelligence (AI), especially in the data-based subfield of machine learning (ML). As a result, ML-based decision support systems (DSSs) are increasingly used in practice today (McKinsey Global Institute, 2021), supporting tasks that were previously performed exclusively by human experts (Shen et al., 2019). This development includes, in particular, high-risk environments such as aviation or healthcare (Lebovitz et al., 2021; Maedche et al., 2019; Sutton et al., 2020). In the field of radiology, multiple clinical ML-based DSSs have recently entered the market (Radboud University Medical Center, 2023) and, in several cases, these systems can outperform expert radiologists in decision accuracy (e.g., Shen et al., 2019). While ML-based DSSs theoretically could automate decision making (e.g., Brynjolfsson & Mitchell, 2017), this is unlikely in the medical domain due to the complexity of the decisions and the stakes of this high-risk environment (Jha & Topol, 2016). Instead, it is likely that ML-based DSSs will support physicians in specific tasks resulting in collaborative human-ML decision making.

Despite the significant improvements of ML systems in recent years, they remain prone to errors (Brynjolfsson & Mitchell, 2017; Russell & Norvig, 2016). During collaborative human-ML decision making in radiology, it is thus essential that the physician is able to detect these instances of error (Goddard et al., 2012). However, this poses various challenges (Fügener et al., 2021; Riedl, 2019): Knowledge degradation or effects like the automation bias (the tendency of humans to over-rely on automated systems, even when presented with contradictory information or evidence; Goddard et al., 2012) might mislead physicians to excessively rely on the ML-based DSS, blindly accepting the system's outputs without consulting their own judgment (Fügener et al., 2021; Goddard et al., 2012). In addition, it is a major concern that ML-generated advice could lead physicians to develop false beliefs (Goddard et al., 2012) and the performance of these experts could degrade in the long term (Fügener et al., 2021). Especially in the medical domain, it is critical that physicians retain and query their own judgment, so that they are able to detect and reject incorrect system recommendations in the future (Lebovitz et al., 2021). While several research efforts aim at improving the capabilities of ML models by integrating human feedback through human-in-the-loop concepts (e.g., Asatiani et al., 2021; Grønsund & Aanestad, 2020; Sturm, Gerlach, et al., 2021), the impact of ML-generated advice on the physician as the decision maker itself is still underexplored, and research now calls for future studies (e.g., Gaube et al., 2023).

Despite their risks, ML-based DSSs offer great potential for human decision makers such as radiologists: Instead of risking a degradation in knowledge and thus performance (Fügener et al., 2021), we argue that the emphasis should shift towards designing ML-based DSSs in a manner



that allows human decision makers to learn from the provided information. ML models recognize patterns based on data and can therefore find solutions to problems on their own (Brynjolfsson & Mitchell, 2017; Russell & Norvig, 2016). In doing so, ML models can offer solutions that humans would not find and contribute knowledge that is complementary to that of humans (e.g., Fügener et al., 2022). Enabling radiologists to learn from the information provided will likely improve their performance over time, or at least prevent degradation (Abdel-Karim et al., 2020; Gaube et al., 2023). This novel perspective is in clear need for future research (e.g., Gaube et al., 2023; Pumplun et al., 2023): Whilst the research area around explainable artificial intelligence (XAI) aims at increasing the explainability of ML-generated output through explanations of the mechanisms underlying the models and thus might enable learning, this is rarely done with the aim to train the human end user in the task itself and more for purposes of making ML more understandable to system developers (Arrieta et al., 2020; Bhatt et al., 2020; Guidotti et al., 2018; Miller, 2019). Since educational research suggests that explanations are critical for effective human learning (Crowley & Siegler, 1999; Fender & Crowley, 2007), the study at hand aims at combining and expanding the findings of these two research streams to explore human learning through the interaction with ML-based DSSs. We therefore also address the question of whether explanations from the field of XAI, being statistical in nature and provided by a machine, can serve a similar purpose as explanations used in education. Rethinking how we design and provide ML-based DSSs would further allow us to align ongoing developments with the Sustainable Development Goals (SDGs) defined by the United Nations (Goralski & Tan, 2020; United Nations, 2015). Early achievement of education-related goals, such as high-quality education, can be supported by systems that promote human learning and can be scaled and deployed across countries.

Several experimental studies in radiology assessed the impact of ML-based DSSs on physicians in diagnosing various cancer types and diseases (Asatiani et al., 2021; Grønsund & Aanestad, 2020; Sturm, Gerlach, et al., 2021). However, these approaches lack a full representation of the dependencies between the design of explainable ML-based DSSs, the performance of the systems, and the individual learning process of a human decision maker. Thus, it is unclear whether a learning process has taken place that affects the accuracy of a diagnosis and the certainty with which it was posed and how these complex processes are intertwined. Moreover, the sparse research on learning in collaboration with ML systems primarily considers decisions that are either correct or incorrect (e.g., Abdel-Karim et al., 2020; Gaube et al., 2023; Jussupow et al., 2021). However, this only considers a subset of the many real-world use cases where decisions can also gradually improve or worsen due to the influence of decision support.

To unravel the full potential of ML-based DSSs to foster human learning, we thus pose the following research questions (RQs): *In the context of radiology, (1) can the interaction with ML-*

*based DSSs promote learning of human decision makers to improve their performance over time, and (2) can explainable design of ML-based DSSs improve potential learning outcomes?*

We investigate both RQs in an experimental study with radiologists who are tasked to segment brain tumors in multimodal magnetic resonance imaging (MRI) scans. By providing different variants of ML-based DSSs for the segmentation of brain tumors, we explore the impacts of high- vs low-performing ML-based decision support, as well as of explainable vs non-explainable design of ML-based decision support on the human decision maker. While little attention has been paid to the quantitative measurement of human learning progress in previous research, we obtain both qualitative and quantitative data to measure the learning progress of physicians. Overall, the results of this study offer empirical insights about how explainability can promote learning from high-performing ML-based decision support and prevent false learning in the case of low-performing support (i.e., prevents the learning of incorrect decisions). Research as well as practice can draw insights from the results of this study for the development of future ML-based DSSs and related artifacts to promote human learning.

## **3.2 Theoretical Background**

This section begins by outlining ML in general, as well as application areas for ML-based DSSs for medical diagnosis and requirements for explainable ML systems from the field of XAI. This is followed by a summary of findings from psychology and educational research on human learning. Lastly, we combine both research streams and derive our study objective.

### *3.2.1 Machine Learning*

While early AI-enabled systems were built on rule-based approaches (Russell & Norvig, 2016), modern AI primarily relies on statistical machine learning approaches (Brynjolfsson & Mitchell, 2017). ML algorithms—a sub-category of AI—have the ability to learn from data by deriving patterns on their own. Once trained, the resulting model can apply these patterns to new data to make predictions (Brynjolfsson & Mitchell, 2017; Mitchell, 1997; Russell & Norvig, 2016). This allows ML models to independently find solutions to problems, rather than requiring a human developer to provide instructions in the form of code (Samuel, 1959). Because ML models can independently find solutions based on data, they can generate new knowledge that humans may not have (Asatiani et al., 2021; Grønsund & Aanestad, 2020; Sturm, Gerlach, et al., 2021). However, the performance of ML models is highly dependent on the data provided, and since ML models rely on statistical patterns, they are prone to errors. This is a serious problem since the inner workings of ML models are often not comprehensible for humans due to their complexity and the use of large amounts of training data (Diakopoulos, 2016; Rudin, 2019). In other words, it is difficult for humans to understand why ML models make certain predictions based on the data

provided. The lack of explainability, as well as the potential for erroneous output are unique to ML-based systems and the main reasons that are preventing the adoption of the technology in many high-risk fields such as medical diagnostics (Lebovitz et al., 2021).

### 3.2.2 *ML-based Decision Support Systems for Medical Diagnostics*

The prediction capabilities of ML models enable ML-based systems to perform tasks that were previously exclusively performed by humans (Maedche et al., 2019; Rai et al., 2019). However, medical decision making remains a very complex field with severe consequences for errors. Thus, human decision makers (e.g., radiologists) are unlikely to be replaced in the near future, but instead they will be supported by ML systems in the form of ML-based DSSs (Jha & Topol, 2016). In practice, the output of ML-based DSSs is usually considered as a form of diagnostic advice that can either be accepted or rejected by the clinician who retains the final decision power (van Leeuwen et al., 2021). Especially in radiology, many ML-based DSSs are already deployed in practice and over 200 AI-based software products are available that have been European Conformity (CE) marked or cleared by the Food and Drug Administration (FDA) for clinical use in Europe and the United States, respectively (Radboud University Medical Center, 2023). In this context, the analysis of image data is a frequent component (Meskó & Görög, 2020) and ML algorithms are already performing with a high level of accuracy (e.g., Jiang et al., 2017). Several ML-based DSSs and prototypes have been developed and tested for radiology, focusing on such areas as diagnosing breast cancer, strokes, or alerting physicians to the detection of other abnormalities on CT images (Calisto et al., 2021; Gaube et al., 2023; Pumplun et al., 2023). Other application areas include the segmentation of abnormal regions on mammogram images, classification of lesions on ultrasound regions, and the segmentation of tumors on MRI scans (Lebovitz et al., 2021).

### 3.2.3 *Explainable Design of ML Systems*

As a response to the increasing complexity of ML models, the research field of XAI has emerged at the intersection of human-computer interaction, computer science, and social science. XAI aims at developing ML system outputs or interpretations that are understandable for humans (Arrieta et al., 2020; Miller, 2019). This is achieved by providing explanations about the processes underlying ML system decisions to make them more tangible, such as the considered features and their respective impact on the decision (Arrieta et al., 2020; Guidotti et al., 2018). Explanations serve as an interface between humans and ML systems that allow humans to comprehend the decision process better (Miller, 2019). Today, various explanatory approaches for the technical implementation of XAI exist (Pumplun et al., 2023). Model explanations provide meta information about the ML model and its development (Diakopoulos, 2016), while global explanations help

users understand the importance of particular features for the overall decision making process (Ghorbani et al., 2019). Local explanations aim to improve understanding of specific ML system outputs, e.g., through pixel assignments, uncertainty estimates, or sensitivity analyses (Guo et al., 2017; Pumplun et al., 2023). However, these explanations are statistical in nature and are different from explanations that are easily understood by the general population (Bhatt et al., 2020).

Explanations can serve different purposes: Oftentimes, they aim at supporting data scientists or developers in understanding ML model behavior during development or are required to pass certain quality assurance tests (Bhatt et al., 2020; Meske et al., 2022). Recently, information systems (IS) research has examined the use of explanations to improve the interaction between human end users, who often have little technical ML knowledge, and ML systems (Asatiani et al., 2021; Gaube et al., 2023; Pumplun et al., 2023). In this case, it is crucial to consider the needs and technical pre-knowledge of the end user (Meske et al., 2022) to avoid the presentation of overly complex explanations (Adadi & Berrada, 2018). In the healthcare context, it is crucial to provide explanations that are understandable to physicians as the primary end users to support medical decision making (Adadi & Berrada, 2018; Bhatt et al., 2020).

#### *3.2.4 Promoting Human Learning Through Machine Learning Systems*

Increased explainability is likely to have several positive effects on human-ML interactions. End users are, for example, more confident to follow the ML system's advice and maintain their domain expertise when the explainability of the ML system is high (Asatiani et al., 2021; Strich et al., 2021; Van den Broek et al., 2021). While XAI further offers the possibility to increase the ability of humans to evaluate ML system performance (Jussupow et al., 2021; Lebovitz et al., 2021), build up trust (Benbya et al., 2021), and communicate in a more effective way with the system (Kane et al., 2021), studies have also suggested that opportunities for learning arise (Abdel-Karim et al., 2020; Gaube et al., 2023; Lebovitz et al., 2021). These learning opportunities arise because ML models can independently find new solution approaches to problems and thus generate new knowledge that can be complementary to human knowledge (Fügener et al., 2022). Unfortunately, the use of ML is also fraught with risks and decision makers may lose the ability to find their own solutions or false learning occurs if incorrect advice is provided, risking a performance degradation in the long-run (Fügener et al., 2021). Providing explanations to make ML advice more human understandable bears the potential to change cognitive processes of humans and thus their beliefs (Bauer, von Zahn, et al., 2021). This could enable human decision makers to better understand advice given by an ML-based DSS and detect incorrect advice to avoid performance degradation. Nevertheless, the explanations from the field of XAI differ from

conventional explanations such as those found in the field of education. How well these arguments provided by machines are accepted needs to be researched.

Recent experimental studies in the field of radiology began to evaluate the impact of ML-based DSSs on the decision making of physicians (Abdel-Karim et al., 2020; Calisto et al., 2021; Gaube et al., 2023; Jussupow et al., 2021; Pumplun et al., 2023). Gaube et al. (2023), for example, evaluate the impact of fully correct, explainable AI-advice on human decision making. Jussupow et al. (2021) also take incorrect advice into account but do not explore the impact of explainable support. However, both and similar studies in the field assume that ML advice as well as the final human decisions are either right (correct) or wrong (incorrect) (Abdel-Karim et al., 2020; Calisto et al., 2021; Gaube et al., 2023; Jussupow et al., 2021; Pumplun et al., 2023). Accepting an ML recommendation that contradicts the physician's own judgment would thus require the physician to fully revise their decision, e.g., changing the diagnosis for a lung disease of a patient to having no lung disease (e.g., Abdel-Karim et al., 2020; Jussupow et al., 2021). In practice, however, diagnostic decisions require a more gradual evaluation such as the segmentation of tumor tissue in medical images instead of simply detecting whether a tumor exists or not. Since the results of the dichotomous experiments are only applicable to a small subset of the complex decisions which radiologists are required to make in the clinical practice, it is essential that we increase our understanding of more gradual diagnostic judgements in practice. Furthermore, such judgements are likely to indicate human learning through a gradual increase or decrease in performance.

The impact of explainable ML-based DSSs on the physician's decision making—and thus human learning—is largely unexplored. For instance, to the best of our knowledge, no previous research investigated whether explainability can reduce false learning from incorrect ML advice. It is a critical challenge to understand how ML-based DSSs should be designed to achieve positive progress in learning for human decision makers such as physicians.

### 3.2.5 Human Learning

Human learning is a complex process that generally involves the acquisition of information through various means such as sensory perception, experience, and instruction. This information-processing conception may include both short- and long-term storage of information (Gagné, 1970). However, it is a challenge to clearly define human learning at its core as well as to establish a metric by which human learning can be measured. (Miller, 1956) studied the amount of information that humans can receive, process, and retrieve. According to Miller (1956), learning comprises the organization of *bits of information* into familiar units, termed chunks of information. Chase and Simon (1973) later developed the *chunking theory*, which states that learning occurs through the accumulation of chunks in long-term memory. Domain experts recognize the familiarity of chunks and establish links to these chunks in their short-term

memory. This recognition of familiar chunks can then lead to skill development and thus a modification of a person's behavioral potential (Chase & Simon, 1973; Miller, 1976). Based on these traditional theories, learning can be defined as the process of acquiring and retaining knowledge in memory (Hunt, 2003). Nevertheless, Hunt (2003) emphasized that a clear distinction between behavioral potential, known as knowledge, and behavior must be made. In contrast to knowledge, which is an invisible and intangible asset for which no measurement standard exists (Sveiby, 1997), performance or behavior can be directly observed and thus measured (Hunt, 2003; Sveiby, 1997). Many studies that aim to observe learning therefore measure behavior with different performance-based metrics (e.g., Calisto et al., 2021; Regueras et al., 2009). These studies do not distinguish between knowledge and behavior which is inaccurate for many situations because not every change in knowledge leads to a behavior change (Hunt, 2003). When measuring learning progress, especially in a high-risk environment such as medical diagnostics, we should distinguish between a person A who comes to an incorrect diagnosis but feels highly uncertain about it and thus might be reluctant to provide it and a person B who is very certain that an incorrect diagnosis is correct and would therefore repeatedly make decisions based on this false belief (cf. Hunt, 2003). Thus, it is essential to determine the type of knowledge that leads to behavior change and hereby distinguish between belief and knowledge. A crucial element of a person's knowledge is the certainty with which they possess it (Hunt, 2003; Quine, 1987). Quine (1987) defines this question of how certain a person's belief must be for it to qualify as usable, behavior-guiding knowledge as the *boundary problem*. In addition, the certainty with which a person possesses knowledge also has a positive impact on the capability to retain that knowledge. Building on the traditional definition of human learning (Chase & Simon, 1973; Miller, 1976) and the definition of usable knowledge (Quine, 1987), Hunt (2003) proposes an *epistemic* method that assesses behavior not only by measuring performance (e.g., the correctness of an answer), but further through taking a person's certainty into account through measuring how sure a person is about a decision. This method of measuring human learning captures the quality of people's real-life performance, as it depends on both the knowledge they hold and the confidence with which they possess it (Hassmen & Hunt, 1994; Hunt, 2003). Our study aims at measuring whether novice radiologists can learn from ML-based DSSs in the context of tumor segmentation. Therefore, we will measure the acquisition of knowledge. Due to the high-stake environment in which we conduct the experiment, it is crucial to distinguish between behavior and knowledge. Thus, we selected the method proposed by Hunt (2003) and captured not only the changes in the radiologists' performance but also in their self-assessed decision confidence.

Explanations are considered crucial for effective human learning since they provide necessary context, background, and understanding to process and integrate new pieces of information into

existing knowledge. It is not surprising that multiple studies suggest that explanations are associated with learning progress (Crowley & Siegler, 1999; Fender & Crowley, 2007). For instance, if explanations are provided for the demonstration of new problem-solving strategies, the ability to transfer strategies to novel problems improves (Brown & Kane, 1988).

### 3.2.6 Hypothesizing Human Learning Through Interaction With ML-based DSSs

Our study links explanations in the context of human learning to the XAI field by exploring the impact of increasing explainability of ML-based DSSs on the learning of human decision makers. In line with previous studies suggesting that correct AI-generated decision support can in general improve the performance of human decision makers (Fügener et al., 2021; Gaube et al., 2023), we formulate **H1**: *Small learning gains can be achieved by interacting with non-explainable high-performing (NEHP) ML-based DSSs.* Despite mixed evidence on the effects of providing explanations or related measures such as certainty on human performance (cf. Fügener et al., 2021), XAI studies suggest that they affect human beliefs (Bauer, von Zahn, et al., 2021) whilst educational research emphasizes that explanations are critical for effective human learning (Fender & Crowley, 2007). Increasing the explainability of ML-based DSSs through explanations should therefore lead to greater learning progress. Thus, we propose **H2**: *Large learning gains can be achieved by interacting with explainable high-performing (EHP) ML-based DSSs.*

Incorrect, or in general low-performing ML-based decision support poses a great risk if decision makers fail to detect system errors (Goddard et al., 2012; Jussupow et al., 2021) risking a performance degradation (Fügener et al., 2021). Since non-explainable system advice is less interpretable for human decision makers—implying less opportunity to detect errors in the decision process (Arrieta et al., 2020)—we formulate **H3**: *False learning is promoted through the interaction with non-explainable low-performing (NELP) ML-based DSSs.* Improving the explainability of ML-based DSSs will likely increase the ability of human decision makers to detect errors (e.g., Sculley et al., 2015). This, in turn, is likely to reduce false learning, resulting in **H4**: *False learning is reduced through the interaction with explainable low-performing (ELP) ML-based DSSs.*

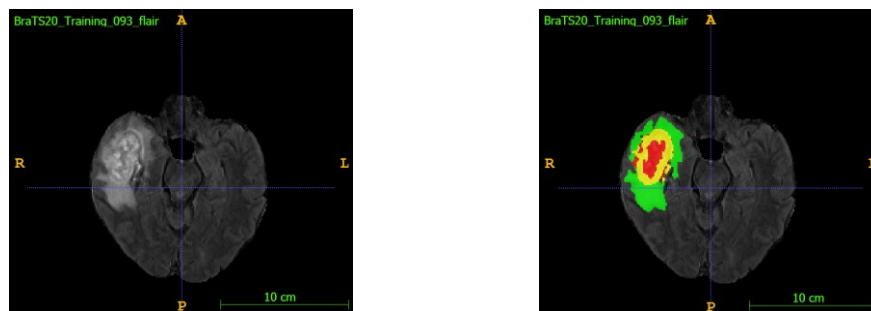
## 3.3 Methodology

To understand whether ML-based DSSs can impact human learning in the context of medical decision making, we designed an experimental setup in which radiologists were asked to segment tumors in MRI scans, before and after receiving decision support. Each radiologist received one of four different types of ML-based DSSs, i.e. a combination of high- vs low-performing and explainable vs non-explainable support. We followed a mixed-method study design: Besides obtaining quantitative, performance-based metrics during the experiment (supported by

subjective ratings of decision confidence), we collected qualitative data through think-aloud protocols and interviews. In-depth interviews with two experienced radiologists provided the necessary domain knowledge to design and validate the experimental setup with clinical expertise in advance.

### 3.3.1 Empirical Context

During pre-study interviews with experienced radiologists, we discussed which task would be most suitable for measuring gradual changes in decision performance and confidence as a result of learning progress in this setting (cf. Hunt, 2003). The radiologists recommended the segmentation of tumors in MRI scans due to their highly challenging nature, especially for novice physicians. Segmentations are particularly challenging for images that contain multiple areas that require interpretation such as a scan showing a brain tumor with perifocal edema<sup>5</sup>. Water-sensitive sequences of MRI scans often result in large, signal-rich areas around the tumors, potentially obscuring the actual, significantly smaller tumor (Unterberg et al., 2004). In such cases, it is highly challenging to distinguish the tumor from the surrounding perifocal edema (Lebovitz et al., 2021). An example of a high-grade glioma (a high-malignant brain tumor) which shows a perifocal edema is provided in Figure 3, along with its segmentation on the right-hand side: Red indicates the segmentation of the necrotic, non-enhancing, whilst yellow marks the Gadolinium (GD)-enhancing tumor (Bakas et al., 2017; Kaggle, 2020; Menze et al., 2015). The edema is colored green. For this example, the optimal segmentation of the tumor core would be the selection of only the yellow and red areas, without including any of the green areas.



**Figure 3: Exemplary MRI scan of a high-grade glioma with edema in FLAIR sequence (Bakas et al., 2017; Kaggle, 2020; Menze et al., 2015)**

Based on insights from our pre-study interviews, we decided to select the segmentation of low-grade gliomas in contrast-enhanced T1 sequences (T1CE) and high-grade gliomas in T2 fluid attenuated inversion recovery (FLAIR) sequences (Figure 3 is an example of these) as the experimental task. MRI scans and corresponding segmentations, which served as the ground truth for the analysis of the data collected in the experiment, were taken from the BraTS20 dataset. All

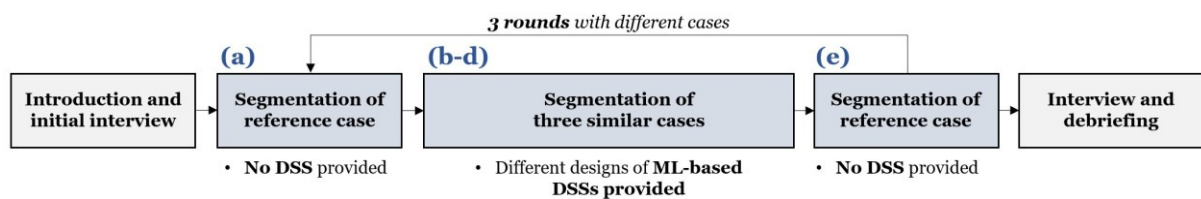
<sup>5</sup> Because of the limited space available in the skull and the often rapid growth of brain tumors, water retention occurs around the tumors, which leads to a so-called perifocal edema (Unterberg et al., 2004).



gliomas in the BraTS 2020 dataset have been manually annotated and the segmentations are approved by experienced neuro-radiologists (Bakas et al., 2017; Kaggle, 2020; Menze et al., 2015).

### 3.3.2 Research Design

The following section outlines the overall research design by first providing a brief overview of the general study procedure before detailing the design of the segmentation tool including the ML-based DSSs and the general experimental setup. Then, we describe which data was collected as well as the statistical methods used to test our experimental hypothesis. The research design was iteratively developed and pre-tested in collaboration with two experienced radiologists, mainly through pre-study interviews and consultations.



**Figure 4: Procedure of the study, which includes pre- and post-experimental interviews (gray), and the segmentation tasks within the experiment (blue)**

**Study Procedure.** The study procedure, outlined in Figure 4, was constant across all participants. Prior to the experimental part of the study, the experimenter conducted a short interview with the participant to collect informed consent and personal information as well as to determine the participant’s experience in general radiology, brain tumors, image segmentation, and diagnostics supported by ML-based DSSs.

Throughout each round of the experiment, each radiologist was instructed to complete five manual tumor segmentations on MRI scans, i.e. the initial (a) and final (e) segmentation of a reference case without decision support and three diagnostically similar segmentation cases (b-d) with ML-based decision support in between (diagnostically similar refers to cases that require similar segmentation considerations as the reference case). The performance of the reference case was used to evaluate whether the segmentation accuracy of the participant changed through the segmentation of similar cases with decision support. The intermittent cases with ML-based decision support provided the participants with the opportunity to segment similar tumors with support through the ML-based DSS. Then, they could apply the gained knowledge—if learning had occurred—to the final reference case. In addition, we asked participants about their confidence in their segmentations of the tumors. We measured the difference in accuracy and confidence between the initial (a) and final segmentation (e) of the reference case to represent learning. Overall, a total of three experimental rounds were performed by all participants (i.e. 15 segmentations), with two rounds containing the segmentation of five similar high-grade gliomas

and one round containing five similar low-grade gliomas. All participants received the same MRI scans for tumor segmentation.

To identify similar cases for the design of the experiment, a team of researchers and an experienced radiologist grouped diagnostically similar cases together, using clinical knowledge and the information provided in the BraTS20 dataset. To further validate the results and to perform a manipulation check (Aronow et al., 2019), the procedure concluded with semi-structured interviews.

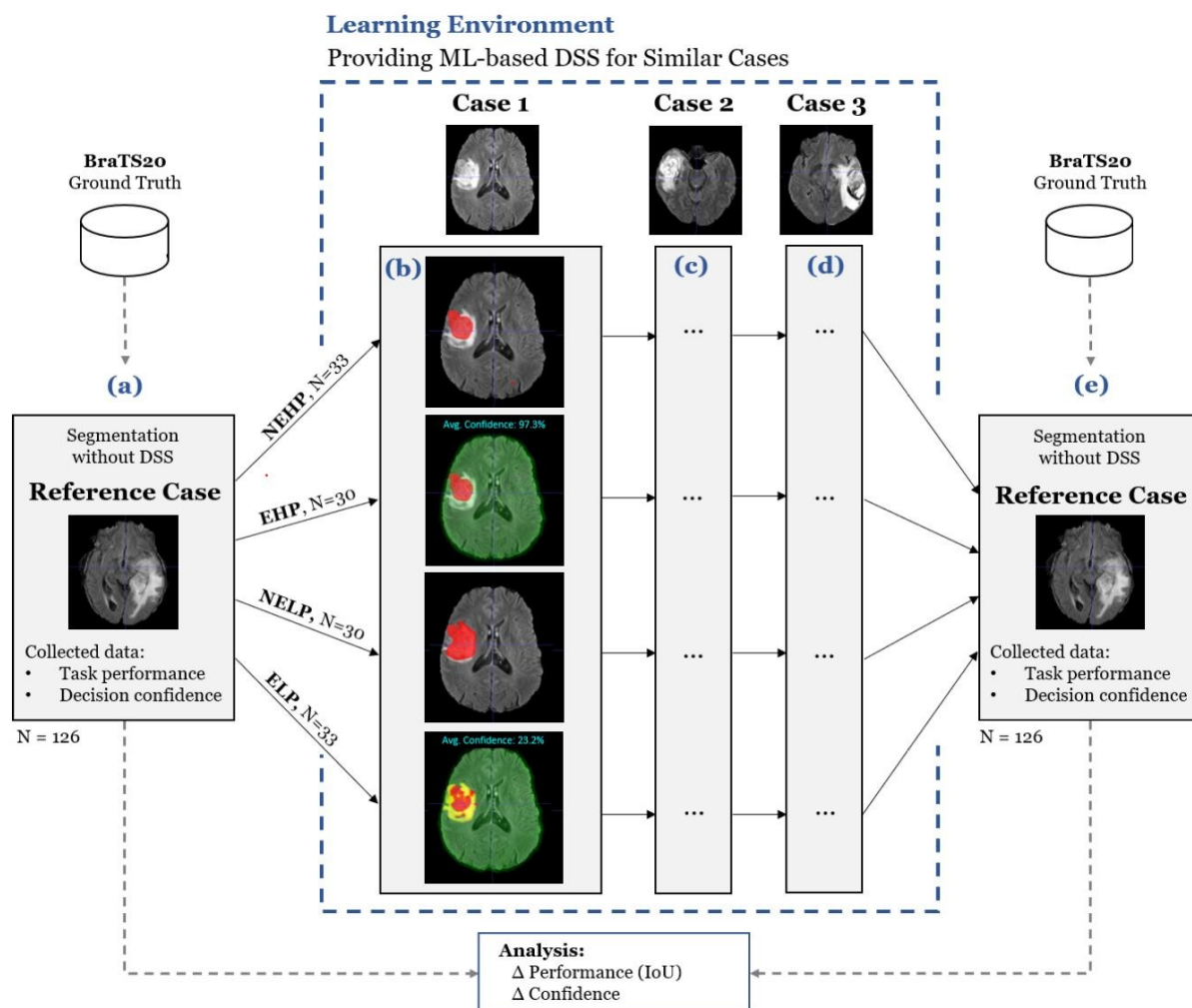
In the following, the between-subject study design, as shown in Figure 5, is described. To gather a baseline assessment, each experimental round started with a reference case (Figure 5 (a)) without any form of decision support. For the segmentation of the similar cases (steps (b), (c) and (d) in Figure 5), the participants were randomly assigned to one of four different types of ML-based decision support with varying levels of performance and explainability to test our hypotheses  $H1$  to  $H4$ . The different types of ML-based DSSs are described in the Experimental Setup section. Once assigned to a group, participants received the same form of ML-based decision support for all cases (b-d) and rounds of the experiment. Since this between-subject study design allowed for an independent evaluation of the impact of each form of decision support, it enabled an unbiased comparison of the impact of the performance and explainability aspects of the ML-based DSS on the participant's learning behavior.

**Experimental Setup.** Each case was presented and completed through a digital interface: The interface included an MRI scan series of a tumor and the participants were able to scroll through the scan along three axes. While all tumors could be analyzed in a 3D viewer, the segmentation task was performed on a single predetermined 2D plane to reduce the technical complexity of the segmentation task as well as to reduce the overall time required for tumor segmentation. For the creation of the interface, we used the interactive software application ITK-SNAP 3.6.0 which provides an intuitive user interface for segmenting structures in 3D medical images (Yushkevich et al., 2006). ITK-SNAP enabled participants to draw a polygon along the tumorous area for fast and easy manual segmentation. For the three similar cases (b-d), the ML-based decision support was provided as a semi-transparent overlay on top of the MRI scan which could be switched on and off by participants during review of the scan and segmentation of the tumorous area.

Four different types of ML-based DSSs were defined to test hypotheses  $H1$  to  $H4$ , respectively: non-explainable high-performing (NEHP), explainable high-performing (EHP), non-explainable low-performing (NELP), and explainable low-performing (ELP). The ML-based decision support variants are constructed through *Accuracy-* and *Explainability Manipulation*:

*Accuracy Manipulation.* Different from prior IS studies in the field of medical decision making (e.g., Jussupow et al., 2021; Pumplun et al., 2023), this study focuses on gradual performance changes

(vs the impact of completely correct or incorrect decision support). The ML-based DSS was either high- or low-performing, with regards to correctly outlining the tumor core. The high-performing ML-based DSSs (NEHP and EHP) provided correct advice for >90% of the tumorous area, thus slightly outperforming prevailing clinical practice (e.g., Jiang et al., 2017). For the low-performing support (NELP and ELP), we reduced the accuracy below <50% so that large portions of tumor were not detected and large portions of healthy tissue were incorrectly classified as tumor.



**Figure 5: Between-subject experimental study design for segmentation of brain tumors and different designs of ML-based DSSs**

*Explainability Manipulation.* The design of our explainable ML-based DSSs (EHP and ELP) is based on the guide for the development of explainable ML-based DSSs by Pumplun et al. (2023) that has been validated through an evaluative study with radiologists. In the explainable decision support conditions, the radiologist is provided with local explanations in form of a confidence heatmap and an uncertainty estimate for the ML-based DSS’s classification of the tumor area. Both methods aim to increase the comprehensibility of the ML advice to enable radiologists to evaluate the system’s performance (cf. Pumplun et al., 2023). The heatmap visualizes the ML model’s confidence for each pixel of the image in three discrete levels: pixels with high confidence for

tumorous tissue are colored red, while areas with high confidence for healthy tissue are colored green. Areas with generally low confidence for either class are colored yellow, signifying uncertainty of the underlying ML model. The uncertainty estimate, i.e. the average confidence of pixels classified as belonging to the tumor, was displayed above the brain scan. Both explainability features are exemplary visualized in Figure 5 (b), EHP and ELP. The non-explainable ML-based DSSs (NEHP and NELP) use a uniform color for all pixels indicated to be tumorous and do not provide an uncertainty estimate, as also shown in Figure 5 (b).

**Data Collection.** Throughout the experiment, think-aloud protocols were recorded (Van Someren et al., 1994). Think-aloud protocols have proven to be highly valuable in IS research as a method to collect qualitative data to study decision making and support systems (e.g., Amershi et al., 2019; Sculley et al., 2015), especially for medical decision making (e.g., Jussupow et al., 2021). Besides qualitative data, quantitative data was collected during each step of the process (a-e). Participants were asked about their confidence in their segmentation of the tumor, i.e. its correctness. This decision-making confidence was assessed using a 5-point Likert scale, ranging from “not sure at all” to “extremely sure” as proposed by Hunt (2003). For each experimental round, the change in decision confidence between the initial and final segmentation of the reference case was derived (see Figure 5), i.e. the delta in decision confidence.

Furthermore, the accuracy of the initial and final segmentation of the reference case was assessed by calculating the intersection over union (IoU). The IoU is the most commonly used standard performance measure for semantic segmentation tasks (Rahman & Wang, 2016; Rezatofighi et al., 2019). This performance measure evaluates the similarity between a segmentation (predicted or manually annotated) for a region A and the ground-truth region B, thus defined as the size of the intersection ( $A \cap B$ ) divided by the union ( $A \cup B$ ) of the two regions (Rahman & Wang, 2016). While participants manually segmented area A, the ground-truth area B was taken from the BraTS20 dataset which provides segmentations that are approved by experienced neuro-radiologists (example in Figure 3) (Bakas et al., 2017; Kaggle, 2020; Menze et al., 2015). The delta IoU between the initial (a) and final (e) reference case was calculated to measure the change in performance in the following way:

$$\Delta IoU = \frac{\text{Area of overlap}_{post}}{\text{Area of union}_{post}} - \frac{\text{Area of overlap}_{pre}}{\text{Area of union}_{pre}} = \frac{A \cap B_{post}}{A \cup B_{post}} - \frac{A \cap B_{pre}}{A \cup B_{pre}}$$

As elaborated in the previous section, we used both, the delta in decision confidence and performance (IoU delta), as equally weighted proxies for our dependent variable of learning, as recommended by Hunt (2003). To conclude each trial, a semi-structured interview was conducted to gain qualitative insights into the individual learning progress and to perform a manipulation check, ensuring that participants were aware of the ML-based decision support and, if present, local explanations (Aronow et al., 2019). Participants were asked to describe their experience

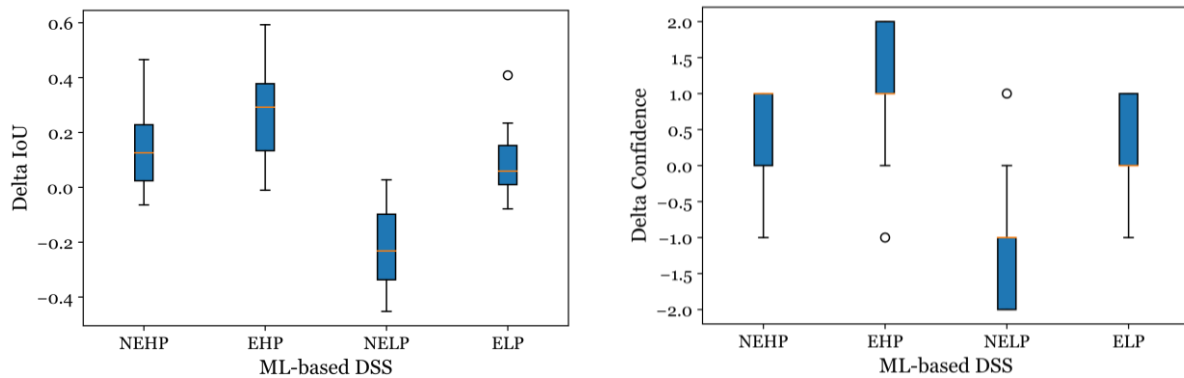
with the ML-based DSS and evaluate how helpful they perceived the given support. The interview guideline was developed in accordance with (Sarker et al., 2013) to collect all necessary information while allowing participants to freely share their experiences from the study.

**Data Analysis.** Interviews and think-aloud protocols were audio recorded and transcribed after study completion. A content analysis was performed to evaluate and structure the participants' statements (Weber, 1990), utilizing the coding process recommended by (Saldana, 2021). The process included attribute, descriptive as well as pattern coding for retrieving descriptive data from the recording, as well as insights into the participants' interaction success and challenges with the ML-based DSSs. Quantitative data - i.e. performance and decision confidence data - was analyzed in SPSS V29. Following an initial data analysis, t-tests for each group and an analysis of variance (ANOVA) were conducted to test the impact of different types of ML-based DSSs on the performance and confidence of radiologists (Girden, 1992; Tabachnick & Fidell, 2006).

**Participants and Execution Statistics.** The application of ML-based DSSs in medical diagnostics will increase in the coming years, thus influencing the future working environment of today's novice physicians. Additionally, ML-based DSSs supporting medical diagnostics are primarily beneficial for physicians with a higher probability for diagnostic errors (Shen et al., 2019). Our pre-study interviews showed that brain tumor segmentations are more challenging for novice physicians, thus providing a greater opportunity to learn and improve. As a result, we primarily recruited novice physicians as study participants to evaluate the impact of ML-based decision support on learning progress. The experiment was conducted between December 2022 and March 2023 at different clinics with radiology departments in Germany. In total, our study included 42 participants (20 female, 22 male), of whom 18 were novice physicians with less than a year of clinical experience (8 medical interns in their first year and 10 residents with less than a year of experience), 16 novice physicians with 1-3 years of clinical experience, and 8 physicians with more than 3 years of experience. All participants work in hospitals in the radiology department at the time of the experiment. Overall, 630 segmentations were obtained throughout the experiment and therefore  $N = 126$  full rounds (step a-e) completed to measure a delta in IoU-based performance and decision confidence. Participants were randomly assigned equally to different ML-based DSS groups ( $N_{NEHP} = 33$ ,  $N_{EHP} = 30$ ,  $N_{NELP} = 30$ ,  $N_{ELP} = 33$ ) and required an average of 6:43 minutes (min = 5:02 minutes; max = 13:24 minutes) to complete a full round of the experiment (five segmentations). Including pre- and post-experimental interviews, the total experiment lasted between 20 and 40 minutes. Three physicians had to pause between two rounds of the experiment, due to medical emergencies at the clinic.

### 3.4 Results

The following section begins by showing the impact of interacting with the different types of ML-based DSSs on the learning progress before comparing the results between groups. Drawing on these insights, we further explore the influence of prior experience on the learning potential offered by different ML-based DSSs. For this between-subject study design, four different variants of ML-based DSSs were presented, namely NEHP, EHP, NELP and ELP decision support. Paired t-tests (Stone, 2010) are conducted to assess the impact of each ML-based DSS type and Cohen's  $d$  is reported as a measure of the effect size (Cohen, 2013). Due to the sample size of 30 or more for each group, the normal distribution of obtained deltas in IoU performance and decision confidence can be assumed, so that paired t-tests can be applied as a robust measure (Stone, 2010). We select  $p < 0.05$  as our significance threshold.



**Figure 6: Measured deltas of pre- and post-experiment IoU-based performance and decision confidence**

Physicians that interacted with the **NEHP** ML-based DSS improved their segmentation performance in the reference case ( $M = 0.14$ ;  $SD = 0.12$ ). Paired t-tests showed significant effects both for IoU ( $t(32) = 6.569$ ,  $p < 0.001$ ,  $d = 1.14$ ) as well as for decision confidence ( $M = 0.61$ ;  $SD = 0.56$ ;  $t(32) = 6.266$ ,  $p < 0.001$ ,  $d = 1.09$ ). These results suggest that learning gains can be achieved through the interaction with NEHP ML-based DSS, supporting **H1**.

In line with **H2**, we observed larger improvements in performance and decision confidence for the participants that interacted with **EHP** ML-based DSS for the segmentation of similar brain tumors. IoU based segmentation performance increased significantly by  $M = 0.27$  ( $SD = 0.16$ ),  $t(29) = 9.252$ ,  $p < 0.001$  with a large effect size,  $d = 1.69$ . The average decision confidence increased by  $M = 1.30$  ( $SD = 0.75$ ,  $t(29) = 9.497$ ,  $p < 0.001$ ), also showing a large effect size ( $d = 1.73$ ). As a result, we conclude that **H2** is supported and EHP decision support led to increased learning.

The changes in performance ( $M = -0.22$ ;  $SD = 0.14$ ) and decision confidence ( $M = -1.10$ ;  $SD = 0.76$ ) of the group that received the **NELP** ML-based DSS showed a significant degradation (performance:  $t(29) = -8.875$ ,  $p < 0.001$ ,  $d = -1.62$ ; decision confidence:  $t(29) = -7.940$ ,  $p < 0.001$ ,

$d = -1.45$ ). These findings support **H3** and the assumption that false learning is promoted by interacting with NELP decision support.

To test **H4**, the fourth group of physicians interacted with **ELP** ML-based DSS. Surprisingly, interacting with ELP ML-based DSS led to a small improvement of IoU in the final reference case segmentation ( $M = 0.08$ ;  $SD = 0.18$ ). While the improvement in segmentation performance ( $t(32) = 4.742$ ,  $p < 0.001$ ,  $d = 0.83$ ) is significant and of large effect size, the decision confidence did only marginally significantly increase ( $M = 0.27$ ;  $SD = 0.11$ ,  $t(32) = 2.502$ ,  $p = 0.09$ ). We conclude that explainability in interactions with low-performing decision support does not only reduce false learning (as suggested in **H4**) but even leads to a small learning gain. Overall, **H1**, **H2**, **H3** and **H4** are supported by the results.

Next, we conducted a one-way ANOVA ( $p < 0.05$ ) to compare the groups and assess differences in the effects of providing explainable and non-explainable, low- and high-performing ML-based DSSs on performance and decision confidence of novice radiologists.

**Segmentation Performance.** First, homogeneity of variances for delta IoU performance scores was asserted using Levene's test which showed that equal variances could be assumed (based on the mean value:  $p = 0.059$ ; based on the median:  $p = 0.092$ ). The ANOVA revealed that changes in IoU performance differed statistically significant for the different forms of ML-based DSSs provided in the learning environment,  $F(3, 122) = 76.705$ ,  $p < 0.001$  with effect size  $\eta^2 = 0.654$ . Tukey post-hoc analysis revealed a significant positive difference ( $p < 0.001$ ) between delta IoU scores of the groups with NEHP and EHP ML-based DSS (difference of means EHP-NEHP: 0.1303, 95%-CI[0.0443, 0.2164]) as well as NELP and ELP ML-based DSS (ELP-NELP: 0.3070, 95%-CI[0.2209, 0.3930]). Interestingly, the Tukey post-hoc analysis showed no significant difference ( $p = 0.309$ ) of delta IoU scores between the group receiving NEHP ML-based DSS and the group receiving ELP ML-based DSS (NEHP-ELP: 0.0560, 95%-CI[-0.0279, 0.1400]). These results show that improving the explainability of a low-performing ML-based DSS has the same learning outcome as a non-explainable ML-based DSS with higher accuracy.

**Decision Confidence.** Before conducting the second ANOVA to evaluate the impact of providing different forms of ML-based DSSs on decision confidence of physicians, the homogeneity of variances for obtained confidence scores was tested using Levene's test. Equal variances for the delta in decision confidence could be assumed (based on the mean value:  $p = 0.741$ ; based on the median:  $p = 0.648$ ). The ANOVA showed that changes in decision confidence of physicians differed significantly between the four groups  $F(3, 122) = 67.396$ ,  $p < 0.001$ ,  $\eta^2 = 0.624$ . A Tukey post-hoc analysis revealed that explainability of ML-based DSS leads to significant differences ( $p < 0.001$ ) in changes in decision confidence (EHP-NEHP: 0.694, 95%-CI[0.25, 1.14]), (ELP-NELP: 1.373, 95%-CI[0.93, 1.82]). The groups that received NEHP ML-based DSS or ELP ML-based DSS show

no significant ( $p = 0.190$ ) differences in their changes of decision confidence (NEHP-ELP: 0.333, 95%-CI[-0.10, 0.77]). The results of both ANOVAs suggest that explainability has a major impact on the learning potential that decision makers can derive from interacting with ML-based DSSs (**H1** vs **H2** and **H3** vs **H4**). In addition, the explainability of low-performing support was able to prevent false learning (i.e. reduce decision confidence and performance, **H4**).

**Prior Experience.** Based on the radiologists' answers (on a 5-point Likert scale) regarding prior clinical experience in the field of brain tumors during the pre-experimental interviews, they were clustered into an experienced ( $N = 58$  rounds, Likert scale rating  $\geq 3$ ) and a non-experienced group ( $N = 68$  rounds, Likert scale rating  $< 3$ ). The Levene test showed that homogeneity of variances for obtained absolute values of delta IoU performance scores ( $p = 0.232$ ) and decision confidence ( $p = 0.127$ ) could be assumed. An ANOVA revealed that the absolute values of change in decision confidence across all variants of decision support, differed significantly between the experienced and non-experienced groups,  $F(1, 124) = 8.3221$ ,  $p = 0.005$ ,  $\eta^2 = 0.063$ . While non-experienced novice physicians on average strongly adapted their decision confidence between the initial (a) and final (e) segmentation of reference cases ( $M = 1.06$ ;  $SD = 0.644$ ), experienced radiologists showed a lower absolute adaption of their decision confidence ( $M = 0.73$ ;  $SD = 0.653$ ). However, the standard deviation in both groups is relatively high. In addition, the absolute deltas in segmentation performance were smaller for physicians with prior experience, but do only marginally significantly differ with  $F(1, 124) = 3.338$ ,  $p = 0.070$ ,  $\eta^2 = 0.026$ , between the non-experienced ( $M = 0.20$ ;  $SD = 0.13$ ) and the experienced ( $M = 0.16$ ;  $SD = 0.15$ ) group.

**A Qualitative Post-Experimental Evaluation.** The think-aloud protocols and post-experimental interviews confirmed that the manipulations of ML-based DSS explainability and performance were successful. All participants stated that AI advice was taken into account and participants receiving EHP or ELP support reported explainability features correctly and described how they were considered. In addition, both variants of the high-performing ML-based decision support received better performance ratings by the physicians.

Multiple physicians that interacted with explainable ML-based DSSs tried to interpret what the underlying reasons for uncertainty of the systems were. Accordingly, one physician (with brain tumor experience) receiving ELP ML-based decision support attempted to evaluate whether his/her own uncertainty was matched by that of the AI: *"Those are really difficult cases, especially due to the blurriness in the area where the edema occurs. It was difficult for me to see where the boundaries of the tumor are. [...] But if we come back to the edema, I noticed that the AI is also uncertain here and highlights large areas—yes—rather yellowish. And especially around the core of the tumor. This is where the difficulty lies in demarcating edema and tumor, but edema doesn't actually belong to it. So, it does make sense that the AI is uncertain but not entirely correct here. I would then segment [...] and here I would delineate the area from the tumor, even if it was not*



*classified as green—so no tumor.*” Several radiologists confirmed that they perceived that they made learning progress, although some found it difficult to describe exactly what was learned: *“I definitely learned something. [...] I think it’s difficult to say what exactly I learned because it’s not like learning vocabularies, but I would say I improved my skills in differentiating between healthy and non-healthy tissue.”* In addition to skill improvement, confidence in one’s decisions was often explicitly perceived as something that changed as a result of interacting with the EHP ML-based DSS: *“It has been a long time since I’ve dealt with brain tumors in any way. But this tool has refreshed my knowledge which I had a bit blurred in my memory. At least I feel more confident, I would say.”*

### 3.5 Discussion

With the increasing use of ML-based DSSs in radiology (Radboud University Medical Center, 2023), collaborative human-AI decision making will become part of the daily clinical routine of radiologists worldwide. While much effort is spent to improve system performance and incorporate human feedback through human-in-the-loop concepts (e.g., Grønsund & Aanestad, 2020), little is still known about the impact of ML-based DSSs on the human decision maker (e.g., Gaube et al., 2023; Pumplun et al., 2023). Despite concerns that the use of ML-based DSSs will result in a degradation in the performance of radiologists as human decision makers (cf. Fügner et al., 2021), opportunities for human learning may arise (e.g., Abdel-Karim et al., 2020). Current research on the matter is scarce and only considers a subset of real-world decisions (e.g., Abdel-Karim et al., 2020; Gaube et al., 2023; Pumplun et al., 2023) whilst not fully representing the dependencies between explainable design, system performance, and the individual human learning process. Our experimental study sheds light on the underexplored area of the impact that the design and performance of ML-based DSSs have on human learning in the context of brain tumor segmentation. We provided radiologists with explainable vs non-explainable decision support that was high- vs low-performing in a between-subject study design. This examined whether radiologists are able to learn from the—highly or less correct—knowledge provided by the system, as well as the impact of explainability on these learning outcomes. In addition to qualitative data, we obtained quantitative data on the learning progress of radiologists.

Our study **contributes to theory** in multiple aspects: First, it demonstrates that high-performing ML-based DSSs are capable of improving the performance and decision confidence of radiologists and thus foster human learning (cf. **H1** and **H2**). However, low-performing decision support that is presented in a non-explainable manner resulted in significant false learning outcomes, i.e. worse performance and decision confidence (**H3**). These findings demonstrate the significant impact that ML-based DSSs can have on the human decision maker. We hope to encourage scholars to consider this important aspect in future studies that poses great opportunities as well as risks to human-DSS teams. Through enriching this discussion with empirical evidence, we can

unlock opportunities to improve the design and development of ML-based artifacts that will have a major impact on various IS research studies. Our findings might be especially informative for scholars interested in extending human-in-the-loops concepts through considering human learning opportunities as further detailed below.

Second, our findings contribute to the growing XAI research stream by adding to the goals that can be achieved through explainable system design: Explanations that increase the explainability of high-performing ML-based DSSs boost the learning potential for human decision makers (**H2**). The significant positive impact of explainability on human learning suggests that increasing the explainability of ML-based DSSs may be more beneficial than further increasing their accuracy. Ensuring a high level of explainability for ML-generated advice is even more important if the ML-based DSS is rather low-performing: Whilst low-performing and non-explainable DSSs lead to false learning and reduced accuracy (**H3**), we show that this can be significantly reduced (**H4**) and in many cases even prevented through increasing the explainability of the ML-based DSS. In addition to detecting false advice, some radiologists were even able to achieve learning gains with low-performing, but explainable ML-based DSSs. This astonishing result is of great importance for the design of ML-based artifacts and future ML adoption strategies. Understanding how harmful effects of ML-based DSSs on physicians (such as false learning) can be prevented is likely to decrease adoption barriers in high-risk areas. By demonstrating that human decision makers can learn from mistakes, even when made by an ML-based DSSs, we provide a new basis for future transdisciplinary research.

Third, our quantitative measurement method empirically demonstrates human learning—a rarity in this research area. Since confidence and performance behave differently in learning environments, this two-sided method to measure learning is essential and can reveal gradual changes in knowledge, which a pure performance-based measurement does not allow for. We offer scholars from the field of ML-based DSSs to adopt our approach to examine how their systems impact performance and decision confidence (and thus learning) of human decision makers. In the context of this study, we were able to link the value of explanations from the XAI domain with the value of explanations from educational research. In doing so, we show that existing theories of human learning need to be extended with respect to ML and our approach provides a common basis to align research from the fields of educational research, psychology, human-computer interaction, and XAI in a transdisciplinary endeavor.

Fourth, the results show that while ML-based decision support offers benefits to decision makers with different levels of experience and can be broadly deployed, learning outcomes decline slightly as experience increases. These results indicate that studies in human learning and XAI need to consider who the relevant end users are and that prior experience will have an impact on study outcomes and should therefore be captured as a control variable.

In addition to theoretical contributions, the study provides **practical contributions**. First, the proposed design of explainable ML-based DSSs for tumor segmentation can be used as a blueprint for the development of clinical systems that aim to support physicians effectively and at low risk. Second, as false learning can be prevented for low-performing but explainable ML-based DSSs, earlier adoption of the technology in high-stake environments might become possible: When implementing ML systems, there is often a chicken-and-egg problem, as there is not enough data to achieve sufficient accuracy during training, and real-world data can only be collected after deployment. Since explainable advice does reduce the risk of false learning, explainable ML-based DSSs can—in some cases—be adopted earlier and increase their accuracy over time. When human experts are involved in the decision making process, explainability of ML-based DSS may have a greater impact on adoption than performance. In addition, we hope to inspire organizations that want to enable continuous learning among their employees or provide quality education to achieve the SDGs to explore and adopt novel approaches such as those presented in this study.

The following paragraph will discuss the **limitations** of our study. Although the experiment was based on real medical cases and diagnostic findings, we only provided the 3D MRI brain scans to the physicians. The physicians did not know the patients' medical history (e.g., smoking, family history) and therefore had to rely solely on the images. In addition, we pre-selected specific difficult patient cases for tumor segmentation. However, in real medical decision making, patients may suffer from more visible tumors or multiple diseases, and physicians may gather more information through face-to-face consultations. Nevertheless, since we used the same cases for all experiments and only manipulated the explainability and performance of the ML-based DSS, we see validity in the results. In addition, the validity of our results is supported by the collaborative development of the study design with experienced radiologists, as well as by the feedback we received from participants in the post-experimental interviews. Lastly, although some results on the impact of ML-based DSSs on human learning can be generalized to other high-risk environments, such as pilots in aviation, biases may occur due to individual challenges in different decision making environments.

We encourage **future research** to apply the results of the study to other application areas. Building on our findings, future studies should further investigate how explanations from the field of XAI should be designed and presented to the end user to best support learning. In addition, studies on the design of these explanations should investigate the impact of the end user's prior knowledge. That humans in general can learn something from mistakes, even if they are made by an AI, is an exciting result of this study. Therefore, we conclude by calling for future research to investigate how explanations should be designed so that decision makers can not only easily identify errors made by AI, but also learn from them.

### **3.6 Conclusion**

In this paper, we proposed four hypotheses to explore if the interaction with ML-based DSSs can promote human learning and which impact explainability has on potential learning outcomes. We developed four ML-based DSSs which are either high- vs low-performing and explainable vs non-explainable to support radiologists in segmenting brain tumors in MRI scans in an experimental study. We evaluated quantitative results on segmentation performance and decision confidence of radiologists as well as qualitative data gained through think-aloud protocols. The evaluation revealed that interaction with high-performing ML-based DSSs fostered human learning, with higher explainability enhancing this effect. In addition, explainability prevented false learning triggered by low-performing ML-based decision support. Interestingly, radiologists were even able to learn from errors made by the low-performing explainable ML-based DSS. These findings provide guidance for the future development of ML-based DSSs, that are particularly beneficial to the human decision maker and provide the opportunity to make even better decisions in the future, especially in high-risk areas such as radiology.

## **4. Paper B: Toward the Sustainable Development of Machine Learning Applications in Industry 4.0**

### **Title**

Toward the Sustainable Development of Machine Learning Applications in Industry 4.0

### **Authors**

Ellenrieder, Sara; Jourdan, Nicolas; Biegel, Tobias; Bretones Cassoli, Beatriz; Metternich, Joachim; and Buxmann, Peter

### **Publication Outlet**

Proceedings of the 31<sup>st</sup> European Conference on Information Systems (ECIS), Kristiansand, Norway, 2023

### **Abstract**

As the level of digitization in industrial environments increases, companies are striving to improve efficiency and resilience to unplanned disruptions through the development of machine learning (ML)-based applications. Still, sustainable deployment and operation beyond proofs-of-concept is a challenging and resource-intensive task in dynamic environments such as industry 4.0, often impeding practical adoption in the long term and thus sustainable ML product development. In this work, we systematically identify these challenges based on the CRISP-ML process model phases by applying a design science research approach. To this end, we conducted 15 interviews with data science practitioners in industry 4.0. Following a qualitative content analysis, design requirements and design principles for the development and sustainable long-term deployment of ML systems are derived to address identified challenges such as robustness to, and management of data drift caused by time-dependencies and machine/product differences, missing metadata, interfaces to other IT systems, expectation management, and MLOps guidelines.

### **Keywords**

Machine learning, industry 4.0, long-term deployment, design science research

## 4.1 Introduction

Machine learning (ML)-based applications increasingly transcend from academia to operationally used applications in a multitude of industry sectors. Nevertheless, sustainable deployment and operation of ML-based applications beyond proofs-of-concept and prototypes is a challenging and resource-intensive task with various possibilities of project failure. Especially the long-term operation of ML applications poses several difficulties that are often not considered in academic research such as decaying accuracy due to the dynamic nature of the environment and the usage by non-experts (Jourdan et al., 2021; Rudin & Wagstaff, 2014; Salama et al., 2021). Still, long-term operation is paramount to justify investments in the development of ML applications and to maximize the benefits that ML offers (McKinsey Global Institute, 2021; Vela et al., 2022). In addition to economic reasons, sustainable product development also requires that the value of a product, such as an ML application, which has consumed large amounts of resources during its development, be maintained over time (Klöpffer, 2003; van Wynsberghe, 2021).

Due to the rising levels of digitization in today's factories, industry 4.0 related applications have received significant attention in this context (Fahle et al., 2020; Wuest et al., 2016). Information technologies such as the Internet of Things (IoT) and the interconnectivity of industry 4.0 transform manufacturing lines into cyber-physical systems (CPS), which yield large amounts of data (Cassoli et al., 2022). Moreover, industry 4.0 related processes such as manufacturing operations are vital parts of their respective companies' value streams and naturally consume large amounts of resources, yielding high optimization potential (Wuest et al., 2016). Thus, companies in industry 4.0 strive to use the available data to improve effectiveness, efficiency and the resilience of their operations against unplanned disturbances to gain competitive advantages. However, in 2021, a study focusing on the manufacturing industry in Germany showed that two-thirds of ML applications developed in the participating companies did not surpass the concept development and prototyping stage, which hints at a significant untapped potential in the industry (Metternich et al., 2021). Overall, ML applications in the scope of industry 4.0 provide a particularly interesting context for our study because, first, ML applications offer high potential in this sector and ongoing investments already call for long-term operation (McKinsey Global Institute, 2021), second, frequent changes occur especially in manufacturing processes (Jourdan et al., 2021) and third, operators that use ML applications in this context usually do not have ML knowledge (Wuest et al., 2016), thus posing special challenges for long-term deployment. In this study, we refer to the specific context of industry 4.0 and can thus investigate how ML applications should be designed to enable long-term deployment in dynamic environments, which in turn has applications in different industries.

Information systems (IS) research on ML applications in organizations also shifted focus to long-term deployment of ML applications and calls for continuous auditing and altering activities for adapting ML models to dynamic changes in problem perception. While several studies highlight the need for continuous maintenance of ML applications after they have been deployed, the integration of human feedback as a solution for model adaptation continues to be a focus of attention (e.g., Asatiani et al., 2021; Grønsund & Aanestad, 2020; Sturm, Gerlach, et al., 2021). However, previous research still thinks too short and intervenes only during the deployment phase of ML, which in practice is often too late to smoothly adapt ML applications to dynamic changes. Many system properties, which are of central importance for the later adaptation of the models, are already determined and implemented in early project phases of ML application development. Thus, we see the need for a holistic approach that takes all phases of ML development and deployment into account and provides clear guidance on how to design ML applications that support the goal of long-term deployment in any ML project, especially in industry 4.0. If ML applications continue to be designed in a way that makes future adaptation difficult, there is a risk that large investments in this area will continue to be associated with little long-term impact. We aim to provide guidance in how to develop ML applications that provide the flexibility and adaptability required for long-term use in dynamic environments. Therefore, we are interested in the predominant technical and organizational challenges of sustainable ML implementation and use the context of industry 4.0 to ask how future research as well as practitioners that aim at designing, developing and deploying ML applications can address these challenges, leading to the following research questions (RQs):

*In the context of industry 4.0, **(RQ1)** which challenges impede the sustainable development and deployment of ML systems, and **(RQ2)** how should ML systems be designed to overcome those challenges in order to ensure sustainable long-term deployment?*

To address our two research questions, we apply a design science research (DSR) approach that allows us to structure the given problem and derive design requirements and principles in order to conceptualize suggested solutions (Kuechler & Vaishnavi, 2008). We start with reviewing the popular Cross Industry Standard Process for Data Mining (CRISP-DM) (Wirth & Hipp, 2000) and the more recent CRISP-ML (Studer et al., 2021) for project management of ML applications in industrial settings, focusing on sustainability and long-term operation support to derive a structure for further analysis. As the main contribution of this study, semi-structured interviews are conducted with 15 data science and machine learning practitioners in the industry 4.0 sector. Challenges identified in these interviews are structured according to the CRISP-ML model and respective solution approaches to avoid these pitfalls and ensure for sustainability of ML applications are derived as design requirements and design principles to ensure ML applications have a long-term impact on industry and society. While many ML projects use the DSR approach,

in the future we would like to enable these projects to incorporate sustainability considerations and maintain the long-term value of their ML application, into which large amounts of resources have been poured during development.

## 4.2 Related Work

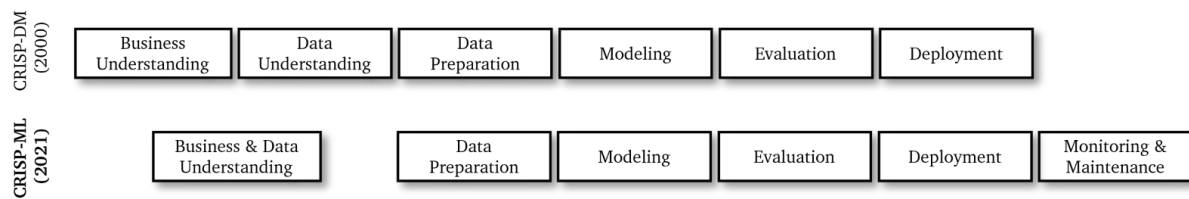
The following section provides an overview of the characteristics of ML and the corresponding process models that structure ML development projects. In addition, related studies that explore the need for continuous adaptation of ML applications in dynamic environments are presented.

### 4.2.1 Definition of ML and Process Models for Structuring ML Projects

Today, artificial intelligence (AI) and especially ML as a subfield of AI is deployed in various industries. In general, ML is defined as an approach that uses learning algorithms to derive patterns from data in order to build models that can solve real-world problems (Brynjolfsson & Mitchell, 2017; Mitchell, 1997; Russell & Norvig, 2016). In addition, different forms of ML have emerged of which supervised, unsupervised and reinforcement learning are the most common ones (Russell & Norvig, 2016). Regardless of the ML form chosen, ML applications use their algorithms to derive solutions to problems based on the data provided. Conversely, this means that the developers of ML systems no longer derive solutions to real-world problems themselves but describe problems based only on data, which is a significant difference from traditional (non-ML) information systems (Mitchell, 1997; Russell & Norvig, 2016). Due to those unique characteristics, the development of ML systems requires a different approach than the development of non-ML systems.

Process models such as CRISP-DM (Wirth & Hipp, 2000) provide guidance and structure to projects developing data mining or machine learning applications. CRISP-DM structures respective projects into six consecutive phases (see Figure 7). In the *Business Understanding* phase, the project's business problem is identified and described by relevant performance metrics and requirements. On this basis, related data mining goals are derived. Relevant available data and meta-data sources are analyzed and characterized during the *Data Understanding* phase. The *Data Preparation* phase involves data cleaning, preprocessing, feature extraction, and selection to build useful datasets for the following stages. In the subsequent *Modeling* phase, machine learning algorithms are applied, primarily involving model selection and parameters training. The *Evaluation* phase typically relies on offline evaluation using held-out test data and suitable performance metrics (Jourdan et al., 2021). If the model fulfills the requirements defined in *Business Understanding*, it is subsequently deployed to production in the last phase (*Deployment*) (Wirth & Hipp, 2000).





**Figure 7: Qualitative comparison of the CRISP-DM (Wirth and Hipp, 2000) model phases to the more recent CRISP-ML (Studer et al., 2021), which provides the structure for the categorization of challenges and derived design requirements.**

The CRISP-DM model assumes that most of the described process, including the final evaluation, is performed offline using static datasets. Furthermore, it does not provide specific guidance for the time after the model is deployed to production. Thus, it does not cover the whole life cycle of an ML application. If a problem occurs after model deployment, a data scientist is expected to fix it. Here, a potential flaw becomes obvious: In dynamic environments which can be seen in industry 4.0, changes in data distributions are to be expected and should thus be addressed preemptively. Not addressing this issue risks performance degradation over time, which leads to false predictions and could cause errors in subsequent systems as well as diminishing trust of the application's users (Asatiani et al., 2021; Yin et al., 2019).

More recently, CRISP-ML has been proposed as a successor to CRISP-DM, intended to fix the aforementioned and other shortcomings (Studer et al., 2021). As one of the core contributions, CRISP-ML adds *Monitoring & Maintenance* as the last phase of the process model, where risks of model degradation in a changing environment are continually assessed, theoretically enabling the deployment of a model in a dynamic environment. The *Monitoring & Maintenance* task has strong connections to the concepts of DevOps and the more recent MLOps, as it touches both, data science as well as IT infrastructure related topics. The problems addressed by MLOps have first been mentioned by Sculley et al. (2015), where special kinds of technical debt are described that are unique to ML projects in contrast to classical software engineering projects. MLOps instruments aim at the standardization and streamlining of machine learning life cycle management (Treveil et al., 2020). Common elements of MLOps cover aspects related to automating, monitoring, testing, managing, and maintaining machine learning models and adjacent code in production through specialized tooling and design patterns (Lakshmanan et al., 2020). In addition to adding *Monitoring & Maintenance*, CRISP-ML merges *Business Understanding* and *Data Understanding* into a single phase, arguing that they are strongly intertwined in practice as business objectives can be derived or changed based on the available data. While CRISP-ML provides a feasible structure for ML projects to integrate post-deployment ML system adaptation activities, it does not provide detailed guidance for sub-processes or system requirements. We draw upon CRISP-ML to structure the results of the expert interviews and the derived design requirements in the phases of the model, with the goal of addressing this distinct lack of guidance.

#### 4.2.2 Sustainable Long-term Deployment of ML Systems

In general, ML can be deployed to solve real-world problems and a set of data is required that represents this problem for training and testing of the ML model (e.g., Russell & Norvig, 2016). Model performance strongly depends on the data quantity and quality (Grover et al., 2018; Smith, 2020). Thus, the data used for ML models must be a good representation of the actual problem at hand (Mitchell, 1997). However, once an ML model is deployed, human perception of a problem may change (Russell & Norvig, 2016) or other factors lead to deviating conditions, risking model performance degradation in the long run (Sculley et al., 2015; Sturm, Gerlach, et al., 2021). Typically, this is caused by a shift or non-stationarity of the data and/or label distribution between the data used to train the model and the data available during inference in practical operation (Žliobaitė, 2010). In industry 4.0 applications, non-stationarity may be caused by various dynamic conditions such as tool and machine wear, changes in product configurations and material properties, changes in upstream processes, changes in factory layout and machine placement, differences in operator preferences and training, seasons, and time of day, environmental conditions such as temperature or humidity, sensor degradation, and data transmission problems (Jourdan et al., 2021). Current IS research on ML applications in organizations also highlights the need for continuous auditing and altering activities to maintain the performance and value of ML applications (e.g., Asatiani et al., 2021; Grønsund & Aanestad, 2020; Sturm, Gerlach, et al., 2021). Multiple studies argue, for example, on the importance of continuously integrating human feedback when ML systems are deployed so that these models remain aligned with humans' dynamically changing understanding of the problem at hand (e.g., Grønsund & Aanestad, 2020; Russell & Norvig, 2016; Stumpf et al., 2009). Such human-in-the-loop patterns allow ML systems to adapt to end users after being deployed in an organization. Closely linking ML systems and end users through collaborative interaction is seen as beneficial for accomplishing a specific task in the moment, as well as increasing the system's accuracy over the long term (e.g., Stumpf et al., 2009). Sturm et al. (2021) refer to activities related to adapting ML systems to the dynamically changing perception of a problem as *reconfiguration* of ML systems. *Reconfiguration* of ML systems may require multiple activities such as adapting criteria for data selection, the learning algorithm itself or settings of hyperparameters (e.g., Amershi et al., 2019; Sculley et al., 2015; Sturm, Gerlach, et al., 2021) and is often described as crucial for continuously ensuring good ML performance (Amershi et al., 2019; Sculley et al., 2015). In addition, an economic life cycle assessment of products, such as ML systems, also requires a positive trade-off of all incurred costs and economic benefits throughout the entire life cycle (Klöpffer, 2003). We argue that ML systems which by their nature consume large amounts of energy and human resources during development, should allow for a long-term deployment to continuously maintain and provide

value. However, in the past ML systems in industry 4.0 have often not progressed beyond the concept development and prototyping stage (Metternich et al., 2021).

IS research recognizes this challenge and calls for methods to continuously adapt ML systems after their deployment (e.g., Audzeyeva & Hudson, 2016; Grønsund & Aanestad, 2020). However, solution approaches only intervene after a developed system has been deployed and are not sufficient to ensure ML systems provide the required adaptability for long-term deployment in dynamic environments such as industry 4.0. Van Wynsberghe (2021) further defines that sustainable AI requires a change in the entire lifecycle of AI products. We address this research call and see the need for a holistic approach that provides clear guidance on how to support long-term development throughout all stages of ML development and deployment. To do so, we aim to identify relevant challenges for the long-term operation of ML systems in the context of industry 4.0 and provide design requirements and principles to enable a sustainable deployment of ML applications.

### 4.3 Research Methodology

#### 4.3.1 Design Science Research Approach

We apply a DSR approach to address challenges for the sustainable long-term deployment of ML applications in industry 4.0. As DSR projects are feasible to find solutions for real-world problems, it allows us to ensure practical relevance as well as scientific rigor (Kuechler & Vaishnavi, 2008). The general design cycle according to Kuechler and Vashnavi (2008) is structured into five distinct phases: *awareness of the problem*, *suggestion*, *development*, *evaluation* and *conclusion*. Our study focuses on the first two phases to identify key challenges, derive design requirements and instantiate a solution by formulating design principles. Within the *awareness of the problem* phase, we combine theoretical input from ML process models with non-theoretical input from experts, thus linking abstract theoretical knowledge with context-specific practical knowledge of industry 4.0 in an interplay (Kuechler & Vaishnavi, 2008; Meth et al., 2015). Based on semi-structured expert interviews, challenges for the development and the sustainable long-term deployment of ML applications in industry 4.0 are identified and general design requirements to address these challenges are derived. In the second phase (*suggestion*), we synthesize the results from the first phase of our DSR approach to develop design principles for solution instantiation. Gregor & Hevner (2013) define three levels of DSR contribution types ranging between specific, limited and less mature knowledge to abstract, complete and mature knowledge. While situated implementations of artifacts are classified as Level 1 contributions, operational design principles and constructs are considered Level 2 contributions, which can be further developed into design theories as Level 3 contributions. We thus provide a Level 2 contribution by transforming abstract

and practical knowledge on long-term deployment of ML applications into operational design principles.

#### 4.3.2 *Semi-structured Expert Interviews*

In this study, we conduct multiple expert interviews to gain practical insights into challenges that arise with the development and long-term deployment of ML models. We further discuss potential solution approaches that are feasible for implementation in continuously evolving industry 4.0 environments. We formulated a semi-structured interview guideline according to Sarker et al. (2013). The semi-structured interview guideline comprises all relevant questions to identify challenges during all phases of the CRISP-ML process model while allowing interviewees to freely share further insights and experiences. Overall, 15 data science and machine learning experts of industry 4.0 start-ups, SMEs, and large companies were interviewed, as shown in Table 2. In addition to the goal of bringing in extensive experience from various companies, experts fill both internal corporate roles as managers and data scientists as well as external consulting roles for ML in industry 4.0, or work as managers of software providers that offer software for ML applications in industry 4.0. All interviewees are involved in ML projects in both the development and deployment phases and therefore allow us to take our holistic approach and highlight the interrelationships of ML system design and long-term deployment in industry. While ML is being deployed in various industries today, we ensured that all interviewees are familiar with the development and deployment of ML models for industry 4.0 applications. Even though most ML applications in this sector have only emerged in the past decade, we were able to interview seven ML experts with more than 15 years of experience. Moreover, 13 interviewees had at least five years of experience in ML development for industry 4.0. All of the participants have already provided expert knowledge to various ML development teams prior to our study. In addition to general experience in ML development in industry 4.0, a majority of the participants (IP1,2,4,5,10,11,13,14) possess expertise in manufacturing quality control in the automotive and aerospace sectors. IP 8 and 9 also bring extensive experience in chemical product manufacturing to the study. IP 3 and 12 primarily gained experience in condition monitoring of machinery plants. After addressing interviewee-related questions to gain insights into their prior experience, current position, and the industries they have been working in, we provided the participants with an overview of the CRISP-DM and CRISP-ML process model. We then inquired whether the presented process models fit the interviewee's current approach of structuring ML projects within their respective companies. Subsequently, in the main part of the expert interviews, we walked through the CRISP-ML phases and specifically asked the interviewees, how these phases are usually implemented in their projects. In each phase, we investigated whether there is already consideration regarding reliability and long-term deployment and which specific challenges arose

in past projects in this regard. Interviews were conducted from February to April 2022 via video call. After mutual agreement, all interviews were recorded and transcribed using the software Amberscript2. On average, an interview lasted 52 minutes, and we were able to reach theoretical saturation during the last three interviews so that no further challenge was mentioned by the remaining interviewees (Flick, 2004).

**Table 2: Overview of interviewee symbols and corresponding industry roles**

<b>Company Internal</b>	<b>External</b>	
Data Scientist: ML Service Provider	Consultant: Operations and Industrial ML	Manager: Industrial ML Software Provider
IP1, IP2	IP7, IP8, IP9, IP10, IP11, IP12, IP13, IP14, IP15	IP3, IP4, IP5, IP6, IP7, IP8, IP9, IP10

To evaluate and categorize the qualitative data gained through the interviews, we performed a content analysis (Myers & Newman, 2007; Weber, 1990). We applied an iterative multi-cycle coding process that comprised two coding cycles following the principles of Saldana (2021). During the first cycle, two types of coding were used. First, attribute coding allowed us to obtain relevant, descriptive information regarding the participants and their respective organizations. Descriptive coding was used next to identify challenges and solution approaches in the participants' statements. The second cycle covered pattern coding, which allowed us to build clusters of similar challenges but especially solution approaches for which we defined a large number of factors throughout the first cycle. In addition, a multi-researcher triangulation was applied. Two authors and one research assistant performed both coding cycles independently and derived an initial framework of identified challenges and the CRISP-ML process model (Saldana, 2021). Before we merged the coding results of the researchers after each interview, the number of matching codes was derived which resulted in an average value of 74.2 %. Afterwards, all results were compared and discussed, and the final framework was derived based on this data analysis which further ensured rigor and trustworthiness.

#### **4.4 Results**

Within the following section, the challenges that were identified during the expert interviews are presented, combined with insights on possible solution approaches that practitioners use to address them. We use the aforementioned categories based on the CRISP-ML process model for categorization. The results are summarized in Table 3. A number of challenges may occur in several phases of the process model. In the following, we assign these challenges to the phases where they were most often mentioned by the interviewed experts.

#### 4.4.1 Challenges for Long-term Operation of ML Systems in Industry 4.0

The majority of interviewees saw the CRISP-ML process model as relatively close to the way they internally structure machine learning projects, even though, at the same time, most of the interviewees have only heard about the predecessor (CRISP-DM) prior to the interview. In addition, a number of interviewees (IP1,4,8,12,13,14,15) mentioned that CRISP-ML, compared to CRISP-DM, fits real projects better, as the additional *Monitoring and Maintenance* phase is crucial for practical deployment. However, interviewees also argued that the practical implementation of this phase is still rare, due to lacking standard solutions and the general novelty character of ML applications in the industry 4.0 sector. For example, an interviewee explained, *“Our ML development pretty much follows the CRISP-DM process model. But if I’m honest, we’re primarily concerned with delivering models, and processes become – let’s say - less standardized after deployment. So we really need something like a monitoring and maintenance phase, but right now I would say we rather intervene in emergencies only”* (IP9).

**Business and Data Understanding.** A frequently mentioned challenge (IP1,4,5,9,10,11), is the expectation management of the shareholders, especially of customers who are not familiar with ML and data science. This challenge may manifest in two very distinct ways, with customers either having unrealistic expectations about the possible performance achievements of ML models (IP5,12,13,15), especially given the available data basis in their use case, or customers being extremely skeptical regarding the use of ML models, not trusting their performance. In the latter case, the usage of either simple or otherwise explainable ML algorithms is seen as a possible solution approach. The alignment of sales and engineering departments is named as another problem in this context as it is hard but critical to align business goals with the uncertainty of ML development, especially regarding specific levels of application performance (IP12). Specifying requirements with respect to robustness is also named as a challenge by the interviewees. A set of operating conditions the application needs to handle must be identified as well as different operating modes of the machine or production line (IP2,6,11). In this context, the integration of domain knowledge is seen as crucial. An interviewee stated, *“We still face this problem that our customers ask us to develop ML systems that should solve all their problems at once, but at the same time it is quite hard for them to define under which conditions this system will be operated in one, two or five years”* (IP11). A fundamental and often encountered problem during *Business and Data Understanding* is missing meta-data in the form of time-series annotations related to, e.g., maintenance activities or machine breakdowns (IP1,3,4,5,6,8,12,13). This data is crucial for data analysis and for the training of supervised ML models. Thus, an alternative approach is the usage of unsupervised ML algorithms, which is not suitable in all use case scenarios though. Further problems with meta-data include privacy issues if the data can be connected to the performance of specific employees (IP12). Poor data quality is another issue (IP11,12,13). At the same time,

interviewees describe the achievable solution quality as mostly dependent on data quality, quantity, and variety (IP4,5,12). Interviewees note that it is hard to estimate the quantity and variety of data at the beginning of the project (IP1,4,6,11,13) and that a first analysis is usually done with smaller datasets. Analysis frequently reveals that more data is needed for a given problem than available (IP1,4,5,6,11,12,13,15), often leading to the cancellation of a project. Interesting events such as machine breakdowns or product defects are usually rare in a dataset (IP5,6,10,15). Furthermore, time dependencies such as seasonality are described as problematic as these usually require an increasing amount of data to be modeled but are a key aspect for long-term operation (IP5,6,11,12,13). One interviewee explained, *“Whatever challenge we face, in the end it’s always about data. Before development, we often don’t know what we need or we do not have the data needed and in the end data drifts and other changes kind of force us to reconsider if our trained models are still useful”* (IP5).

**Data Preparation.** While the *Data Preparation* phase itself is a frequently mentioned general challenge as it takes a significant portion of the total project time, it was not specifically mentioned as impeding the sustainability or long-term operation of the ML models by the interviewees.

**Modeling.** Interviewees see the tradeoff between simple and complex ML models as a particular challenge during the modeling phase. While complex models such as deep neural networks may provide high capacities to model complicated functions, they are often more prone to overfitting, which is detrimental to their robustness (IP4,11,15). It was further mentioned that those simpler models should be preferred if it has no impact on performance or lower performance is acceptable for the given use case as they commonly provide better explainability (IP5,6,8,10) and *“customer acceptance is much better for simple and interpretable models”* (IP6). Such a model selection strategy is referred to as Occam’s razor which was explicitly mentioned by two experts (IP11,15). Interviewees further mentioned that model-agnostic techniques for explainable AI may be used to provide explainability of complex models. However, they increase overall system complexity and the results are often hard to interpret in practice (IP8,10). Besides that, one interviewee mentioned that they almost *“exclusively use unsupervised approaches in manufacturing projects, because labeled data is rarely available in sufficient quality and quantity”*, and target variables such as fault types may change over time (IP4). A frequently mentioned general challenge related to modeling (IP2,4,5,11,13), is the scalability of models to different instances of machines or production lines. It is indicated, that even if a customer utilizes multiple instances of the exact same machine or production line, a model that was trained on data of either one of the instances or a held-out instance on a test bench, does not perform well if deployed to other instances, as the data distribution already differs too strongly. Possible reasons are sensor placement, differences in manufacturing processes as well as human operator preferences. This problem is described as especially severe in predictive maintenance applications which are strongly dependent on the

machine specifics (IP5,9,13,15). Possible strategies for mitigation of this problem include transfer-learning or fine-tuning of a trained model using smaller datasets of the target machine or production line (IP9,13).

**Evaluation.** While most interviewees mention that evaluation of the ML models is commonly performed using an offline held-out test set it is often noted that this kind of evaluation is not sufficient for real-world applications. Interviewees suggested that the model's robustness needs to be explicitly tested, by focusing on edge cases in the input data, which can be found using statistical analysis of the datasets (IP2,5,6,11). Test cases might then be generated using Monte-Carlo simulation or design of experiments (IP2,6). In addition to this offline evaluation, an interviewee mentioned *"For us, it is crucial to evaluate the model before and after deployment and we were surprised how many problems can only be revealed after deployment. You simply cannot detect all these issues with a static test dataset beforehand"* (IP4). Corresponding issues include data quality problems and non-stationarity of the data that was not noticed during development (IP1,13). If possible, a pre-existing application for the use case or frequent manual checks should be run in parallel in the test phase for output validation (IP9,10,14,15).

**Deployment.** During deployment, primarily organizational and IT infrastructure-related challenges were mentioned by the interviewees. In the scope of this study, they are of special interest, as these are typically issues that do not arise in academic work but are crucial for practitioners. Especially challenging is the interfacing of the application to various subsystems or data acquisition devices (IP5,6,11,13). An interviewee explained *"You know, we follow all these theoretical guidelines, but once you get close to deploy such ML systems, it is always a mess. Hundreds of different systems must interface with our application. It's exhausting and I don't see our customers taking on this task and better yet doing it repeatedly on their own with their current setup"* (IP11). A possible solution approach can be the usage of standardized interfaces such as Open Platform Communications - Unified Architecture (OPC-UA). Furthermore, it is critical to automate and document the deployment process in such a way that it is repeatable in the monitoring and maintenance phase (IP2,5), building on the principles of DevOps and MLOps. This issue was emphasized in the case of deployment on edge devices.

**Monitoring and Maintenance.** As *Monitoring and Maintenance* is a key phase of the CRISP-ML model with regard to sustainable deployment, the majority of mentioned challenges are of this category. Most interviewees agreed that this phase is paramount for long-term deployment, as *"ML models only provide value in situations that are shown in the dataset"* (IP2). In this context, an interviewee referred to deployed ML models as *"only ever meta-stable"* (IP12). When something changes or the data distribution is non-stationary, the performance of the ML model will most likely strongly degrade and the model requires updating (IP1,4,5,6,9,10,11,12,13,14,15). Four, often experienced sources of change were mentioned in the interviews: Sensor drift (IP1,4,5,6),



seasonalities or time dependencies not identified during development (IP5,11,15), changes in the configuration of the machine or the product (IP1,3,12,13) and network or hardware problems that render their respective data sources invalid or non-available (IP6,12). This yields two additional challenges: The detection of a change, as well as the adjustment of the processing pipeline, i.e., updating of the model, because drifts and changes impact sensor data which is used as an input for the ML model in operation.

A number of interviewees mentioned that it is primarily important to monitor the properties of the input data (IP2,6,8). Statistical measures can be used to capture properties of the training data that are then compared to the live data during operation. Suitable measures include distribution distance metrics or simple thresholds. The complexity of the monitoring task rises with the number of data sources. Automatic checks within the processing pipeline are often used to detect the described changes. In addition to the input data, it was mentioned that the model error rate should be monitored too, as monitoring of the input data is only a proxy to this quantity (IP1,2,11,14). An interviewee mentioned *“Even though it is still a challenging task, we are already able to partially automate some monitoring activities. But if we want to update a model, this is a manual and often tedious task and we really need to improve this”* (IP2). In addition to issues related to data science or machine learning, interviewees describe several challenges that relate to IT infrastructure in this phase. Model updates can be a challenging task since it requires access to datasets, training infrastructure, and the deployed model which often runs on an edge device. While cloud computing solutions such as Microsoft Azure or Amazon Web Services (AWS) offer standardized pipelines for data collection, management, model training, and cloud deployment, edge deployment is still an issue. In this context, data security and privacy are seen as additional challenges (IP5,6,11,12,13) that prevent customers from relying on cloud solutions. If automatic deployment is not an option, data scientists require physical access to the computing devices. Again, the interviewees indicated that a large variety of data sources to a model and thus connected subsystems increase complexity. Extensive logging during operation is therefore important to quickly analyze application errors and find their root causes (IP1,6). Although challenging, Monitoring and Maintenance of ML models is seen as a viable addition to the business model of ML solution providers as mentioned both by manufacturing company internal and external interviewees (IP1,6,8,10,11,12,13,14,15) as it *“provides a constant revenue stream whereas, you know, for this prior development process, we usually agree on project-based fixed-term work and payment”* (IP15). At the same time, this may provide another challenge, as *Monitoring and Maintenance* activities are often not covered by the initial development contracts. Multiple interviewees stated that robustness and sustainable, long-term use of ML applications have only recently become a focus of their work, as the productive usage of ML applications has only slowly become a reality in the last years and the number of deployed models is still relatively small

(IP1,12,14,15). An associated challenge that was mentioned by the interviewees is the lack of a standard / best-practice solution for *Monitoring and Maintenance* of ML applications in industrial environments (IP1,2,6,12,14). A manager of a large ML software provider stated *“Never change a running system – that’s something I hear quite often from our customers in this context and it’s really slowing us down. I’m convinced that we will provide services to maintain ML systems in the future, but to do that we need to reduce the complexity of these systems. [...] So I advocate for future projects to keep ML system maintenance in mind from the beginning and consider that in, yeah, pretty much every future system design”* (IP13).

#### 4.4.2 Deriving Design Requirements and Principles for a Solution Instantiation

Following the identification of challenges through expert interviews, we derive solution approaches as design requirements and structure them along the CRISP-ML process model. Overall, we pose nine requirements that address challenges mentioned by interviewees as outlined in Table 3. First, uncertainty that is often associated with ML development projects, raises concerns of customers and customer expectation management remains a challenge today. To align expectations, address data privacy concerns and integrate domain knowledge in the very beginning of ML development, we propose **DR1: Collaboratively align system requirements with stakeholders**. Second, data preparation remains a tedious process and especially labeling of data requires significant resources which would also be required for future retraining of algorithms as a reconfiguration activity for long-term use of ML systems. Algorithms from the field of unsupervised learning that do not require labeled data for training were mentioned as feasible to ease this process. We therefore propose **DR2: Use unsupervised ML algorithms, if possible**. Third, variances in operating conditions and other influencing factors should be considered from the outset to cope with future data shifts or time dependencies. To ensure robustness of models in the long term and ensure necessary data variety, we propose **DR3: Design for a comprehensive range of operating conditions and their boundaries**. Fourth, while performance of ML systems was not mentioned as a major challenge by any of the interviewees, explainability of models and limitations in capacity were often seen as a challenge for developers that hinders maintenance of models in the future and further risks customer acceptance. A trade-off is seen between simple, explainable models and complex models with potentially higher performance. We thus recommend to follow Occam’s razor during model selection and resort to simpler models if the use case allows for it, while resorting to explicit techniques for explainable AI if complex models are required. This yields **DR4: Use simple models and explainable AI-techniques**.

**Table 3: Challenges and corresponding solution approaches identified as design requirements**

CRISP-ML Phase	Challenges	Design Requirements
<i>Business and Data Understanding</i>	Customer expectation management	<b>DR1</b> Collaboratively align system requirements with stakeholders
	Customer acceptance	
	Domain knowledge integration	<b>DR2</b> Use unsupervised ML algorithms, if possible
	Data privacy and security	
	Limited quantity and quality of data	<b>DR3</b> Design for a comprehensive range of operating conditions and their boundaries
	Limited availability of meta-data or labels	
	Requirements definition regarding robustness and data variety	
Time dependencies and seasonalities in data		
<i>Modeling</i>	Trade-off between simple and complex models: Capacity, explainability, computational requirements	<b>DR4</b> Use simple models and explainable AI-techniques
	Scalability of models to different instances of machines or production lines	<b>DR5</b> Use transfer learning techniques such as domain adaptation
<i>Evaluation</i>	Offline evaluation not sufficient for real-world applications	<b>DR6</b> Continuously and frequently evaluate model performance
	Unnoticed non-stationarities in the data	
<i>Deployment</i>	Interfacing and integration to various IT-systems such as MES and ERP	<b>DR7</b> Maximize use of standardized interfaces
	Retrofit of sensors complicated due to required line (re-)certification / qualification	
	Time consuming deployment processes	<b>DR8</b> Maximize automation of testing and deployment, e.g. MLOps and DevOps
	Automated deployment to edge devices	
<i>Monitoring and Maintenance</i>	Data drift due to seasonality, time dependencies, configuration changes, network and hardware problems	<b>DR9</b> Monitor model confidence and input data using statistical measures
	Change detection	
	Model updates	
	Missing standardized solutions, lacking automation of tasks	

Fifth, to allow for scalability of ML systems to e.g. other machines or plants in industry 4.0 environments or cope with varying operating conditions, ML models require adaption in the future. We therefore propose **DR5**: *Use transfer learning techniques such as domain adaptation*. Sixth, non-stationarities in data often cause offline model performance evaluations to deviate from evaluations conducted after deployment. To detect declining model performance in time during operation before models are no longer usable, we propose **DR6**: *Continuously and frequently evaluate model performance*. The deployment of ML systems often involves a difficult and time consuming process of integration and interfacing with IT systems to access the required data streams. To ensure continuous development of ML systems and maintain their value over time, models will be deployed regularly during the Monitoring & Maintenance phase of CRISP-ML and this process should be as simple and automated as possible. MLOps and DevOps were

mentioned as promising solution approaches to closely link machine learning related continuous development activities and IT operations of organizations in industry 4.0. We thus propose **DR7: Maximize use of standardized interfaces**, and **DR8: Maximize automation of testing and deployment**, e.g. through MLOps and DevOps techniques. Lastly, transfer learning techniques can be used to adapt ML systems after deployment (see **DR5** for the *Modeling* phase) to cope with data drifts, time dependencies, configuration changes or any other aberrant conditions. However, organizations in industry 4.0 that want to leverage ML systems over longer periods of time need to continuously monitor input data and model performance to decide when system updates are needed. Therefore, we propose **DR9: Monitor model confidence and input data using statistical measures**.

Following the derivation of general design requirements, we formulated overarching action-oriented design principles as outlined in Figure 8. DPs are assigned to specific DRs which were derived from the identified challenges. However, as suggested by Hevner et al. (2004), design principles can serve on the one hand as an actionable blueprint for e.g. a prototypical implementation of ML artefacts, but on the other hand can also be used as testable hypotheses for future research work. In addition, DPs support an holistic approach to ML system development and deployment that closely links all phases of the CRISP-ML process model. To address the core problem of ML applications being deployed in dynamic environments such as industry 4.0, where problem perception and other operating conditions may continuously change, ML applications require flexibility to be deployed in and adapted to varying conditions. We therefore propose **DP1: Continuously provide ML applications with sufficient training and testing data that covers all relevant operating conditions and evaluative capabilities to verify both databases are aligned with real-world dynamics**. To detect deviating conditions which could lead to a declining performance of ML models over time, systems require self-monitoring capabilities and human-machine interfaces that allow system operators to understand emerging discrepancies. To address these challenges, we propose **DP2: Provide ML applications with continuous mechanism that detects data drifts or changes in the certainty of the model and capabilities to transparently present these dynamics to the responsible system user**. Besides the adaption to dynamic environmental conditions, ML systems that required large resources for development, should allow for scalability to maximize the provided value over time. Therefore, by building on techniques from the domain adaptation area, we propose **DP3: Provide ML applications with capabilities that allow for scaling their area of operation and adapt them to dynamic environmental changes**. In the past, much attention was paid to the development of complex algorithms to maximize ML model performance. However, our interview study showed that model performance was of less interest and the explainability of models has emerged as a major challenge today, on the one hand for customer acceptance but also in the long run for the maintenance of ML systems. Oftentimes,

maintenance or reconfiguration activities (e.g. retraining of ML models) will be performed by people other than the original system developers. We therefore propose **DP4**: *Provide ML applications with capabilities to meet defined performance while resorting to simple models to ensure that decisions are presented in a way that is understandable to the responsible system user.* Although monitoring and maintenance of ML systems has only recently become a focal point in many industry 4.0 projects, related activities such as retraining and deployment of updated models should be performed frequently in the future. To facilitate this process, automation and user support are of key interest. We therefore derive **DP5**: *Provide ML applications with mechanisms for automatic testing and deployment and provide features that, if necessary, alert the responsible system user.* Interviewees frequently mentioned the tedious process of integrating ML systems into the existing IT landscape of industry 4.0 environments. Scalability as well as the continuous integration of newly acquired data streams should become a key capability of ML systems to allow for a more sustainable long-term deployment which further supports customer acceptance. Thus, ML projects should strive for standardized interfaces to ease this process and we propose **DP6**: *Build ML applications upon standardized interfaces for maximum compatibility and use simple ML models to facilitate the definition and integration of evolving data requirements.*

→	Design Principles	
DR3 DR5 DR6	<b>DP1</b>	Continuously provide ML applications with sufficient training and testing data that covers all relevant operating conditions and evaluative capabilities to verify both databases are aligned with real-world dynamics.
DR6 DR9	<b>DP2</b>	Provide ML applications with a continuous mechanism that detects data drifts or changes in the certainty of the model and capabilities to transparently present these dynamics to the responsible system user.
DR3 DR5	<b>DP3</b>	Provide ML applications with capabilities that allow for scaling their area of operation and adapt them to dynamic environmental changes.
DR1 DR2 DR4	<b>DP4</b>	Provide ML applications with capabilities to meet defined performance while resorting to simple models to ensure that decisions are presented in a way that is understandable to the responsible system user.
DR2 DR8	<b>DP5</b>	Provide ML applications with mechanisms for automatic testing and deployment and provide features that, if necessary, alert the responsible system user.
DR1 DR4 DR7	<b>DP6</b>	Build ML applications upon standardized interfaces for maximum compatibility and use simple ML models to facilitate the definition and integration of evolving data requirements.

**Figure 8: Design principles for ML applications suitable for long-term deployment in dynamic environments**

## 4.5 Discussion and Conclusion

Due to the increasing availability of data in the industry 4.0 sector, ML applications slowly transcend from academia to practical applications. However, the level of adoption and actual reliance on these applications is still low and questions about reliability, sustainable long-term use and robustness remain largely unanswered. Research on ML applications in organizations highlight the need for continuous auditing and altering activities to enable ML systems to provide long-term value (e.g., Asatiani et al., 2021; Grønsund & Aanestad, 2020; Sturm, Gerlach, et al., 2021). We follow a DSR approach to identify which challenges impede practitioners in the sustainable long-term deployment of ML in the context of real-world industry 4.0 applications with a focus on monitoring and maintenance of ML systems and thus address **RQ1**. The identified challenges cover most phases in the lifecycle of an ML model, starting with the initial project specifications and ending with the continuous monitoring and maintenance. Semi-structured expert interviews were performed and besides challenges, corresponding solution approaches were inductively derived and transformed into tangible design requirements for ML system to support long-term operation. In addition, action-oriented design principles were formulated that link all phases of development and deployment to support a holistic approach, thus addressing **RQ2**. IS researchers and ML system developers can use DRs and DPs derived in this study for guidance in how to design ML systems that allow for a more sustainable long-term use and maintain their value over time. As suggested by (Kuechler & Vaishnavi, 2008), we will continue our DSR approach by instantiating an ML system as a suggested solution within an industry 4.0 environment based on our DPs. Consequently, we will evaluate the extent to which the consideration of our DPs supports the adaption of our ML system to dynamic environmental changes by simulating long-term deployment through various data drifts. Feedback will be incorporated into several design cycles.

The presented research yields several **contributions** for academia as well as practitioners. First, we provide a structured overview of the challenges currently hindering long-term deployment of ML systems in dynamic environments such as industry 4.0. On the one hand, it is crucial to understand and contextualize these challenges in order to align research work more closely with the needs of practitioners. On the other hand, this deeper understanding enables us to redesign the way we develop ML applications today to ensure sustainable long-term use. Second, we extend research on sustainable ML which primarily focuses on sustainable use cases for ML to sustainable ML product development that focuses on maintaining the value of ML applications over time. Broad applicability of the provided challenge-design-requirement framework in the industry 4.0 sector creates a common starting point for targeted design of sustainable IS artifacts in the future. Finally, our study addresses the call for future work in IS research on adapting ML systems to dynamic changes in problem perception (e.g., Grønsund & Aanestad, 2020). We extend that call

for research by showing the multitude of potential challenges in dynamic environments such as industry 4.0 and provide a holistic approach that considers all phases of ML system development and deployment. Furthermore, the derived DRs and DPs provide a basis for future research in the field of sustainable long-term deployment of ML as well as clear guidance for researchers and practitioners to integrate sustainability into the core of every ML system and to create acceptance for resource-intensive technology development that enables the efficiency of industrial processes to keep pace with the rapid developments in the interconnected world and to secure competitive advantages in the long term.

At the same time, this study faces **limitations** that need to be considered when using the results. First, the number of interviewees is limited, which may bias the results towards certain subsectors of industry 4.0. However, due to the vast experience, diversity and internationality of the interviewees, we are confident in the validity of the results of the study. Second, due to the qualitative nature of this study, the severity of the challenges cannot be easily quantified and thus the challenges cannot be prioritized. Lastly, the derived DRs and DPs have not yet been evaluated. This study yields ample opportunities for future work, both regarding the limitations of the study itself as well as in addressing the identified challenges and derived DRs and DPs. While we plan to implement and test the DRs and DPs in a prototype ML application as part of a follow-up study, we also invite **future research** to evaluate our design guidelines in a wide variety of dynamic environments. Moreover, it is to be expected that challenges will differ to a certain degree, depending on the specific industry that is analyzed, which should be examined in separate studies to refine DRs and DPs. While derived DRs and DPs can support and facilitate the process of maintaining an ML application over time, organizations still need to provide an infrastructure for adapting ML models to, for example, data drifts, and train their workforce in MLOps and DevOps techniques, as well as in new areas of research such as domain adaptation.

Interestingly, lacking performance of existing ML algorithms was never mentioned as a challenge during the interviews, thereby highlighting the importance of holistic research that focuses on integration and long-term deployment of ML in contrast to focusing on optimizing benchmark dataset scores or developing more complex algorithms.

#### 4.6 Acknowledgements

The research leading to these results has received funding from the European Commission's H2020 program under grant agreement number 958357 (InterQ) and the German Federal Ministry of Labor and Social Affairs and the European Union through the European Social Fund Plus (ESF Plus) as part of their joint support for the "Future Centers" program.

## **5. Paper C: Pilots and Pixels: A Comparative Analysis of Machine Learning Error Effects on Aviation Decision Making**

### **Title**

Pilots and Pixels: A Comparative Analysis of Machine Learning Error Effects on Aviation Decision Making

### **Authors**

Ellenrieder, Sara; Ellenrieder, Nils; Hendriks, Patrick; and Mehler, Maren F.

### **Publication Outlet**

Proceedings of the 32<sup>nd</sup> European Conference on Information Systems (ECIS), Paphos, Cyprus, 2024

Awarded with the Claudio Ciborra First Runner Up Award

### **Abstract**

Despite immense improvements in machine learning (ML)-based decision support systems (DSSs), these systems are still prone to errors. For use in high-risk environments such as aviation it is critical, to find out what costs the different types of ML error cause for decision makers. Thus, we provide pilots holding a valid flight license with explainable and non-explainable ML-based DSSs that output different types of ML errors while supporting the visual detection of other aircraft in the vicinity in 222 recorded scenes of flight simulations. The study reveals that both false positives (FPs) and false negatives (FNs) detrimentally affect pilot trust and performance, with a more pronounced effect observed for FNs. While explainable ML output design mitigates some negative effects, it significantly increases the mental workload for pilots when dealing with FPs. These findings inform the development of ML-based DSSs aligned with Error Management Theory to enhance applications in high-stakes environments.

### **Keywords**

Machine learning error, explainable artificial intelligence, human-AI interaction, aviation decision making



## 5.1 Introduction

Advances in machine learning (ML) have driven the development of increasingly sophisticated ML-based decision support systems (DSS) that enable human decision makers to gain valuable insights in complex situations, ultimately improving their ability to make more informed and data-driven decisions (e.g., Berente et al., 2021; Jussupow et al., 2021; Sturm et al., 2023). As ML-based DSSs have demonstrated remarkable capabilities, at times surpassing human experts in specific tasks (e.g., Shen et al., 2019), these systems are now increasingly being adopted in high-stakes environments (Lebovitz et al., 2022; Maedche et al., 2019; Sutton et al., 2020). While the potential of ML-based DSSs in healthcare has been explored for some time (Gaube et al., 2021; Jha & Topol, 2016; Jussupow et al., 2021), their potential is also gaining recognition in aviation, with the European Union Aviation Safety Agency (EASA) taking a proactive step by releasing an artificial intelligence (AI) roadmap in May 2023, outlining the goal of securing official approval for ML-based systems to assist human decision makers in aviation operations by 2025 (European Union Aviation Safety Agency, 2023).

Nevertheless, there is a concern that decision makers may hesitate to adopt approved ML-based DSSs and utilize the provided advice (e.g., Sturm et al., 2023) or that their use may introduce additional risks (Fügener et al., 2021; Riedl, 2019). For example, the increasing complexity of ML algorithms poses a significant challenge as it often renders the output of these algorithms incomprehensible for humans (e.g., Arrieta et al., 2020; Berente et al., 2021; Castelvechi, 2016). In addition, even high-performing ML-based systems are prone to errors (Russell & Norvig, 2021), which presents a substantial concern for practical applications. Decision makers, such as pilots, may fail to recognize limitations of the model and consequently erroneous system output (Jussupow et al., 2021) or their trust in ML-based decision support systems erodes (Gurney et al., 2022), potentially causing them to exclude these systems from their decision making processes. Overall, there is a need to focus on efficient error management to ensure that these systems can positively contribute to safety-critical decisions in aviation and other high-stake environments (cf. Green & Swets, 1966). Emerging approaches from the field of explainable AI (XAI), for instance, aim to make the system's output more understandable for human decision makers (Arrieta et al., 2020; Reyes et al., 2020), with studies showing that this aids in recognizing erroneous output (Abdel-Karim et al., 2020).

However, existing research primarily addresses the general impact of incorrect support without further differentiating error types and predominantly examines the effects on decision makers' accuracy and performance (Abdel-Karim et al., 2020; Gaube et al., 2023). Nevertheless, the differentiated examination of the costs of different ML error types on the decision maker is of crucial importance, since ML developers must determine the sensitivity of the systems and with

this design choice have a significant influence on which ML error type will occur more frequently and which less frequently (Wenkel et al., 2021). Existing research cannot yet provide ML developers with efficient error management guidance for the different types of ML errors in detection models, which are widely used for decision support in high-stake domains (e.g., Jussupow et al., 2021; Lebovitz et al., 2021). To provide guidance, a more in-depth examination of the effects of different types of errors and their associated costs on the decision maker is required. For example, there is a significant difference between an ML-based DSS in aviation falsely detecting an aircraft in vicinity when the airspace is actually clear—a false positive (FP) error—versus failing to detect an existing aircraft in vicinity and not alerting the pilot—a false negative (FN) error. Both types of errors can impact pilots' performance and trust. Moreover, explanations from the XAI field can assist in error detection (Abdel-Karim et al., 2023, 2020; Gaube et al., 2023; Pumplun et al., 2023; Sturm, Gerlach, et al., 2021) but may also increase cognitive resources required for information processing (Pumplun et al., 2023), potentially affecting mental workload. In addition, XAI approaches could have different effects for different types of errors. There is a lack of knowledge about the costs incurred on the decision maker for different types of errors and whether and how XAI approaches mitigate negative effects. In addition, this knowledge is crucial for ML developers who want to adhere to the Error Management Theory (EMT), which states that systems should be biased in a way that minimizes the type of error associated with the greatest long-term costs (Green & Swets, 1966; Haselton & Nettle, 2006; Johnson et al., 2013). To enable the design of ML-based DSSs in line with EMT in the future, our study integrates theories and approaches from error management, human-AI interaction, and XAI to address the following research questions:

*In the context of General Aviation, (1) how do FP and FN errors caused by ML-based DSSs affect the performance, trust and perceived mental workload of human decision makers, and (2) can the explainable design of ML-based DSSs have a positive effect on the costs incurred?*

We investigate these research questions through an online experiment with pilots holding valid flight licenses. To conduct the experiment, we initially recorded various flight scenes in a crowded airspace using the Microsoft Flight Simulator and developed six different variants of ML-based decision support to support pilots in visually detecting other aircraft in the vicinity. These ML-based DSSs challenge decision makers with different error types and levels of explainability. Overall, the following ML-based DSSs were developed: a non-explainable false negative (NEFN), an explainable false negative (EFN), a non-explainable error-free (NEEF), an explainable error-free (EEF), a non-explainable false positive (NEFP), and an explainable false positive (EFP) ML-based DSS. We collect quantitative data on the pilots' performance, mental workload, and system trust, analyzing differences between the groups. This research offers insights into the costs associated with different types of ML errors and provides a foundation for developing future ML-

based DSSs in line with EMT. By conducting this experiment with real-world pilot participants, we aim to contribute valuable knowledge to enhance the deployment of ML-based DSSs in aviation and other high-stake domains.

## 5.2 Theoretical Background

The following section offers a conceptual overview of ML and its application in decision support. It further outlines error types, along with insights from the XAI domain to render ML results comprehensible to users. Lastly, the section examines how ML systems affect user performance, trust, and mental workload before hypotheses for human-AI interaction with erroneous ML-based DSSs are derived.

### 5.2.1 Machine Learning

The fundamental principle of ML involves the use of algorithms that can independently find patterns in data, allowing them to solve problems without the need for providing explicit solutions or decision rules (Brynjolfsson & Mitchell, 2017; Russell & Norvig, 2021). Once trained, these algorithms can recognize patterns, classify data, make predictions, or take actions based on the data they have been exposed to (Mitchell, 1997; Russell & Norvig, 2021). Since ML algorithms derive solutions on their own without using human-coded instructions (Samuel, 1959), they can generate insights that are complementary to human knowledge (Fügener et al., 2022). While this provides opportunities to boost decision makers' performance (e.g., Abdel-Karim et al., 2020; Gaube et al., 2023), challenges arise with the use of ML-based DSSs. ML models are based on statistical patterns and are susceptible to errors. This is particularly problematic given the often opaque nature of these models that hinders humans to understand why and how decisions are made (Berente et al., 2021; Rudin, 2019). As a result, the lack of interpretability and the potential for unpredictable outputs are major obstacles preventing the widespread use of ML-based systems in high-risk areas such as healthcare or aviation (Lebovitz et al., 2021). Therefore, instead of automating decisions, ML-based systems should primarily be used as decision support systems that assist human decision makers (European Union Aviation Safety Agency, 2023; Jha & Topol, 2016). Nevertheless, there is a risk in collaborative decision making that decision makers such as pilots may fail to detect system errors, and their performance may deteriorate (Fügener et al., 2021; Jussupow et al., 2021), or they may not consider system advice (Sturm et al., 2023).

### 5.2.2 ML Error Types and Error Management Theory

Detection models, which are commonly deployed in high-risk environments (Lebovitz et al., 2021), have the ability to categorize data into classes. In medical decision making such an ML-based DSS could, for example, advise a radiologist whether the tissue on a scan is tumorous or

healthy (Pumplun et al., 2023). In the context of decision making in aviation, ML models could use image data to detect if other aircraft are in close proximity. Although ML-based DSSs have improved significantly in recent years (Shen et al., 2019), they are still error-prone (Mitchell, 1997; Russell & Norvig, 2021). In the case of detection models, two types of errors can occur here, namely FPs and FNs (Goutte & Gaussier, 2005; Padilla et al., 2020). An FP occurs when the system classifies data as belonging to the positive class when it should have actually been classified as belonging to the negative class (Padilla et al., 2020), e.g., falsely detecting an aircraft in an empty airspace. An FN occurs when the system incorrectly classifies data as belonging to the negative class (Swets et al., 2000), e.g., not detecting an aircraft in the surrounding airspace. The given example shows that the implications of these error types are significant, especially in high-risk environments (Luce & Kahn, 1999). While improving accuracy of ML-based DSSs has remained in the focus in recent years (e.g., Roy et al., 2022), research has also shown that it is still critical to understand the implications of ML errors for efficient error management (Johnson et al., 2013). Benefits and costs of correct and incorrect decisions need to be understood in order to judge which type of error is preferable to the other (Swets et al., 2000).

ML developers can prioritize reducing the rate of one error type over the other through adjusting the confidence threshold of a model and thus changing its sensitivity (Padilla et al., 2020). The confidence threshold is a critical value that sets the minimum confidence level required for a model's detection to be deemed valid (Wenkel et al., 2021). To increase sensitivity of the ML model, the confidence threshold is lowered and FNs are reduced. However, increasing sensitivity can increase the rate of FPs. On the other hand, decreasing sensitivity and thus raising the confidence threshold leads to a higher occurrence of FNs but might reduce the rate of FPs. From a technical perspective, it is thus easy to influence which of the different types of errors occurs more frequently when deploying an ML-based DSS (Padilla et al., 2020).

From the error management perspective, on the other hand, complex questions about the costs of the different error types need to be answered before the sensitivity of such systems should be adjusted (Swets et al., 2000). Once it is accepted that errors can occur, it is necessary to understand why people and especially responsible decision makers react differently to different types of errors (Haselton & Nettle, 2006). In this context, Haselton and Nettle (2006) introduce the Error Management Theory (EMT) for judgments under uncertainty, which states that if the costs of false negative and false positive errors are asymmetric over evolutionary time, "humanly engineered systems" should be biased to make the less costly errors. While introducing this bias may result in a higher error rate overall, it minimizes the more costly error and therefore minimizes the overall cost as well (Green & Swets, 1966; Johnson et al., 2013). In addition, the introduction of this bias into systems follows human psychology, which states that adaptations in decision making have evolved through natural selection to make predictable mistakes (Haselton

& Nettle, 2006). Therefore, it is crucial to fully understand the costs of error in real-world scenarios (Johnson et al., 2013). Often, in high-risk environments, it is easy to assess that FNs (e.g., overlooking an aircraft in nearby airspace) pose the greater risk. Nevertheless, it is a great concern to understand the psychological costs of FPs as well (Luce & Kahn, 1999) in order to fully assess the costs and implement a meaningful system bias in terms of EMT (Haselton & Nettle, 2006; Wardle & Pope, 1992) for ML-based DSSs. For example, an overly sensitive ML system that detects an infinite number of non-existent aircraft in the airspace will not add value to a pilot in the long run and thus comes at a large cost as well. IS research, which recently explored the impact of incorrect ML-based decision support on the decision maker, does not differentiate between the different types of errors and thus cannot improve our understanding of these different error costs or solely focus on the impact of incorrect ML-based DSSs on human performance (e.g., Jussupow et al., 2021; Pumplun et al., 2023).

### 5.2.3 *Explainable Design of ML Systems*

While ML models are becoming more and more complex, the field of XAI has emerged to address arising challenges. The primary aim of XAI is to render ML system outputs and interpretations understandable for humans (Arrieta et al., 2020; Miller, 2019). This can aid data scientists or developers in error detection and improvement of ML model performance, be necessary for legal or regulatory compliance, or enhance the interaction between end users and ML systems (Bhatt et al., 2020; Pumplun et al., 2023; Reyes et al., 2020). To do so, explanations are provided that inform about the inner workings of ML systems, shedding light on factors such as the considered features and their respective impact on the ML decision making process (Arrieta et al., 2020; Bhatt et al., 2020). Presently, various technical approaches to XAI exist and three explanation categories can be defined (Pumplun et al., 2023): First, model explanations offer meta-information about the ML development process and the ML model itself (Cai et al., 2019). Second, global explanations aid users in grasping the importance of specific features for the decision of the ML model (Ghorbani et al., 2019). Lastly, local explanations aim to enhance human understanding of specific ML system outputs, for example, through confidence estimates (Guo et al., 2017; Pumplun et al., 2023).

However, research in the IS field is now focusing on the impact that especially local explanations have on the interaction between end users and ML systems (e.g., Fügener et al., 2021; Pumplun et al., 2023). Here it is particularly important that the explanations are presented in such a way that the end user, who often has little or no knowledge of ML, can understand them (Bhatt et al., 2020). Improved explainability can lead to several positive outcomes such as fostering user confidence in following ML system recommendations while maintaining their domain expertise (Asatiani et al., 2021; Strich et al., 2021). Moreover, XAI offers the potential for human decision makers to

improve performance (e.g., Gaube et al., 2023), build up trust (Benbya et al., 2021), learn from the knowledge provided by the ML system (Abdel-Karim et al., 2020) and detect instances of ML error (e.g., Fügener et al., 2021). In this context, however, it is important to keep in mind that end users must make some cognitive effort to understand the explanations offered and to take them into account in their decision making process (Arrieta et al., 2020; Pumplun et al., 2023), especially since they are statistical in nature (Bhatt et al., 2020).

#### *5.2.4 Collaborative Decision Making: Performance, Trust, and Mental Workload*

IS research to date is divided on the impact of ML-based DSSs on human **performance**. There are studies that show that people make better decisions and improve their performance with the help of ML-based DSSs (e.g., Abdel-Karim et al., 2020; Gaube et al., 2023). Nevertheless, ML errors thus pose a risk to the overall decision performance (Fügener et al., 2021; Jussupow et al., 2021). While higher explainability can help human decision makers to recognize ML errors (e.g., Fügener et al., 2021), previous studies do not show what influence different types of errors have on performance and which types of errors are more easily detected by humans.

Since ML-based DSSs are prone to errors (e.g., Berente et al., 2021), end users need to evaluate when to adopt or reject system advice, and well-calibrated **trust** is thus a critical aspect for successful human-AI interaction (Gurney et al., 2022). The concept of trust within the research field of IS and in relation to the use of technology has been extensively studied (Glikson & Woolley, 2020; McKnight et al., 2011; Thiebes et al., 2021). Trust is characterized as a party's readiness to be open to potential risks from the actions of another, with the expectation that the latter will act in a way that is important to the trustor, regardless of the trustor's ability to oversee or control the other party (Mayer et al., 1995). While this definition of trust was primarily used to describe interactions between individuals, research has also applied this concept of trust to the interplay between humans and technologies in recent years (Glikson & Woolley, 2020; Gurney et al., 2022; McKnight et al., 2011). For example, Gurney et al. (2022) define trust in AI as "the degree to which a person feels that they can rely on the AI to reduce vulnerability and/or uncertainty in a given situation or instance" (Gurney et al., 2022), p.23). For measuring trust in technology, functionality, reliability, and helpfulness should be considered (McKnight et al., 2011). Functionality pertains to the belief in a technology's capability to accomplish a task for which it was designed. Reliability is the belief in a technology's consistent and stable operation, enhancing trust in its performance. Helpfulness is the belief that the technology offers meaningful assistance to users, aiding them in reaching their objectives (McKnight et al., 2011). Although it is obvious that errors in ML systems can have a negative impact on trust, we know little about the extent to which FNs and FPs differ in their influence.

For humans to achieve a certain level of performance on a task, **mental workload** is required. Mental workload has widely been studied in psychology and is often defined as the costs incurred by an individual while accomplishing a task at a certain performance level (Hart & Staveland, 1988; Sweller et al., 1998). The concept of mental workload is critical for system designers, as humans are unable to perform tasks accurately and reliably with all available resources and maintain high performance without incurring physical, mental, or emotional costs such as fatigue, stress, or accidents (Hart, 2006; Liu & Wickens, 1994; Xi et al., 2023). While the concept of mental workload was initially given great importance in the aviation sector (Hart, 2006), rapid technological progress in many areas relating to information systems has made it necessary to take mental workload into account in order to realize productivity and efficiency benefits. Today, system designers aim to understand how to design and refine systems to ensure their intended benefits remain unaffected by excessive workloads during operation. Nevertheless, as the complexity or difficulty of a task grows, the perceived workload also intensifies. If this surpasses the acceptable level, it will result in a decline in performance (Xi et al., 2023). Research in the field of XAI has also recognized that the inclusion of explanations for ML output—besides the benefits—can also incur costs, as additional information needs to be considered in the decision making process, which requires cognitive resources (e.g., Pumplun et al., 2023). In addition to the influence of explanations on the mental workload, we argue that the influence of different types of errors (FNs and FPs) on the mental workload should also be investigated in order to understand which costs are associated with ML errors and how we can design explainable ML systems that minimize overall costs in the sense of EMT.

### 5.2.5 *Hypothesizing Human Interaction With Erroneous ML-based DSSs*

While prior research has already provided important insights into the impact of incorrect ML-based DSSs on human performance (e.g., Jussupow et al., 2021; Fügener et al., 2021), there is a lack of knowledge regarding the influence of different error types. In line with previous findings, we hypothesize that incorrect (FP and FN) ML-based DSSs will negatively affect not only the performance but also the trust of decision makers such as pilots and pose **H1**: *Interaction with incorrect ML-based DSSs is associated with lower levels of decision maker performance and trust*. In addition, we hypothesize that omissions (FNs) have a more substantial impact on pilots than over-detections (FPs) and propose **H1.1**: *This negative effect of erroneous ML-based DSSs is larger for FNs compared to FPs*.

Research has also shown that explainability of ML-based DSSs output, achieved through local explanations, affects user trust (Benbya et al., 2021; de Zoeten et al., 2023; Glikson & Woolley, 2020) and decision maker performance (e.g., Abdel-Karim et al., 2020). Building on these findings, we anticipate that the decline in trust will be less severe when local explanations are provided

and propose **H2**: *Lower levels of decision maker performance and trust that results from interaction with incorrect ML-based DSSs can be improved through explainability.* In addition, local explanations for ML output can be provided for FPs but not for FNs, thereby increasing the pilot's capability to interpret system advice that contains FPs and we thus pose that **H2.1**: *This positive effect of explainability is larger for FPs compared to FNs.*

Explanations for outputs of ML-based DSSs are typically complex and statistical in nature (Bhatt et al., 2020), demanding that decision makers, such as pilots in our study, expend additional cognitive effort for processing the information (Lebovitz et al., 2022; Pumplun et al., 2023), potentially affecting their mental workload. We therefore propose **H3**: *Interacting with explainable ML-based DSSs is associated with higher levels of mental workload.* As the occurrence of FPs increases the amount of information provided to the pilot, we hypothesize that **H3.1**: *This negative effect of increased mental workload is larger for FPs compared to FNs.*

### 5.3 Methodology

To understand the impact of the different types of errors and the explainability of the ML output on the decision maker, we invited pilots (who hold a valid flight licence) to participate in an online experiment. During the experiment, we showed the pilots a video of a challenging flight scene and provided a certain variant of decision support to help the pilots visually identify other aircraft in airspace. This particular task and the designs of the decision support systems were derived in pre-study interviews with three experienced commercial pilots.

#### 5.3.1 Empirical Context

The three experienced pilots from our pre-study interviews indicated that they see great potential to incorporate ML-based systems into aviation decision making, particularly in the general aviation sector. Here, the pilot's sole decision is often relied upon, instead of relying on redundant instruments. In general aviation, which includes private and recreational flying, pilots commonly operate under visual flight rules (VFR), if weather conditions are clear enough. In VFR conditions, pilots are required to rely on visual reference for maintaining control and navigating the aircraft. Pilots may use navigation aids, however, not all aircraft are equipped with such systems to support situational awareness. Thus, it is paramount for collision avoidance that pilots see and avoid other aircraft, helicopters, parachutes, animals etc. in vicinity and maintain visual separation from them (Civil Aviation Safety Authority, 2023). In addition, it is mandatory to report all incidents that led to unexpected close proximity of aircraft in flight, also known as airprox or near miss. Airprox boards and aviation safety agencies regularly publish these reports (Civil Aviation Safety Authority, 2023), which outline the safety risks that can be associated with visual see and avoid tactics under VFR conditions (e.g., UK Airprox Board, 2021). Based on these insights



from pre-study interviews and airprox reports, we decided to select the visual detection of other aircraft in vicinity during flight under VFR conditions as an appropriate and challenging task for our experiment, for which we provide different variants of ML-based decision support.

To analyze and understand the failure modes of visual detection systems in real-world scenarios, we implemented and trained a state-of-the-art object detection model on real flight scenes from the Amazon Prime Air Airborne Object Tracking (AOT) Challenge dataset. The dataset was released in 2021 and features 164 hours of labelled flight sequences showing in-flight encounters with other aircraft, helicopters, birds and drones, captured by two aircraft that were equipped with high-resolution cameras (Amazon Prime Air, 2021). We choose the recently proposed YOLOv8<sup>6</sup> object detection model for implementation as it achieves state-of-the-art performance on detection tasks with high computational efficiency. The object detector was trained to recognize airplanes and helicopters on a subset of the dataset. It reaches a final precision<sup>7</sup> of 96% and recall<sup>2</sup> of 77% for the class airplanes for the YOLOv8 default threshold. The failure modes of the trained detector were consequently evaluated on a test split of the dataset. Object detector neural networks provide confidence scores for their detections and as described in Section 2.2, a confidence threshold must be selected, at which a detection from the neural network is considered a relevant, real object (aircraft or helicopter) in the application context. The choice of the threshold is a significant design decision as it influences the error behavior of the application with low thresholds allowing more FPs and high thresholds yielding more FNs (Wenkel et al., 2021). Table 4 outlines the percentage decrease in FPs and increase in FNs as the confidence threshold of the trained object detection model is increased, highlighting the relevance of understanding the costs of both error types to determine an appropriate confidence threshold and thus bias the ML-based DSS.

**Table 4: Changes in FPs and FNs for increasing the confidence thresholds of the YOLOv8 model trained on the AOT dataset from 20% to 40%, 40% to 60%, and 60% to 80%**

Confidence threshold:	20% → 40%	40% → 60%	60% → 80%
$\Delta\% FP$	- 35%	- 56%	- 99%
$\Delta\% FN$	+ 3%	+ 5%	+ 21%

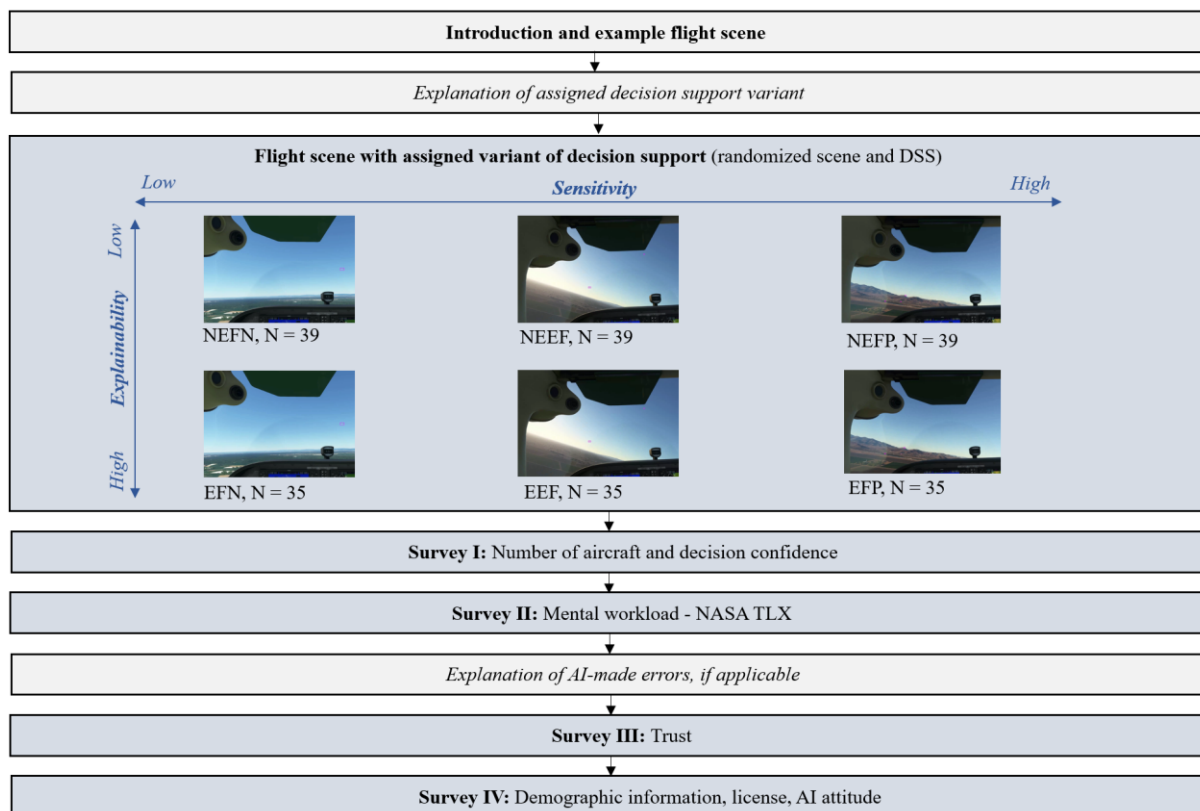
### 5.3.2 Study Procedure and Data Collection

The overall experimental design was iteratively improved and pre-tested with three experienced pilots and one IS researcher. All participants (N = 74) in our experiment followed the same study procedure and were tasked to visually detect and count aircraft encounters in different flight

<sup>6</sup> <https://github.com/ultralytics/ultralytics>

<sup>7</sup> precision  $P = \frac{TP}{TP+FP}$ ; recall  $R = \frac{TP}{TP+FN}$  with true positive (TP)

sequences. They first read an introduction regarding VFR and airprox incidents and watched an exemplary flight sequence showing crowded airspace to get used to the field of view from the cockpit of the Cessna 172 from which the flight scene was recorded using Microsoft Flight Simulator (MSFS) 2020 40th Anniversary Edition (Microsoft Flight Simulator, 2023). They were then randomly split into two groups, one of which received explainable ML-based DSSs and the other non-explainable DSSs (a design description is provided in the following section). Pilots were given a brief description of how each DSS would detect and indicate to the pilot other aircraft in the airspace. Each pilot interacted sequentially with three systems in a randomized order that outputted either fully correct support, FPs, or FNs under the assigned output design requirements. After watching one flight scene and interacting with one variant of ML-based DSS, each pilot reported which aircraft they visually detected and individually completed the surveys described below.



**Figure 9: Design and procedure of the between-subject online experiment for detecting aircraft in vicinity with different variants of erroneous ML-based DSSs**

Next, they completed a NASA Task Load Index (NASA-TLX) (Hart & Staveland, 1988) questionnaire which has widely been applied to measure mental workload of pilots in aviation (Hart, 2006), but has also proven highly valuable in IS research (e.g., Dang et al., 2020; Xi et al., 2023). The multidimensional rating scale NASA-TLX comprises six bipolar dimensions, namely mental demand, physical demand, temporal demand, performance, effort and frustration level, to assess the cost to a decision maker of performing a task at a given level of performance (Hart & Staveland, 1988). While the original NASA-TLX allows system users to rate all dimensions of

mental workload on a 0 to 100 scale, research later adopted a 7-point Likert scale and suggested to exclude physical demand as a dimension of mental workload when it becomes negligibly small (Chen et al., 2009; Dang et al., 2020). We followed these suggestions when implementing the NASA-TLX questionnaire.

Next, the pilots received a report if the ML-based DSS has made any errors and additional information about the type of error that occurred before system trust was assessed through a third questionnaire. This ensures that decision makers are aware of whether errors are made, in order to measure their impact on trust. Pilots were asked to assess trust in the technology along three dimensions, namely functionality, reliability and helpfulness on a 7-point Likert scale as proposed by (McKnight et al., 2011). Finally, demographic data was requested, as well as previous experience in flight hours, type of flight license, a personal risk assessment on airprox incidents, and attitude toward the use of AI.

### 5.3.3 Experimental Setup

To test our hypotheses, six different ML-based DSSs were developed. In pre-study interviews, pilots argued that flight sequences from the AOT dataset could not be used for an online experiment with real pilots because all video data consists of grayscale images and colored footage would be necessary for a realistic setup. Thus, we selected appropriate aircraft encounters from the AOT dataset and recreated them with a group of pilots in a multiplayer session in Microsoft Flight Simulator (Microsoft Flight Simulator, 2023) to simulate crowded airspace in traffic patterns around Frankfurt-Egelsbach Airport (ICAO-Code: EDFE) and Phoenix Goodyear Airport (ICAO-Code: KGYR). All video data was recorded from the cockpit of a Cessna 172 in MSFS and two to three additional aircraft were present in all scenes. Next, all images of a flight scene were labelled in accordance with the following design guidelines:

*Explainability Manipulation:* To assess the impact that explainability of erroneous ML output has on the decision maker, two different levels of explainability of the ML-based DSSs were defined. First, a non-explainable decision support design was developed which only displays bounding boxes around detected aircraft without further information. All bounding boxes are colored in magenta which is commonly used to display system managed information in cockpits, as exemplary shown in Figure 10. Second, a more explainable decision support design was developed in accordance with the design guideline by (Pumplun et al., 2023) which aims at minimizing the cognitive effort induced and EASA's recently published concept paper that aims to guide the development of explainable ML applications in aviation: "Assuming that the decisions, actions, or diagnoses provided by an AI-based system may not always be fully reliable, the AI-based system should compute a level of confidence in its outputs" (European Union Aviation Safety Agency, 2024, p. 92). The explainable variants of ML-based decision support thus provide a local

explanation by indicating a confidence score of the ML model for each detected aircraft. More complex explanations from the XAI field were explicitly excluded because pilots need to make quick decisions under time pressure, and interpreting complex, statistical outputs is disadvantageous in this context. While many studies on ML image analysis in high-risk environments use a color scale (from green to red) to display varying confidence scores to the end user (e.g., Pumplun et al., 2023), pilots in pre-study interviews have strongly discouraged this design to avoid violating common color coding conventions for cockpit design<sup>8</sup>. Instead, we decided that for pixels assigned to the aircraft class with low confidence, the corresponding bounding box would be displayed in a low color density, while all detections made with high confidence would have a magenta bounding box with high color density.



**Figure 10: Exemplary MSFS2020 scene with detected aircraft in magenta bounding box (NEEF)**

*Sensitivity Manipulation:* To assess the costs incurred to the decision maker in case of erroneous ML-advice, we selected three levels of sensitivity for the development of ML-based DSSs. First, an ML-based DSS with a low confidence threshold which did not detect all aircraft in vicinity (a false negative). Second, an ML-based DSS with an optimal threshold (control group–error-free support cannot be guaranteed in real-world use) which provides error-free decision support and detects all aircraft in vicinity. Lastly, an ML-based DSS with a high confidence threshold which challenges the decision maker with a false positive, thus detecting a non-existing aircraft in vicinity.

Overall, our explainability and sensitivity manipulation results in three levels of sensitivity and two levels of explainability. Therefore, six different ML-based DSSs were developed, namely a non-explainable false negative (NEFN), an explainable false negative (EFN), a non-explainable error-free (NEEF), an explainable error-free (EEF), a non-explainable false positive (NEFP), and an explainable false positive (EFP) ML-based DSS (as shown in Figure 9).

<sup>8</sup> According to the color convention by the Federal Aviation Administration, red, amber and yellow are used for alerting flight crew (FAR §25.1322e) to non-normal conditions. Other aircraft detected in airspace but not falling within the specified minimum distances do not warrant an alert requiring immediate action. Therefore, this color selection is waived and all bounding boxes are colored in magenta using different color densities.

### 5.3.4 Data Analysis and Pilot Statistics

Quantitative data was analyzed using SPSS V29. Subsequently, an analysis of variance (ANOVA) was conducted to examine the differences of the NEFN, EFN, NEEF, EEF, NEFP and EFP ML-based DSS on pilots' performance, trust, and mental workload (Girden, 1992; Tabachnick & Fidell, 2006). The online experiment was conducted between July and September 2023. Overall, 222 flight scenes were watched by 74 pilots. The pilots were randomly divided into two groups, one of which received explainable ML-based decision support and the other non-explainable ML-based decision support. Each pilot interacted with three systems in a randomized order that outputted either fully correct support, FPs, or FNs under the assigned output design requirements ( $N_{NEFN} = 39$ ,  $N_{EFN} = 35$ ,  $N_{NEEF} = 39$ ,  $N_{EEF} = 35$ ,  $N_{NEFP} = 39$ ,  $N_{EFP} = 35$ ). All pilots hold valid flight licenses (LAPL, PPL, CPL, ATPL, MPL, SPL, TMG, UL flight licenses were reported), have accumulated an average of 2322.61 flight hours before the experiment, and required an average of 13:26 minutes to complete the experiment. Overall, 9.46% of participating pilots were female and 90.54% were male, reflecting the low percentage of women in the aviation industry (Bartels, 2018).

## 5.4 Results

In the following section, the quantitative results of the online experiment are presented and then discussed. Mental workload (Cronbach's alpha 0.903) and trust (Cronbach's alpha 0.934) of pilots was measured with multiple items and dimensions of the NASA-TLX were weighted equally. We performed an explorative data analysis and tested for homogeneity of variances before conducting a one-way ANOVA ( $p < .05$ ) to examine the groups and assess differences in the effects of presenting different variants of the ML-based decision support. Given that each group has a sample size of more than 30, it is reasonable to assume a normal distribution for the obtained quantitative data (Stone, 2010). Homogeneity of variances was asserted using Levene's Test which showed that equal variances could not be assumed (Levene's Test,  $p < .05$ ) and we thus performed a robust Welch-ANOVA and Games-Howell post-hoc tests.

The Welch-ANOVA, as outlined in Table 5, revealed that **performance** differed statistically significant for pilots interacting with different design variants of ML-based decision support. Games-Howell post-hoc analysis further outlines that pilots' performance is significantly ( $p < .001$ ) lower in groups interacting with non-explainable decision support variants that output FNs or FPs (difference of means NEFP-NEEF: -0.795, 95%-CI[-1.00, -0.59] and NEFN-NEEF), which supports performance related statements of **H1**. In addition, pilots incorrectly included both types of errors in their decision making to the same extent for non-explainable ML-based DSSs (NEFN and NEFP). Thus, **H1.1** is not supported by the results of erroneous and non-explainable ML-based DSSs. Nevertheless, overlooking an aircraft (FN) compared to detecting a non-existent aircraft in

the airspace (FP) is associated with worse risks in real-world application. In addition, the pilots were either completely correct in their assessment or made the same mistakes as the system.

**Table 5: Performance, mental workload and trust assessment following interaction with ML-based DSSs and results of the conducted Welch-ANOVA**

ML-based DSS	N	Performance (0/1) Mean   Std	NASA-TLX* Mean   Std	Trust* Mean   Std
NEFP	39	0.08   0.27	4.17   0.77	4.52   0.89
NEEF	39	0.87   0.34	4.15   0.82	5.78   0.56
NEFN	39	0.08   0.27	4.18   0.87	2.24   1.28
EFP	35	0.66   0.48	5.21   1.39	5.61   0.59
EEF	35	0.91   0.28	4.45   0.91	5.81   0.67
EFN	35	0.11   0.32	4.49   1.01	3.14   0.89
Welch-ANOVA		F(5, 99.55) = 66.11, p < .001, $\eta^2 = 0.561$	F(5, 99.68) = 98.36, p < .001, $\eta^2 = 0.125$	F(5, 99.34) = 3.89, p = .003, $\eta^2 = 0.732$

*\*7-point Likert scale*

In addition, the post-hoc analysis reveals that explainable design of ML-based decision support significantly ( $p < .001$ ) reduced the performance decrease for FPs (difference of means EFP-NEFP: 0.580, 95%-CI[0.31, 0.85]), which shows that **H1.1** is at least supported for explainable ML-based DSSs and supports performance related statements in **H2**. However, explainable design does not significantly reduce the performance decrease caused by FNs (difference of means EFN-NEFN: 0.037, 95%-CI[-0.17, 0.24] and thus, the positive impact of explainability is larger for FPs (which supports **H2.1**).

Besides performance, pilots' **trust** further differed significantly between groups receiving different variants of ML-based DSSs. For non-explainable design of the decision support, post-hoc analysis reveals that trust significantly ( $p < .001$ ) decreases for both error types (difference of means NEFP-NEEF: -1.263, 95%-CI[-1.76, -0.77] and NEFN-NEEF: -3.538, 95%-CI[-4.20, -2.88]) which supports trust related statements of **H1** for non-explainable design. In addition, trust decreases significantly ( $p < .001$ ) more if pilots are challenged with FNs compared to FPs (difference of means NEFN-NEFP: -2.276, 95%-CI[-3.01, -1.54]) which supports **H1.1**.

**Table 6: Hypotheses testing and results from the online experiment**

Hypothesis	Decision Support Variant	Games-Howell p	Outcome
<b>H1: Incorrect</b> ML-based DSS → Performance and trust decrease	NEFP vs NEEF	Performance: $p < .001$ Trust: $p < .001$	Supported for FPs
	NEFN vs NEEF	Performance: $p < .001$ Trust: $p < .001$	Supported for FNs
<b>H1.1:</b> Performance and trust decrease <u>FNs &gt; FPs</u>	NEFN vs NEFP	Performance: $p > .05$ Trust: $p < .001$	Not supported for performance*; Supported for trust
<b>H2: Explainable</b> , incorrect ML-based DSS → Reduces performance and trust decrease	EFP vs NEFP	Performance: $p < .001$ Trust: $p < .001$	Supported for FPs
	EFN vs NEFN	Performance: $p > .05$ Trust: $p = .009$	Performance: Not supported for FNs Trust: Supported for FNs
<b>H2.1:</b> Reducing performance and trust decrease <u>FPs &gt; FNs</u>	EFP vs EEF	Performance: $p > .05$ Trust: $p > .05$	Supported for performance and trust
	EFN vs EEF	Performance: $p < .001$ Trust: $p < .001$	
<b>H3: Explainable</b> , incorrect ML-based DSS → Mental workload increase <b>H3.1:</b> Mental workload increase: <u>FPs &gt; FNs</u>	EFP vs NEFP	$p = .004$	<b>H3:</b> Supported for FPs; Not supported for FNs <b>H3.1:</b> Supported
	EFN vs NEFN	$p > .05$	

\*The magnitude of the performance drop is similar if the criticality of FNs and FPs is rated equally.

Post-hoc analysis further shows that the loss in trust can be significantly reduced through explainability (difference of means EFP-NEFP: 1.095, 95%-CI[0.59, 1.60]) for FPs ( $p < .001$ ) and FNs ( $p = .009$ ) (EFN-NEFN: 0.899, 95%-CI[0.15, 1.64]) which supports trust related statements of **H2**. In addition, for the explainable design, no significant ( $p = .772$ ) difference in trust is observed for the group receiving FPs in the output compared to the error-free variant (EFP-EEF: -0.200, 95%-CI[-0.64, 0.24], while it remains significant ( $p < .001$ ) for the group receiving FNs compared to the error-free variant, which supports trust related statements of **H2.1**.

The Welch-ANOVA confirmed that **mental workload** differed significantly between treatment groups. We further assessed the impact of explainability by comparing treatment groups receiving explainable vs. non-explainable ML-based decision support through a post-hoc analysis. Results outlined that mental workload did significantly ( $p = .004$ ) increase for pilots that received explainable ML-based decision support and were confronted with false positive errors compared to the non-explainable ML-based DSS, (EFP-NEFP: 1.03, 95%-CI[0.25, 1.83]). For error-free support variants ( $p = .701$ ) and support variants that outputted FNs ( $p = .718$ ), no significant difference in the mental workload for explainable output design was found. Thus, **H3** is only supported for FPs and explainable design poses a negative impact on pilots' mental workload for

ML-based decision support that erroneously outputs FPs. However, it does not impact mental workload in case of error-free output or FNs which thus supports **H3.1**.

## 5.5 Discussion

In recent years, ML-based systems have exhibited exceptional performance, offering significant potential for applications in high-stake environments such as aviation (European Union Aviation Safety Agency, 2023). Particularly under VFR conditions, ML-based DSSs could enhance the early and reliable detection and tracking of other aircraft in crowded airspace. However, the inherent susceptibility of ML systems to errors (Berente et al., 2021) poses risks for decision makers (Fügener et al., 2021; Jussupow et al., 2021). Current research on the impact of erroneous ML recommendations on decision making is limited, lacking a distinction between types of errors, namely false positives and false negatives, and therefore fails to offer an exhaustive analysis of the costs associated with each. Our study addresses this gap by examining the effects of FPs and FNs on the performance, trust, and perceived workload of pilots, aiming to inform future ML-based DSS design in line with Error Management Theory (**RQ1**). In addition, we explore how explainable ML design influences pilot interaction and the associated costs, again differentiating between error types (**RQ2**). To this end, we developed six variants of an ML-based DSS, featuring explainable and non-explainable designs, and providing either FP, FN, or error-free support to aid 74 pilots during an online experiment in detecting other aircraft in 222 recorded flight scenes that were recorded in MSFS2020.

Our study makes several **theoretical contributions**. Firstly, we empirically establish a significant performance decline of pilots when interacting with incorrect ML-based DSS outputs for both types of errors—FPs and FNs—in detecting the correct number of aircraft in the airspace (**H1 – Performance**). Contrary to our hypothesis, pilots were unable to discern more FPs compared to FNs when interacting with non-explainable ML-based DSSs (**H1.1 – Performance**). While the amount of deviation (+/-) was identical for both error types, our results highlight the impact each error type has on decision making performance. Our findings underscore the necessity for a qualitative evaluation of the individual risks associated with FPs and FNs in the development of ML-based DSSs. This is critical as there are significant implications whether too many (FPs) or too few (FNs) aircraft are detected, illustrating the importance of EMT in design considerations.

Furthermore, our results regarding trust demonstrate that both types of errors lead to a significant loss of trust (**H1 – Trust**), with this loss being substantially greater in the case of FNs (**H1.1 – Trust**). This demonstrates the need for a “humanly engineered bias” (Haselton & Nettle, 2006) by adjusting the system’s confidence threshold and thus its sensitivity. It further emphasizes the importance of trust management for the development of ML-based DSSs. Finally, the differential loss of trust between error types highlights the need to investigate the different



effects of each error type for ML-based DSSs. This differentiation is crucial for the development of systems that not only effectively support collaborative human-AI decision making but also maintain their human trust in ML-based systems.

Secondly, our study reveals that an explainable design of ML-based DSS output positively influences the detection of FPs and significantly mitigates the resulting loss of trust (**H2 – FP**). Conversely, providing explanations for FNs does not lead to a significant improvement in error detection by pilots, although it still reduces the trust deficit (**H2 – FN**). Overall, explanations from the field of XAI have a significantly greater, positive effect on the costs of FPs compared to FNs (**H2.1**). This distinction is again of crucial importance for the design of systems according to the principles of EMT. The goals pursued by XAI research become increasingly important as we bias ML systems towards FPs and thereby develop more sensitive systems. Our results integrate error management research with the XAI research area and illustrate that a combined view is essential to best support human decision makers and minimize the impact of erroneous output. The results argue not only for the inclusion of explainability features in the design of ML-based DSSs, but also for a differentiated approach to dealing with different ML error types.

Thirdly, the results from the online experiment indicate that the mental workload of pilots significantly increases with the provision of local explanations for ML outputs in the case of FPs, but not for FNs (**H3** and **H3.1**). The presence of FPs leads to an increased amount of information that pilots must process under time pressure, which is exacerbated by explainable design as pilots have to interpret additional data, such as confidence scores. Our study thus exposes a trade-off that future research on developing explainable ML-based DSSs should consider: XAI approaches can enhance the detection of FPs and reduce loss of trust for FPs compared to FNs. This could lead ML developers to bias systems towards favoring the occurrence of FPs over FNs and thus develop sensitive ML-based DSSs. However, this approach, may jeopardize the decision maker's mental workload, particularly when additional explanations are included. This highlights the complex balancing act required in designing ML-based DSSs that need to be sensitive enough to be effective without overburdening the human decision maker.

In addition, we provide several **practical contributions**. Firstly, our trained YOLOv8 object detector and the output design implemented in MSFS flight scenes serves as a blueprint for the development of explainable ML-based cockpit systems that aid pilots in detecting other aircraft under VFR conditions. This design incorporates pilot requirements gathered from pre-study interviews and is further validated by feedback obtained through free-text fields in our online experiment, confirming the system's utility and user-friendliness from the pilots' perspective. Secondly, our study provides a basis for new adoption strategies for ML-based DSSs in high-stake environments. Our results suggest that in scenarios with low mental workload, sensitive but explainable ML-based DSSs can be deployed early, as they enable better detection of FPs and thus

reduce the risk of performance degradation and loss of trust on the decision maker's side. Thirdly, our research emphasizes the critical importance of distinguishing between different types of errors for practical application. Organizations aiming to ensure that complex ML-based DSSs are taken into consideration by decision makers over the long term, should adjust the sensitivity of these systems based on end user needs and risks associated with the different error types.

## 5.6 Limitations and Future Research Directions

Our study, while offering valuable insights, is not without its **limitations**. The number of participating pilots was limited, and the pilots reviewed the flight scenarios via video recordings rather than through real-time testing in flight simulators and long-term effects of the different ML-based DSSs cannot be explored through this study. Additionally, the pilots focused solely on the detection of other aircraft, whereas actual flight demands simultaneous management of multiple tasks. Despite this, we believe our results hold validity, as experienced pilots selected and recorded the flight scenes. However, there is a risk that the measured costs incurred on pilots may be significantly higher in real-world applications. Moreover, our participant demographics reflect the persistently low representation of women among pilots in the industry, which we acknowledge as a limitation. When designing future ML systems, the requirements of female pilots must be taken into account just as much as those of their male colleagues. **Future research** should undertake a combined investigation into how explanations from the field of XAI ought to be designed, dependent on the sensitivity of ML-based DSSs, to foster an optimal cost trade-off for decision makers. In this context, the impact of global explanations to further improve explainability should be considered and special emphasis should be placed on the examination of mental workload in high-stake environments. Building on our findings, guidelines for ML developers should be established, delineating how ML systems should be biased according to the principles of EMT to minimize the overall costs incurred by different ML error types. Furthermore, the results should be validated in real flight simulators, and the requirements should be tested with a more diverse group of pilots over a longer period of time to ensure that the findings are robust and applicable across a broader demographic spectrum. Overall, this study provides exciting insights into the effects of different ML error types (FPs and FNs) on pilots and can serve as a basis for aligning the development of ML-based DSSs more closely with the needs of decision makers in the future.

## **6. Paper D: Design for Acceptance and Intuitive Interaction: Teaming Autonomous Aerial Systems with Non-experts**

### **Title**

Design for Acceptance and Intuitive Interaction: Teaming Autonomous Aerial Systems with Non-experts

### **Authors**

Ellenrieder, Sara; Mehler, Maren F.; and Turan Akdag, Merve

### **Publication Outlet**

Proceedings of the 27<sup>th</sup> Pacific Asia Conference on Information Systems (PACIS), Nanchang, China, 2023

### **Abstract**

In recent years, rapid developments in artificial intelligence (AI) and robotics have enabled transportation systems such as delivery drones to strive for ever-higher levels of autonomy and improve infrastructure in many industries. Consequently, the significance of interaction between autonomous systems and humans with little or no experience is steadily rising. While acceptance of delivery drones remains low among the general public, a solution for intuitive interaction with autonomous drones to retrieve packages is urgently needed so that non-experts can also benefit from the technology. We apply a design science research approach and develop a mobile application as a solution instantiation for both challenges. We conduct one expert and one non-expert design cycle to integrate necessary domain knowledge and ensure acceptance of the artifact by potential non-expert users. The results show that teaming of non-experts with complex autonomous systems requires rethinking common design requirements, such as ensuring transparency of AI-based decisions.

### **Keywords**

Human-autonomy teaming, design science research, autonomous drones, app design

## 6.1 Introduction

Rapid advances in the field of robotics and artificial intelligence (AI) boosted the development of autonomous transportation systems in recent years. In particular drones, also known as unmanned aerial systems (UAS) or unmanned aerial vehicles (UAVs), offer a promising solution for low-emission, autonomous transportation of various goods and multiple successful test flights have already been conducted to deliver several medications, vaccines or even defibrillators to people in need (Krey, 2018; Scott & Scott, 2017). For example, during test flights in Sweden in December 2021, a drone successfully transported an automated external defibrillator (AED) to a 71-year-old who suffered an out-of-hospital cardiac arrest. A bystander quickly administered cardiopulmonary resuscitation using the AED before emergency medical services arrived, saving the patient's life (Hicks, 2022). Besides applications in healthcare, drone service providers and manufacturers strive to offer existing services in other industries today and food or general parcel deliveries have been tested by global players such as Amazon PrimeAir (Amazon, 2016) and Google Wing (Levin, 2016) as well as entrepreneurial startups (Giones & Brem, 2017; Heunemann, 2022). In addition to expanding application areas, the increasing autonomy of transportation systems such as delivery drones enables scalability and may also allow inexperienced individuals to use these technologies to their own advantage in the future (Hicks, 2022; Moshref-Javadi & Winkenbach, 2021; Pasztor & Ferek, 2021).

However, even if delivery drone systems are operated autonomously to allow for large scale operations that provide value to the general public, humans still need to interact with these systems to retrieve parcels and take advantage of the benefits that a modern drone delivery network can offer. In this case, the human does not control the drone and many decision-making powers are transferred to the drone system. Nevertheless, interaction is still required and effective teaming between the human and the autonomous system is thus critical to success (McNeese et al., 2019, 2021). In many growing application areas of delivery drones, this will lead to untrained humans with little or no knowledge of AI, robotics or autonomous systems to interact and team up with autonomous drones with increasing frequency in the upcoming years. Besides being non-experts, studies show that people oftentimes oppose the use of drones in general and acceptance of this technology remains a major societal challenge today (Eißfeldt et al., 2020; Eißfeldt & End, 2020; Rice et al., 2018). The research field of human-autonomy teaming, which has grown in recent years as machines have become more capable, is already investigating how attributes such as situational awareness (Demir et al., 2017; Endsley, 2018) or trust (McNeese et al., 2019, 2021) influence effective teaming between humans and autonomous systems. Nonetheless, research in the field of human-autonomy teaming (McNeese et al., 2021) and human-drone collaboration (Dolata & Aleya, 2022) calls for further work and we aim to take a step forward by providing guidance on how to implement this special form of interaction in real-world

use cases and, in particular, to consider the role of non-experts in this context to enable scalability and foster acceptance in the future. Hence, we aim to answer the following research question: *In the context of autonomous delivery drones, how should a mobile application be designed to promote the acceptance and use of autonomous delivery services and to enable non-experts to interact intuitively in such human-autonomy teams?*

We apply a design science research (DSR) approach to develop a mobile application that enables non-expert users to interact with autonomous drone systems in order to safely retrieve a delivery and thus have access to modern service offerings. While intuitive interaction design is a focus of the design project, we specifically aim to increase acceptance of autonomous delivery drone technology among non-experts. We follow the approach of Kuechler & Vaishnavi (2008) and perform two design cycles, an expert and a non-expert design cycle to derive design requirements (DRs) and principles (DPs) that are structured along the unified theory of acceptance and use of technology (UTAUT) model (Venkatesh et al., 2003) and instantiate a solution in form of a mobile app prototype afterwards. The implementation of our artifact in the form of an app allows to provide access to autonomous transport systems to a wide range of end users and in particular to non-experts (Pitt et al., 2011). Within the first expert design cycle, semi-structured interviews are performed with experts that work at different drone manufacturers as well as delivery drone pilots to incorporate the necessary domain knowledge. Based on the interview results, DRs and DPs are instantiated in a click prototype and evaluated in expert focus groups. After analyzing the results of the expert design cycle, a non-expert design cycle explores how non-expert users of different ages and experience evaluate the instantiated solution to derive a final design that incorporates domain knowledge and satisfies the needs of end users while fostering acceptance of autonomous delivery drones. Overall, we focus on fully autonomous drones with vertical take-off and landing (VTOL) capabilities because they offer great flexibility in adjusting their trajectory and have been successfully deployed in real-world delivery flights with non-experts in the past (e.g., Hicks, 2022). Focusing on fully autonomous drones allows us to explore the interaction between non-experts and autonomous systems without risking interference of other human intermediators such as human pilots.

## **6.2 Theoretical Background**

The following section reviews existing research in human-autonomy teaming and shows how our study aims to extend this relatively young area of research. In addition, the UTAUT model, which is used as the kernel theory in our DSR approach, is outlined.

### 6.2.1 *Interaction for Human–Autonomy Teaming*

The field of human-machine interaction has grown considerably in recent years, leading to the emergence of multiple sub-fields. Today, research on human-machine interaction also covers the fields of human-machine teaming as well as human-autonomy teaming. In general, this research stream explores interactions between a human and a machine working in interdependent roles to achieve a common goal (McNeese et al., 2021, 2018). However, in the case of human-autonomy teaming, the machine has the ability and authority to make decisions independently and is not supervised in those decisions by the human (Demir et al., 2017; Endsley, 2018; McNeese et al., 2018). The latter area of research has received less attention in the past because limited machine capabilities did not allow for the required level of autonomy. However, autonomous machines will be increasingly used in the coming years and research in the area of human-autonomy teaming will increase in practical relevance (Bradshaw et al., 2004; McNeese et al., 2019). McNeese et al. (2021) argue that we need to rethink human-machine interaction if humans and machines do not interact as supervisors and subordinates anymore. The authors state that we are currently at an *inflection point*, which is why we need to transfer concepts from the field of human-machine interaction and insights from teams that have exclusive human members into the field of human-autonomy teaming. They conducted an experiment to analyze the role of trust in human-autonomous teams, and also studied the context of remotely piloted aircraft operating in the role of autonomous team members. Here, the aircraft interacts with, but is not controlled by, the human team member (McNeese et al., 2021). A study by Yuan Zhang & Jessie Yang (2017) showed that uncertainty in human-autonomy teams will lead to a higher perceived workload and those teams will, in general, be able to perform fewer tasks at the same time. Modeling team interactions and incorporating human cognition into the design of the autonomous agent are crucial aspects to be considered for effective human-autonomy teaming in dynamic environments (Gutzwiller et al., 2018; Klein et al., 2004; Parasuraman et al., 2000). While some studies exist that explore the impact of characteristics such as trust (McNeese et al., 2019, 2021) or situational awareness (Demir et al., 2017; Endsley, 2018) on human-autonomy teaming and thus already allow researchers to draw new insights from empirical work, research calls for additional work (McNeese et al., 2021). Rapid developments in the field of autonomous systems reinforce the need for research to also address the implementation of such human-autonomy teams. Furthermore, the higher level of autonomy should especially enable non-experts to use such systems in the future.

### 6.2.2 *Designing Human–Drone Collaboration*

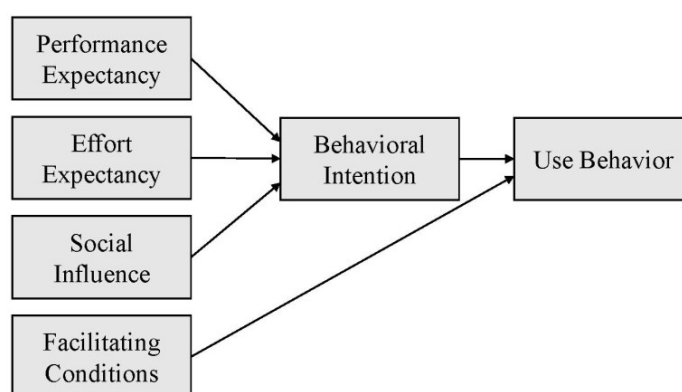
As for autonomous systems, delivery drones have already shown great potential to improve logistic networks in various industries today (e.g., Mao et al., 2019; Scott & Scott, 2017). However,

many use cases require non-experts to interact with these autonomous systems in the future, and the general public remains opposed to the use of drones as a means of transportation (Eißfeldt et al., 2020; Eißfeldt & End, 2020; Rice et al., 2018). Intuitive interaction for non-experts that further promotes acceptance of this technology is urgently needed to make the benefits of delivery drones accessible to the general public (Ellenrieder et al., 2023). In addition, interacting with a drone as an autonomous agent provides special challenges due to the physical embodiment of the agent in an autonomously controlled, flying object (e.g., McNeese et al., 2019, 2021). While the IS community is primarily concerned with the advantages and disadvantages of various drone applications, there is a lack of approaches to designing human-drone collaboration (Dolata & Aleya, 2022). To address this shortcoming, Dolata & Aleya (2022) propose a taxonomy of dimensions and characteristics that are relevant for designing human-drone collaboration. However, they do not provide requirements for the design of artifacts in this field. We preselect outdoor delivery flights in the proposed context dimension, untrained operator's skillset in the social dimension, single drone flights that interact with one human in the mutual interactions dimension and lastly fully autonomous drones with light to heavy payload in the technical component dimension (Dolata & Aleya, 2022). Other characteristics such as unidirectional or bidirectional communication between drone and human will be defined in the design cycles. The taxonomy outlines that social and organizational aspects must be considered for the design of human-drone collaboration instead of solely focusing on technical requirements. While Dolata & Aleya (2022) already argue that the operator's skillset (expert vs non-expert) is a relevant social dimension for the design of human-drone collaboration (Alex & Vijaychandra, 2016; Allen & Mazumder, 2020; Dolata & Aleya, 2022), they do not mention the influence of the operator's acceptance of the technology or the operator's environment. However, existing research primarily targets the use of drones for emergency operations which is associated with higher levels of acceptance (Aydin, 2019). In the context of our paper, we aim to bridge the gap between non-experts and autonomous drones by providing clear design requirements that combine domain knowledge from experts with user requirements from non-experts and promote acceptance of the technology.

### 6.2.3 *Unified Theory of Acceptance and Use of Technology*

The UTAUT model was derived in 2003 by Venkatesh et al. (2003) and is useful for understanding which factors affect acceptance of a technology and thus influence the likelihood of success for that technology. In addition, the model can be used to derive strategies for targeting potential users who are less inclined to adopt the new technology (Venkatesh et al., 2003, 2016). Performance expectancy, effort expectancy, social influence and facilitating conditions are the four core determinants of intention and usage that are considered in the UTAUT model, as

pictured in Figure 11 (Venkatesh et al., 2003). In general, the performance expectancy of users can be defined as the extent to which a person believes that using the system will help them improve their job performance. In addition to performance expectancy, effort expectancy also has a significant influence at the beginning of use, but it decreases over time. Effort expectancy is defined as the degree of ease that is associated with using the technology or system. In addition, social influence affects the intention to use a new technology and is defined as the extent to which an individual perceives that significant others believe that he or she should use the new technology. Social influence is also found to be significant in the beginning when experience is still low. Lastly, facilitating conditions are defined as the extent to which a person believes that resources such as organizational and technical infrastructure are in place to support the use of the system (Brown & Venkatesh, 2005; Venkatesh et al., 2003, 2012).



**Figure 11: The UTAUT model with its four direct determinants of user acceptance and usage behavior (Venkatesh et al., 2003)**

The relationship between these four independent variables and the dependent variables of behavioral intention and usage are moderated by gender, age, experience and voluntariness of use (Venkatesh et al., 2003, 2016). Venkatesh et al. (2003) showed, for example, that the effect of performance expectancy is stronger for men and younger workers while social influence has a stronger effect on women, older workers and those with less experience. The unified theory of acceptance and use of technology will serve as a kernel theory in our DSR approach. The UTAUT model has proven to be particularly useful compared to other technology acceptance models because it directly incorporates social influences, which are currently considered one of the biggest challenges facing the drone industry (e.g., Eißfeldt et al., 2020; Tan et al., 2021). In addition, we decided to use UTAUT instead of the extended UTAUT2 model (Venkatesh et al., 2012) since constructs such as price value and experience and habit of the UTAUT2 model are not feasible for our use case. Non-experts do not hold prior drone experience or regularly interact with fully autonomous systems and the current tradeoff of perceived benefits of drones and costs cannot be compared as costs are expected to decrease significantly in the future as the autonomy



level increases and economies of scale can be realized through mass production of autonomous drones in the future.

### 6.3 Design Science Research Approach

We apply a design science research approach by Kuechler & Vaishnavi (2008) to develop a mobile application that enables non-experts to interact with autonomous drone systems in human-autonomy teams while fostering acceptance of this technology. We contribute to the growing science of design for IT artifacts by deriving DRs and DPs for human-autonomous drone teams as well as creating and evaluating the instantiated solution in multiple design cycles in this study. The general design cycle as defined by Kuechler & Vaishnavi (2008) comprises five distinct phases, namely *awareness of the problem*, *suggestion*, *development*, *evaluation* and *conclusion*. We perform two design cycles (as outlined in Figure 12), an expert design cycle and a non-expert design cycle to ensure the instantiated solution is built on domain knowledge and fosters acceptance by end users. Thus, we performed 10 semi-structured expert interviews within the *awareness of the problem* phase from which we derived initial DRs and overarching action-oriented DPs during the *suggestion* phase of the first design cycle. We formulated a semi-structured interview guideline according to Sarker et al. (2013) which addressed all independent variables of the UTAUT model and allowed interviewees to freely share their experiences and ideas. At the beginning of each interview, we outlined the use case of delivering parcels with autonomous VTOLs to the general public, most of whom can be defined as non-experts in drone technology. Afterwards, we collected interviewee-related information such as their current position and prior experience in the drone industry. We then asked experts to share their opinions and insights on why drone acceptance remains low and what risks the general public typically sees with test flights before discussing how experts envision a solution approach in the form of a mobile application that would allow non-experts to interact with these systems.

General DSR Project Phases	First Design Cycle	Second Design Cycle	Third Design Cycle
Awareness of Problem	Expert interviews	Analysis of initial evaluation	Analysis and synthetization of results from expert and non-expert design cycles and final refinement of principles and artifact. Experimental evaluation with drone experts and non-expert users.
Suggestion	Derivation of initial design principles	Refinement of design principles	
Development	Instantiation of design principles in mock-ups	Instantiation of design principles in click prototype	
Evaluation	Artifact evaluation in expert focus group	Artifact evaluation in non-expert focus groups	
Conclusion	Analysis of expert focus group results	Analysis of non-expert focus groups results	
	<b>Expert Design Cycle</b>	<b>Non- Expert Design Cycle</b>	<b>Mixed Expert and Non-Expert Design Cycle</b>

**Figure 12: DSR project structure (cf. Kuechler & Vaishnavi, 2008) including expert, non-expert and mixed design cycles. This study focuses on the cycles highlighted in grey.**

Experts work in different areas within the drone industry and provide in-depth knowledge on safety-related aspects of drone deliveries as well as current AI-based capabilities of (partially) autonomous drone systems. We followed a theoretical sampling approach and initially invited drone pilots (E1 - E3) to our interviews to gain insights from today’s users who already interact with drones. Drone pilots, who still control some processes manually today, are already preparing for their new role in the coming years, in which they will primarily monitor a fleet of autonomous drones and intervene only in exceptional situations. Thus, they enriched this study with real-world experiences with the challenges of human-autonomy teaming. However, drone pilots mentioned that interfaces to guide interaction between drones and end users are usually designed for pilots and the design of interfaces for non-experts is still a research and development (R&D) topic in many companies. Thus, we invited experts (E4 - E8) from R&D and product development departments of different drone manufacturers in the second round. Our focus on autonomous delivery drones is associated with special risks since drones are required to carry payload and transfer parcels from the drone to the non-expert. A payload specialist (E9) was thus invited for an interview once we identified these risks. Lastly, we invited a drone program manager (E10) who gained prior experience in R&D as well as drone customer service and operations management to ensure all relevant information was being obtained. The program manager confirmed the derived DRs and no additional DRs were identified during the last interview. Although the civilian delivery drone industry is relatively young, seven out of ten respondents, already had five or more years of experience in this industry. Interviews were conducted from April to August 2022 via online meetings and recorded following mutual agreement. Interviews lasted 58 minutes on average and were subsequently transcribed.

**Table 7: Roles of expert interview participants**

Case	Work Title/ Area of Expertise	Years of Experience
E1	Drone Pilot	6 years
E2	Drone Pilot	5 years
E3	Drone Pilot	3 years
E4	R&D, Drone Manufacturer	5 years
E5	Product Development, Drone Manufacturer	2 years
E6	Product Development, Supplier Drone Technology	9 years
E7	Product Development, Supplier Drone Technology	5 years
E8	Product Development, Drone Manufacturer	11 years
E9	Payload Specialist	4 years
E10	Program Manager, Drone Manufacturer	12 years

We performed a content analysis according to Weber (1990) for the evaluation of the qualitative data gained and conducted two coding cycles that comprised descriptive and pattern coding (Saldana, 2021). Following the *suggestion* phase, we developed mock-ups of a suggested solution instantiation using the wireframing tool Balsamiq which considered all derived DRs and DPs during the *development* phase (Balsamiq, 2022). All experts listed in Table 7 were invited to a focus group study that lasted 78 minutes to evaluate the prototypical implementation of the DRs and DPs. Following an analysis of the focus group study results, a second design cycle to involve potential non-expert users was conducted and initial DPs were refined before a clickable prototype was developed using Figma (Figma, 2022). For *evaluation* of the non-expert design cycle, we conducted a combined exploratory and confirmatory focus group study which is outlined in detail in the evaluation section (Tremblay et al., 2010). Overall, experts that participated in the first design cycle did not participate in the second design cycle. Our focus group studies aim at obtaining in-depth understanding of how a selected group of individuals evaluates the artifact (O.Nyumba et al., 2018) in terms of promoting acceptance and intention to use the technology. The approach allows us to refine our design principles iteratively before obtaining a statistically representative sample of a broader population in the final and still outstanding mixed-expert and non-expert design cycle. The three non-expert focus groups, which were conducted in October 2022, lasted for 66 minutes on average to ensure all DRs, DPs and instantiated prototype features were evaluated in detail. Audio recordings of all sessions were transcribed afterwards. We applied a mixed content analysis in accordance with Morgan & Scannell (1998) to obtain both qualitative and quantitative information in our focus group studies. We allowed for a maximum of eight participants for non-expert groups because they may require a different amount of explanations which becomes difficult to manage in larger groups or leads to the formation of

subgroups that have independent discussions (Krueger, 2014). Following the expert and non-expert design cycle, results were discussed and summarized during the *conclusion* phase and we prepared the outstanding final design cycle for which more information is provided in the outlook on future research.

## 6.4 Results

The following section describes the results collected and synthesized in all phases of the design cycles. For the derivation of the DRs and DPs, the focus is on the inductively obtained results from the expert interviews. Subsequently, the final prototype is presented, which already integrates the feedback from the expert evaluation, before the results of the non-expert focus group evaluation are outlined.

### 6.4.1 Awareness of the Problem

As stated in the introduction and background section of this study, current information systems (IS) research examines what challenges arise and need to be addressed in the area of human-autonomy teaming (e.g., Gutzwiller et al., 2018; Parasuraman et al., 2000). Autonomous aircraft, which can be a physical embodiment of an autonomous agent, recently received attention from this research stream (McNeese et al., 2021). To advance this research, we address the specific design of human interaction with autonomous delivery drones. In doing so, we consider two specific aspects that we identified in expert interviews as key challenges for practical applications in the coming years, namely: the interaction of non-experts with complex, autonomous drone systems and the low acceptance of this technology among non-experts (E1,3,4,5,8,9,10). In terms of the relevance of the topic, one interviewee stated: *“What we’ve seen in rapid technology development is quite impressive. But in the past, when we conducted test flights, only trained people interacted with our drones to, for example, retrieve a package. That’s definitely going to change now as we look to grow our drone fleet in many different areas and the drone has more and more capabilities”* (E2). Many use cases of drone delivery services require non-experts to interact with drone systems even though they do not control actions of the drone such as its movement in airspace (E3,4,6,8,9). Thus, interaction should work without much effort and should not require any prior knowledge (E5,7,9). While this poses a great challenge, an interviewee also argued that *“bringing drones closer to the general public may finally help us to foster acceptance”* (E7). Raising awareness of the benefits of drone technology in modern urban areas and areas with poor infrastructure is a key task that providers of these services have addressed in previous pilot projects (E2,3,10). Delivery flights directly to customers now offer the possibility of *“first-hand experience”* (E4) and it will be particularly important at the outset to create a convenient experience for the end user here (E1,5). Several interviewees argued that deriving and following

design suggestions from drone experts will be key due to the complexity of the systems and the associated safety risk if used incorrectly (E1,2,6,9). Bringing together the requirements of experts and non-experts for the interaction of autonomous drone systems and translating them into a common design proposal represents a core challenge that we address in this study. One interviewee even stated: *“Systems in aviation have actually always been operated by experts. However, with increasing autonomy, people will have to interact with more and more systems themselves in the future, and it will be our goal to find solutions in close cooperation between experts and people without prior knowledge”* (E10).

#### 6.4.2 Suggestion

To ensure our artifact fosters acceptance of autonomous delivery drones among non-expert users, we structure the derivation of design requirements along the four independent variables of our kernel theory the UTAUT model, namely performance expectancy, effort expectancy, social influence, and facilitating conditions (Venkatesh et al., 2003). Figure 13 provides an overview of all derived design requirements. All sub-requirements were derived based on challenges or solution approaches that were explicitly mentioned by the experts and could be identified during the coding process.

**Performance expectancy** is defined as the extent to which a person believes that using the system will enhance their job performance. It has the highest impact on the intention to use a technology as shown by (Venkatesh et al., 2003). Multiple interviewees argued that acceptance for delivery drones remains low, primarily because the general public is unaware of the benefits drones can offer in terms of delivery times and emissions (E1,2,4,10). *“We need to get to the point where we are really transparent about the positive impact that this technology could have. People need to be able to critically question what the advantages or disadvantages of delivery by a drone are – especially if it’s not manually controlled”* (E8). Apart from the performance characteristics of this autonomous means of transport, non-experts are usually not familiar with the process that will be followed during delivery and what interaction is required (E2,3). An interviewee further stated: *“Users must understand that it is not just autonomous drone features that they can rely on. It is also their active participation in the interaction that influences the overall performance.”* (E1). We therefore propose **(DR1) Highlight improved performance characteristics**: Interacting with the app should increase performance expectancy and reassure users that the capabilities of autonomous drones will improve delivery performance. Sub-requirements include **(DR1.1)** providing information on delivery time, emissions, and costs, **(DR1.2)** a performance comparison with all applicable other modes, and **(DR1.3)** providing a forecast of the delivery process to the end user, including upcoming interactions.

Design Requirements	
<b>Performance expectancy</b>	
<b>DR1</b>	Highlight improved performance characteristics
1.1	Provide information on delivery time, emissions, costs
1.2	Compare performance with other means of transport
1.3	Provide forecast of delivery process
<b>Effort expectancy</b>	
<b>DR2</b>	Enable effortless interaction
2.1	Provide a design that is familiar to delivery application users
2.2	Provide intuitive guidance through minimized options
2.2	Outline future interactions before user decides to engage
<b>Social influence</b>	
<b>DR3</b>	Communicate social acceptance
3.1	Provide information on other users within the social system
3.2	Allow sharing of experiences
3.3	Outline the contribution to community objectives
3.4	Offer the possibility of communication with other users
<b>Facilitating conditions</b>	
<b>DR4</b>	Provide safe interaction control
4.1	Enable transparent, bidirectional feedback
4.2	Provide decision support for safety-related interaction
4.3	Transfer final decision-making authority to the end user

**Figure 13: Design requirements derived from expert interviews**

**Effort expectancy** is defined by the ease of use of the system. The determinant was found to be especially significant during the first temporal use (Tremblay et al., 2010). Multiple interviewees argued that effort expectancy will be one of the key challenges in teaming up non-experts with autonomous systems (E3,5,6,7). In addition, one interviewee argued: *“There are so many preconceptions about how complicated it is to control an autonomous system because the systems are so complex now. But non-experts won’t control our drone systems. Rather, they will only interact with the system in certain situations.”* (E10). To ensure intuitive use, the possibilities for the end user to shape the interaction with the autonomous system should be limited (E1,6,8,10). The non-expert user must be prepared for every interaction and, if possible, feel reminded of already familiar environments and tasks (E4,6,8). Drone pilots and others who interact with drones in the aviation industry must undergo a knowledge assessment to obtain the required licenses. While in the case of autonomous drones, no special skills are required, a self-assessment is still beneficial to strengthen the user’s confidence that no skills are required that they do not possess (E1,2). We therefore propose **(DR2) Enable effortless interaction**: The app should enable intuitive interaction with autonomous drones to reduce complexity and maximize perceived ease of use. Respective sub-requirements ensure that **(DR2.1)** the design of the artifact reminds the user of

familiar delivery applications, **(DR2.2)** user options for shaping the interaction are minimized, and **(DR2.3)** emerging interactions are communicated to the user early on.

**Social influence** is defined as the extent to which the user perceives that significant others believe that he or she should use the new system. User behavior is thus influenced by the user's belief in how a technology may enhance his or her status in his or her social system (Venkatesh et al., 2003). An interviewee described the current situation as follows: *"If you have, say, an electric car. You would probably be proud and talk about it with neighbors and friends. However, drone technology is known to be used by the military, and people are often biased toward civilian applications as well, especially in urban areas. In our pilot test flights, we realized that sharing experiences between communities had the most positive effect. In fact, people were even excited after learning about the positive experiences and improved infrastructure of similar communities."* (E1). First and foremost, respondents mentioned that an app needs to connect users with each other to convey that the technology is already being used by a variety of similar users, especially non-experts. This can reduce the fear of rejection by the social environment and has already been experienced as beneficial during real test flights (E1,2,3,9,10). In addition to encouragement from the social environment, it is important to convey to the user the contribution they can make to society by choosing a low-emission mode of transport. Highlighting the contribution to emissions targets can further encourage users to share their experiences with others (E4,10). We therefore propose **(DR3) Communicate social acceptance:** The app should communicate to the user that the use of autonomous drones is gaining social acceptance and supports societal goals such as achieving the SDGs. Corresponding sub-requirements define **(DR3.1)** to provide information about other users within the app user's social system and **(DR3.2)** to enable experience sharing among users. In addition, **(DR3.3)** the artifact should clearly demonstrate how users can contribute to emission goals or other social goals by using the new technology.

**Facilitating conditions** are defined as the extent to which a person believes that an organizational and technical infrastructure is in place to support the use of the technology. If facilitating conditions are supporting the use of a technology, the user would have a positive perception of behavioral control (Venkatesh et al., 2003). While this could also involve available resources or compatibility issues, facilitating conditions for human-autonomy interaction are rather understood in the context of behavioral control to address questions such as: How much control should the human have and how much autonomy should the drone have? In the interviews, the experts argued that the facilitating conditions should first be determined by experts to ensure safety (E1,2,4,5,6,8). However, they also mentioned that input from non-experts will be crucial to ensure that users feel comfortable and can interact with the technology without being overwhelmed (E2,5,8,10). A drone pilot further stated: *"As a drone pilot, I already feel comfortable letting the system make certain decisions on its own. However, I would like to receive*

*information about why the system reacts in a certain way. We should also provide easy-to-understand feedback to people who have no prior knowledge. [...] And even if I'm monitoring multiple drones at once, I always want to be able to abort a mission. This decision should remain possible for every user in every planned interaction.*" (E2). In addition, the experts made it clear that the app must offer the possibility of requesting help from experts (E6,7). It is difficult to define in advance all the situations that may occur in the real world during the interaction between a non-expert and the autonomous drone system. In all safety-critical interactions such as the acceptance of a dropped package, the user should therefore be provided with decision support and the opportunity to provide feedback (E1,8,9,10). Thus, we propose **(DR4) Provide safe interaction control**: The app should allow the user to access the necessary infrastructure to safely control the interaction with the autonomous delivery system. The sub-requirements specify **(DR4.1)** that bidirectional feedback should be possible that is transparent but understandable to users. In addition, **(DR4.2)** decision support should be provided for all safety-related interactions, **(DR4.3)** with the final decision-making authority to enter or exit an interaction remaining with the human.

#### 6.4.3 Deriving Design Principles

Before the development of a solution instantiation, we derive design principles in accordance with (Gregor et al., 2020) which provide the basis for all features that are integrated into our artifact. To promote informed decisions about the use of the emerging technology, non-expert users should understand what autonomous delivery drones can offer in terms of performance metrics such as delivery times or emissions, and what impact this will have on the social environment of users (see **DR1** and **DR3**). We thus define **DP1**: *For an intuitive mobile application to inform and encourage, the user should be able to evaluate the emissions-related environmental impact of the autonomous transportation system and compare the system's performance with other transportation modalities to assess whether the user's performance expectations are being met.* The user's effort to perform a given task in cooperation with the autonomous drone system depends on the provided guidance and preparation, which must be explicitly designed for non-experts (see **DR2** and **DR4**). We therefore derive **DP2**: *For an intuitive mobile application to guide the user throughout the interactive process, the user should be able to inform and prepare for all upcoming decision-making activities to avoid frustration and overwhelm.* While respondents frequently mentioned that a non-expert will not control the delivery drone system, the ability to actively shape the interaction between the human and autonomous system will become a key characteristic (see **DR1**, **DR2**, and **DR4**). Thus, we propose **DP3**: *For an intuitive mobile application to support and enable active engagement of the user, the application should mediate feedback between the autonomous system and end user and provide decision support to the user for initiating or aborting interactive processes.* Even though users do not supervise the autonomous



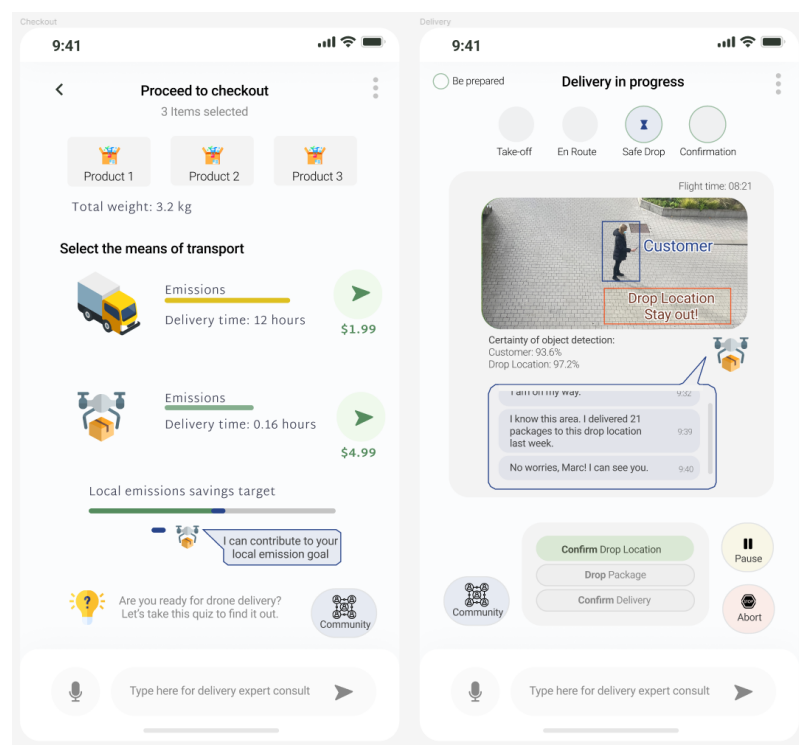
system, the experts see ensuring situational awareness as highly relevant to user-friendliness and safety (see **DR2** and **DR4**). We propose **DP4**: *For an intuitive mobile application to ensure situational awareness of the user for safe collection of deliveries, the user should be able to access the current and upcoming operational status of the autonomous drone system in a transparent manner without having to navigate through menus.* Finally, according to experts, user acceptance is strongly influenced by the social acceptance of the technology by the general public. In addition to the exchange between the user and other users from his/her social environment, the offer of support by trained, human experts can also promote acceptance (see **DR3** and **DR4**). We therefore derive **DP5**: *For an intuitive mobile application to promote usage and relationship building, users in the same neighborhood should be able to connect, share experiences and provide external human support for human-autonomy interaction.*

#### 6.4.4 Development

In the first step, we derived design features (DFs) based on the defined requirements and principles and used Balsamiq to build a sketched artifact in form of a mobile application (Balsamiq, 2022). Following an expert evaluation, we then developed the clickable prototype in Figma as pictured in Figure 14 (Figma, 2022). This version of the artifact allows us to vividly present how the interaction between users and the autonomous delivery drone system is designed within our non-expert design cycle. The artifact primarily aims to allow non-experts to test DFs on their own smartphone and get a realistic sense of the interaction possibilities and challenges. Two scenarios are selected to provide some insights into the suggested solution instantiation. The first scenario requires the user to accept the technology and directly interact with it following this decision in the second scenario. The left screenshot of the click prototype in Figure 14 shows a scenario that occurs before the user selects the desired mode of transport for a delivery (DF2).

The performance features are shown below a familiar shopping cart display (DF1). As a gamification element, the user is also offered a small quiz (DF3, DF11) to playfully test and extend his or her drone knowledge and find out what knowledge and skills are required. To give the feeling of two-way communication in all scenarios, the user can write with a human drone expert in the chat (DF10) and receives regular updates from a small drone icon (DF5). In the figure on the right, which shows a scenario shortly before the drone drops the goods, these interaction options are also available. In addition, the user is shown the entire delivery process to prepare him or her for the interaction at an early stage (DF4). To increase situational awareness, an AI-assisted bird's-eye view allows the user to understand what the autonomous system can detect (DF7). Experts explained that even for pure interaction, it will be crucial to understand that the drone detects e.g. objects and with what degree of certainty. Detected objects such as a feasible

drop location and the human itself are marked by a 2D-bounding box and a certainty for object detection is provided below. In the interaction panel at the bottom of the right screenshot (DF6), the user can only pause or permanently cancel the interaction. Further design options are only made available in the appropriate interaction phase and are grayed out beforehand to prepare the user in advance for future interactions. Finally, the user can access the community platform within this panel to get support from their social circle or share experiences (DF9).



## Design Features

**DF1:** Performance comparison

**1.1** Delivery characteristics

**1.2** Regional goal achievement

**DF2:** Means of transport selection

**DF3:** Capability self-assessment

**DF4:** Operational phases forecast

**DF5:** Proactive drone messenger

**DF6:** Interaction control panel

**DF7:** AI-enhanced bird view

**DF8:** Navigation

**DF9:** Community platform

**9.1** Regional flight history

**9.2** Milestone overview

**9.3** Community FAQs

**9.4** Community messenger

**DF10:** Expert support

**DF11:** Gamification

**Figure 14: Instantiated solution for teaming non-experts with autonomous drone delivery systems and respective design features (icons made by Freepik, Pixel Perfect and Satawatdesign from Flaticon, 2022)**



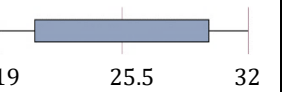
### 6.4.5 Evaluation and Discussion of Results

The instantiation of derived DPs in the form of the clickable prototype was evaluated through three focus groups that were conducted in October 2022 with potential non-expert users (Tremblay et al., 2010). Participants with extensive experience with drones, such as using drones for filming, were excluded from the focus groups to ensure evaluation by non-experts. In addition, participants had to be at least 18 years old. During the exploratory phase, we gathered insights into the participants' attitudes toward delivery drones and the use of autonomous systems as well as their knowledge base. After a brief introduction about delivery drones and the collection of personal data, we presented a video of a drone delivery and showed how our artifact would shape the interactive process using an Apple iPhone12. Non-experts could then click independently through the functions of our prototype. We aimed to understand how non-experts evaluate and interact with our artifact in terms of performance and effort expectancy, as well as social influence

and facilitating conditions. Finally, we gained insights into how non-experts evaluate derived DPs and interaction with the prototype in order to assess whether the artifact promotes acceptance and use of the emerging technology. Our sample with  $N = 22$  participants (50% male, 50% female) represents the group of potential users ranging between 19 and 81 years. All participants owned a smartphone and reported familiarity with delivery apps. Overall, participants indicated a low level of experience with drone technology (on a 7-point Likert scale) as shown in Table 8.

After a brief introduction to the use of delivery drones as a means of transportation, the majority of participants in all focus groups expressed concerns, such as privacy risks and the risk of accidents. Several participants stated that they would feel watched if drones flew over private property. In addition, concerns arose regarding noise pollution. However, during the simulated delivery, participants also expressed excitement and quickly engaged with the artifact themselves. When we asked participants what they liked most about the app, they all mentioned its intuitive use: *"I don't really have an idea how such drones work, but this is like ordering something on Amazon."* (P3G2), *"it is easy to use"* (P6G1), *"I feel like I always know what to do next"* (P3G3). Effort expectancy was assessed to be low overall and several participants mentioned that the self-assessment was very helpful before beginning the delivery process. In addition, the majority of the participants see the bird-eye view in which the drone camera footage is shared as generally beneficial to *"understand what the drone can see and is doing"* (P4G2). Nevertheless, when we asked participants to tell us about what they disliked most about the app, they mainly complained about the features that aim to increase the transparency of AI-based decisions and display e.g., with which certainty the drone currently detects objects: *"This is just confusing. The drone says it detects me with 93 something percent certainty. What does this mean now? Can it see me now or not? [...] And then there is also nothing I can do about it. I would prefer the system to only tell me things it is sure about."* (P6G3). While participants mentioned that they need more information to understand this data, it also became clear that sharing this information in such a human-autonomy team is not very helpful because non-experts cannot interpret this information and act on it by, for example, taking control of the drone system. In addition, such statements clearly show why it is crucial to involve both experts and non-experts to design any collaboration between complex, autonomous systems and non-experts. During the first design cycle, experts have repeatedly emphasized that it is important to be able to evaluate the reliability of a system. Even if the user no longer actively controls the system, this would be helpful for creating trust. This view is influenced by the experts' previous experience of system-human interaction, which is strongly characterized by human control over the system. However, evaluation by non-experts in the second design cycle shows that known design principles for human-autonomy teaming need to be revised.

**Table 8: Non-expert focus group study sample**

	Group 1 (G1)		Group 2 (G2)		Group 3 (G3)	
<b># of Participants (P)</b>	8		8		6	
<b>Age</b>	 25 36 81		 22 29.5 53		 19 25.5 32	
<b>Gender (m/f)</b>	62.5% / 37.5%		37.5% / 62.5%		50.0% / 50.0%	
<b>Prior Drone Experience</b>	Mean: 1.2	SD: 0.43	Mean: 1.5	SD: 0.50	Mean: 1.8	SD: 0.69

The proactive drone messenger, symbolized by a small drone icon sending messages, was rated as encouraging and helpful in providing needed information at the right time. However, non-experts pointed out that they need to be fully focused as soon as direct interaction with the drone system becomes necessary. The number of short messages received via the drone messenger should be kept to a minimum in these situations. All participants mentioned that they need a performance comparison to make an informed decision about the optimal means of transport. Since they considered different benefits such as faster delivery times or lower emissions to be more important, the comparison of performance characteristics should therefore also be individualized and adapted to the goals of the users. In addition, participants wished the app communicated more clearly that their presence was expected during delivery, as they cannot predict what interaction will be required before they need to make the decision for or against drone delivery. While only a minority of the participants indicated they saw *“no need to share something like this with their social network”* (P5G2), many participants said they were concerned about disturbing their neighbors with delivery drones and it would be good to know if this technology was already being used in their region. In addition, participants indicated that they would share their experiences with the drone app with family and friends, resulting in a positive overall assessment of the community platform features.

Many participants argued that intention to use increased following the interaction with the prototype. Participants stated that gaining more knowledge on drones and to be able to better assess effort and performance expectancy were primary reasons. This finding is supported by other studies which showed a positive correlation between public drone knowledge and acceptance of the emerging technology (Aydin, 2019; Eißfeldt et al., 2020). Relationships in the UTAUT model are moderated by gender, age, experience, and voluntariness of use (Venkatesh et al., 2003). While we only consider voluntary use by non-experienced users, we observed and gathered qualitative insights into the impact of gender and age. Older participants indicated that they were hesitant to use the new technology primarily because they expected *“a lot of effort”* (P3G2) during the delivery process and were unsure if the app would support them sufficiently to complete their tasks. Lower familiarity with delivery apps was also cited as an influencing factor.

Younger study participants, on the other hand, reported that they enjoy testing out new technologies and have little fear of being overwhelmed. While a majority of female participants expressed concern about whether their social environment would support the use of delivery drones, this statement was only made by one male, older study participant. Apart from perceptions of social influence, no other gender differences were found.

While our study focuses on delivery drones, we derived general design principles for non-experts to interact with any autonomous transportation system and discussed these principles in our focus group studies following the demonstration of the artifact. **DP1**, which aims to inform the user on the system's performance and environmental impact and encourage use, was overall found to be *"very important"* (P1G1), *"crucial for such new technologies that I haven't tested out before"* (P4G1), the *"primary reason I would decide to use it"* (P2G3). However, participants disagreed on whether they primarily wanted persuasive arguments or preferred neutral information that required individual interpretation. **DP2**, which aims to guide the user in all interaction and decision-making activities and enable intuitive use, was evaluated to make users feel *"comfortable"* (P4G3) and like they *"have done it before"* (P5G1). However, non-experts strongly argued that intuitiveness depends on their level of knowledge and insights that might be helpful to a drone pilot were found to be confusing such as displaying the certainty in object detection tasks. **DP3** and **DP5**, which state that the app should be equipped with functions for communication and relationship building, can provide autonomous systems *"with a human component"* (P1G3). Overall, the communicative form of interaction was perceived by the study participants as very easy and pleasant. However, individual participants stated that they would not contact expert support early on because they *"did not want to ask stupid questions"* (P5G1). While experts described this service as particularly safety-critical, the support could also be labeled as ordinary customer service to lower this inhibition threshold. Regarding **DP4**, which aims to increase situational awareness, participants emphasized that they were primarily focused on the current interaction and therefore the bird's eye view was particularly helpful. Overall, **DP4** was rated as *"absolutely necessary"* (P2G3). Some participants further argued that the app should dynamically adapt to situations to improve situational awareness and one participant said: *"I tried this app for the first time. When I interact with the drone, I want to know what's going on at that second. I can't check all the other things at that moment, like the upcoming process stages"* (P5G3). In addition to these insights into the lively discussion about the suitability of derived DPs, Table 9 highlights other identified design aspects that found wide acceptance during the focus group study. In addition, opportunities for further improvement of the artifact are shown which will be addressed during the upcoming third design cycle.

**Table 9: Results of focus group discussions by non-experts on derived design principles**

Design Principles	Identified Design Aspects with High User Acceptance	Identified Opportunities for Improvement
<b>DP1:</b> Inform and encourage use	<ul style="list-style-type: none"> <li>• Providing neutral information on emissions and performance</li> <li>• Comparison with other transport modalities</li> </ul>	<ul style="list-style-type: none"> <li>• Customization to personal performance goals</li> <li>• Avoiding persuasion to use the technology (e.g., remove drone icon with proactive usage prompt)</li> </ul>
<b>DP2:</b> Guide interaction	<ul style="list-style-type: none"> <li>• Relying on familiar delivery application design</li> <li>• Early communication of upcoming interactive events that require user action</li> </ul>	<ul style="list-style-type: none"> <li>• Minimization of information about system functions that are beyond the user's control</li> <li>• Clear communication of the autonomy level of the drone</li> </ul>
<b>DP3:</b> Support active engagement	<ul style="list-style-type: none"> <li>• Receiving regular messages from the system, even if no interaction is required</li> <li>• Suggest user interactions that are structured in a meaningful way according to the time sequence</li> </ul>	<ul style="list-style-type: none"> <li>• Minimization of system messages in situations that require high cognitive effort</li> </ul>
<b>DP4:</b> Ensure situational awareness and safety	<ul style="list-style-type: none"> <li>• Transparently displaying a summary of the delivery process</li> <li>• Making the perspective of the autonomous system understandable to the user (e.g., bird eye view)</li> </ul>	<ul style="list-style-type: none"> <li>• Identification of situations that may require the system to have final/shared decision-making authority, rather than always relying on the user (e.g., complete failure of the user, detected unauthorized use by minors)</li> </ul>
<b>DP5:</b> Promote relationship building	<ul style="list-style-type: none"> <li>• Sharing their own successes in using technology and learning from others</li> </ul>	<ul style="list-style-type: none"> <li>• Ensuring that only users with sufficient competence offer their support to other users in the community area</li> </ul>

Finally, we asked participants if they considered the drone to be a team member. Participants found it “weird” (P4G3) to consider an object as a team member. Even though the drone makes many decisions on its own, participants rather viewed it as “a supporting tool” (P8G2). Nevertheless, one participant stated: “I wouldn’t say that the drone was my team partner, but I wouldn’t call our delivery guy my team partner either. So, it can also be more due to the task. If I work very well with someone regularly, I develop a team feeling over time. If a drone brought me my daily newspaper every morning, I might develop – yeah - something like a team feeling after a while. For example, my son has such a vacuum cleaner robot and even gives it a name. I find that strange, but here you could perhaps already see the robot as a team partner” (P7G1). Although it is not expected to be a viable business model to deliver printed daily newspapers by drones, this statement clearly shows that human-autonomy teams will have to develop over a longer time and also that the developed artifact will fulfil many goals only after a longer period of use.

The qualitative evaluation of the artifact is very feasible to iteratively revise DPs and improve interface design. However, for the final and outstanding third design cycle, our evaluation strategy

primarily aims to obtain quantitative data from a larger sample of potential non-expert users and experts to rigorously test the joint value of derived DPs. In addition, we designed an artifact for fully autonomous VTOL drones to deliver parcels. While some DRs and DPs, such as DR1 and DR3, can be applied to the design of artifacts for semi-autonomous drones that interact with non-expert users, DR4 would require adaptation for such use cases. This also applies to all DPs which guide the interaction in human-autonomy teams.

## 6.5 Conclusion, Limitations, and Directions for Future Research

Due to technological advances in the fields of robotics and artificial intelligence, many systems reach for higher and higher levels of autonomy today. Especially delivery drone technology has matured in recent years and a variety of civilian use cases that also foster the creation of sustainable cities have emerged. To enable non-experts with little to no knowledge of autonomous systems such as drones, to also benefit from the emerging technology, we apply a DSR approach to derive a solution instantiation in two consecutive expert and non-expert design cycles. We are responding to the call to expand the growing research field of human-autonomy teaming (McNeese et al., 2021) by addressing implementation in the context of a real-world use case. We develop a mobile application for non-experts to safely and intuitively interact with autonomous delivery drones and promote acceptance and use of the technology, thus addressing our research question. Following the second design cycle, we plan a final mixed design cycle with experts and non-experts to test the app in a real-world experiment.

Our study provides several **theoretical contributions**. First, we present a methodological approach that allows non-experts and experts to be involved in the design process on an equal basis and without mutual interference during the initial derivation of DRs and DPs and initial evaluation. This approach creates a common ground for future research which aim to make complex autonomous systems accessible to the general public. Incorporating domain knowledge and translating it into a design intuitive enough to be understood by non-experts remains a major challenge in many growing research areas, and we provide a foundation on which targeted future research can be built. Second, our study demonstrates an approach that not only places user satisfaction at the center of the design process but also promotes general technology acceptance. Many AI-based technologies are viewed controversially, and drone technology research is just one example that can benefit from this approach. Third, we provide several insights on human-autonomy teaming in the context of non-expert use which goes beyond existing research. We show that the interaction of non-experts with autonomous systems requires an iterative design approach and many existing design requirements need to be reconsidered for human-autonomy teams as already suggested by (McNeese et al., 2021). To provide a concrete example, the transparency features of AI-based autonomous system decisions were found to be rather

unhelpful, because users have too little background knowledge of AI, have to trust system decisions, and can no longer control systems. These findings pave the way for transferring and adapting knowledge from the human-machine interaction domain to human-autonomy teaming. Fourth, we contextualize general DPs and develop a solution instantiation for non-experts to interact with autonomous systems. Thus, we contribute by guiding researchers and IS designers on how to design easily understandable artifacts that interface with autonomous systems. Lastly, we outline how UTAUT can be applied as a kernel theory in the DSR approach to develop artifacts for human-AI teaming. Practical insights from experts confirm the impact of all constructs of UTAUT on the intention to use autonomous delivery drones. While UTAUT has been shown to be very suitable, we have also identified factors, such as privacy, that have not yet been adequately addressed by the theory and provide a basis for further research.

This step toward implementing human-autonomy teams in a real-world application also holds several **practical contributions**. First, our solution instantiation vividly demonstrates how broad the application area of delivery drones can be if the autonomy level is sufficient to allow interaction with non-experts and thus the general population. Second, iteratively derived DRs and DPs also clearly show how important it is to involve this new end user group in the development work of manufacturers and service providers at an early stage. With our study, we also call on the industry to address the challenges in the area of human-autonomy teaming. While the drone industry has already identified this challenge (Ellenrieder et al., 2023), it is expected that this topic will also become more relevant in other industries as autonomy levels increase. Lastly, the civil drone industry can build on the knowledge gained and use our artifact as a basis to develop optimal solutions tailored to different use cases and drone types. In the upcoming third design cycle, drone manufacturers and service providers will have the opportunity to test our artifact to transfer the results into practice.

However, our contributions are also subject to **limitations**. The number of semi-structured expert interviews is limited and results may be biased toward popular fields of application such as deliveries in the healthcare sector. In addition, we focus on the specific use case of parcel delivery by VTOL drones which offer a high degree of flexibility in adapting to human actions and the applicability of our DRs, DPs and the developed artifact cannot be guaranteed for all drone types and application domains. Moreover, it is important to note that qualitative research results were obtained in this study and statistically significant results will have to be added through follow-up studies. Lastly, focus group participants' opinions were obtained after a demonstration of the technology at the prototype stage. Interaction with large delivery drones that appear to move self-directed in the airspace is likely to have an impact on participants' subjective assessment in the real world. Both, the contributions and limitations of this study provide a broad spectrum for **future research**. Besides obtaining statistically significant results, it will be of great relevance to



conduct experiments under real conditions and refine design principles in the future. While acceptance of delivery drone technology can be fostered through targeted artifact design, future research and practice will need to take further approaches to ensure that this technology can continue to improve infrastructures and make future cities more sustainable. Delivery drones have evolved rapidly in recent years, and this application domain has provided a great context for our and previous studies investigating human-autonomy teams. We encourage future research to apply the findings to other autonomous and especially semi-autonomous systems. While UTAUT served as a feasible kernel theory to design our artifact for human-autonomy teaming, we see a need for future research to extend this model by also incorporating privacy aspects that arise for all levels of system autonomy. Lastly, drone technology has rapidly improved in recent years and poses many opportunities to improve e.g., infrastructure and access to healthcare services. However, their actual impact on the sustainability of future cities remains controversial, and several delivery drone programs have been discontinued in recent years. Future research should critically discuss sustainability aspects of the technology, and research on the design of human-autonomy interaction should focus precisely on the application areas that have a major impact on sustainability.

## **7. Paper E: The Influence of Effort on the Perceived Value of Generative AI: A Study of the IKEA Effect**

### **Title**

The Influence of Effort on the Perceived Value of Generative AI: A Study of the IKEA Effect

### **Authors**

Mehler, Maren F.\*; Ellenrieder, Sara\*; and Buxmann, Peter

*\*shared first authorship*

### **Publication Outlet**

Proceedings of the 32nd European Conference on Information Systems (ECIS), Paphos, Cyprus, 2024

### **Abstract**

While the use of generative artificial intelligence (GenAI) aims to automate human tasks, psychology research shows how crucial human effort is for the appreciation of the final results. The so-called “IKEA effect” refers to the increased valuation individuals attribute to self-created products. However, the potential implications of this effect for GenAI have remained unexplored. This study delves into the presence of the IKEA effect in GenAI, specifically focusing on image creation. Through an online experiment involving 174 participants in Germany, we observed that participants valued images higher if more human effort was invested during collaborative co-creation with GenAI. Our findings indicate a significant presence of the IKEA effect, although existing GenAI research primarily focuses on the automation of processes. This discovery emphasizes the importance of understanding user psychology and also offers valuable insights for designing and leveraging GenAI applications.

### **Keywords**

IKEA effect, generative artificial intelligence, online experiment, human-AI collaboration

## 7.1 Introduction

Cognitive biases have been shown to influence decision-making in various instances (e.g., Hristov et al., 2022; Ni et al., 2019), including the IKEA effect, which states that people tend to value self-assembled physical objects more than identical objects that were assembled by others (Norton et al., 2012). Attributing a higher perceived value to objects into which personal effort has been poured during development or assembling already significantly impacts how companies shape customer experiences today (Franke et al., 2010). Adidas, for instance, enables their customers to customize clothing through their online platform (Adidas, 2023). Similar design choices for customers are offered by the popular backpack brand Fjällräven (Fjällräven, 2023), and Build-A-Bear provides customers with the chance to design stuffed animals online and even participate in the stuffing process at their physical stores (Build-A-Bear, 2023). As a result, customers are evolving into active co-creators rather than merely passive consumers of products (Mochon et al., 2012).

Advancements in data availability, training techniques, and scale of model parameters have made Generative Artificial Intelligence (GenAI) models more versatile, capable, and applicable to a wide range of tasks and domains such as text or image creation (Dwivedi et al., 2023). Recent industry reports show that modern GenAI promises to automate tasks that currently consume 60 to 70 percent of employees' daily work hours. This heightened potential for technical automation primarily stems from the advancements in GenAI's proficiency in comprehending natural language—a critical requirement for tasks constituting 25 percent of the overall work hours (McKinsey & Company, 2023). The automation goals pursued with the use of intelligent technologies are a paradox with the trend of allowing customers to participate in the product development process.

While the primary goals of the deployment of (generative) AI are the automation of tasks and minimization of human effort to successfully accomplish tasks (Berente et al., 2021; Brynjolfsson & Mitchell, 2017; Russell & Norvig, 2021), challenges for the success of this technology arise as psychology and behavioral research has already taught us that “labour leads to love” (Norton et al., 2012). While many studies have proven the IKEA effect for physical objects such as origami, food, or furniture (e.g., Dohle et al., 2014; Ling et al., 2020; Mochon et al., 2012; Norton et al., 2012), we see a clear lack of research for non-physical objects or content in general. Thus, we do not yet understand to what extent the IKEA effect also affects the perceived value of non-physical goods. At the same time, the research stream of human-AI collaboration investigates how humans and intelligent machines can work together synergistically to enhance problem-solving and decision-making processes (e.g., Asatiani et al., 2021; Sculley et al., 2015; Sturm, Gerlach, et al., 2021; Sturm, Koppe, et al., 2021). However, several unique challenges and characteristics make it uncertain

whether and how the IKEA effect could manifest. In previous research on the IKEA effect, humans could always estimate what activity and skill was needed to assemble or create goods such as furniture or food. Missing transparency (often known as black box behavior) in the AI algorithms increases the difficulty of comprehending how the AI makes decisions or creates results (Bauer, Hinz, et al., 2021; Dwivedi et al., 2023). It is also unclear which influence humans have on the results through prompting and how the AI processes human inputs; only the results can be evaluated (Dwivedi et al., 2023; OpenAI, 2023).

In this context, it is crucial to understand the impact that the takeover of human tasks by AI will have on the perceived value of the final results. Regarding the IKEA effect, it can be assumed that the perceived value of AI-generated content and users' behavioral intention to use GenAI tools increase, especially when they have the opportunity to put effort into the collaboration. Thus, we seek to answer the following research questions: *(1) Does human effort invested in collaboration with generative AI promote overvaluation of AI-generated solutions, and (2) does this heightened perceived value also increase the behavioral intention to use GenAI technology?*

To answer these research questions, we conducted an online experiment with 174 people who interacted with GenAI tools to perform work tasks with a high vs low level of effort in the collaborative creation of content. We hereby assess the impact that effort poured into the collaboration with the AI has on the perceived value of the generated content and the behavioral intention to use the technology. We contribute to research by examining how humans and AI should collaborate in the future to value the content created through collaboration. The results of our online experiment reveal that the IKEA effect is prevalent in the collaborative creation of content by humans and GenAI. Humans tend to overvalue AI-generated content if effort is invested into collaboration with the technology. Future research can build on this study to derive design guidelines for GenAI tools and strategies for human-AI collaboration that ensure users value results.

## **7.2 Theoretical Background**

The following section outlines the unique characteristics of GenAI. It then provides an overview of cognitive biases, focusing on the IKEA effect and its impact on the perceived value of self-created objects so we can hypothesize the impact of the IKEA effect on human-AI collaboration afterward.

### *7.2.1 Artificial Intelligence*

Berente et al. (2021) describe AI as the frontier of emerging technologies that is focused on human intelligence for complex decision-making. There are several subcategories of AI, such as machine

learning (ML), which are learning algorithms that recognize patterns in data to make decisions or predictions (Brynjolfsson & Mitchell, 2017; Mitchell, 1997; Russell & Norvig, 2021). Or, more recently, the advances in GenAI such as ChatGPT, Google's Bard, or other Large Language Models (LLMs), which differ from ML in creating new data from recognized patterns and not only analyze existing data (Dwivedi et al., 2023; Teubner et al., 2023). However, these language models are not the only novelty in this area. AI image creation tools like Stable Diffusion, Dall-E 2, or Midjourney also prove the capabilities of AI as generated images are very realistic and appealing (Göring et al., 2023). The use of ChatGPT is currently being discussed in radiology (Rao et al., 2023), for cybersecurity (Prasad et al., 2023), and in a wide range of business areas. Although such systems can already facilitate work today, especially in the future (Dwivedi et al., 2023), caution must be taken that jobs are not destroyed and that users accept the outcomes of AI.

AI is characterized, in particular, by the three properties autonomous, inscrutable, and self-learning (Berente et al., 2021). In addition, the potential for bias in training data, the ownership of the training data as well as the output, and the potential of wrong outputs are concerns in GenAI (Lund et al., 2023). In particular, the "black box" characteristic of AI makes it difficult for humans to follow the decision of AI and thus understand and adapt the given output (Bauer, Hinz, et al., 2021). This characteristic remains applicable in GenAI as well (Dwivedi et al., 2023). The research stream on human-AI collaboration investigated the influence of machines and humans working together (e.g., Boyacı et al., 2024; Fügenger et al., 2022). It was shown that humans want to decide rationally but often cannot, e.g., due to a misjudgment of the task's difficulty (Fügenger et al., 2022). The research also shows that people are reluctant to perform simple tasks and enjoy demanding tasks themselves, which makes delegation to AI problematic (Fügenger et al., 2022). In addition, it was demonstrated that when the results of decisions differ between humans and AI, people have different approaches to dealing with it. For example, experienced physicians tend to ignore AI advice, while novice physicians question their own decisions and are less satisfied with the AI system (Jussupow et al., 2021). It is also essential that human knowledge is not lost in the interaction between humans and AI but should be actively incorporated into decision-making (Fügenger et al., 2021). Hence, there should be a focus on developing systems that allow humans and AI to work together rather than AI exclusively reaching final decisions (Abdel-Karim et al., 2020).

### 7.2.2 *Cognitive Bias*

Humans often deviate from optimal decisions since always thinking and acting rationally is impossible. This is due to several factors: People usually cannot process the entire amount of information or do not have it available; when processing the information, incorrect conclusions can be drawn, for example, due to interpretation influenced by emotions or existing prejudices;

or heuristics are used when making quick decisions (Kahnemann, 2012). These influences on rational decisions are called cognitive biases. They can be grouped into four categories: First, *information and perception biases* can occur, such as the anchoring bias, which refers to the tendency to rely on the first information that comes to mind when making decisions (Kahnemann, 2012). Second, *decision biases* exist, such as the overconfidence bias, which is the tendency to overestimate one's own judgment. It leads to a misplaced sense of certainty in decision-making, making decisions riskier (Odean, 1998). In addition, *social and emotional biases* are prevalent, such as the self-serving bias, which is the tendency to make decisions that benefit one's own interests (Miller & Ross, 1975). Lastly, *technology and change biases* affect decisions, such as the status quo bias, which describes the tendency to prefer the current state of affairs and resist change. This can lead to rejecting new technologies, ideas, or practices even if they offer improvements (Kim & Kankanhalli, 2009).

Especially in IS research, there is a lot of literature on cognitive biases in decision-making (e.g., Hristov et al., 2022; Ni et al., 2019) or biases in the adoption of emerging technologies (e.g., Balakrishnan et al., 2021; Frank et al., 2023; Piehlmaier, 2022). For example, Phillips-Wren et al. (2019) show that overconfidence bias is an inhibitor in adopting and using decision aids. However, Piehlmaier (2022) finds that overconfident investors are more likely to use Robo advisors. In consumer adoption, another exemplary bias, the negativity bias, is evident (Frank et al., 2023) or individuals do not always choose the best algorithm for their decisions (Dietvorst & Bharti, 2020). In addition, Kim and Kankanhalli (2009) illustrate that user resistance to information systems is related to status quo bias. Similarly, Balakrishnan et al. (2021) show status quo bias as a factor in accepting AI-powered voice assistance. Furthermore, a focus on organizations can often be found. In this regard, Ni et al. (2019) show that anchoring bias occurs in a corporate context when making decisions using a BI system. More generally, Hristov et al. (2022) show what possible cognitive biases occur within decisions about performance management systems. Once biases are known, they can be prevented or exploited to force rational or reasonable decisions.

### 7.2.3 Understanding the IKEA Effect

One bias that can be best categorized as a *decision bias* and is widely studied to explore the impact of user participation on the success of products or services is the so-called IKEA effect. The IKEA effect states that people value objects higher if they assemble or create them on their own (Norton et al., 2012). Norton et al. (2012) named this phenomenon after the Swedish manufacturer whose products are particularly likely to involve a high level of assembly effort. Higher effort means a person must invest more work in a task (see Marsh et al., 2022). While the IKEA effect is unsurprising for some products such as art, studies have demonstrated the IKEA effect for a

variety of physical objects such as food, Lego, clothing, or simple IKEA cardboard boxes (e.g., Dohle et al., 2014; Ling et al., 2020; Mochon et al., 2012; Norton et al., 2012; Radtke et al., 2019). In fact, the value of self-created objects is estimated even higher than the value of objects created by experts (Norton et al., 2012) and the IKEA effect also remains prevalent in collaborative creation (Marsh et al., 2022). In addition, the effort that goes into the creation of the objects does not have to be associated with fun for consumers to overvalue their creations (Mochon et al., 2012; Norton et al., 2012). However, successfully completing a task is necessary for the IKEA effect to emerge (Norton et al., 2012).

Evidence that “labour leads to love” (Norton et al., 2012) can also be derived from the success of past product launches. When instant cake mixes were introduced to minimize the manual labor required for baking, the initial success failed as the preparation of a cake was now perceived as too simple. The recipes were changed so that the addition of eggs was required. Although other influences may have been at work here, the subsequent adoption success is often attributed to the IKEA effect, as the customer puts additional work into the product (Norton et al., 2012; Shapiro, 2004). Today, multiple companies take advantage of the IKEA effect to shape customer experience by providing opportunities to customize and create products online through configurators, tool-kits and choice menus (Franke et al., 2010). Therefore, customers can increasingly be described as co-creators instead of passive recipients of goods (Mochon et al., 2012). Providing ideas, thoughts, feelings, and, most importantly, actions to participate in the co-creation process leads to a higher perceived value, which has been shown to increase the willingness to pay (WTP) for these products (Mochon et al., 2012; Norton et al., 2012).

Although the bias has been successfully demonstrated in many studies (e.g., Dohle et al., 2014; Ling et al., 2020; Marsh et al., 2022; Norton et al., 2012), uncertainty exists regarding the actual cause. Psychology and behavioral research suggest three primary mechanisms that cause the IKEA effect: (a) signal of competence, (b) effort justification, and (c) ownership (Marsh et al., 2018). First, self-created objects can express the competence of the creator and can be used as a trophy to show off, increasing the perceived value of the object for the creator (Mochon et al., 2012). Second, within the concept of effort justification, creations reflect the investment of effort rather than signal competence (Norton et al., 2012). Thus, the increased value of a created object may reflect the effort invested. Lastly, the creation of an object can lead to ownership claims in some scenarios or promote the sense of ownership of the person who created the object (Kanngiesser et al., 2010; Marsh et al., 2022). Since people tend to place a higher value on their personal belongings compared to equivalent items that they do not own, this can further enhance the IKEA effect (Kahneman et al., 1990). Nevertheless, there is disagreement about the underlying mechanisms and it remains difficult to adequately explain the IKEA effect (Marsh et al., 2018).

A related research stream can also be found in the area of mass customization (MC), where customers are provided with easy-to-use configurators to design products themselves online, which are then produced by the manufacturer (Franke et al., 2010; Ling et al., 2020). Research on MC toolkits, however, is based on two assumptions: firstly, that *preference fit* is the essential benefit for customers, while *design effort* represents costs for the customer and should therefore be kept to a minimum (Franke et al., 2010; Randall et al., 2007). These two goals are consequently in conflict with each other, and MC toolkits should balance them optimally to gather enough information from customers to customize products according to their preferences while keeping the effort low. However, the IKEA effect contradicts these assumptions, suggesting that regardless of personal preferences, effort alone leads to a perceived increase in value (Norton et al., 2012). The so-called “I designed it myself” effect has also been shown in the context of MC toolkits, indicating that a feeling of accomplishment arises from creating a self-designed product, challenging the existing concept (Franke et al., 2010). While it is important in research on virtual co-creation tools for customers to contribute their own preferences in the development process, the IKEA Effect also occurs independently of this, even with simple products like IKEA cardboard boxes (Norton et al., 2012).

Furthermore, to the best of our knowledge, the IKEA effect has only been studied for physical products (e.g., Dohle et al., 2014; Norton et al., 2012) even though labor can also be invested into the creation of non-physical products or general content. In addition, the IKEA effect has been explored by comparing if people value their own creations more than similar objects created by others. However, the creation of physical as well as non-physical objects and content can also be performed by machines. Thus, we are interested if this will affect the perceived value of the results and whether the human effort involved in creating content or objects through extensive prompting will continue to be of great importance in the future.

#### 7.2.4 Hypothesizing the IKEA Effect in Human-AI Collaboration

Our study combines insights from psychology on the IKEA effect (e.g., Norton et al., 2012; Radtke et al., 2019) with the research fields of human-AI collaboration in the context of GenAI and suggests that contrary to current efforts to replace human activities by GenAI (Dwivedi et al., 2023; Teubner et al., 2023), humans need to contribute effort to collaborative development in order to fully value results (e.g., Marsh et al., 2022). Although the black box nature of GenAI makes it difficult for humans to assess how much of an impact they have on the final outcome and how the AI works (Berente et al., 2021; Dwivedi et al., 2023; Lund et al., 2023), we hypothesize that the IKEA effect will occur, increasing WTP as shown in other IKEA effect studies on physical products (e.g., Liu et al., 2023; Marsh et al., 2022; Norton et al., 2012). Norton et al. (2012) showed, for example, that the WTP for origami symbols was significantly higher for self-created symbols



compared to ones created by experts. In addition to studies showing that WTP increases for self-made objects, Walasek et al. (2017) further demonstrated that the price at which one would sell self-made products (in this case, assembling various science kits) is also significantly higher compared to the identical objects assembled by someone else. The fact that effort can increase WTP has already been shown for various physical products (e.g., Marsh et al., 2022; Walasek et al., 2017) and can also apply to content collaboratively created with GenAI if a lot of personal effort has been put into the collaboration. Moreover, a debate on WTP for GenAI arose with the introduction of ChatGPT (Capgemini, 2023), which we hope to contribute to. Thus, we propose:

**H1:** *Willingness to pay for AI-generated solutions increases when effort is put into collaborating with generative AI.*

Other studies have also shown that people often find it difficult to describe why they prefer a self-created product over the same product created by someone else and that this is often described as a personal feeling of *liking* or *appreciating* the self-created product or solution more (Dohle et al., 2014; Liu et al., 2023; Norton et al., 2012). Since we expect personal feelings to arise even with solutions that can be created by GenAI, such as images or texts, we propose:

**H2:** *Appreciation for AI-generated solutions increases when effort is put into collaborating with generative AI.*

Although research around the IKEA effect measures the perceived value of self-created products through WTP or abstract constructs such as *Liking*, *Niceness*, or *Appreciation* (e.g., Dohle et al., 2016; Liu et al., 2023; Norton et al., 2012), marketing and IS research proposes four dimensions that are positively associated with the overall *perceived value*, namely quality value, emotional value, value-for-money and social value (Turel et al., 2007). To measure whether the perceived value of AI-generated solutions increases as a result of investing effort, we build on the hypotheses of Turel et al. (2007). Within the realm of service-oriented marketing, research has demonstrated that superior quality assessments contribute to greater overall value (Baker et al., 2002; Brady & Robertson, 1999). Second, emotional components such as joy can enhance the formation of an overall value that individuals perceive, and third, users are price sensitive when evaluating the trade-off value-for-money (Turel et al., 2007). Lastly, individuals have the potential to improve their self-concept through the utilization of modern technologies, such as GenAI. This is because such technology can be perceived as cutting-edge and innovative, thereby signaling the user's affiliation with a specific social class (Schewe & Dillon, 1978). Thus, social value is also positively associated with the overall perceived value (Turel et al., 2007). Consistent with the IKEA effect (Norton et al., 2012), we propose that overvaluation of AI-generated solutions occurs when humans invest effort in collaborative development. This results in a higher perceived value (**H3**)

which is driven by quality value (**H3.a**), emotional value (**H3.b.**), value-for-money (**H3.c**), and social value (**H3.d**):

**H3:** *Perceived value of AI-generated solutions increases when effort is put into collaborating with generative AI.*

**H3.a:** *Quality value of AI-generated solutions increases when effort is put into collaborating with generative AI.*

**H3.b:** *Emotional value of AI-generated solutions increases when effort is put into collaborating with generative AI.*

**H3.c:** *Value-for-money of AI-generated solutions increases when effort is put into collaborating with generative AI.*

**H3.d:** *Social value of AI-generated solutions increases when effort is put into collaborating with generative AI.*

Based on this, Turel et al. (2007) demonstrated—without reference to the IKEA effect—that higher perceived value is a key determinant of behavioral usage intentions. They conclude that only if the benefits of a technology are clear, it will be utilized. Similarly, Kamtarin (2012) demonstrated that perceived value as a positive effect on online purchase intention. As we expect that people perceive a higher value of content generated through the interaction with GenAI, they will learn about the benefits and thus have a higher intention to reuse it. Therefore, we hypothesize that with a higher perceived value of AI-generated content, behavioral usage intentions for generative AI will increase:

**H4:** *Behavioral intention to use generative AI increases when effort is put into collaborating with the technology.*

### 7.3 Methodology

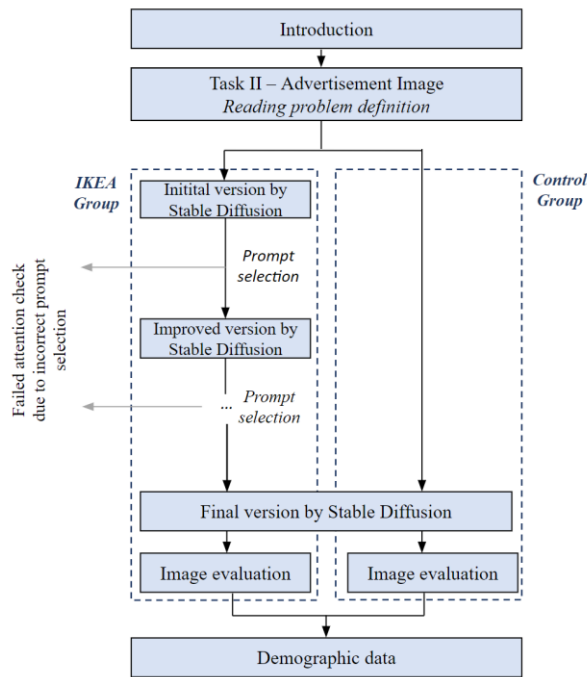
To answer our research question of whether the IKEA effect increases the perceived value of content created in collaboration with GenAI tools, such as ChatGPT or Stable Diffusion, we conducted an experiment, which is often utilized in IS research on human-AI collaboration (e.g., Fügner et al., 2022) as well as IKEA studies (e.g., Norton et al., 2012). Our goal is to investigate whether the IKEA effect exists for GenAI, meaning that people overvalue AI-generated content when they have put their own effort into a collaborative development process. We defined our target group as (working) individuals in Germany who might use GenAI now or in the future for their daily life/ work.

### 7.3.1 *Research Design and Measures*

For the experiment, we developed two tasks to collaboratively generate content with AI: one task where participants had to generate a text about a company mission statement for a drone start-up in the healthcare sector using ChatGPT and one task where the participants had to generate an image for advertising sustainably packaging apples using Stable Diffusion. GenAI can be used to create different types of content such as audio, code, images, text, simulations and videos. IKEA studies have already shown that the IKEA effect can occur with physical objects when creating pictures (Raghoobar et al., 2017). We decided to choose images and text for the overall design of the experiment because, firstly, both are content that humans can generate without the addition of technology and it is easier to assess the effort and value of the end result without specialist knowledge. Secondly, there are powerful solutions for generating both types of content on the market that are already widely used in everyday professional life. We have explicitly excluded code here, as functionality and lean implementation are the main focus when generating code and individual preferences are often given less importance. For both tasks we created an IKEA group, which allowed participants to put effort into the collaboration with the AI, and a control group respectively, which was confronted with the final results of the GenAI tools immediately after reading the problem definition. Participants were randomly assigned to the IKEA group for one task and then assigned to the control group for the other task. Therefore, each participant solved one task (image or text) for the IKEA group and the other task for the control group. Task I (text) and task II (image) and thus the sequence of participation in the IKEA and control group were randomized in order for each participant. The interaction part for the IKEA group of both tasks was structured the same (see Figure 15a).

First, participants were given the assignment, and then they were given a first, poor fitting output from the respective AI. Then, they were asked to adapt it which requires effort and were given four input options—two of them suggested improvements in relation to the problem definition, and the other two did not. Exemplary input options for the second round of the image task are provided in Figure 15b. The next output would better fulfil the problem definition but would still miss some crucial details. Thus, the participants could improve the output again. If they chose one option that did not make sense, they failed the attention check and were excluded from the experiment. Overall, they adapted the output three times until a final output was given. In addition, both tasks (text and image generation) were implemented in a version where no input selection for iteratively adapting the output was possible for the control group. Instead, the identical, final AI output was shown to the participants immediately after reading the problem definition. All images and text excerpts provided to participants throughout the experiment were created with Stable Diffusion and ChatGPT respectively, and pre-selected for this experiment by IS researchers to ensure equal experimental conditions for all participants.

The questionnaire was set up with Unipark, a survey tool emphasizing privacy and allowing us to include many different types of questions. After agreeing to a privacy policy and reading their rights according to GDPR, the experiment started with information on the task. Then, the participants were randomly assigned to one of the two groups for Task I. After each task, the participants were asked follow-up questions to measure the IKEA effect. We chose to leverage existing constructs and therefore conducted a structured literature review on the “IKEA effect” following (Brocke et al., 2009). Consequently, we first defined the review scope. Here, we focus on research methods, especially their used constructs to measure the IKEA effect. Our goal is to summarize them. The organization is conceptual and the perspective neutral. Our audience are general and specialized scholars, and our coverage is representative. In the second step, we conceptualized the topic by reading some general literature on the IKEA effect (e.g., Marsh et al., 2022; Norton et al., 2012). Here, we also set our inclusion criteria—a quantitative study was conducted with a measurement of the IKEA effect; and exclusion criteria—no measurement of the IKEA effect. Third, we conducted the literature search. To remain as broad as possible and also include papers, especially from IS, we used the search term “IKEA effect” in WebOfScience and AIS eLibrary. This resulted in 133 and 122 papers, respectively, with one duplicate. We also conducted a backward search to find the first paper and especially the definition of the IKEA effect (resulting in 247 publications) and a forward search which resulted in one additional paper (Turel et al., 2007). Fourth, we analyzed and synthesized the literature by first excluding all papers which did not include a quantitative measurement of the IKEA effect according to their title and abstract, resulting in 33 papers. We then carefully extracted the constructs used in the papers to measure the IKEA effect.

**Example task:**

You asked the AI to adjust the image and focus only on one type of fruit. The AI generated the following image:



Again, you can further modify the image. **Please specify how you want the AI to change the image.**

Please select the most suitable customization for your company and use case.

- “Please add additional vegetables to the image.”
- “Please make it explicit that the plastic is recyclable.”
- “Please cut the apples.”
- “Please generate less plastic by not packaging the apples individually.”

**Figure 15: a) Experimental setup for the image task and b) example task with AI collaboration. Participants followed the same procedure for the text generation task.**

This resulted in measuring the willingness to pay (one item “Are you willing to pay for the output of the AI? If yes, how much?”; Norton et al., 2012; Marsh et al., 2022; Liu et al., 2023), the appreciation (one item on a 7-point Likert scale: “How much do you appreciate the AI-generated output?”; Radtke et al., 2019), the perceived value (four constructs each four to five items on a 7-point Likert scale; e.g., “The quality of the AI-generated output is convincing.”; Turel et al., 2007), and the behavioral intention to use the technology (two items on a 7-point Likert scale; e.g., “Assuming that I have access to this GenAI, I would use it in the future.”; Turel et al., 2007). In addition, the effort is measured to prove effort manipulation between the groups was successful (one item on a 7-point Likert scale: “How much effort did you invest in creating the output with GenAI?”; Raghoobar et al., 2017). We adapted the constructs slightly to our context and translated them into German. Finally, the demographic data (age and gender) and control variables (AI experience (five items on a 7-point Likert scale, e.g., “Compared to most people, I know more about AI”) following (Flynn & Goldsmith, 1999) and AI attitude (seven items on a 7-point Likert scale, e.g., “A majority of society will benefit from AI in the future”) following Schepman & Rodway, (2020)) were asked. In addition, we ensured privacy, chose high-quality constructs, and randomized the options for the participants to prevent common method bias (CMB). Also, during the experiment, validity was ensured by using existing constructs and testing the survey and

reliability by using multi-item constructs if possible. In addition, we will check for both later in the analysis.

### 7.3.2 Data Collection and Sample

Before administering the experiment, we conducted a pretest with four experienced IS researchers. We changed some wording to make the questions easier to understand and improved some spelling mistakes. In addition, we asked three potential participants from our target group to review the experiment to get insights into unclear task descriptions and improvements. Afterwards, our target group was contacted through Prolific, a market research institute, and paid to participate in our experiment 16 €/hr. We therefore calculated the sample size following the formula that assuming the effect is small ( $d = 0.35$ ), we require 72 participants to show the effect at 90% for both groups. Thus, we contacted 200 people in our target group, assuming that at least 10% would fail the attention check and choose an option to improve the text or image that does not make sense.

Overall, the experiment was expected to take 15 minutes, and the participants required, on average, 9:34 minutes. 26 participants failed the attention check and were excluded. Other data was not excluded as the attention check was comprehensive enough, every question was marked as mandatory and no straight-liners were detected. Thus, we had a final sample size of 174 participants from Germany, of whom 89 were assigned to the IKEA group and 85 to the control group for the text task and vice versa for the image task. In our sample, most participants ( $n = 128$ ) were between 18 and 33 years old (74%), 32 participants were between 34 and 44 years old (18%) and the rest above 45. The gender was more balanced. We had 87 females, 84 males, and three participants identifying as divers. 68 participants are working full-time, 40 part-time, 5 are not in paid work (e.g., homemaker or retired) and 15 are unemployed but job-seeking. The other participants did not reveal their working status. Finally, participants reported an average prior AI experience (measured with five items on a 7-point Likert scale, Cronbach's alpha = .932) of 4.07 ( $SD = 1.45$ ) and an average attitude towards AI (measured with seven items on a 7-point Likert scale, Cronbach's alpha = .874) of 5.19 ( $SD = 1.27$ ).

## 7.4 Data Analysis and Results

We begin with checking the effort manipulation because only a perceived increase of effort can cause the IKEA effect (Mochon et al., 2012; Norton et al., 2012). For both tasks, the participants were asked about their own perceived effort (on a 7-point Likert scale) in creating the text and the image. Regarding the task of creating a company mission statement for a start-up in the healthcare sector using ChatGPT, the effort manipulation was not successful, as the IKEA group perceived the effort of the collaboration through four rounds of prompting was only slightly

higher on average ( $M = 2.77, SD = 1.06$ ) than the control group ( $M = 2.68, SD = 1.15$ ),  $t(169.02)^9 = -0.56, p = .290$ , with an effect size of  $d = 0.08$ , indicating a very small effect. The effort manipulation regarding the image creation for advertising sustainably packaging apples using Stable Diffusion on the other hand was successful, as the IKEA group reported significant higher perceived effort on average ( $M = 3.14, SD = 1.22$ ) than the control group ( $M = 2.52, SD = 1.17$ ),  $t(170.81) = 3.41, p < .001$ , with an effect size of  $d = 0.52$ , indicating a medium effect. Thus, in the following, we will only present the results of the image creation task for which effort manipulation was successful and not the text creation task.

To assess convergent validity, we evaluated that all latent variables are above the recommended threshold of .5 for average variance extracted (AVE) and composite reliability and Cronbach's alpha exceeded the threshold of .7 (Hair et al., 2017). Table 10 further outlines that all item loadings surpass the threshold of .7 and all constructs fulfil reliability and convergent validity.

**Table 10: Assessment of reliability and convergent validity (values are 1.000 for one-dimensional constructs)**

	<b>Factor Loadings</b>	<b>Composite Reliability</b>	<b>AVE</b>	<b>Cronbach's Alpha</b>
<b>Perceived Effort</b>	.768-.839	.785	.647	.850
<b>Appreciation</b>	1.000	1.000	1.000	1.000
<b>Quality Value</b>	.926-.968	.971	.894	.963
<b>Emotional Value</b>	.829-.921	.946	.779	.938
<b>Value-for-Money</b>	.879-.943	.951	.830	.950
<b>Social Value</b>	.761-.849	.891	.671	.894
<b>Behavioral Intention to Use</b>	.948-.955	.950	.905	.938

The results of our discriminant validity analysis are shown in Table 11. We verified that the square root of AVE (pictured on the diagonal; is 1.000 for one-dimension constructs) is greater than the interconstruct correlations (Gefen et al., 2000) and thus conclude that all constructs indicate sufficient discriminant validity (Fornell & Larcker, 1981).

First, homogeneity of variances was asserted using Levene's test. If equal variances could be assumed, unpaired t-tests were performed to test for differences between the IKEA group that invested effort into collaboration through the selection of appropriate prompts and the control

<sup>9</sup> As shown later, we first performed Levene's test for homogeneity of variances. If homogeneity of variances cannot be assured, instead of a t-test a robust Welch test is performed. To ensure significant p-values the degrees of freedom are automatically adapted in the Welch test.

group. Otherwise, Welch tests were performed. Due to the sample size of 30 or more for each group, the normal distribution of the data obtained can be assumed so that unpaired t-tests can be used as a robust measure (Stone, 2010).

**Table 11: Discriminant validity**

	Appreciation	Quality Value	Emotional Value	Value-for-Money	Social Value	Behavioral Intention to Use
Appreciation	<b>1.000</b>					
Quality Value	.871	<b>.945</b>				
Emotional Value	.801	.856	<b>.882</b>			
Value-for-Money	.694	.744	.762	<b>.911</b>		
Social Value	.519	.526	.574	.524	<b>.819</b>	
Behavioral Intention to Use	.713	.657	.713	.661	.600	<b>.952</b>

**H1: Willingness to pay.** We found a significant difference in WTP (yes or no) for the group that collaboratively created images through prompting and the control group,  $t(172) = 3.98, p < .001, d = 0.48$ . While this effect is of small size, we further explored differences in the WTP for a monthly license. The results showed that the average price (in Euro) participants were willing to pay was significantly higher in the group that iteratively interacted with GenAI ( $M = 18.58, SD = 62.58$ ) compared to the control group ( $M = 5.01, SD = 10.68$ ),  $t(172) = 2.02, p = .045$  with an effect size of  $d = 0.30$ , indicating a medium effect. Thus, we conclude that **H1** is supported by the findings of this study, and WTP for images generated by AI is significantly higher if effort is invested into collaborating with the GenAI.

**H2: Appreciation.** As shown by Norton et al. (2012), it is to be expected that with higher effort, the *appreciation* or *liking* of created objects increases. The findings of this study show that this bias also applies to the collaborative creation of content, such as images with a GenAI. Participants of the IKEA group reported a significantly higher average appreciation of the final image ( $M = 4.88, SD = 1.35$ ) than the control group ( $M = 3.71, SD = 1.60$ ),  $t(172) = 5.18, p < .001$ , with an effect size  $d = 0.79$ , indicating a medium effect, even though they evaluated the same image. Therefore, we successfully demonstrated the IKEA effect in our online experiment, and the result supports **H2**.

**H3: Perceived value.** While appreciation or liking of an object or content would rather be described as a personal feeling, perceived value, which is a key determinant for behavioral



intentions, can be measured more objectively through four constructs (quality, emotional, value-for-money, and social value) as proposed by Turel et al. (2007). We first examine H3.a through H3.d before deriving an overall assessment of whether perceived value increases with effort.

**H3.a: Quality value.** Participants of the IKEA group report a significantly higher quality value for the generated solution ( $M = 4.77$ ,  $SD = 1.32$ ) than the control group ( $M = 3.70$ ,  $SD = 1.55$ ),  $t(172) = 4.90$ ,  $p < .001$ . In addition, this effect is  $d = 0.74$ , indicating a medium effect. We thus conclude that **H3.a** is supported, and the quality value of an AI-generated solution increases if effort is put into collaborative creation.

**H3.b: Emotional value.** Besides the quality value, participants of the IKEA group further reported a significantly higher emotional value, which they attached to the created final image ( $M = 4.94$ ,  $SD = 1.19$ ) than the control group ( $M = 3.76$ ,  $SD = 1.40$ ),  $t(169.71) = 5.97$ ,  $p < .001$ ,  $d = 0.91$ , indicating a large effect. This finding supports **H3.b** and the emotional value significantly increased with effort in our experiment.

**H3.c: Value-for-money.** We further asked all participants how they would rate the value-for-money if a monthly license would cost 15 euros. Value-for-money received significantly better ratings by the IKEA group ( $M = 4.46$ ,  $SD = 1.48$ ) than the control group ( $M = 3.50$ ,  $SD = 1.48$ ),  $t(171.61) = 4.29$ ,  $p < .001$ ,  $d = 0.65$ , indicating a medium effect which supports **H3.c**.

**H3.d: Social value.** Lastly, the social value of the created solution was assessed. Participants of the IKEA group attributed a significantly higher social value to their creation ( $M = 4.21$ ,  $SD = 1.23$ ) than the control group ( $M = 3.72$ ,  $SD = 1.26$ ),  $t(171.95) = 2.80$ ,  $p = .006$ ,  $d = 0.39$ , indicating a small effect. Therefore, in conclusion, we can say that the social value has also increased with effort, and the finding thus supports **H3.d**.

**Table 12: Hypotheses testing**

Hypothesis	t-Value	p-Value	Cohen's d	Outcome
H1: Effort—WTP	2.02	.045	0.30	Supported
H2: Effort—Appreciation	5.18	< .001	0.79	Supported
H3: Effort—Perceived Value	-	-	-	Supported by H3.a - H3.d
H3.a: Effort—Quality Value	4.90	< .001	0.74	Supported
H3.b: Effort—Emotional Value	5.97	< .001	0.91	Supported
H3.c: Effort—Value-for-Money	4.29	< .001	0.65	Supported
H3.d: Effort—Social Value	2.80	.006	0.39	Supported
H4: Effort—Behavioral Intention to Use	3.88	< .001	0.59	Supported

Since all four core determinants of perceived value are significantly higher for the IKEA group that invested effort in generating the image with the AI, we conclude that perceived value increases with effort, and **H3** is supported.

**H4: Behavioral intention to use.** Perceived value is seen as a driver for the adoption of a system or a general technology (e.g., Turel et al., 2007). We therefore examined if the behavioral intention to use the technology (GenAI for image creation) increases if participants put effort into collaborating with the technology. Interestingly, the behavioral intention to use GenAI also significantly increased for the IKEA group and thus with effort ( $M = 4.95$ ,  $SD = 1.52$ ) compared to the control group ( $M = 4.08$ ,  $SD = 1.45$ ),  $t(170.49) = 3.88$ ,  $p < .001$ ,  $d = 0.59$ , indicating a medium effect. We thus conclude that **H4** is supported.

## 7.5 Discussion

Previous research on the IKEA effect does not address whether the IKEA effect can also be observed for non-physical content and, in particular, GenAI. Especially as the results of AI are unpredictable and, to some extent, cannot be generated by humans in the same way (Berente et al., 2021; Dwivedi et al., 2023), findings from the literature are not transferable. With our experiment, we aim to determine if the IKEA effect can be found in order to be considered in the design and adoption of AI systems. Although AI seeks to automate processes, customer experience strategies of companies such as Adidas show that the contribution of effort by the customer or end user is essential so that collaboration produces valuable results, solutions, or products.

In order to determine the IKEA effect in collaborative content creation with GenAI, we divided our online experiment participants into two groups. The IKEA group was able to co-create solutions in an iterative development process that required human effort using the AI tool. In contrast, the control group received the generated AI solution immediately. Participants were randomly assigned to one of the two groups (IKEA or control group) for the image task and then assigned to the other group for the text task. The tasks were presented in a way that both groups received the identical final output, regardless of whether the task was solved for the IKEA or the control group. Unfortunately, we did not notice any difference in effort in the text generation task, which led to the IKEA effect not being evident, a finding that has also been made for physical products (Mochon et al., 2012; Norton et al., 2012). Based on feedback from the participants, we assume that the difference between the effort collaborating with AI and receiving the final output was not observed here is simply due to the fact that it is significantly more exhausting for humans to read and evaluate texts than images. Therefore, we focused our evaluation on the image generation task where effort manipulation was successful.

Various methods can be found in the literature to measure the IKEA effect. One of the most popular is the measurement of the WTP. We found a strong significant difference between the IKEA and the control group, so that H1 is supported. Thus, we can explain to a certain degree that the WTP for tools like ChatGPT is so high, because people can invest effort in the collaboration and create their own output. Likewise, we can assume H2—the appreciation of the output since a significant difference exists in the fact that the participants prefer the final version, where they themselves have invested more effort in the development process. We believe this can be due to the labor invested, causing either a signal of competence, effort justification, or feelings of ownership. Finally, we measured the perceived value of generated content through the four constructs: quality value, emotional value, value-for-money, and social value proposed by Turel et al. (2007). Also, these four constructs show a significant difference between the IKEA and the control group, so we can answer our first research question with *human effort invested in collaboration with GenAI increases the perceived value of AI-generated solutions*. Furthermore, RQ2, whether *perceived value also increases the behavioral intention to use GenAI technology* can be confirmed with H4. The behavioral intention to use GenAI tools, as in our example Stable Diffusion, can thus increase over time if the IKEA effect is exploited.

### 7.5.1 Contributions

With our experiment, we are able to make several theoretical contributions: First, by confirming the hypotheses, we can show that the IKEA effect can occur not only for physical products such as origami, food, or furniture (e.g., Dohle et al., 2014; Ling et al., 2020; Mochon et al., 2012; Norton et al., 2012), but also for content produced by GenAI. As (generative) AI has unique characteristics (Bauer, Hinz, et al., 2021; Dwivedi et al., 2023), this effect is not self-evident. However, a sufficient difference in perceived effort must exist for the IKEA effect to manifest in using GenAI. The IKEA effect only occurs when people perceive that they have achieved and contributed to something. This can most likely also be generalized to former AI use and leads to our second theoretical contribution. With the ability of GenAI to create new content instead of solely making decisions, we show with our findings that human-AI collaboration needs to be rethought, and we add to the goals that should be pursued in human-AI collaboration. Humans now find themselves in a new role: Rather than just receiving final content or decisions, they can participate in the outcome and co-develop the result with their input. While it remains unclear what effect a prompt has on the output of GenAI, we show that collaboration is essential for human end users to value results. Here, new XAI approaches may gain importance, which face new challenges, especially with regard to image generation for example. This also changes the way humans are considered in research: With regard to the black box character, it is important to explore how humans can cooperate with AI although they may not fully understand the functionality. Even though bias is

often perceived as negative, thirdly, we believe that the IKEA effect can also be leveraged. To increase the value of the outcome for users, they could be integrated artificially into the process. The IKEA effect can be utilized by developing suitable interfaces and interaction opportunities. Furthermore, our research offers practical contributions: First, we can find a higher WTP and behavioral intention to use for GenAI with human participation. This can be exploited by the providers of such services. It may also explain to some extent the high WTP for the monthly license of ChatGPT and the rapidly growing number of users. Therefore, (generative) AI providers should adapt strategies of customizing from brands like Adidas and Build-A-Bear instead of exclusively selling automation solutions to customers. These remain necessary for many tasks, especially when monotonous or dangerous tasks can be substituted. Still, also due to legal reasons, creative work or decisions in high-stakes environments are likely to remain with humans in the foreseeable future. But our findings are also valuable for the users of GenAI. Organizations planning to implement such solutions should primarily choose tasks that require collaborative approaches. Otherwise, employees might fear losing their jobs and the acceptance of these solutions could decrease significantly. By exploiting the IKEA effect, we anticipate the opposite: In collaboration between humans and machines, the perceived value of the end product increases so that the behavioral intention to use increases, which can lead to higher acceptance. Finally, we can recommend potential users of GenAI to give it a try and discover new ways of facilitating work by contributing to and shaping the outcome of AI.

### 7.5.2 *Limitations and Future Research*

While our study has added surprising insights into the realm of the IKEA effect in the context of GenAI, a few limitations must be acknowledged and further provide a basis for future research. First, the online experiment was conducted exclusively with German participants, and respective cultural norms, attitudes, and experiences may influence individual perceptions and biases. In addition, the majority of participants were between 18 and 33 years old, and the results on the IKEA effect may not be transferable to children and older people. While the results are thus relevant for most of the current users of GenAI, it is still interesting to understand how and whether the IKEA effect affects human-AI collaboration at different stages of users' lives. Second, our experimental design did not allow participants to have a "real" interaction with Stable Diffusion. While this was intentionally done to ensure uniformity in the results across participants and to trigger the IKEA effect, it is worth noting that the interaction was simulated. Nevertheless, the feedback from participants indicates that they believed they were indeed interacting with an actual GenAI, which speaks to the validity of our design. Third, the experiment focused on image generation. The realm of GenAI spans far beyond image creation, including outputs like code, music, etc. The occurrence of the IKEA effect in these contexts poses a topic for future research.

Lastly, while the phenomenon of the IKEA effect was observed in the context of GenAI, the reasons for its occurrence remain unclear. Previous research on the IKEA effect has pointed to factors like the signal of competence, effort justification, and feelings of ownership as potential underlying causes. Also, simple factors such as the amount of time spent on a task could influence the perceived effort. Future research could delve deeper into understanding the specifics of why this effect is manifested, especially in the context of AI-generated content. In addition, future research should investigate whether the IKEA effect, in the context of GenAI, varies across different cultures. Moreover, our findings underline the potential influence of the IKEA effect on user perceptions. Future research should focus on how this effect can be factored into the design of GenAI tools. Incorporating psychological insights might enhance the perceived value of AI-generated solutions and human-AI collaboration in the future. In conclusion, we encourage scholars to build on this work, further unraveling the intricate relationship between humans and GenAI.

## **7.6 Acknowledgement**

This research and development project is/was funded by the German Federal Ministry of Education and Research (BMBF) within the “Innovations for Tomorrow’s Production, Services, and Work” Program (funding number 02L19C150) and implemented by the Project Management Agency Karlsruhe (PTKA). The authors are responsible for the content of this publication.

## 8. Discussion of Contributions and Conclusion

To effectively use ML systems to augment human capabilities and enhance decision making in high-risk domains such as healthcare or aviation, challenges in collaboration between ML systems and human decision makers must be overcome (e.g., Fügener et al., 2022; Lebovitz et al., 2022; Maedche et al., 2019; Seeber et al., 2020). It is crucial to recognize that humans will remain an integral part of the decision-making chain in high-risk domains, and it is increasingly important to develop an understanding of the impact of ML systems on human decision makers (e.g., Gaube et al., 2023; Jussupow et al., 2021; Lebovitz et al., 2021; Maedche et al., 2019; Seeber et al., 2020). In this context, research calls for exploring novel approaches on how humans can learn from ML systems (e.g., Abdel-Karim et al., 2023, 2020; Gaube et al., 2023; Pumplun et al., 2023; Sturm, Gerlach, et al., 2021), which can generate new knowledge during training (e.g., Brynjolfsson & Mitchell, 2017; Fügener et al., 2021). In addition, research should consider not only humans as learners, but also the ML systems themselves, which are trained on static datasets and therefore must continuously adapt to dynamic problem perceptions and deployment environments (e.g., Asatiani et al., 2021; Sculley et al., 2015; Sturm, Gerlach, et al., 2021; Sturm, Koppe, et al., 2021). However, there is a lack of guidance on how to design ML systems that can continuously learn to maintain or even improve their performance. Finally, characteristics such as error-proneness, inscrutability, and shifting roles in collaboration with the increasing autonomy of ML systems pose significant challenges for effective teaming between humans and ML systems (e.g., Berente et al., 2021; McNeese et al., 2018; Seeber et al., 2020; Vössing et al., 2022).

This dissertation aims to expand our understanding of the learning potential and effective collaboration between humans and ML systems by examining these research gaps from three perspectives: (1) the *human*, (2) the *ML system*, and (3) the *collaborative perspective*. The theoretical and practical contributions of the five papers included in this dissertation are discussed below. Future research directions and limitations are provided in the respective papers.

### 8.1 Theoretical Contributions

Paper A addresses RQ1 and explores whether human decision makers, such as radiologists, can learn from collaborating with ML systems, thus responding to several research calls (e.g., Abdel-Karim et al., 2023, 2020; Gaube et al., 2023; Lebovitz et al., 2021; Pumplun et al., 2023; Sturm, Gerlach, et al., 2021). The mixed-methods study demonstrates that radiologists can improve their performance and decision confidence through collaboration with high-performing ML systems. However, the study also shows that collaboration with low-performing ML systems poses the risk of false learning, which can lead to a decline in the performance of human decision makers. Paper A can serve as a basis for expanding human-in-the-loop concepts (e.g., Fügener et al., 2021;

Grønsund & Aanestad, 2020; Seidel et al., 2018; Sturm, Koppe, et al., 2021) and informs about both opportunities and risks for decision makers. In addition, the study contributes to the growing XAI research stream (e.g., Bauer, Hinz, et al., 2021; Lebovitz et al., 2022; Miller, 2019; Pumplun et al., 2023; Reyes et al., 2020) by showing how explainable output design of ML systems can significantly improve learning outcomes and prevent false learning in the case of incorrect system output. Interestingly, some radiologists were even able to learn from the ML systems' mistakes if they were presented in an explainable manner. The findings pave the way for the development of new adoption strategies for ML systems in high-risk domains by preventing harmful effects on the decision maker through explainable output design. In addition, the study adds to the goals that should be pursued with XAI methods and links the value of explanations in educational research (Crowley & Siegler, 1999; Fender & Crowley, 2007) with explanations from the XAI field (Bhatt et al., 2020; Meske et al., 2022). Finally, the study shows that the results are broadly applicable, as both highly experienced and novice radiologists can learn from ML systems, and only a small decrease in learning success is observed with increasing experience. Consequently, XAI research should prioritize the identification of individual end user needs and consider their prior experience.

Paper B turns to the ML system perspective, addressing RQ2. The DSR study investigates how ML systems should be designed and developed along the CRISP-ML(Q) process model (Studer et al., 2021) to flexibly adapt to changes in their environment and learn from new insights and changing problem perceptions of human decision makers after deployment (e.g., Asatiani et al., 2021; Grønsund & Aanestad, 2020; Sturm, Gerlach, et al., 2021). This enables maintaining ML system performance and addresses several IS research calls (e.g., Grønsund & Aanestad, 2020; Sturm, Gerlach, et al., 2021). The study identifies challenges for the long-term use of ML systems and underscores the need for continuous system maintenance. Based on this, design requirements and principles for all phases of the CRISP-ML(Q) process are derived, serving as a blueprint for the development of sustainable ML systems suitable for long-term use. The holistic approach also shows that it is not sufficient to intervene only after an ML system has been deployed. Instead, continuous learning of ML systems and adaptation to the problem perception of human decision makers requires that these needs be considered early in the ML system development process. This study thus provides a further basis for the expansion of human-in-the-loop approaches, that intervene only after deployment, although future changes in human problem perception should also be taken into account during the development of ML systems.

Papers C to E address RQ3 and offer several theoretical contributions for the design of effective human-ML system collaboration. Paper C provides insights into how different types of errors (FNs and FPs) of ML systems affect human decision makers. Previous research has examined the impact of erroneous output from ML systems on humans, but has not differentiated between error types

(e.g., Abdel-Karim et al., 2020; Fügener et al., 2021; Jussupow et al., 2021; Lebovitz et al., 2022). Here, paper C provides new insights, showing that both types of errors negatively affect decision maker performance and trust, but that these negative effects are more severe for FNs in the context of the experiment. Providing explanations from the XAI field to improve the explainability of ML system output can mitigate these effects, particularly for FPs. At the same time, however, the mental workload of decision makers increases when processing additional information provided by explanations. By illustrating the cost to the human decision maker of different ML errors, this can serve as a basis for aligning ML systems according to error management theory to minimize the most costly error type (Haselton & Nettle, 2006; Johnson et al., 2013). The findings further indicate that a trade-off must be found in the sensitivity and explainability of ML systems to enable effective and high-performing collaboration without overburdening the human decision maker. Paper D further highlights how the increasing autonomy of ML systems also enables non-experts to safely and intuitively collaborate with the technology, contributing to the growing research field of human-autonomy teaming (e.g., McNeese et al., 2021, 2018; Seeber et al., 2020; Zercher et al., 2023). The study also underscores the need to rethink the design requirements for the explainability of ML systems at increasing levels of autonomy and collaboration with non-experts. Finally, acceptance issues for human-ML collaboration remain a significant challenge (Pumplun et al., 2021; Seeber et al., 2020). In this context, paper D shows how the iterative, user-centered design of ML artifacts based on the UTAUT theory can promote end-user acceptance. Paper E extends the findings on the impact of the level of autonomy of ML systems on the success of collaboration, clarifying that people who invest effort in the collaborative generation of content perceive a higher value in the created solutions. Known as the IKEA effect (e.g., Norton et al., 2012), the influence of effort invested in creation on perceived value has been demonstrated for a variety of physical products (e.g., Dohle et al., 2014; Ling et al., 2020; Mochon et al., 2012; Norton et al., 2012). The study shows that despite the unique characteristics of (generative) ML systems (Bauer, Hinz, et al., 2021; Dwivedi et al., 2023), the effort that people put into collaboration is also essential for valuing solutions and content developed in collaboration with ML systems. The IKEA effect should be considered when designing collaboration between humans and ML systems to ensure that the results are appreciated and utilized. Ultimately, the results demonstrate why designing effective collaboration is preferable to an automation approach in certain areas.

Overall, the papers in this dissertation extend our understanding of the impact of ML systems on human decision makers and how this can be effectively leveraged in collaboration by focusing on the explainability of ML systems and the potential for both humans and ML systems to learn from each other through collaboration in the long term. The findings provide a broad foundation for designing more effective collaboration between humans and ML systems and lowering the barriers to adoption of ML systems, even in high-risk areas.



## 8.2 Practical Contributions

The five papers included in this dissertation provide not only theoretical, but also practical contributions that enable organizations to design ML systems and their collaboration with humans to make better decisions in the long run. In addition, a foundation is laid to enable continuous improvement of both humans and ML systems through collaboration.

Paper A addresses RQ1 and demonstrates how the deliberate use of explainable design in ML systems can enable human decision makers to learn from these systems and recognize ML errors. Using tumor segmentation in radiology as an example, the study provides a blueprint for the design of an ML-based DSS that can significantly improve learning outcomes for radiologists during collaboration and prevent false learning by providing local output explanations. ML developers can build on the design and insights to develop ML systems that can be adopted by organizations in high-risk areas at an earlier stage: While insufficient accuracy often prevents the adoption of ML systems in fields such as healthcare or aviation due to the severe consequences of incorrect decisions, real data can only be collected after deployment to improve performance over time. Our study shows how this chicken-and-egg problem can be overcome in practice by using explainable design to enable human decision makers to detect ML errors and thus allow for early, low-risk adoption in practice. In addition to developing early adoption approaches for explainable ML systems, organizations can use the insights to strategically deploy ML systems to create new learning opportunities for their human members. For example, ML systems could be used specifically as a training tool for novice employees for tasks that require humans in the decision-making chain in the long term.

Beyond the learning potential for humans in collaboration, Paper B addresses the need for ML systems to learn from humans and their dynamic environments over time in order to maintain or even improve their performance. In response to RQ2, Paper B outlines the challenges for the long-term deployment of ML systems. These challenges illustrate why many ML systems remain in the prototype phase and ML projects often fail in long-term deployment (e.g., Deloitte, 2020; Metternich et al., 2021). However, the long-term deployment of ML systems is crucial for a sustainable ML lifecycle, as the development process is often associated with high resource consumption (Klöpffer, 2003; van Wynsberghe, 2021). The derived DRs and DPs for developing and operating ML systems that are adaptive and can continuously learn from their environment provide organizations with guidance for more sustainable ML system development. Paper B also shows that it is essential for organizations to consider future changes, e.g., in human problem perception or environmental conditions that the systems will be exposed to, even in the early development phases of ML systems. In addition, the challenge-design-requirement framework presented can help organizations ensure that ML systems can maintain their performance,

enabling safe deployment even in high-risk areas, and that the competitive advantages associated with the use of ML systems can be leveraged over the long term.

Papers C, D, and E, on the other hand, offer several practical contributions for designing effective collaboration between humans and ML systems. Paper C demonstrates the impact of different ML error types on the performance, trust, and mental workload of human decision makers. The study emphasizes that explanations for ML output can mitigate performance and trust degradation in the case of FPs, but also notes that additional explanations can lead to an increase in mental workload. Organizations can use these insights to optimize the sensitivity of ML systems to minimize the frequency of more costly ML errors (FPs vs FNs), depending on the task and end-user needs. This also allows the introduction of early adoption approaches in high-risk environments: Especially in situations with low mental workload, sensitive and explainable ML systems can be deployed early, as FPs are better recognized and the risk of performance degradation and loss of trust is significantly reduced. Ultimately, the study provides a blueprint for developing an explainable ML-based cockpit system that can assist pilots in detecting other aircraft, and whose design can be used by organizations and decision makers in the general aviation field. The increasing level of autonomy of ML systems (e.g., Berente et al., 2021; McNeese et al., 2018) also enables collaboration with non-experts. Paper D presents DRs and DPs for the design of an artifact for human-autonomy teaming in the aviation industry and involves non-experts as a user group in an iterative design process. The findings inform ML developers and organizations that want to deploy ML systems in collaboration with non-experts about how to adapt approaches to improve explainability depending on the user group and level of autonomy. Furthermore, the DRs and DPs address the acceptance problems in the ML field (e.g., Pumplun et al., 2021; Seeber et al., 2020) and show how the design of an artifact can be leveraged as an acceptance-promoting interface between humans and ML systems. In addition, paper E provides practical suggestions for effective human-ML system collaboration and demonstrates that people rate the perceived value of collaboratively generated content significantly higher when a lot of personal effort is invested in the collaboration. The study also finds a significantly higher WTP for ML systems that require a high amount of personal effort on the part of the user to collaborate. Providers of ML systems and services can use these insights to develop new collaboration strategies that increase their customers' WTP. Moreover, the study shows that organizations should use ML systems for task areas that require collaborative work to improve the valuation of the created solutions. This approach may also be advantageous over an automation approach in some areas.

Overall, the papers in this dissertation provide practical guidance on how to design ML systems so that both humans and ML systems can learn from each other in collaboration. The approaches

presented, especially from the XAI domain, allow human decision makers to benefit from the collaboration and facilitate the development of low-risk early adoption strategies for ML systems.

### 8.3 Concluding Remarks

The deployment of ML systems in high-risk areas has the potential to support human decision makers in a variety of complex tasks to improve decision performance. However, it is essential to deepen the understanding of the mutual influence between ML systems and human decision makers in order to implement effective collaboration (e.g., Abdel-Karim et al., 2023; Gaube et al., 2023; Jussupow et al., 2021; Lebovitz et al., 2021; Maedche et al., 2019; Pumplun et al., 2023; Seeber et al., 2020). This dissertation includes a mixed-methods study, two qualitative DSR studies, and two quantitative online experiments in the contexts of healthcare, aviation, and industry 4.0 to investigate the potentials and risks that arise in the collaboration between humans and ML systems. A key objective is to explore the potential for mutual learning to enhance the capabilities and skills of both parties in the collaboration. In this context, approaches from the XAI field are examined, and the developed design frameworks provide guidance for incorporating the findings on learning potential, the influence of explainability, and occurring cognitive biases advantageously in the design of ML systems and in their collaboration with humans. In addition to identifying learning potential, the study also shows how XAI approaches can significantly reduce the risks posed to human decision making by the unique characteristics of ML systems, such as inscrutability and opacity (Asatiani et al., 2021; Berente et al., 2021; Lebovitz et al., 2022). Based on the findings, avenues for future research can be derived. While the papers already present specific suggestions for future research, overarching directives for future research for the three perspectives considered in this dissertation are outlined in the following. First, the *human perspective* (RQ1) shows that humans can learn from ML systems and that explanations from the XAI field can improve learning outcomes and prevent false learning. Future research can build on these findings to expand our understanding of the potential of XAI approaches for end users and to develop concrete design guidelines for explanations that promote human learning. While papers A and C provide local explanations for end users such as radiologists and pilots to improve explainability, future research should also incorporate global and model explanations (Pumplun et al., 2023; Rai et al., 2019). The development of design guidelines for XAI approaches should also take into account the area of application and the user's prior knowledge, and explore how the collaboration experience can be customized for the end user's individual learning needs. Ultimately, insights into the influence of ML systems and their explainability on cognitive processes in decision making need to be expanded (Abdel-Karim et al., 2023; Bauer, Hinz, et al., 2021). The experiments on human learning in this dissertation focus primarily on the influence

on short-term memory. Future research should build on this to investigate the influence of long-term collaboration with ML systems in this context.

Second, from an *ML system perspective* (RQ2), a holistic approach was presented that identifies challenges and associated DPs for all phases of the ML system development process to enable long-term deployment and maintenance of ML systems that can learn from humans and their environment. This approach shows that human-in-the-loop concepts (e.g., Grønsund & Aanestad, 2020; Seidel et al., 2018) should be expanded, as the prerequisites for flexible adaptability in deployment must be created in the early development stages of systems, and current concepts that focus exclusively on ongoing operations fall short. Furthermore, it is important for future research to investigate how human problem perception changes and how dynamic changes in the ML system environment affect system performance, in order to define future system requirements more proactively. In addition, the future design of interfaces will be central to informing human decision makers about changes that may affect system performance, and enabling them to provide continuous feedback to ML systems in the context of collaboration.

From a *collaborative perspective* (RQ3), it was shown that ML errors can lead to performance and trust losses on the part of the human decision maker. Although explainable output design can mitigate these effects, it is important to develop strategies for rebuilding trust after an ML system error and to explain what led to the error. In addition, appropriate measures should be derived that enable humans and ML systems to learn from each other's occasional incorrect decisions. Furthermore, the mental workload of end users should be increasingly taken into account in the development of explainable ML systems and in the selection of suitable use cases in organizations. IS researchers should investigate what types and complexities of tasks are suitable for collaboration with ML systems, and how and what kind of explanations for the output can be provided without overburdening the decision maker. How to adapt and provide explanations for ML performance depending on the situation should also be explored in future research. Moreover, non-experts, who will increasingly use ML systems without domain or ML knowledge in the coming years, should also be considered as an end-user group. Finally, this dissertation shows that cognitive biases affect not only the collaboration with ML systems, but also the valuation of jointly generated solutions. Future research is needed to further explore the causes of these cognitive biases and how they can be beneficially incorporated into collaboration.

In high-risk sectors, organizations often face challenges when adopting ML systems. In this dissertation, I aim to lay the groundwork for the safe and reliable deployment of this technology by deepening our understanding and promoting effective collaboration between humans and ML systems. In particular, I want to highlight the significant benefits that explainable ML systems offer to end users. It is my intention that this work will inspire further research into the role of

---

explainability as a catalyst for the wider adoption of ML systems, ensuring that these technologies not only meet technical expectations, but are also closely aligned with end-user needs.

## References

- Abdel-Karim, B. M., Pfeuffer, N., Carl, K. V., & Hinz, O. (2023). How AI-based Systems Can Induce Reflections: The Case of AI-augmented Diagnostic Work. *MIS Quarterly*, 47(4), 1395–1424. <https://doi.org/10.25300/MISQ/2022/16773>
- Abdel-Karim, B. M., Pfeuffer, N., Rohde, G., & Hinz, O. (2020). How and What Can Humans Learn from Being in the Loop? *KI - Künstliche Intelligenz*, 34(2), 199–207. <https://doi.org/10.1007/s13218-020-00638-x>
- Adadi, A., & Berrada, M. (2018). Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access*, 6, 52138–52160. <https://doi.org/10.1109/ACCESS.2018.2870052>
- Adidas. (2023). *Personalisable*. adidas DE. <https://www.adidas.de/personalisable> (accessed on Oct 21, 2023)
- Ahsen, M. E., Ayvaci, M. U. S., & Raghunathan, S. (2019). When Algorithmic Predictions Use Human-Generated Data: A Bias-Aware Classification Algorithm for Breast Cancer Diagnosis. *Information Systems Research*, 30(1), 97–116. <https://doi.org/10.1287/isre.2018.0789>
- Alavi, M., Leidner, D. E., & Mousavi, R. (2024). Knowledge Management Perspective of Generative Artificial Intelligence. *Journal of the Association for Information Systems*, 25(1), 1–12. <https://doi.org/10.17705/1jais.00859>
- Alex, C., & Vijaychandra, A. (2016). Autonomous Cloud Based Drone System for Disaster Response and Mitigation. *2016 International Conference on Robotics and Automation for Humanitarian Applications (RAHA)*, 1–4. <https://doi.org/10.1109/RAHA.2016.7931889>
- Allen, R., & Mazumder, M. (2020). Toward an Autonomous Aerial Survey and Planning System for Humanitarian Aid and Disaster Response. *2020 IEEE Aerospace Conference*, 1–11. <https://doi.org/10.1109/AERO47225.2020.9172766>
- Amazon. (2016). *Amazon PrimeAir*. <https://www.amazon.com/-/de/Amazon-Prime-Air/b?ie=UTF8&node=8037720011> (accessed on Mar 10, 2022)

- Amazon Prime Air. (2021). *Airborne Object Tracking Challenge*.  
<https://www.aicrowd.com/challenges/airborne-object-tracking-challenge> (accessed on Sep 19, 2023)
- Amershi, S., Begel, A., Bird, C., DeLine, R., Gall, H., Kamar, E., Nagappan, N., Nushi, B., & Zimmermann, T. (2019). Software Engineering for Machine Learning: A Case Study. *2019 IEEE/ACM 41st International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP)*, 291–300. <https://doi.org/10.1109/ICSE-SEIP.2019.00042>
- Arkes, H. R., & Mellers, B. A. (2002). Do Juries Meet Our Expectations? *Law and Human Behavior*, 26(6), 625–639. <https://doi.org/10.1023/a:1020929517312>
- Arnott, D. (2006). Cognitive Biases and Decision Support Systems Development: A Design Science Approach. *Information Systems Journal*, 16(1), 55–78.  
<https://doi.org/10.1111/j.1365-2575.2006.00208.x>
- Aronow, P. M., Baron, J., & Pinson, L. (2019). A Note on Dropping Experimental Subjects who Fail a Manipulation Check. *Political Analysis: An Annual Publication of the Methodology Section of the American Political Science Association*, 27(4), 572–589.  
<https://doi.org/10.1017/pan.2019.5>
- Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges Toward Responsible AI. *An International Journal on Information Fusion*, 58, 82–115.  
<https://doi.org/10.1016/j.inffus.2019.12.012>
- Asatiani, A., Malo, P., Nagbøl, P. R., Penttinen, E., Rinta-Kahila, T., & Salovaara, A. (2021). Sociotechnical Envelopment of Artificial Intelligence: An Approach to Organizational Deployment of Inscrutable Artificial Intelligence Systems. *Journal of the Association for Information Systems*, 22(2), 325–352. <https://doi.org/10.17705/1jais.00664>
- Audzeyeva, A., & Hudson, R. (2016). How to Get the Most From a Business Intelligence Application During the Post Implementation Phase? Deep Structure Transformation at a U.K. Retail Bank. *European Journal of Information Systems*, 25(1), 29–46.  
<https://doi.org/10.1057/ejis.2014.44>
- Aydin, B. (2019). Public Acceptance of Drones: Knowledge, Attitudes, and Practice. *Technology in Society*, 59, 101180. <https://doi.org/10.1016/j.techsoc.2019.101180>
- Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J. S., Freymann, J. B., Farahani, K., & Davatzikos, C. (2017). Advancing the Cancer Genome Atlas Glioma MRI Collections With

- Expert Segmentation Labels and Radiomic Features. *Scientific Data*, 4, 170117.  
<https://doi.org/10.1038/sdata.2017.117>
- Baker, J., Parasuraman, A., Grewal, D., & Voss, G. B. (2002). The Influence of Multiple Store Environment Cues on Perceived Merchandise Value and Patronage Intentions. *Journal of Marketing*, 66(2), 120–141. <https://doi.org/10.1509/jmkg.66.2.120.18470>
- Balakrishnan, J., Dwivedi, Y. K., Hughes, L., & Boy, F. (2021). Enablers and Inhibitors of AI-Powered Voice Assistants: A Dual-Factor Approach by Integrating the Status Quo Bias and Technology Acceptance Model. *Information Systems Frontiers*, 1–22.  
<https://doi.org/10.1007/s10796-021-10203-y>
- Balsamiq. (2022). *Balsamiq Wireframes - Industry Standard Low-Fidelity Wireframing Software*.  
<https://balsamiq.com/wireframes/> (accessed on Oct 21, 2022)
- Bartels, T. (2018). *30 Jahre Frauen im Cockpit der Lufthansa*. Stern.de.  
<https://www.stern.de/reise/follow-me/30-jahre-frauen-im-cockpit-der-lufthansa-8216866.html> (accessed on Sep 26, 2023)
- Bauer, K., Hinz, O., van der Aalst, W., & Weinhardt, C. (2021). Expl(AI)n It to Me – Explainable AI and Information Systems Research. *Business & Information Systems Engineering*, 63(2), 79–82. <https://doi.org/10.1007/s12599-021-00683-2>
- Bauer, K., von Zahn, M., & Hinz, O. (2021). Expl(AI)ned: The Impact of Explainable Artificial Intelligence on Users' Information Processing. *SAFE Working Paper No. 315*. SAFE Working Paper No. 315. <https://doi.org/10.2139/ssrn.3872711>
- Bauer, K., von Zahn, M., & Hinz, O. (2023). Expl(AI)ned: The Impact of Explainable Artificial Intelligence on Users' Information Processing. *Information Systems Research*, 34(4), 1582–1602. <https://doi.org/10.1287/isre.2023.1199>
- Benbya, H., Pachidi, S., & Jarvenpaa, S. L. (2021). Artificial Intelligence in Organizations: Implications for Information Systems Research. *Journal of the Association for Information Systems*, 22(2). <https://doi.org/10.17705/1jais.00662>
- Benbya, H., Strich, F., & Tamm, T. (2024). Navigating Generative Artificial Intelligence Promises and Perils for Knowledge and Creative Work. *Journal of the Association for Information Systems*, 25(1), 23–36. <https://doi.org/10.17705/1jais.00861>
- Benjamins, S., Dhunoo, P., & Meskó, B. (2020). The State of Artificial Intelligence-based FDA-Approved Medical Devices and Algorithms: An Online Database. *NPJ Digital Medicine*, 3(1), 118. <https://doi.org/10.1038/s41746-020-00324-0>



- Berente, N., Gu, B., Recker, J., & Santhanam, R. (2021). Managing Artificial Intelligence. *MIS Quarterly*, 45(3), 1433–1450. <https://doi.org/10.25300/MISQ/2021/16274>
- Bhatt, U., Xiang, A., Sharma, S., Weller, A., Taly, A., Jia, Y., Ghosh, J., Puri, R., Moura, J. M. F., & Eckersley, P. (2020). Explainable Machine Learning in Deployment. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 648–657. <https://doi.org/10.1145/3351095.3375624>
- Boyacı, T., Canyakmaz, C., & de Véricourt, F. (2024). Human and Machine: The Impact of Machine Input on Decision Making Under Cognitive Limitations. *Management Science*, 70(2), 1258–1275. <https://doi.org/10.1287/mnsc.2023.4744>
- Bradshaw, J. M., Acquisti, A., Allen, J., Breedy, M., Bunch, L., Chambers, N., Galescu, L., Goodrich, M., Jeffers, R., Johnson, M., Jung, H., Kulkarni, S., Lott, J., Olsen, D., Sierhuis, M., Suri, N., Taysom, W., Tonti, G., Uszok, A., & Van Hoof, R. (2004). Teamwork-Centered Autonomy for Extended Human-Agent Interaction in Space Applications. *AAAI 2004 Spring Symposium*, 22–24.
- Brady, M. K., & Robertson, C. J. (1999). An Exploratory Study of Service Value in the USA and Ecuador. *International Journal of Service Industry Management*, 10(5), 469–486. <https://doi.org/10.1108/09564239910289003>
- Brasse, J., Broder, H. R., Förster, M., Klier, M., & Sigler, I. (2023). Explainable Artificial Intelligence in Information Systems: A Review of the Status Quo and Future Research Directions. *Electronic Markets*, 33(1), 26. <https://doi.org/10.1007/s12525-023-00644-5>
- Brocke, J., Simons, A., Niehaves, B., Niehaves, B., & Reimer, K. (2009). Reconstructing the Giant: On the Importance of Rigour in Documenting the Literature Search Process. *ECIS 2009 Proceedings*, 161. <https://aisel.aisnet.org/ecis2009/161>
- Brown, A. L., & Kane, M. J. (1988). Preschool Children Can Learn to Transfer: Learning to Learn and Learning From Example. *Cognitive Psychology*, 20(4), 493–523. [https://doi.org/10.1016/0010-0285\(88\)90014-x](https://doi.org/10.1016/0010-0285(88)90014-x)
- Brown, S. A., & Venkatesh, V. (2005). A Model of Adoption of Technology in the Household: A Baseline Model Test and Extension Incorporating Household Life Cycle. *MIS Quarterly*, 29(3), 11. <https://aisel.aisnet.org/misq/vol29/iss3/11/>
- Brynjolfsson, E., & McAfee, A. (2014). *The Second Machine Age: Work, Progress, and Prosperity in a Time of Brilliant Technologies*. W. W. Norton & Company.
- Brynjolfsson, E., & Mitchell, T. (2017). What Can Machine Learning Do? Workforce Implications. *Science*, 358(6370), 1530–1534. <https://doi.org/10.1126/science.aap8062>

- Build-A-Bear. (2023). *The-Bear-Builder*.  
[https://www.buildabear.co.uk/on/demandware.store/Sites-buildabear-uk-Site/en\\_GB/BearBuilder-Show?step=chooseFriends](https://www.buildabear.co.uk/on/demandware.store/Sites-buildabear-uk-Site/en_GB/BearBuilder-Show?step=chooseFriends) (accessed on Oct 21, 2023)
- Cai, C. J., Jongejan, J., & Holbrook, J. (2019). The Effects of Example-based Explanations in a Machine Learning Interface. *Proceedings of the 24th International Conference on Intelligent User Interfaces*, 258–262. <https://doi.org/10.1145/3301275.3302289>
- Calisto, F. M., Santiago, C., Nunes, N., & Nascimento, J. C. (2021). Introduction of Human-Centric AI Assistant to Aid Radiologists for Multimodal Breast Image Classification. *International Journal of Human-Computer Studies*, 150, 102607.  
<https://doi.org/10.1016/j.ijhcs.2021.102607>
- Capgemini. (2023). *Why Consumers Love Generative AI*. [https://prod.ucwe.capgemini.com/wp-content/uploads/2023/06/GENERATIVE-AI\\_Final\\_WEB\\_060723.pdf](https://prod.ucwe.capgemini.com/wp-content/uploads/2023/06/GENERATIVE-AI_Final_WEB_060723.pdf) (accessed on Nov 01, 2023)
- Cassoli, B., Jourdan, N., Nguyen, P. H., & Sen, S. (2022). Frameworks for Data-Driven Quality Management in Cyber-Physical Systems for Manufacturing: A Systematic Review. *Procedia CIRP*, 112, 567-572. <https://doi.org/10.1016/j.procir.2022.09.062>
- Castelvecchi, D. (2016). Can We Open the Black Box of AI? *Nature*, 538(7623), 20–23.  
<https://doi.org/10.1038/538020a>
- Chase, W. G., & Simon, H. A. (1973). The Mind's Eye in Chess. In W. G. Chase (Ed.), *Visual Information Processing* (pp. 215–281). Academic Press. <https://doi.org/10.1016/B978-0-12-170150-5.50011-1>
- Chen, N., Guimbretière, F., Sun, L., Czerwinski, M., Pangaro, G., & Bathiche, S. (2009). Hardware Support for Navigating Large Digital Documents. *International Journal of Human-Computer Interaction*, 25(3), 199–219. <https://doi.org/10.1080/10447310802629819>
- Cheng, J.-Z., Ni, D., Chou, Y.-H., Qin, J., Tiu, C.-M., Chang, Y.-C., Huang, C.-S., Shen, D., & Chen, C.-M. (2016). Computer-Aided Diagnosis with Deep Learning Architecture: Applications to Breast Lesions in US Images and Pulmonary Nodules in CT Scans. *Scientific Reports*, 6, 24454. <https://doi.org/10.1038/srep24454>
- Civil Aviation Safety Authority. (2023). *Visual Flight Rules Guide*.  
<https://www.casa.gov.au/resources-and-education/publications-and-resources/industry-guides-and-publications/pilot-guides/visual-flight-rules-guide#Feedback> (accessed on Aug 03, 2023)

- Cohen, J. (2013). *Statistical Power Analysis for the Behavioral Sciences*, (Vol. 73, p. 680). Routledge. <https://doi.org/10.2307/2286629>
- Crowley, K., & Siegler, R. S. (1999). Explanation and Generalization in Young Children's Strategy Learning. *Child Development*, *70*(2), 304–316. <https://doi.org/10.1111/1467-8624.00023>
- Dang, Y., Zhang, Y., Brown, S. A., & Chen, H. (2020). Examining the Impacts of Mental Workload and Task-Technology Fit on User Acceptance of the Social Media Search System. *Information Systems Frontiers*, *22*(3), 697–718. <https://doi.org/10.1007/s10796-018-9879-y>
- Das, T. K., & Teng, B.-S. (1999). Cognitive Biases and Strategic Decision Processes: An Integrative Perspective. *The Journal of Management Studies*, *36*(6), 757–778. <https://doi.org/10.1111/1467-6486.00157>
- de Zoeten, M., Ernst, C. P. H., & Rothlauf, F. (2023). A Matter of Trust: How Trust in AI-Based Systems Changes During Interaction. *AMCIS 2023 Proceedings*, *15*, 1491. [https://aisel.aisnet.org/amcis2023/sig\\_odis/sig\\_odis/15](https://aisel.aisnet.org/amcis2023/sig_odis/sig_odis/15)
- Deloitte. (2020). *AI Enablement on the Way to Smart Manufacturing*. <https://www2.deloitte.com/content/dam/Deloitte/cn/Documents/cip/deloitte-cn-cip-ai-manufacturing-application-survey-en-200116.pdf> (accessed on Apr 23, 2024)
- Demir, M., McNeese, N. J., & Cooke, N. J. (2017). Team Situation Awareness Within the Context of Human-Autonomy Teaming. *Cognitive Systems Research*, *46*, 3–12. <https://doi.org/10.1016/j.cogsys.2016.11.003>
- Diakopoulos, N. (2016). Accountability in Algorithmic Decision Making. *Communications of the ACM*, *59*(2), 56–62. <https://doi.org/10.1145/2844110>
- Dietvorst, B. J., & Bharti, S. (2020). People Reject Algorithms in Uncertain Decision Domains Because They Have Diminishing Sensitivity to Forecasting Error. *Psychological Science*, *31*(10), 1302–1314. <https://doi.org/10.1177/0956797620948841>
- Dohle, S., Rall, S., & Siegrist, M. (2014). I Cooked It Myself: Preparing Food Increases Liking and Consumption. *Food Quality and Preference*, *33*, 14–16. <https://doi.org/10.1016/J.FOODQUAL.2013.11.001>
- Dohle, S., Rall, S., & Siegrist, M. (2016). Does Self-Prepared Food Taste Better? Effects of Food Preparation on Liking. *Health Psychology: Official Journal of the Division of Health Psychology, American Psychological Association*, *35*(5), 500–508. <https://doi.org/10.1037/HEA0000315>

- Dolata, M., & Aleya, K. B. (2022). Morphological Analysis for Design Science Research: The Case of Human-Drone Collaboration in Emergencies. *The Transdisciplinary Reach of Design Science Research*, 17–29. [https://doi.org/10.1007/978-3-031-06516-3\\_2](https://doi.org/10.1007/978-3-031-06516-3_2)
- Dwivedi, Y. K., Kshetri, N., Hughes, L., Slade, E. L., Jeyaraj, A., Kar, A. K., Baabdullah, A. M., Koohang, A., Raghavan, V., Ahuja, M., Albanna, H., Albashrawi, M. A., Al-Busaidi, A. S., Balakrishnan, J., Barlette, Y., Basu, S., Bose, I., Brooks, L., Buhalis, D., ... Wright, R. (2023). “So What if ChatGPT Wrote It?” Multidisciplinary Perspectives on Opportunities, Challenges and Implications of Generative Conversational AI for Research, Practice and Policy. *International Journal of Information Management*, 71, 102642. <https://doi.org/10.1016/j.ijinfomgt.2023.102642>
- Eißfeldt, H., Vogelpohl, V., Stolz, M., Papenfuß, A., Biella, M., Belz, J., & Kügler, D. (2020). The Acceptance of Civil Drones in Germany. *CEAS Aeronautical Journal*, 11(3), 665–676. <https://doi.org/10.1007/S13272-020-00447-W/TABLES/5>
- Eißfeldt, H., & End, A. (2020). Investigating Attitudes Towards Drone Delivery. *Proceedings of the Human Factors and Ergonomics Society*, 64(1), 169–173. <https://doi.org/10.1177/1071181320641042>
- Ellenrieder, S., Jourdan, N., & Reuter-Oppermann, M. (2023). Delivery Drones - Just a Hype? Towards Autonomous Air Mobility Services at Scale. *Proceedings of the 56nd Hawaii International Conference on System Sciences*.
- Endsley, M. R. (2018). Situation Awareness in Future Autonomous Vehicles: Beware of the Unexpected. *Congress of the International Ergonomics Association*, 303–309. [https://doi.org/10.1007/978-3-319-96071-5\\_32](https://doi.org/10.1007/978-3-319-96071-5_32)
- European Commission. (2023). *AI Act*. Shaping Europe’s Digital Future. <https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai> (accessed on Feb 28, 2024)
- European Union Aviation Safety Agency. (2023). *EASA Artificial Intelligence Roadmap 2.0 - A Human-Centric Approach to AI in Aviation*. <https://www.easa.europa.eu/en/document-library/general-publications/easa-artificial-intelligence-roadmap-20> (accessed on Oct 15, 2023)
- European Union Aviation Safety Agency. (2024). *EASA Concept Paper: Guidance for Level 1 & 2 Machine Learning Applications* (No. 02). <https://www.easa.europa.eu/en/document-library/general-publications/easa-artificial-intelligence-concept-paper-issue-2> (accessed on March 21, 2024)

- Fahle, Prinz, & Kuhlenkötter. (2020). Systematic Review on Machine Learning (ML) Methods for Manufacturing Processes—Identifying Artificial Intelligence (AI) Methods for Field Application. *Procedia CIRP*, 93, 413–418. <https://doi.org/10.1016/j.procir.2020.04.109>
- Fender, J. G., & Crowley, K. (2007). How Parent Explanation Changes What Children Learn From Everyday Scientific Thinking. *Journal of Applied Developmental Psychology*, 28(3), 189–210. <https://doi.org/10.1016/j.appdev.2007.02.007>
- Figma. (2022). *Figma: The Tool for Collaborative Interface Design*. Figma. <https://www.figma.com> (accessed on Oct 21, 2022)
- Fjällräven. (2023). *Kånken Me*. <https://www.fjallraven.com/us/en-us/bags-gear/kanken/kanken-bags/kanken-me> (accessed on Oct 31, 2023)
- Flick, U. (2004). Triangulation in Qualitative Research. *A Companion to Qualitative Research*, 3, 178–183.
- Flynn, L. R., & Goldsmith, R. E. (1999). A Short, Reliable Measure of Subjective Knowledge. *Journal of Business Research*, 46(1), 57–66. [https://doi.org/10.1016/S0148-2963\(98\)00057-5](https://doi.org/10.1016/S0148-2963(98)00057-5)
- Fornell, C., & Larcker, D. (1981). Evaluating Structural Equation Models With Unobservable Variables and Measurement Error. *JMR, Journal of Marketing Research*, 18, 39–50. <https://doi.org/10.1177/002224378101800104>
- Frank, D.-A., Chrysochou, P., & Mitkidis, P. (2023). The Paradox of Technology: Negativity Bias in Consumer Adoption of Innovative Technologies. *Psychology & Marketing*, 40(3), 554–566. <https://doi.org/10.1002/mar.21740>
- Franke, N., Schreier, M., & Kaiser, U. (2010). The “I Designed It Myself” Effect in Mass Customization. *Management Science*, 56(1), 125–140. <https://doi.org/10.1287/mnsc.1090.1077>
- French, A. M., Storey, V. C., & Wallace, L. (2023). The Impact of Cognitive Biases on the Believability of Fake News. *European Journal of Information Systems*, 1–22. <https://doi.org/10.1080/0960085X.2023.2272608>
- Fügener, A., Grahl, J., Gupta, A., & Ketter, W. (2021). Will Humans-in-The-Loop Become Borgs? Merits and Pitfalls of Working with AI. *MIS Quarterly*, 45(3). <https://ssrn.com/abstract=3879937>
- Fügener, A., Grahl, J., Gupta, A., & Ketter, W. (2022). Cognitive Challenges in Human-AI Collaboration: Investigating the Path Towards Productive Delegation. *Information Systems Research*, 33(2), 678–696. <https://doi.org/10.1287/isre.2021.1079>

- Gagné, R. M. (1970). Some New Views of Learning and Instruction. *Phi Delta Kappan*, 51(9), 468–472. <http://www.jstor.org/stable/20372726>
- Gaube, S., Suresh, H., Raue, M., Lermer, E., Koch, T. K., Hudecek, M. F. C., Ackery, A. D., Grover, S. C., Coughlin, J. F., Frey, D., Kitamura, F. C., Ghassemi, M., & Colak, E. (2023). Non-task Expert Physicians Benefit From Correct Explainable AI Advice When Reviewing X-Rays. *Scientific Reports*, 13(1), 1383. <https://doi.org/10.1038/s41598-023-28633-w>
- Gaube, S., Suresh, H., Raue, M., Merritt, A., Berkowitz, S. J., Lermer, E., Coughlin, J. F., Guttag, J. V., Colak, E., & Ghassemi, M. (2021). Do as AI Say: Susceptibility in Deployment of Clinical Decision-Aids. *NPJ Digital Medicine*, 4(1), 31. <https://doi.org/10.1038/s41746-021-00385-9>
- Gefen, D., Straub, D., & Boudreau, M.-C. (2000). Structural Equation Modeling and Regression: Guidelines for Research Practice. *Communications of the Association for Information Systems*, 4, 7. <https://doi.org/10.17705/1cais.00407>
- Ghorbani, A., Wexler, J., Zou, J. Y., & Kim, B. (2019). Towards Automatic Concept-Based Explanations. *Advances in Neural Information Processing Systems*, 32. <https://proceedings.neurips.cc/paper/2019/hash/77d2afcb31f6493e350fca61764efb9a-Abstract.html>
- Giones, F., & Brem, A. (2017). From Toys to Tools: The Co-evolution of Technological and Entrepreneurial Developments in the Drone Industry. *Business Horizons*, 60(6), 875–884. <https://doi.org/10.1016/j.bushor.2017.08.001>
- Girden, E. R. (1992). *ANOVA: Repeated Measures* (Vol. 84). SAGE.
- Glikson, E., & Woolley, A. W. (2020). Human Trust in Artificial Intelligence: Review of Empirical Research. *Academy of Management Annals*, 14(2), 627–660. <https://doi.org/10.5465/annals.2018.0057>
- Goddard, K., Roudsari, A., & Wyatt, J. C. (2012). Automation Bias: A Systematic Review of Frequency, Effect Mediators, and Mitigators. *Journal of the American Medical Informatics Association: JAMIA*, 19(1), 121–127. <https://doi.org/10.1136/amiajnl-2011-000089>
- Goralski, M. A., & Tan, T. K. (2020). Artificial Intelligence and Sustainable Development. *The International Journal of Management Education*, 18(1), 100330. <https://doi.org/10.1016/j.ijme.2019.100330>
- Göring, S., Rao, R. R. R., Merten, R., & Raake, A. (2023). Analysis of Appeal for Realistic AI-Generated Photos. *IEEE Access*, 11, 38999–39012. <https://doi.org/10.1109/ACCESS.2023.3267968>

- Goutte, C., & Gaussier, E. (2005). A Probabilistic Interpretation of Precision, Recall and F-Score, with Implication for Evaluation. *Advances in Information Retrieval*, 345–359.  
[https://doi.org/10.1007/978-3-540-31865-1\\_25](https://doi.org/10.1007/978-3-540-31865-1_25)
- Green, D. M., & Swets, J. A. (1966). *Signal Detection Theory and Psychophysics* (Vol. 1). New York: Wiley.
- Gregor, S., Chandra Kruse, L., Seidel, S., & Others. (2020). Research Perspectives: The Anatomy of a Design Principle. *Journal of the Association for Information Systems*, 21(6).  
<https://doi.org/10.17705/1jais.00649>
- Gregor, S., & Hevner, A. R. (2013). Positioning and Presenting Design Science Research for Maximum Impact. *MIS Quarterly*, 37(2), 337–355.  
<https://www.jstor.org/stable/43825912>
- Grønsund, T., & Aanestad, M. (2020). Augmenting the Algorithm: Emerging Human-In-The-Loop Work Configurations. *The Journal of Strategic Information Systems*, 29(2), 101614.  
<https://doi.org/10.1016/j.jsis.2020.101614>
- Grover, V., Chiang, R. H. L., Liang, T.-P., & Zhang, D. (2018). Creating Strategic Business Value from Big Data Analytics: A Research Framework. *Journal of Management Information Systems*, 35(2), 388–423. <https://doi.org/10.1080/07421222.2018.1451951>
- Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A Survey of Methods for Explaining Black Box Models. *ACM Comput. Surv.*, 51(5), 1–42.  
<https://doi.org/10.1145/3236009>
- Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On Calibration of Modern Neural Networks. *Proceedings of the 34th International Conference on Machine Learning*, 1321–1330.  
<https://proceedings.mlr.press/v70/guo17a.html>
- Gurney, N., Pynadath, D. V., & Wang, N. (2022). Measuring and Predicting Human Trust in Recommendations from an AI Teammate. *Artificial Intelligence in HCI*, 22–34.  
[https://doi.org/10.1007/978-3-031-05643-7\\_2](https://doi.org/10.1007/978-3-031-05643-7_2)
- Gutzwiller, R. S., Espinosa, S. H., Kenny, C., & Lange, D. S. (2018). A Design Pattern for Working Agreements in Human-Autonomy Teaming. *Advances in Human Factors in Simulation and Modeling*, 12–24. [https://doi.org/10.1007/978-3-319-60591-3\\_2](https://doi.org/10.1007/978-3-319-60591-3_2)
- Hair, J. F., Jr, Hult, G. T. M., Ringle, C. M., & Sarstedt, M. (2017). *A Primer on Partial Least Squares Structural Equation Modeling (PLS-SEM): Vol. 2nd edn.* SAGE Publications Inc.

- Hart, S. G. (2006). Nasa-Task Load Index (NASA-TLX); 20 Years Later. *Proceedings of the Human Factors and Ergonomics Society Meeting*, 50(9), 904–908.  
<https://doi.org/10.1177/154193120605000909>
- Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. In P. A. Hancock & N. Meshkati (Eds.), *Advances in Psychology*, 52, 139–183. North-Holland. [https://doi.org/10.1016/S0166-4115\(08\)62386-9](https://doi.org/10.1016/S0166-4115(08)62386-9)
- Haselton, M. G., & Nettle, D. (2006). The Paranoid Optimist: An Integrative Evolutionary Model of Cognitive Biases. *Personality and Social Psychology Review: An Official Journal of the Society for Personality and Social Psychology, Inc*, 10(1), 47–66.  
[https://doi.org/10.1207/s15327957pspr1001\\_3](https://doi.org/10.1207/s15327957pspr1001_3)
- Hassmen, P., & Hunt, D. P. (1994). Human Self-Assessment in Multiple-Choice Testing. *Journal of Educational Measurement*, 31(2), 149–160. <https://doi.org/10.1111/j.1745-3984.1994.tb00440.x>
- Heunemann, F. (2022). *Rewe beteiligt sich an Wingcopter*. Frankfurter Allgemeine Zeitung GmbH.  
<https://www.faz.net/aktuell/rhein-main/wirtschaft/rewe-investiert-in-drohnenhersteller-wingcopter-18117047.html> (accessed on Oct 12, 2022)
- Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design Science in Information Systems Research. *MIS Quarterly*, 28(1), 75–105. <https://doi.org/10.2307/25148625>
- Hicks, J. (2022). *Drone Carrying a Defibrillator Saves Its First Heart Attack Patient in Sweden*. The Verge. <https://www.theverge.com/2022/1/5/22868777/everdrone-drone-defibrillator-health-tech-sweden> (accessed on May 04,2022)
- Hristov, I., Camilli, R., & Mechelli, A. (2022). Cognitive Biases in Implementing a Performance Management System: Behavioral Strategy for Supporting Managers' Decision-Making Processes. *Management Research Review*, 45(9), 1110–1136.  
<https://doi.org/10.1108/MRR-11-2021-0777>
- Hunt, D. P. (2003). The Concept of Knowledge and How to Measure It. *Journal of Intellectual Capital*, 4(1), 100–113. <https://doi.org/10.1108/14691930310455414>
- Huyen, C. (2022). *Designing Machine Learning Systems*. O'Reilly Media, Inc.
- Jha, S., & Topol, E. J. (2016). Adapting to Artificial Intelligence: Radiologists and Pathologists as Information Specialists. *JAMA: The Journal of the American Medical Association*, 316(22), 2353–2354. <https://doi.org/10.1001/jama.2016.17438>



- Jiang, F., Jiang, Y., Zhi, H., Dong, Y., Li, H., Ma, S., Wang, Y., Dong, Q., Shen, H., & Wang, Y. (2017). Artificial Intelligence in Healthcare: Past, Present and Future. *Stroke and Vascular Neurology*, 2(4), 230–243. <https://doi.org/10.1136/svn-2017-000101>
- Johnson, D. D. P., Blumstein, D. T., Fowler, J. H., & Haselton, M. G. (2013). The Evolution of Error: Error Management, Cognitive Constraints, and Adaptive Decision-Making Biases. *Trends in Ecology & Evolution*, 28(8), 474–481. <https://doi.org/10.1016/j.tree.2013.05.014>
- Jordan, M. I., & Mitchell, T. M. (2015). Machine Learning: Trends, Perspectives, and Prospects. *Science*, 349(6245), 255–260. <https://doi.org/10.1126/science.aaa8415>
- Jourdan, N., Sen, S., Husom, E. J., Garcia-Ceja, E., Biegel, T., & Metternich, J. (2021). On The Reliability Of Machine Learning Applications In Manufacturing Environments. In *presented at the Neural Information Processing Systems (NeurIPS 2021): Workshop on Distribution Shifts*. arXiv. <http://arxiv.org/abs/2112.06986>
- Jussupow, E., Spohrer, K., Heinzl, A., & Gawlitza, J. (2021). Augmenting Medical Diagnosis Decisions? An Investigation into Physicians' Decision-Making Process with Artificial Intelligence. *Information Systems Research*, 32(3), 713–735. <https://doi.org/10.1287/isre.2020.0980>
- Kaggle. (2020). *Brain Tumor Segmentation (BraTS2020)*. <https://www.kaggle.com/datasets/awsaf49/brats2020-training-data> (accessed on May 03, 2022)
- Kahneman, D., Knetsch, J. L., & Thaler, R. H. (1990). Experimental Tests of the Endowment Effect and the Coase Theorem. *The Journal of Political Economy*, 98(6), 1325–1348. <https://doi.org/10.1086/261737>
- Kahnemann, D. (2012). *Thinking Fast and Slow*. New York: Farrar, Straus and Giroux.
- Kamtarin, M. (2012). The Effect of Electronic Word of Mouth , Trust and Perceived Value on Behavioral Intention From the Perspective of Consumers. *International Journal of Academic Research in Economics and Management Sciences*, 1(4), 56–66.
- Kane, G. C., Young, A. G., Majchrzak, A., & Ransbotham, S. (2021). Avoiding an Oppressive Future of Machine Learning: A Design Theory for Emancipatory Assistants. *MIS Quarterly*, 45(1), 371–396. <https://doi.org/10.25300/misq/2021/1578>
- Kanngiesser, P., Gjersoe, N., & Hood, B. M. (2010). The Effect of Creative Labor on Property-Ownership Transfer by Preschool Children and Adults. *Psychological Science*, 21(9), 1236–1241. <https://doi.org/10.1177/0956797610380701>

- Kim, H.-W., & Kankanhalli, A. (2009). Investigating User Resistance to Information Systems Implementation: A Status Quo Bias Perspective. *MIS Quarterly*, 33(3), 567–582. <https://doi.org/10.2307/20650309>
- Klein, G., Woods, D. D., Bradshaw, J. M., Hoffman, R. R., & Feltovich, P. J. (2004). Ten Challenges for Making Automation a “Team Player” in Joint Human-Agent Activity. *IEEE Intelligent Systems*, 19(06), 91–95. <https://doi.org/10.1109/MIS.2004.74>
- Klöpffer, W. (2003). Life-Cycle Based Methods for Sustainable Product Development. *International Journal of Life Cycle Assessment*, 8(3), 157–159. <https://doi.org/10.1007/BF02978462>
- Krey, M. (2018). Drones in Healthcare: Application in Swiss Hospitals. *Proceedings of the 51st Hawaii International Conference on System Sciences*, 3081–3089. <http://hdl.handle.net/10125/50277>
- Krueger, R. A. (2014). *Focus Groups: A Practical Guide for Applied Research*. SAGE Publications.
- Kuechler, B., & Vaishnavi, V. (2008). On Theory Development in Design Science Research: Anatomy of a Research Project. *European Journal of Information Systems*, 17(5), 489–504. <https://doi.org/10.1057/ejis.2008.40>
- Lakshmanan, V., Robinson, S., & Munn, M. (2020). *Machine Learning Design Patterns*. O'Reilly Media, Inc.
- Lebovitz, S., Levina, N., & Lifshitz-Assa, H. (2021). Is AI Ground Truth Really True? The Dangers of Training and Evaluating AI Tools Based on Experts' Know-What. *MIS Quarterly*, 45(3), 1501–1526. <https://doi.org/10.25300/MISQ/2021/16564>
- Lebovitz, S., Lifshitz-Assaf, H., & Levina, N. (2022). To Engage or Not to Engage with AI for Critical Judgments: How Professionals Deal with Opacity When Using AI for Medical Diagnosis. *Organization Science*, 33(1), 126–148. <https://doi.org/10.1287/orsc.2021.1549>
- Levin, A. (2016). *Alphabet and Chipotle Are Bringing Burrito Delivery Drones to Campus*. <https://www.bloomberg.com/news/articles/2016-09-08/burrito-by-drone-coming-to-campus-in-test-of-alphabet-s-delivery> (accessed on Mar 14, 2022)
- Ling, I. L., Liu, Y. F., Lin, C. W., & Shieh, C. H. (2020). Exploring IKEA Effect in Self-Expressive Mass Customization: Underlying Mechanism and Boundary Conditions. *Journal of Consumer Marketing*, 37(4), 365–374. <https://doi.org/10.1108/JCM-09-2017-2373/FULL/PDF>
- Liu, T., Vickers, B. D., & Seidler, R. D. (2023). Neural Correlates of Overvaluation and the Effort to Save Possessions in a Novel Decision Task: An Exploratory fMRI Study. *Frontiers in Psychology*, 14, 1059051. <https://doi.org/10.3389/FPSYG.2023.1059051/BIBTEX>

- Liu, Y., & Wickens, C. D. (1994). Mental Workload and Cognitive Task Automaticity: An Evaluation of Subjective and Time Estimation Metrics. *Ergonomics*, *37*(11), 1843–1854. <https://doi.org/10.1080/00140139408964953>
- Luce, M. F., & Kahn, B. E. (1999). Avoidance or Vigilance? The Psychology of False-Positive Test Results. *The Journal of Consumer Research*, *26*(3), 242–259. <https://doi.org/10.1086/209561>
- Lund, B. D., Wang, T., Mannuru, N. R., Nie, B., Shimray, S., & Wang, Z. (2023). ChatGPT and a New Academic Reality: Artificial Intelligence - Written Research Papers and the Ethics of the Large Language Models in Scholarly Publishing. *Journal of the Association for Information Science and Technology*, *74*(5), 570–581. <https://doi.org/10.1002/asi.24750>
- Maedche, A., Legner, C., Benlian, A., Berger, B., Gimpel, H., Hess, T., Hinz, O., Morana, S., & Söllner, M. (2019). AI-Based Digital Assistants. *Business & Information Systems Engineering*, *61*(4), 535–544. <https://doi.org/10.1007/s12599-019-00600-8>
- Mao, R., Du, B., Sun, D., & Kong, N. (2019). Optimizing a UAV-based Emergency Medical Service Network for Trauma Injury Patients. *IEEE International Conference on Automation Science and Engineering, 2019-August*, 721–726. <https://doi.org/10.1109/COASE.2019.8843138>
- Marsh, L. E., Gil, J., & Kanngiesser, P. (2022). The Influence of Collaboration and Culture on the IKEA Effect: Does Cocreation Alter Perceptions of Value in British and Indian Children? *Developmental Psychology*, *58*(4), 662. <https://doi.org/10.1037/dev0001321.supp>
- Marsh, L. E., Kanngiesser, P., & Hood, B. (2018). When and How Does Labour Lead to Love? The Ontogeny and Mechanisms of the IKEA Effect. *Cognition*, *170*, 245–253. <https://doi.org/10.1016/J.COGNITION.2017.10.012>
- Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An Integrative Model Of Organizational Trust. *Academy of Management Review*, *20*(3), 709–734. <https://doi.org/10.5465/amr.1995.9508080335>
- McCarthy, J., Minsky, M. L., Rochester, N., & Shannon, C. E. (2006). A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence, August 31, 1955. *AI Magazine*, *27*(4), 12–12. <https://doi.org/10.1609/aimag.v27i4.1904>
- McKinney, S. M., Sieniek, M., Godbole, V., Godwin, J., Antropova, N., Ashrafiyan, H., Back, T., Chesus, M., Corrado, G. S., Darzi, A., Etemadi, M., Garcia-Vicente, F., Gilbert, F. J., Halling-Brown, M., Hassabis, D., Jansen, S., Karthikesalingam, A., Kelly, C. J., King, D., ... Shetty, S. (2020). International Evaluation of an AI System for Breast Cancer Screening. *Nature*, *577*(7788), 89–94. <https://doi.org/10.1038/s41586-019-1799-6>

- McKinsey & Company. (2023). *The State of AI in 2023: Generative AI's Breakout Year*.  
<https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai-in-2023-generative-ais-breakout-year> (accessed on Feb 17, 2024)
- McKinsey Global Institute. (2021). *The state of AI in 2021*.  
<https://www.mckinsey.com/capabilities/quantumblack/our-insights/global-survey-the-state-of-ai-in-2021>
- McKnight, D. H., Carter, M., Thatcher, J. B., & Clay, P. F. (2011). Trust in a Specific Technology: An Investigation of its Components and Measures. *ACM Trans. Manage. Inf. Syst.*, 2(2), 1–25.  
<https://doi.org/10.1145/1985347.1985353>
- McNeese, N. J., Demir, M., Chiou, E., Cooke, N., & Yanikian, G. (2019). Understanding the Role of Trust in Human-Autonomy Teaming. *Proceedings of the 52nd Hawaii International Conference on System Sciences*. <https://hdl.handle.net/10125/59466>
- McNeese, N. J., Demir, M., Chiou, E. K., & Cooke, N. J. (2021). Trust and Team Performance in Human–Autonomy Teaming. *International Journal of Electronic Commerce*, 25(1), 51–72.  
<https://doi.org/10.1080/10864415.2021.1846854>
- McNeese, N. J., Demir, M., Cooke, N. J., & Myers, C. (2018). Teaming With a Synthetic Teammate: Insights into Human-Autonomy Teaming. *Human Factors*, 60(2), 262–273.  
<https://doi.org/10.1177/0018720817743223>
- Menze, B. H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., Burren, Y., Porz, N., Slotboom, J., Wiest, R., Lanczi, L., Gerstner, E., Weber, M.-A., Arbel, T., Avants, B. B., Ayache, N., Buendia, P., Collins, D. L., Cordier, N., ... Van Leemput, K. (2015). The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS). *IEEE Transactions on Medical Imaging*, 34(10), 1993–2024. <https://doi.org/10.1109/TMI.2014.2377694>
- Meske, C., Bunde, E., Schneider, J., & Gersch, M. (2022). Explainable Artificial Intelligence: Objectives, Stakeholders, and Future Research Opportunities. *Information Systems Management*, 39(1), 53–63. <https://doi.org/10.1080/10580530.2020.1849465>
- Meskó, B., & Görög, M. (2020). A Short Guide for Medical Professionals in the Era of Artificial Intelligence. *NPJ Digital Medicine*, 3, 126. <https://doi.org/10.1038/s41746-020-00333-z>
- Meth, H., Mueller, B., & Maedche, A. (2015). Designing a Requirement Mining System. *Journal of the Association for Information Systems*, 16(9), 2. <https://doi.org/10.17705/1jais.00408>
- Metternich, J., Biegel, T., Bretones Cassoli, B., Hoffmann, F., Jourdan, N., Rosemeyer, J., Stanula, P., & Ziegenbein, A. (2021). Künstliche Intelligenz zur Umsetzung von Industrie 4.0 im

- Mittelstand: Expertise des Forschungsbeirats der Plattform Industrie 4.0, München.  
*Acatech-Deutsche Akademie Der Technikwissenschaften.*
- Mezulis, A. H., Abramson, L. Y., Hyde, J. S., & Hankin, B. L. (2004). Is There a Universal Positivity Bias in Attributions? A Meta-Analytic Review of Individual, Developmental, and Cultural Differences in the Self-Serving Attributional Bias. *Psychological Bulletin*, *130*(5), 711–747.  
<https://doi.org/10.1037/0033-2909.130.5.711>
- Microsoft Flight Simulator. (2023). *Microsoft Flight Simulator - The Next Generation of One of the Most Beloved Simulation Franchises.* Microsoft Flight Simulator.  
<https://www.flightsimulator.com/> (accessed on Sep 19, 2023)
- Miller, D. T., & Ross, M. (1975). Self-Serving Biases in the Attribution of Causality: Fact or Fiction? *Psychological Bulletin*, *82*(2), 213–225. <https://doi.org/10.1037/h0076486>
- Miller, G. A. (1956). The Magical Number Seven Plus or minus Two: Some Limits on Our Capacity for Processing Information. *Psychological Review*, *63*(2), 81–97.  
<https://doi.org/10.1037/h0043158>
- Miller, J. G. (1976). The Nature of Living Systems. *Behavioral Science*, *21*(5), 295–319.  
<https://doi.org/10.1002/bs.3830210502>
- Miller, T. (2019). Explanation in Artificial Intelligence: Insights From the Social Sciences. *Artificial Intelligence*, *267*, 1–38. <https://doi.org/10.1016/j.artint.2018.07.007>
- Mirbabaie, M., Brünker, F., Möllmann Frick, N. R. J., & Stieglitz, S. (2022). The Rise of Artificial Intelligence – Understanding the AI Identity Threat at the Workplace. *Electronic Markets*, *32*(1), 73–99. <https://doi.org/10.1007/s12525-021-00496-x>
- Mitchell, T. M. (1997). *Machine learning* (Vol. 1). McGraw-hill New York.
- Mochon, D., Norton, M. I., & Ariely, D. (2012). Bolstering and Restoring Feelings of Competence via the IKEA Effect. *International Journal of Research in Marketing*, *29*(4), 363–369.  
<https://doi.org/10.1016/J.IJRESMAR.2012.05.001>
- Morgan, D. L., & Scannell, A. (1998). *Planning Focus Group as Qualitative Research.* Sage Publications. <https://dx.doi.org/10.4135/9781412984287>.
- Moshref-Javadi, M., & Winkenbach, M. (2021). Applications and Research Avenues for Drone-Based Models in Logistics: A Classification and Review. *Expert Systems with Applications*, *177*, 114854. <https://doi.org/10.1016/j.eswa.2021.114854>
- Myers, M. D., & Newman, M. (2007). The Qualitative Interview in IS Research: Examining the Craft. *Information and Organization*, *17*(1), 2–26.  
<https://doi.org/10.1016/j.infoandorg.2006.11.001>

- Ni, F., Arnott, D., & Gao, S. (2019). The Anchoring Effect in Business Intelligence Supported Decision-Making. *Journal of Decision Systems*, 28(2), 67–81.  
<https://doi.org/10.1080/12460125.2019.1620573>
- Norton, M. I., Mochon, D., & Ariely, D. (2012). The IKEA Effect: When Labor Leads to Love. *Journal of Consumer Psychology*, 22(3), 453–460.  
<https://doi.org/10.1016/J.JCPS.2011.08.002>
- Odean, T. (1998). Volume, Volatility, Price, and Profit When All Traders Are Above Average. *The Journal of Finance*, 53(6), 1887–1934. <https://doi.org/10.1111/0022-1082.00078>
- O.Nyumba, T., Wilson, K., Derrick, C. J., & Mukherjee, N. (2018). The Use of Focus Group Discussion Methodology: Insights From Two Decades of Application in Conservation. *Methods in Ecology and Evolution / British Ecological Society*, 9(1), 20–32.  
<https://doi.org/10.1111/2041-210x.12860>
- OpenAI. (2023). *Language Models can Explain Neurons in Language Models*.  
<https://openai.com/research/language-models-can-explain-neurons-in-language-models> (accessed on Oct 30, 2023)
- Padilla, R., Netto, S. L., & da Silva, E. A. B. (2020). A Survey on Performance Metrics for Object-Detection Algorithms. *2020 International Conference on Systems, Signals and Image Processing (IWSSIP)*, 237–242. <https://doi.org/10.1109/IWSSIP48289.2020.9145130>
- Panigutti, C., Hamon, R., Hupont, I., Fernandez Llorca, D., Fano Yela, D., Junklewitz, H., Scalzo, S., Mazzini, G., Sanchez, I., Soler Garrido, J., & Gomez, E. (2023). The Role of Explainable AI in the Context of the AI Act. *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 1139–1150.  
<https://doi.org/10.1145/3593013.3594069>
- Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2000). A Model for Types and Levels of Human Interaction With Automation. *IEEE Transactions on Systems, Man, and Cybernetics*, 30(3), 286–297. <https://doi.org/10.1109/3468.844354>
- Pasztor, A., & Ferek, K. S. (2021). *FAA Approves First Fully Automated Commercial Drone Flights*. The Wall Street Journal. <https://www.wsj.com/articles/faa-approves-first-fully-automated-commercial-drone-flights-11610749377> (accessed on Nov 14, 2022)
- Phillips-Wren, G., Jefferson, T., & McKniff, S. (2019). Cognitive bias and decision aid use under stressful conditions. *Journal of Decision Systems*, 28(2), 162–184.  
<https://doi.org/10.1080/12460125.2019.1643695>

- Piehlmaier, D. M. (2022). Overconfidence and the Adoption of Robo-Advice: Why Overconfident Investors Drive the Expansion of Automated Financial Advice. *Financial Innovation*, 8(1), 14. <https://doi.org/10.1186/s40854-021-00324-3>
- Pitt, L. F., Parent, M., Junglas, I., Chan, A., & Spyropoulou, S. (2011). Integrating the Smartphone Into a Sound Environmental Information Systems Strategy: Principles, Practices and a Research Agenda. *The Journal of Strategic Information Systems*, 20(1), 27–37. <https://doi.org/10.1016/j.jsis.2010.09.005>
- Prasad, S. G., Sharmila, V. C., & Badrinarayanan, M. K. (2023). Role of Artificial Intelligence based Chat Generative Pre-trained Transformer (ChatGPT) in Cyber Security. *2023 2nd International Conference on Applied Artificial Intelligence and Computing (ICAAIC)*, 107–114. <https://doi.org/10.1109/ICAAIC56838.2023.10141395>
- Pumplun, L., Fecho, M., Wahl, N., Peters, F., & Buxmann, P. (2021). Adoption of Machine Learning Systems for Medical Diagnostics in Clinics: Qualitative Interview Study. *Journal of Medical Internet Research*, 23(10), e29301. <https://doi.org/10.2196/29301>
- Pumplun, L., Peters, F., Gawlitza, J. F., & Buxmann, P. (2023). Bringing Machine Learning Systems into Clinical Practice: A Design Science Approach to Explainable Machine Learning-Based Clinical Decision Support Systems. *Journal of the Association for Information Systems*, 24(4), 953–979. <https://doi.org/10.17705/1jais.00820>
- Quine, W. V. O. (1987). *Quiddities: An Intermittently Philosophical Dictionary*. Harvard University Press.
- Radboud University Medical Center. (2023). *Find the Artificial Intelligence Based Software for Radiology That You Are Looking For*. AI for Radiology. <https://grand-challenge.org/aiforradiology/> (accessed on Feb 09, 2023)
- Radtke, T., Liszewska, N., Horodyska, K., Boberska, M., Schenkel, K., & Luszczynska, A. (2019). Cooking Together: The IKEA Effect on Family Vegetable Intake. *British Journal of Health Psychology*, 24(4), 896–912. <https://doi.org/10.1111/BJHP.12385>
- Raghoobar, S., van Kleef, E., & de Vet, E. (2017). Self-Crafting Vegetable Snacks: Testing the IKEA-Effect in Children. *British Food Journal*, 119(6), 1301–1312. <https://doi.org/10.1108/BFJ-09-2016-0443>
- Rahman, M. A., & Wang, Y. (2016). Optimizing Intersection-Over-Union in Deep Neural Networks for Image Segmentation. *Advances in Visual Computing*, 234–244. [https://doi.org/10.1007/978-3-319-50835-1\\_22](https://doi.org/10.1007/978-3-319-50835-1_22)

- Rai, A., Constantinides, P., & Sarker, S. (2019). Next Generation Digital Platforms: Toward Human-AI Hybrids. *MIS Quarterly*, 43(1), iii-ix. <http://wrap.warwick.ac.uk/113653>
- Randall, T., Terwiesch, C., & Ulrich, K. T. (2007). Research Note—User Design of Customized Products. *Marketing Science*, 26(2), 268–280. <https://doi.org/10.1287/mksc.1050.0116>
- Rao, A., Kim, J., Kamineni, M., Pang, M., Lie, W., & Succi, M. D. (2023). Evaluating ChatGPT as an Adjunct for Radiologic Decision-Making. *MedRxiv : The Preprint Server for Health Sciences*. <https://doi.org/10.1101/2023.02.02.23285399>
- Regueras, L. M., Verdu, E., Munoz, M. F., Perez, M. A., de Castro, J. P., & Verdu, M. J. (2009). Effects of Competitive E-Learning Tools on Higher Education Students: A Case Study. *IEEE Transactions on Education*, 52(2), 279–285. <https://doi.org/10.1109/TE.2008.928198>
- Reyes, M., Meier, R., Pereira, S., Silva, C. A., Dahlweid, F.-M., von Tengg-Kobligk, H., Summers, R. M., & Wiest, R. (2020). On the Interpretability of Artificial Intelligence in Radiology: Challenges and Opportunities. *Radiology. Artificial Intelligence*, 2(3), e190043. <https://doi.org/10.1148/ryai.2020190043>
- Rezatofghi, H., Tsoi, N., Gwak, J., Sadeghian, A., Reid, I., & Savarese, S. (2019). Generalized Intersection Over Union: A Metric and a Loss for Bounding Box Regression. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 658–666. <https://doi.org/10.48550/arXiv.1902.09630>
- Rice, S., Tamilselvan, G., Winter, S. R., Milner, M. N., Anania, E. C., Sperlak, L., & Marte, D. A. (2018). Public Perception of UAS Privacy Concerns: A Gender Comparison. *Journal of Unmanned Vehicle Systems*, 6(2), 83–99. <https://doi.org/10.1139/juvs-2017-0011>
- Riedl, M. O. (2019). Human - Centered Artificial Intelligence and Machine Learning. *Human Behavior and Emerging Technologies*, 1(1), 33–36. <https://doi.org/10.1002/hbe2.117>
- Rosenfeld, A., & Richardson, A. (2019). Explainability in Human-Agent Systems. *Autonomous Agents and Multi-Agent Systems*, 33(6), 673–705. <https://doi.org/10.1007/s10458-019-09408-y>
- Roy, A. M., Bose, R., & Bhaduri, J. (2022). A Fast Accurate Fine-Grain Object Detection Model Based on YOLOv4 Deep Neural Network. *Neural Computing & Applications*, 34(5), 3895–3921. <https://doi.org/10.1007/s00521-021-06651-x>
- Rudin, C. (2019). Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead. *Nature Machine Intelligence*, 1(5), 206–215. <https://doi.org/10.1038/s42256-019-0048-x>



- Rudin, C., & Wagstaff, K. L. (2014). Machine Learning for Science and Society. *Machine Learning*, 95(1), 1–9. <https://doi.org/10.1007/s10994-013-5425-9>
- Russell, S., & Norvig, P. (2016). *Artificial Intelligence: A Modern Approach* (3rd ed.). Boston: Addison Wesley.
- Russell, S., & Norvig, P. (2021). *Artificial Intelligence, Global Edition a Modern Approach*. Pearson Deutschland.
- Salama, K., Kazmierczak, J., & Schut, D. (2021). Practitioners Guide to MLOps: A Framework for Continuous Delivery and Automation of Machine Learning. Google Cloud White Paper.
- Saldana, J. (2021). *The Coding Manual for Qualitative Researchers*. SAGE.
- Sambasivan, N., Kapania, S., Highfill, H., Akrong, D., Paritosh, P., & Aroyo, L. M. (2021). “Everyone Wants to Do the Model Work, Not the Data Work”: Data Cascades in High-Stakes AI. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–15. <https://doi.org/10.1145/3411764.3445518>
- Samuel, A. L. (1959). Some Studies in Machine Learning Using the Game of Checkers. *IBM Journal of Research and Development*, 3(3), 210–229. <https://doi.org/10.1147/rd.33.0210>
- Sarker, S., Xiao, X., & Beaulieu, T. (2013). Guest Editorial: Qualitative Studies in Information Systems: A Critical Review and Some Guiding Principles. *MIS Quarterly*, 37(4), iii–xviii. <http://www.jstor.org/stable/43825778>
- Schepman, A., & Rodway, P. (2020). Initial Validation of the General Attitudes Towards Artificial Intelligence Scale. *Computers in Human Behavior Reports*, 1, 100014. <https://doi.org/10.1016/j.chbr.2020.100014>
- Schewe, C. D., & Dillon, W. R. (1978). Marketing Information System Utilization: An Application of Self-Concept Theory. *Journal of Business Research*, 6(1), 67–79. [https://doi.org/10.1016/0148-2963\(78\)90020-6](https://doi.org/10.1016/0148-2963(78)90020-6)
- Schuetz, S., & Venkatesh, V. (2020). The Rise of Human Machines: How Cognitive Computing Systems Challenge Assumptions of User-System Interaction. *Journal of the Association for Information*, 21(2), 460–482. <https://papers.ssrn.com/abstract=3680306>
- Scott, J., & Scott, C. (2017). Drone Delivery Models for Healthcare. *Proceedings of the 50th Hawaii International Conference on System Sciences (2017)*. <https://doi.org/10.24251/HICSS.2017.399>
- Sculley, Holt, & Golovin. (2015). Hidden Technical Debt in Machine Learning Systems. *Advances in Neural Information Processing Systems*, 28.

- Seeber, I., Bittner, E., Briggs, R. O., de Vreede, T., de Vreede, G.-J., Elkins, A., Maier, R., Merz, A. B., Oeste-Reiß, S., Randrup, N., Schwabe, G., & Söllner, M. (2020). Machines as Teammates: A Research Agenda on AI in Team Collaboration. *Information & Management*, 57(2), 103174. <https://doi.org/10.1016/j.im.2019.103174>
- Seidel, S., Berente, N., Lindberg, A., Lyytinen, K., & Nickerson, J. V. (2018). Autonomous Tools and Design: A Triple-Loop Approach to Human-Machine Learning. *Communications of the ACM*, 62(1), 50–57. <https://doi.org/10.1145/3210753>
- Shapiro, L. (2004). *Something From the Oven: Reinventing Dinner in 1950s America*. New York: Viking.
- Shen, J., Zhang, C. J. P., Jiang, B., Chen, J., Song, J., Liu, Z., He, Z., Wong, S. Y., Fang, P.-H., & Ming, W.-K. (2019). Artificial Intelligence Versus Clinicians in Disease Diagnosis: Systematic Review. *JMIR Medical Informatics*, 7(3), e10010. <https://doi.org/10.2196/10010>
- Siemon, D. (2022). Elaborating Team Roles for Artificial Intelligence-based Teammates in Human-AI Collaboration. *Group Decision and Negotiation*, 31(5), 871–912. <https://doi.org/10.1007/s10726-022-09792-z>
- Silva, C., & Ribeiro, B. (2011). Purging False Negatives in Cancer Diagnosis Using Incremental Active Learning. *Intelligent Data Engineering and Automated Learning - IDEAL 2011*, 394–402. [https://doi.org/10.1007/978-3-642-23878-9\\_47](https://doi.org/10.1007/978-3-642-23878-9_47)
- Smith, G. (2020). Data Mining Fool's Gold. *Journal of Information Technology*, 35(3), 182–194. <https://doi.org/10.1177/0268396220915600>
- Sovrano, F., Sapienza, S., Palmirani, M., & Vitali, F. (2022). Metrics, Explainability and the European AI Act Proposal. *J*, 5(1), 126–138. <https://doi.org/10.3390/j5010010>
- Stone, E. (2010). T Test, Paired Samples. *Encyclopedia of Research Design*.
- Strich, F., Mayer, A.-S., & Fiedler, M. (2021). What Do I Do in a World of Artificial Intelligence? Investigating the Impact of Substitutive Decision-Making AI Systems on Employees' Professional Role Identity. *Journal of the Association for Information Systems*, 22(2), 304–324. <https://doi.org/10.17705/1jais.00663>
- Studer, S., Bui, T. B., Drescher, C., Hanuschkin, A., Winkler, L., Peters, S., & Müller, K.-R. (2021). Towards CRISP-ML(Q): A Machine Learning Process Model With Quality Assurance Methodology. *Machine Learning and Knowledge Extraction*, 3(2), 392–413. <https://doi.org/10.3390/make3020020>
- Stumpf, S., Rajaram, V., Li, L., Wong, W.-K., Burnett, M., Dietterich, T., Sullivan, E., & Herlocker, J. (2009). Interacting Meaningfully With Machine Learning Systems: Three Experiments.

- International Journal of Human-Computer Studies*, 67(8), 639–662.  
<https://doi.org/10.1016/j.ijhcs.2009.03.004>
- Sturm, T., Gerlach, J. P., Pumplun, L., Mesbah, N., Peters, F., Tauchert, C., Nan, N., & Buxmann, P. (2021). Coordinating Human and Machine Learning for Effective Organizational Learning. *MIS Quarterly*, 45(3). <https://doi.org/10.25300/MISQ/2021/16543>
- Sturm, T., Koppe, T., Scholz, Y., & Buxmann, P. (2021). The Case of Human-Machine Trading as Bilateral Organizational Learning. *ICIS 2021 Proceedings*.  
[https://aisel.aisnet.org/icis2021/ai\\_business/ai\\_business/3/](https://aisel.aisnet.org/icis2021/ai_business/ai_business/3/)
- Sturm, T., Pumplun, L., Gerlach, J. P., Kowalczyk, M., & Buxmann, P. (2023). Machine Learning Advice in Managerial Decision-Making: The Overlooked Role of Decision Makers' Advice Utilization. *The Journal of Strategic Information Systems*, 32(4), 101790.  
<https://doi.org/10.1016/j.jsis.2023.101790>
- Sutton, R. T., Pincock, D., Baumgart, D. C., Sadowski, D. C., Fedorak, R. N., & Kroeker, K. I. (2020). An Overview of Clinical Decision Support Systems: Benefits, Risks, and Strategies for Success. *NPJ Digital Medicine*, 3, 17. <https://doi.org/10.1038/s41746-020-0221-y>
- Sveiby, K. E. (1997). *The New Organizational Wealth: Managing & Measuring Knowledge-based Assets*. Berrett-Koehler Publishers.
- Sweller, J., van Merriënboer, J. J. G., & Paas, F. G. W. C. (1998). Cognitive Architecture and Instructional Design. *Educational Psychology Review*, 10(3), 251–296.  
<https://doi.org/10.1023/A:1022193728205>
- Swets, J. A., Dawes, R. M., & Monahan, J. (2000). Psychological Science Can Improve Diagnostic Decisions. *Psychological Science in the Public Interest: A Journal of the American Psychological Society*, 1(1), 1–26. <https://doi.org/10.1111/1529-1006.001>
- Tabachnick, B. G., & Fidell, L. S. (2006). *Experimental Designs Using ANOVA*. Brooks/Cole.
- Tan, L. K. L., Lim, B. C., Park, G., Low, K. H., & Seng Yeo, V. C. (2021). Public Acceptance of Drone Applications in a Highly Urbanized Environment. *Technology in Society*, 64, 101462.  
<https://doi.org/10.1016/j.techsoc.2020.101462>
- Teubner, T., Flath, C. M., Weinhardt, C., van der Aalst, W., & Hinz, O. (2023). Welcome to the Era of ChatGPT et al. *Business & Information Systems Engineering*, 65(2), 95–101.  
<https://doi.org/10.1007/s12599-023-00795-x>
- Thiebes, S., Lins, S., & Sunyaev, A. (2021). Trustworthy Artificial Intelligence. *Electronic Markets*, 31(2), 447–464. <https://doi.org/10.1007/s12525-020-00441-4>

- Todd, P., & Benbasat, I. (1991). An Experimental Investigation of the Impact of Computer Based Decision Aids on Decision Making Strategies. *Information Systems Research*, 2(2), 87–115. <https://doi.org/10.1287/isre.2.2.87>
- Tremblay, M. C., Hevner, A. R., & Berndt, D. J. (2010). The Use of Focus Groups in Design Science Research. In A. Hevner & S. Chatterjee (Eds.), *Design Research in Information Systems: Theory and Practice* (pp. 121–143). Springer US. [https://doi.org/10.1007/978-1-4419-5653-8\\_10](https://doi.org/10.1007/978-1-4419-5653-8_10)
- Treveil, M., Omont, N., Stenac, C., Lefevre, K., Phan, D., Zentici, J., Lavoillotte, A., Miyazaki, M., & Heidmann, L. (2020). *Introducing MLOps*. O'Reilly Media, Inc.
- Turel, O., Serenko, A., & Bontis, N. (2007). User Acceptance of Wireless Short Messaging Services: Deconstructing Perceived Value. *Information & Management*, 44(1), 63–73. <https://doi.org/10.1016/J.IM.2006.10.005>
- UK Airprox Board. (2021). *Analysis of Airprox in UK Airspace* (No. 37). <https://www.airproxboard.org.uk/media/lisbn3ze/book-37-ukab-annual-report-2021.pdf> (accessed on Aug 18, 2023)
- United Nations. (2015). *Take Action for the Sustainable Development Goals*. United Nations Sustainable Development; United Nations: Sustainable Development Goals. <https://www.un.org/sustainabledevelopment/sustainable-development-goals/> (accessed on Jun 10, 2022)
- Unterberg, A. W., Stover, J., Kress, B., & Kiening, K. L. (2004). Edema and Brain Trauma. *Neuroscience*, 129(4), 1021–1029. <https://doi.org/10.1016/j.neuroscience.2004.06.046>
- Van den Broek, E., Sergeeva, A., & Huysman, M. (2021). When the Machine Meets the Expert: An Ethnography of Developing AI for Hiring. *MIS Quarterly*, 45(3), 1557–1580. <https://doi.org/10.25300/MISQ/2021/16559>
- van Leeuwen, K. G., Schalekamp, S., Rutten, M. J. C. M., van Ginneken, B., & de Rooij, M. (2021). Artificial Intelligence in Radiology: 100 Commercially Available Products and Their Scientific Evidence. *European Radiology*, 31(6), 3797–3804. <https://doi.org/10.1007/s00330-021-07892-z>
- Van Someren, M., Barnard, Y. F., & Sandberg, J. (1994). The Think Aloud Method: A Practical Approach to Modelling Cognitive. *London: Academic Press*, 11, 29–41. <http://hdl.handle.net/11245/1.103289>
- van Wynsberghe, A. (2021). Sustainable AI: AI for Sustainability and the Sustainability of AI. *AI and Ethics*, 1(3), 213–218. <https://doi.org/10.1007/s43681-021-00043-6>

- Vela, D., Sharp, A., Zhang, R., Nguyen, T., Hoang, A., & Pianykh, O. S. (2022). Temporal Quality Degradation in AI Models. *Scientific Reports*, *12*(1), 11654. <https://doi.org/10.1038/s41598-022-15245-z>
- Venkatesh, V., Morris, M. G., Davis, G. B., & Davis, F. D. (2003). User Acceptance of Information Technology: Toward a Unified View. *MIS Quarterly*, *27*(3), 425–478. <https://doi.org/10.2307/30036540>
- Venkatesh, V., Thong, J. Y. L., & Xu, X. (2012). Consumer Acceptance and Use of Information Technology: Extending the Unified Theory of Acceptance and Use of Technology. *MIS Quarterly*, *36*(1), 157–178. <https://doi.org/10.2307/41410412>
- Venkatesh, V., Thong, J. Y. L., & Xu, X. (2016). 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. <https://ssrn.com/abstract=2800121>
- Vössing, M., Kühl, N., Lind, M., & Satzger, G. (2022). Designing Transparency for Effective Human-AI Collaboration. *Information Systems Frontiers*, *24*(3), 877–895. <https://doi.org/10.1007/s10796-022-10284-3>
- Walasek, L., Rakow, T., & Matthews, W. J. (2017). When Does Construction Enhance Product Value? Investigating the Combined Effects of Object Assembly and Ownership on Valuation. *Journal of Behavioral Decision Making*, *30*(2), 144–156. <https://doi.org/10.1002/BDM.1931>
- Wardle, J., & Pope, R. (1992). The Psychological Costs of Screening for Cancer. *Journal of Psychosomatic Research*, *36*(7), 609–624. [https://doi.org/10.1016/0022-3999\(92\)90051-3](https://doi.org/10.1016/0022-3999(92)90051-3)
- Weber, R. P. (1990). *Basic Content Analysis*. SAGE.
- Wenkel, S., Alhazmi, K., Liiv, T., Alrshoud, S., & Simon, M. (2021). Confidence Score: The Forgotten Dimension of Object Detection Performance Evaluation. *Sensors*, *21*(13). <https://doi.org/10.3390/s21134350>
- Wirth, & Hipp. (2000). 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*, *1*, 29–39.
- Wuest, T., Weimer, D., Irgens, C., & Thoben, K.-D. (2016). Machine Learning in Manufacturing: Advantages, Challenges, and Applications. *Production & Manufacturing Research*, *4*(1), 23–45. <https://doi.org/10.1080/21693277.2016.1192517>

- Xi, N., Chen, J., Gama, F., Riar, M., & Hamari, J. (2023). The Challenges of Entering the Metaverse: An Experiment on the Effect of Extended Reality on Workload. *Information Systems Frontiers*, 25(2), 659–680. <https://doi.org/10.1007/s10796-022-10244-x>
- Yin, M., Wortman Vaughan, J., & Wallach, H. (2019). Understanding the Effect of Accuracy on Trust in Machine Learning Models. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–12. <https://doi.org/10.1145/3290605.3300509>
- Yuan Zhang, M., & Jessie Yang, X. (2017). Evaluating Effects of Workload on Trust in Automation, Attention Allocation and Dual-Task Performance. *Proceedings of the Human Factors and Ergonomics Society*, 61(1), 1799–1803. <https://doi.org/10.1177/1541931213601932>
- Yushkevich, P. A., Piven, J., Hazlett, H. C., Smith, R. G., Ho, S., Gee, J. C., & Gerig, G. (2006). User-Guided 3D Active Contour Segmentation of Anatomical Structures: Significantly Improved Efficiency and Reliability. *NeuroImage*, 31(3), 1116–1128. <https://doi.org/10.1016/j.neuroimage.2006.01.015>
- Zercher, D., Jussupow, E., & Heinzl, A. (2023). When AI joins the Team: A Literature Review on Intragroup Processes and their Effect on Team Performance in Team-AI Collaboration. *ECIS 2023 Research Papers*, 307. [https://aisel.aisnet.org/ecis2023\\_rp/307/](https://aisel.aisnet.org/ecis2023_rp/307/)
- Žliobaitė, I. (2010). Learning under Concept Drift: an Overview. In *arXiv [cs.AI]*. arXiv. <https://doi.org/10.48550/arXiv.1010.4784>

