JULIUS-MAXIMILIANS-UNIVERSITÄT WÜRZBURG

FACULTY OF BUSINESS MANAGEMENT AND ECONOMICS

# Algorithmic Decision-Making Facilities: Perception and Design of Explainable AI-based Decision Support Systems

Inauguraldissertation

to achieve the academic degree

Doctor rerum politicarum (Dr. rer. pol.)

by:

Lukas-Valentin Herm, M.Sc.

First reviewer:
**Prof. Dr. Christian Janiesch**
Chair of Enterprise Computing
Technische Universität Dortmund

Second reviewer:
**Prof. Dr. Axel Winkelmann**
Chair of Business Management and Business Information Systems
Julius-Maximilians-Universität Würzburg

| | |
|---|---|
| Date of submission: | 10.05.2023 |
| Date of defense: | 28.07.2023 |

# Acknowledgments

*"We can only see a short distance ahead, but we can see plenty there that needs to be done."*
Alan Turing

When I began my doctoral studies, I was not aware of how and where the journey would take me, but I knew that it would require a great deal of effort, consistency, and dedication, something that could only be maintained with a decent research group. This cumulative doctoral thesis was consequently only possible with the motivation and support of the many amazing people I have met along my scientific journey in the past four years.

First, I would like to thank my mentor and first supervisor, Christian Janiesch, who guided and supported me through the jungle of academic publishing by constantly pushing the boundaries of what I deemed possible. His dedication, knowledge, and persistence prompted me to pursue higher standards making this thesis achievable. I would also like to thank my second supervisor, Axel Winkelmann, for his support during my doctoral studies and the warm welcome I received when moving to his chair. In hindsight, working at both chairs helped me to grow professionally and personally while constantly challenging myself.

Furthermore, my special thanks go to Jonas Wanner, who encouraged me to contemplate an academic career and, thereon, to focus on this doctoral journey. Looking back, it is incredible how many days and nights we worked, despaired, and cried together to meet (conference) deadlines; we always set our targets much too late. Thank you so much, buddy! Also, I would like to thank all my co-authors, particularly Kai Heinrich, Franz Seubert, and Theresa Steinbach, for the fruitful and constructive discussions during our research. As I became part of the Chair of Business Management and Business Information Systems, I had the pleasure of meeting awesome colleagues. Together, we not only had fun at work but also spent nights out, forced us into sports competitions, and went on vacations. Over the past two years, you have filled this time with unforgettable memories. To my fellow researchers - Nicolas Neis, Christoph Tomitza, Myriam Schaschek, Fabian Gwinner, Lisa Straub, Christian Zeiß, and Norman Pytel - thank you for brightening my path, even when days were dreary!

Beyond all academic contributors, I am more than grateful to my entire family and all of my friends for always inspiring me and backing me up. Last but most importantly, I would like to express my deepest gratitude to my fiancée Marie-Theres, who has been with me from the beginning and has always stood by my side. Without your constant love, support, and encouragement, this journey would not have been doable at all!

Thank you, all!

# Abstract

Recent computing advances are driving the integration of artificial intelligence (AI)-based systems into nearly every facet of our daily lives. To this end, AI is becoming a frontier for enabling algorithmic decision-making by mimicking or even surpassing human intelligence. Thereupon, these AI-based systems can function as decision support systems (DSSs) that assist experts in high-stakes use cases where human lives are at risk. All that glitters is not gold, due to the accompanying complexity of the underlying machine learning (ML) models, which apply mathematical and statistical algorithms to autonomously derive nonlinear decision knowledge. One particular subclass of ML models, called deep learning models, accomplishes unsurpassed performance, with the drawback that these models are no longer explainable to humans. This divergence may result in an end-user's unwillingness to utilize this type of AI-based DSS, thus diminishing the end-user's system acceptance.

Hence, the explainable AI (XAI) research stream has gained momentum, as it develops techniques to unravel this black-box while maintaining system performance. Non-surprisingly, these XAI techniques become necessary for justifying, evaluating, improving, or managing the utilization of AI-based DSSs. This yields a plethora of explanation techniques, creating an XAI jungle from which end-users must choose. In turn, these techniques are preliminarily engineered by developers for developers without ensuring an actual end-user fit. Thus, it renders unknown how an end-user's mental model behaves when encountering such explanation techniques.

For this purpose, this cumulative thesis seeks to address this research deficiency by investigating end-user perceptions when encountering intrinsic ML and post-hoc XAI explanations. Drawing on this, the findings are synthesized into design knowledge to enable the deployment of XAI-based DSSs in practice. To this end, this thesis comprises six research contributions that follow the iterative and alternating interplay between behavioral science and design science research employed in information systems (IS) research and thus contribute to the overall research objectives as follows: First, an in-depth study of the impact of transparency and (initial) trust on end-user acceptance is conducted by extending and validating the unified theory of acceptance and use of technology model. This study indicates both factors' strong but indirect effects on system acceptance, validating further research incentives. In particular, this thesis focuses on the overarching concept of transparency. Herein, a systematization in the form of a taxonomy and pattern analysis of existing user-centered XAI studies is derived to structure and guide future research endeavors, which enables the empirical investigation of the theoretical trade-off between performance and explainability in intrinsic ML algorithms, yielding a less gradual trade-off, fragmented into three explainability groups. This includes an empirical investigation on end-users' perceived explainability of post-hoc explanation types, with local explanation types performing best. Furthermore, an empirical investigation emphasizes the correlation between comprehensibility and explainability, indicating almost significant (with outliers) results for the assumed correlation. The final empirical investigation aims at researching XAI explanation types on end-user cognitive load and the effect of cognitive load on end-user task performance and task time, which also positions local explanation types as best and demonstrates the correlations between cognitive load and task performance and, moreover, between cognitive load and task time. Finally, the last research paper utilizes i.a. the obtained knowledge and derives a nascent design theory for XAI-based DSSs. This design theory encompasses (meta-) design requirements, design principles, and design features in a domain-independent and interdisciplinary fashion, including end-users and developers as potential user groups. This design theory is ultimately tested through a real-world instantiation in a high-stakes maintenance scenario.

From an IS research perspective, this cumulative thesis addresses the lack of research on perception and design knowledge for an ensured utilization of XAI-based DSS. This lays the foundation for future research to obtain a holistic understanding of end-users' heuristic behaviors during decision-making to facilitate the acceptance of XAI-based DSSs in operational practice.

# Zusammenfassung

Jüngste technische und algorithmische Fortschritte treiben die Integration von Systemen auf der Basis von künstlicher Intelligenz (KI) in nahezu alle Bereiche unseres täglichen Lebens voran. Inzwischen sind diese Systeme in der Lage, menschliche Intelligenz anhand von algorithmischer Entscheidungsfindung nachzuahmen und sogar zu übertreffen. Insbesondere können KI-basierte Systeme als Entscheidungsunterstützungssysteme (Decision Support Systems - DSS) dienen und damit Domänenexperten in hochsensiblen Anwendungsfällen helfen, bei denen Menschenleben auf dem Spiel stehen.

Dies resultiert in komplexen Modellen des maschinellen Lernens (ML), welche mathematische und statistische Algorithmen benutzen, um nichtlineares Entscheidungswissen automatisch abzuleiten. Besonders eine Unterklasse von ML-Modellen, die sogenannten Deep-Learning-Modelle (DL-Modelle), erreichen eine unübertroffene Leistung. Sie haben allerdings den Nachteil, dass sie für den Menschen nicht mehr nachvollziehbar sind. Diese Divergenz kann jedoch dazu führen, dass Endanwender nicht bereit sind, diese Art von KI-basierten DSS zu benutzen. Dadurch wird die Akzeptanz solcher Systeme beeinträchtigt.

Um dieses Problem anzugehen, ist der Forschungszweig der erklärbaren KI (Explainable Artificial Intelligence - XAI) entstanden. Darin werden Techniken und Methoden entwickelt, die das wahrgenommene Blackbox-Verhalten dieser Modelle aufbrechen. Die XAI-Methoden können für KI-basierte DSS eingesetzt werden und ermöglichen es, Entscheidungen und Modelle zu rechtfertigen, zu bewerten, zu verbessern und zu verwalten. Dieser Ansatz resultiert jedoch in einer Vielzahl von Erklärungstechniken, aus denen die Anwender eine passende Erklärung wählen müssen. Gleichzeitig werden diese Methoden zurzeit primär von Entwicklern für Entwickler konzipiert, ohne, dass dabei ermittelt wird, ob eine tatsächliche Eignung für den Endanwender gewährleistet ist. Im Umkehrschluss ist daher unbekannt, wie sich das mentale Modell eines Endanwenders im Umgang mit solchen Erklärungstechniken verhält.

Die vorliegende kumulative Dissertation thematisiert dieses Forschungsdefizit, indem die Wahrnehmung des Endanwenders im Umgang mit intrinsischen ML- und Post-Hoc-XAI-Erklärungen untersucht wird. Die gewonnenen Erkenntnisse werden in gestaltungsorientiertes Wissen synthetisiert, um den Einsatz von XAI-basierten DSS in der Praxis zu ermöglichen. Zu diesem Zweck umfasst die Dissertation sechs Forschungsbeiträge. Diese richten sich nach dem für den Bereich Information Systems (IS) typischen alternierendem Zusammenspiel zwischen verhaltenswissenschaftlicher und designorientierter Forschung und tragen wie folgt zu den übergeordneten Forschungszielen bei:

Zu Beginn erfolgt durch Erweiterung und Validierung des Modells Unified Theory of Acceptance and Use of Technology eine Untersuchung des Einflusses von Transparenz und (initialem) Vertrauen auf die Akzeptanz der Endanwender. Die Studie zeigt einen starken, aber indirekten Effekt beider Faktoren auf die Systemakzeptanz und liefert damit die wissenschaftliche Bestätigung für weitere Forschungsinitiativen. Diese Arbeit konzentriert sich insbesondere auf diesen übergeordneten Einflussfaktor Transparenz. Darauf aufbauend wird eine Systematisierung in Form einer Taxonomie und Analyse bestehender nutzerzentrierter XAI-Studien durchgeführt, um zukünftige Forschungsbestrebungen zu strukturieren. Diese Systematisierung ermöglicht anschließend empirische Untersuchungen weiterer Einflussfaktoren auf die Endanwenderwahrnehmung. Zunächst wird eine Untersuchung des theoretischen Zielkonflikts zwischen Leistung und Erklärbarkeit in intrinsischen ML-Algorithmen vorgenommen, welche eine dreiteilige Anordnung empirisch bestätigt. Ebenso erfolgt eine empirische Untersuchung der vom Endanwender wahrgenommenen Erklärbarkeit von Post-Hoc-Erklärungstypen, wobei hier lokale Erklärungstypen am besten abschneiden. Anschließend wird der Zusammenhang zwischen Verständlichkeit und Erklärbarkeit betrachtet, wobei sich eine überwiegend (mit Ausreißern) signifikante Korrelation aufzeigen lässt. Der letzte Teil der empirischen Untersuchungen widmet sich dem Einfluss von XAI-Erklärungstypen auf die kognitive Belastung und die Auswirkung dieser Belastung auf die Aufgabenleistung und -zeit des Endanwenders. Hier zeigt sich, dass lokale Erklärungstypen ebenfalls verhältnismäßig am besten abschneiden und die Korrelationen zwischen kognitiver Belastung und Aufgabenleistung sowie kognitiver Belastung und Aufgabenzeit gegeben sind. Der letzte Forschungsbeitrag fügt u. a. die Ergebnisse zusammen und leitet daraus eine Design-Theorie für XAI-basierte DSS ab. Diese Design Theorie umfasst (Meta-)Design-Anforderungen, Design-Prinzipien und Design-Merkmale in einer domänenunabhängigen und interdisziplinären Art und Weise, welche den Einbezug sowohl von Endanwendern als auch von Entwicklern als potenzielle Nutzergruppen ermöglicht.

Aus der Perspektive der IS Disziplin widmet sich diese kumulative Dissertation dem Mangel an Forschung zu Wahrnehmung und Designwissen für eine gesicherte Nutzung von XAI-basierten DSS. Damit legt sie den Grundstein für zukünftige Forschung, um ein ganzheitliches Verständnis des heuristischen Verhaltens der Endanwender während der Entscheidungsfindung zu erlangen und somit die Akzeptanz von XAI-basierten DSS in der betrieblichen Praxis zu fördern.

# Contents

# List of Figures

# List of Tables

# List of Abbreviations

| | |
|---|---|
| AB | Ability Beliefs |
| AI | Artificial Intelligence |
| ANN | Artificial Neural Networks |
| ANOVA | Analysis of Variance |
| ATT | Attitude Towards Technology |
| AVE | Average Variance Extracted |
| BI | Behavioral Intention |
| C-MAPSS | Commercial Modular Aero-Propulsion System Simulation |
| CA | Cronbach's Alpha |
| CFT | Cognitive Fit Theory |
| CNN | Convolutional Neural Network |
| CQ | Control Question |
| CR | Composite Reliability |
| DF | Degree of Freedom |
| DF | Design Feature |
| DL | Deep Learning |
| DP | Design Principle |
| DR | Design Requirement |
| DSR | Design Science Research |
| DSS | Decision Support System |
| EE | Effort Expectancy |
| EIS | Explainable Intelligent System |
| FL | Fornell-Larcker |
| GDPR | General Data Protection Regulation |
| HCI | Human Computer Interaction |
| I4.0 | Industry 4.0 |
| IS | Information Systems |
| MDR | Meta Design Requirement |

| | |
|---|---|
| ME | Mental Efficiency |
| ML | Machine Learning |
| PE | Performance Expectancy |
| PM | Process Mining |
| RO | Research Objective |
| RQ | Research Question |
| RUL | Remaining Useful Lifetime |
| SHAP | Shapley Additive Explanations |
| ST | System Transparency |
| SVM | Support Vector Machine |
| TAM | Technology Acceptance Model |
| TP | Trust Propensity |
| UI | User Interface |
| UTAUT | Unified Theory of Acceptance and Use in Information Technology |
| XAI | Explainable Artificial Intelligence |

# 1 Introduction

## 1.1 Research Motivation

Artificial intelligence (AI) is a leading advancement in computer technology that is currently expanding the application possibilities for data-driven problem-solving (Berente et al., 2021; Janiesch, Zschech, et al., 2021). Notably, AI-based decision support systems (DSSs) are now capable of performing in high-stakes workspaces (Herm, Steinbach, et al., 2022). Moreover, they can enable professionals and experts to rely almost entirely on AI-based DSSs to perform tasks, even when human lives are at risk (McKinney et al., 2020).

While in the context of information systems (IS) research, the umbrella term AI is defined as a technology-independent, evolving frontier of computational capabilities to mimic or surpass human intelligence, its implementation occurs primarily through machine learning (ML) algorithms (Berente et al., 2021; Russell & Norvig, 2021). This involves deploying statistical and mathematical algorithms to automatically derive nonlinear decision-making knowledge via data analysis without the necessity of conscious programming (Janiesch, Zschech, et al., 2021). In addition, previous research endeavors have centered on surmounting mathematical constraints to improve the performance of the generated ML modes by increasing their inherent algorithmic complexity (Arrieta et al., 2020). In particular, a subclass of ML algorithms, called deep learning (DL), exceeds established ML models in terms of model performance due to their deep neural network structure (Janiesch, Zschech, et al., 2021).

In turn, these deeply integrated ML cores in contemporary AI-based DSSs are no longer traceable by humans. That is, while their decision support quality is constantly increasing, humans are not able to comprehend why an AI-based DSS recommends a decision or how the integrated ML model operates, resulting in a perceived black-box (Dwivedi et al., 2023). Despite the fact that users subconsciously attribute anthropomorphic traits to these systems and thus assume them to be capable of efficiently handling complex tasks, the absence of explanation causes an information asymmetry between the user and the AI-based DSS (Pfeuffer et al., 2019). This may create system acceptance barriers and thus hinder initial user trust development toward the system (McKnight et al., 2002; Shin et al., 2020). Thereon, this can lead to algorithm aversion, which represents the unwillingness of users to rely on and accept a high-performance AI-based DSS within a professional workspace (Berger et al., 2021). This becomes critical since a hybrid intelligence can only be effective if a user relies on the recommendations of an AI-based DSS (Dellermann et al., 2019; Wanner, Herm, et al., 2022a).

To address this information asymmetry, the emerging research stream of explainable AI (XAI) develops techniques to transform the inner decision logic of a black-box model into an explanation comprehensible to humans while maintaining the ML model's performance level (Arrieta et al., 2020; Dwivedi et al., 2023). Although this research stream has gained significant momentum in recent years, drawbacks remain (Meske et al., 2022). Herein, numerous explanation techniques have materialized to cover the different ML algorithm types, task types, and explanation scopes, creating an XAI jungle from which one must select a suitable XAI technique (Das & Rad, 2020; Herm, Wanner, et al., 2022). Moreover, these developed techniques are mostly mathematically driven and have been engineered by developers for developers, resulting in inappropriate explanations for end-users (van der Waa et al., 2021).

To this extent, it remains unknown how end-users perceive these XAI explanations in terms of distinct factors such as transparency, explainability, comprehensibility, or cognitive load, and to what magnitude explanations increase end-user's initial trust and thus system acceptance (Herm, Wanner, et al., 2022). These findings can contribute to the design of a sound knowledge transfer from an explainer (XAI-based DSS) to the explainee (end-user) (Miller, 2019). In light of current XAI developments, it becomes necessary to conduct interdisciplinary research involving research disciplines such as psychology and cognitive science to conduct end-user-centered evaluations rather than focusing solely on technically driven evaluations (Mohseni et al., 2021). Conversely, research also lacks knowledge on how to integrate XAI explanations into a developer-centered XAI-based DSS to facilitate the evaluation, improvement, and management of integrated ML models (Herm, Steinbach, et al., 2022). Bearing this in mind, it is unclear how user perceptions should be incorporated into design knowledge to develop a profound XAI-based DSS that fulfills the needs and requirements of multiple XAI stakeholders and thus also facilitate acceptance. From an IS research perspective, combining behavioral science and design science research (DSR) seems necessary to discuss the role of AI-based explanations (Gregor & Benbasat, 1999; Shin, 2020a) and also facilitate explainable algorithmic decision-making in operational practice (Meske et al., 2022).

Against this background, this cumulative thesis utilizes this IS research procedure to address the lack of knowledge in end-user perception of XAI-based explanations and further the deficiencies in design knowledge to ensure the application of XAI-based DSSs in practice. To accomplish this, the topics of perception and design are combined into a structured research approach by defining two overarching research objectives (RO), of perception (RO1) and design (RO2), for XAI-based DSSs. Furthermore, these ROs are subdivided into sub-ROs that are addressed by six closely related scientific research contributions. The remainder of this section is structured as follows. First, Subsection 1.2 provides a conceptual background that serves as the theoretical foundation. Subsection 1.3 complements this by framing the thesis in an IS-related methodological context. Furthermore, Subsection 1.4 comprehensively describes the ROs and an overview of the included research contributions. Subsection 1.5 places the research contributions in the holistic context of the author's XAI-related publications. Finally, Subsection 1.6 discusses the results of this thesis, before a conclusion and outlook are provided in Subsection 1.7.

## 1.2 Conceptional Background

This section serves as a conceptual background for the rest of the thesis to facilitate an understanding of the theoretical foundation. According to the findings of Gunning et al. (2019), Miller (2019), Herm, Steinbach, et al. (2022), and Wanner, Herm, et al. (2022a), Figure 1.1

provides an overview of an interaction process occurring when using an XAI-based DSS. Building on this, the remainder of the section provides a comprehensive overview of the associated topics of DSSs, AI, user-centric explanations, and XAI.

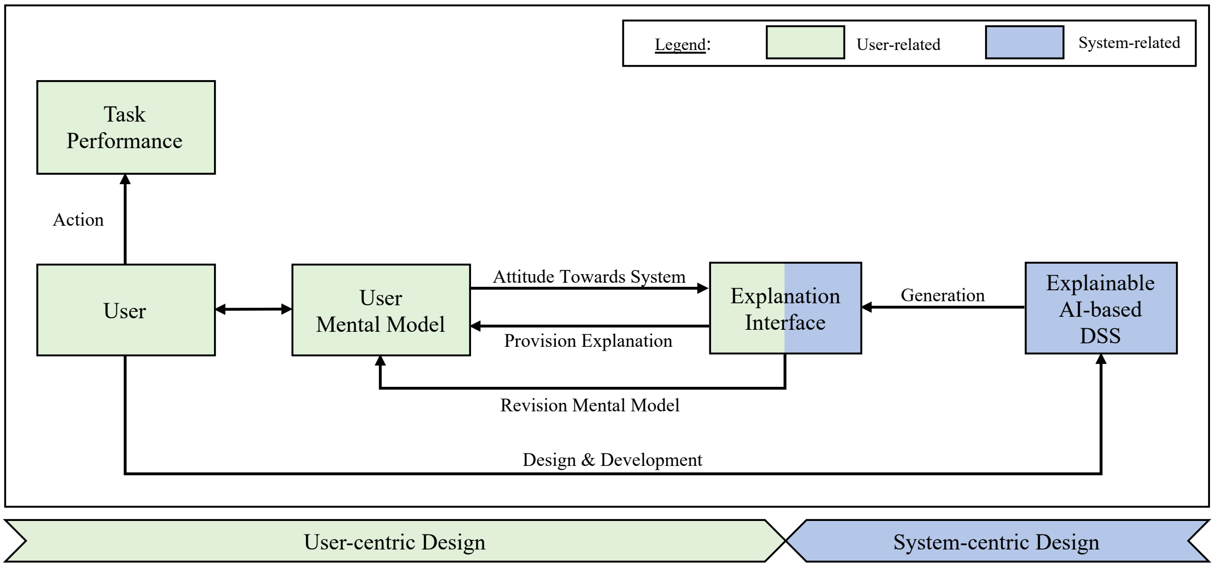


Figure 1.1 User-System Interaction Process

Within a decision scenario, two distinct design components interact with each other. The first is the user-centric one, which embraces the perceptions and actions of a user (e.g., end-user or developer) when processing a task, and the second is the system-centric one, which encompasses the actions that an XAI-based DSS performs during operation (Meske et al., 2022). The former involves a user and the associated mental model receiving an explanation from an XAI-based DSS to perform a task (Gunning, 2019). By receiving an explanation, the user develops an attitude toward the XAI-based DSS (e.g., algorithmic aversion or appreciation) (Wanner, Herm, et al., 2022a) based on their own perception and thus revises their mental model (Berger et al., 2021; Miller, 2019). The latter involves the XAI-based DSS. Here, an XAI-based DSS recommends actions for a decision problem and thus generates an explanation for its reasoning. Due to the strong interaction of both design components, it is essential to incorporate the user's design requirements into the XAI-based DSS to provide an adequate explanation and thus promote the user's acceptance toward the XAI-based DSS (Herm, Steinbach, et al., 2022).

**Decision Support Systems.** Due to the increasing amount of available information and the complexity of use cases, the application of DSSs has gained momentum, especially in the 1970s to 1980s (Liu et al., 2008). Since then, they have supported users, empowering them to make informed and efficient decisions on unstructured or semi-structured problems (Power, 2008), distinguishing them from expert systems designed to replace the actual user (Turban & Watkins, 1986). This becomes particularly necessary when a decision problem arises within a constrained situation requiring an action where multiple decision alternatives exist. This applies to not only new or inexperienced employees in a field but also experienced and qualified employees (Power, 2008).

In this context, a DSS is characterized as any computer-based IS that contribute to the decision-making process by providing recommendations, additional information, and optimization strategies to users that enable them to conduct their work (Arnott & Pervan, 2008). That is, this interactive computer-based system interacts with a user through a human-computer interface. Preliminary research distinguishes between five different types of DSS application developments, with data-driven DSSs (e.g., executive information systems) equipped with analytic processing capabilities valued with the highest functionality due to their ability to analyze complex data (Shim et al., 2002).

In the early days of DSS development, software engineers handcrafted the decision logic within a DSS. In doing so, experienced staff transferred their knowledge into these computationally readable decision rules (Sprague, 1980). With increasing monitoring of real-world events, these rule sets are becoming more complex, making it nearly impossible for software engineers to handcraft them for complex situations (Herm, Steinbach, et al., 2022). AI-based algorithms concurrently provide the foundation for the autonomous creation of complex decision rules through computational breakthroughs that empower users to make complex decisions even in unstructured, time-constrained, or highly sensitive situations, which results in a hybrid intelligence (Dellermann et al., 2019; Janiesch, Zschech, et al., 2021). From an IS research perspective, integrating AI-based knowledge generation and human decision-making into IS enhances organizations' effectiveness and strategic alignment. These types of AI-based DSSs are also called intelligent systems (Gregor & Benbasat, 1999; Mohseni et al., 2021). Within this context, relying on academic principles and design knowledge ensures an effective and valuable intelligent system application in practice (Demigha, 2021).

**Artificial Intelligence.** In recent IS literature, the term AI is not linked to a specific set of technologies but is instead seen as a *"frontier of computational advancements that references human intelligence in addressing ever more complex decision-making problems"* (Berente et al., 2021, p. 5). In this regard, recent AI-based applications reflect not only human-like intelligence but also surpass human capabilities, creating superhuman cognitive capabilities (Herm, Heinrich, et al., 2023; Janiesch, Zschech, et al., 2021). Non-surprisingly, AI-based DSSs are frequently applied (Meske et al., 2022).

While in the early stages of AI research, these inference models, known as symbolic AI, were built by hand (Goodfellow et al., 2016), recent groundbreaking computational advances have enabled the automatic generation of decision knowledge by using mathematical and statistical algorithms (Thiebes et al., 2021). To this end, ML algorithms have received increasing attention in the last few decades (Goodfellow et al., 2016; Russell & Norvig, 2021).

Here, different types of ML-based algorithms have emerged that can be subdivided into three application types: *i)* reinforcement learning, a stateful agent-based approach to interact within an environment (e.g., games or autonomous driving) by performing actions and getting feedback from it, to optimize an overall goal; *ii)* unsupervised learning, an approach to detect patterns in unstructured and unlabeled data (e.g., cluster analysis, dimension reduction); and *iii)* supervised learning, an approach to learn functions that associate input data with output labels to predict new unlabeled data (e.g., classification or regression) (Goodfellow et al., 2016).

Accompanying the opportunities and benefits of ML that impact people's lives throughout all domains, challenges emerge that impede sound ML application (Berente et al., 2021). This also includes the trade-off between performance and explainability, which assumes a linear correlation between the performance (e.g., accuracy) and explainability of an ML model. That is, ML models with less complex decision logic and thus high explainability tend to perform worse than ML models with more complex decision logic but lower explainability (Dam et al., 2018; Gunning et al., 2019). The scholarly literature assumes that this behavior results due to the mathematical concept of the applied ML algorithm. That is, single tree-based (e.g., decision trees) or coefficient-based (e.g., linear regression) models seems to be more traceable for users compared to mathematically more complex models such as support vector machines (Gunning et al., 2019). Furthermore, a subclass of ML algorithms called DL is classified as having the highest complexity but the lowest explainability. DL models consist of multiple layers forming artificial neural networks (ANN). As their architectural structure becomes increasingly complex, the internal decision logic becomes untraceable for humans (Janiesch, Zschech, et al., 2021). In recent applications, DL models have been able to generate precise results, even in high-stakes use cases such as medicine (Dwivedi et al., 2023). However, their application in practice is hampered, as users might not be permitted to use such systems due to regulations or may be unwilling to do so due to concerns of ambiguity and uncertainty while decision-making (Epley et al., 2007; Goodman & Flaxman, 2017). Ultimately, this diminishes the user's acceptance of ML-based systems in practice (Wanner, Herm, et al., 2022a).

**User-Centric Explanations.** Following Miller (2019)'s explanation theory, explanations are products of cognitive and socio-knowledge transfer processes. That is, an explanation describes the outcome of an interactive knowledge transfer from an explainer (e.g., XAI-based DSS) to an explainee (e.g., end-user), with the resulting product indicating how well an explainee receives an explanation (Miller, 2019). A not well designed transfer process consequently results in an information asymmetry between the explainee and the XAI-based DSS (Pfeuffer et al., 2019). As IS research centers on human-technology interaction, the role of AI explanations is widely discussed (Bauer et al., 2021). Unsurprisingly, IS research is currently exploring the potential benefits, pitfalls, and designs of (X)AI-based explanations (Gregor & Benbasat, 1999; Meske et al., 2022).

However, since the perception of explanations is highly dependent on the user's preferences, the process of knowledge transfer and the resulting product vary between individuals (Chromik & Butz, 2021; Herm, Steinbach, et al., 2022). This behavior is embodied in the user's mental model. It reflects the user's understanding of how the system (e.g., XAI-based DSS) may operate and what impact an action will have on this system (Hoffman et al., 2018). Due to the complexity of real-world scenarios, this mental model functions by using a heuristic behavior while iteratively adjusting the attitude toward this system (Kenny et al., 2021). As a consequence, an explanation is a complex product consisting of multiple dimensions, such as information requirements, information access, the pragmatic goals of the explainee, beliefs, emotions, intentions, cognitive, and social aspects (Miller, 2019). Hence, an explanation is not perceived as satisfactory if the explanation is misunderstood or not comprehended as relevant by the explainee (Hilton, 1996).

The goodness of an explanation defines the perceived degree to which one processes and transforms information into logical decision chains (Hoffman et al., 2018). In turn, an inadequately perceived explanation leads to trust issues and thus reduces the overall acceptance (Wanner, Herm, et al., 2022a). This behavior is also referred to as algorithm aversion. Conversely, increased acceptance is termed algorithmic appreciation (Berger et al., 2021; Shin et al., 2020). Significantly, when it comes to new technologies, building initial trust is hampered due to unknown risk factors, as users may fear uncertainties (Epley et al., 2007). Although users suspect no ill will from the system, the capabilities and abilities of the system are still unknown to them, making it difficult to build initial trust (Dam et al., 2018; McKnight et al., 1998). In this context, psychology and social sciences have developed numerous techniques to evaluate how users perceive explanations (Hoffman et al., 2018). These evaluations are preliminary conducted through qualitative (e.g., interviews, discussions) and quantitative (e.g., questionnaires) evaluations. Similarly, socio-technical AI research has derived several measurement constructs, including transparency, explainability, comprehensibility, cognitive load, time, and confidence, to scrutinize an explanation (Herm, Wanner, et al., 2022).

**Explainable Artificial Intelligence.** Accordingly, explaining the logic of ML-based models is of paramount importance (Lebovitz et al., 2021). The multidisciplinary XAI research stream aims to develop transfer techniques to overcome the black-box behavior of ML-based models, thereby providing tractable explanations for the model's decision logic (Arrieta et al., 2020). Moreover, recent regulatory policies, such as the General Data Protection Regulation (GDPR), are forcing the development of XAI techniques (Goodman & Flaxman, 2017). In this sense, XAI-based DSS becomes necessary to justify, evaluate, improve, and manage AI-based systems, especially in high-stakes use cases. Unsurprisingly, utilizing XAI is essential for different types of user groups, such as regulators, developers, managers, and end-users (Meske et al., 2022).

Explanation techniques have consequently been devised for a subset of ML models (model agnostic) or specific ML model types (model specific) for different data formats (e.g., tables, text, or images) and data task types (classification, regression). Furthermore, two distinct approaches exist to illuminate a model's decision logic. The first describes intrinsically explainable models that are comprehensible to humans due to the nature of the employed algorithms, and the second post-hoc models approximating actual ML models (Das & Rad, 2020; Speith, 2022). Here, techniques such as shapely values or sparse linear models are used to analyze the actual model and develop a separate one that is interpretable to humans but potentially comprises a simplified decision logic compared to the original (Lundberg et al., 2020; Speith, 2022). The purpose is to provide explanations that describe the internal decision logic of the model (global) or a particular prediction (local) (Arrieta et al., 2020). In addition, five explanation types have emerged to deliver explanations to explainees with differing objectives.

First, *how* explanations provide holistic descriptions of how the ML model's inner decision logic functions. Second, *how-to* explanations describe how hypothetical adjustments to the input data modify the outcome of the ML model (counterfactual explanation). Third, *what-else* explanations aim to provide instances for the input data of the ML model that result in similar output to the ML model (explanation by example). Fourth, *why* explanations describe why predictions were made for the input data of the ML model based on features relevant to the model's decision logic. Fifth, *why-not* explanations describe why the input features are not relevant to the ML model's decision logic for certain outputs (contrastive explanation) (Herm, Heinrich, et al., 2023; Mohseni et al., 2021).

This variety of explanation types, combined with the possible set of ML models, data formats, and task types to be explained, leads to a plethora of explanation possibilities and thus to countless potential XAI applications from which to pick (Herm, 2023a). To this end, an XAI jungle is created, complicating the XAI selection and development process (Herm, Wanner, et al., 2022). Moreover, developers generate these XAI applications predominantly for developers (van der Waa et al., 2021). As a consequence, research and practice lack an understanding of how actual end-users perceive these explanations when it comes to factors such as transparency, explainability, comprehensibility, or cognitive load, making how well an end-user performs when using these explanations unknown (Herm, Wanner, et al., 2022; Shin, 2021).

To this end, recent IS research proposes several research challenges. From a behavioral science perspective, this includes open research topics on how the different explanation types influence the end-user's perception, what an appropriate explanation should look like, and how these explanations can increase or inhibit the end-user's trust in AI and thus increase system acceptance. From a DSR perspective, research should address the lack of interdisciplinary design principles for different stakeholders to develop an appropriate XAI-based system (e.g., XAI-based DSS). Ultimately, it becomes evident that a combination of behavioral science and DSR is essential to support socio-technical XAI research within IS research (Ågerfalk et al., 2022; Meske et al., 2022; Venkatesh, 2022).

## 1.3 Methodological Background

This section provides a brief overview of the methodology applied within this cumulative thesis. This includes the overarching methodology of IS research and the included research paradigms of behavioral science and DSR.

**Overview of Information Systems Research.** IS research aims to gather knowledge for the development, use, and impact of the IS used within an organization to improve its effectiveness and efficiency (Silver et al., 1995). For this purpose, the IS research framework uses the paradigms of behavioral science and DSR to guide IS research projects (Hevner et al., 2004). See Figure 1.2 for an overview of the framework.

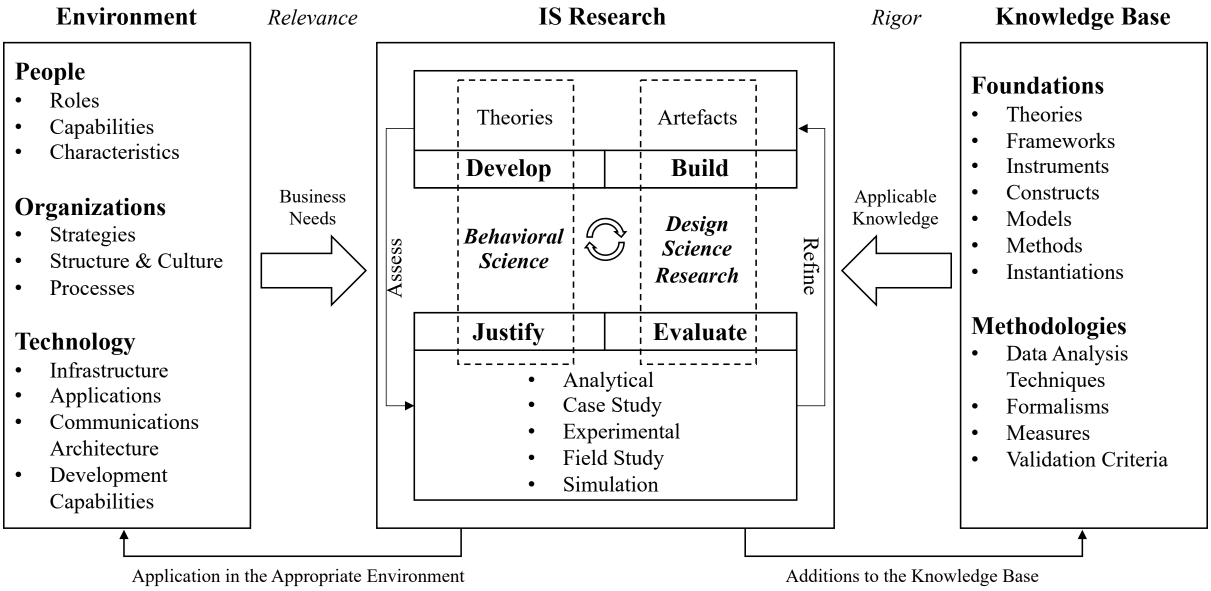


Figure 1.2 Interaction of Behavioral Science and Design Science Research in Information Systems Research Framework according to Hevner et al. (2004)

Within this framework, the environment influences IS research by defining a problem space for IS research projects. This environment exposes business needs that arise due to business- or people-related problems (e.g., people, organizations, or technology) (Simon, 2019). Concurrently, the knowledge base delivers applicable knowledge (e.g., foundations or methodologies) to support IS research. In turn, IS output contributes by integrating applications in the environment and knowledge contributions in the knowledge base. IS research is consequently characterized by the interplay between the relevance of a problem and the rigor of the knowledge base (Hevner et al., 2004). The quality of IS research must be measured according to both theoretical and practical impacts (Baskerville et al., 2018). Furthermore, IS research is subdivided into two corresponding but distinct major research paradigms: behavioral science and DSR (Winter, 2008). While behavioral science aims to explain and predict organizational and human behavior to derive cause-effect relationships, DSR aims to derive artifacts for solving defined problems (Hevner et al., 2004). Thus, IS research attempts to combine the empirical knowledge of behavioral science (truth) with the creativity and precision of DSR (utility). Consequently, *"Truth informs design and utility informs theory"* (Hevner et al., 2004, p. 80). That is, the empirical knowledge gained from theories should contribute to the creation of artifacts, and in turn, the DSR artifact instantiation should postulate empirical knowledge to derive additional theories (Hevner & March, 2003).

**Behavioral Science.** Behavioral science roots in the natural sciences and represents one of the research paradigms in the dual nature of IT research (Hevner et al., 2004). It aims to explain how and why objects behave the way they do to develop concepts that can be used to characterize observed processes. That is, behavioral science targets developing and justifying theories (e.g., principles, laws) to explain or predict organizational and human phenomena (e.g., user perception) (Achinstein, 1968; Österle et al., 2010). In particular, it focuses on the interaction between humans and technology by analyzing, designing, implementing, managing, and using IS (Hevner et al., 2004). New theories evolve, facilitating more profound and comprehensive explanations of observed phenomena (March & Smith, 1995). To measure the quality of these theories, their explanatory power is assessed by their ability to predict future occurrences. From an applied research perspective, behavioral science projects predominantly employ hypothetico-deductive methods (Holmström et al., 2009). For this, theories such as the unified theory of acceptance and use of technology (UTAUT) (Venkatesh et al., 2003) and the theory of mind (Malle, 2004) are transposed to new application domains (e.g., AI) to derive hypotheses and subsequently assess them with empirical data. Ultimately, this research outcome impacts DSR, and vice versa (Hevner et al., 2004; Holmström et al., 2009).

**Design Science Research.** DSR, the second research paradigm within the dual nature of IT research (Hevner et al., 2004), is a pragmatic research paradigm that centers on artifacts with the intention of creating innovative IT-related artifacts to solve real-world problems (March & Storey, 2008; Winter, 2008). Unlike behavioral science, it has its roots in engineering and artificial sciences (Simon, 2019). While a kernel theory is required to initialize a DSR project, it draws from evaluated and modified behavioral science theories (Kuechler & Vaishnavi, 2008). During the execution of a DSR project, the research process may consist of multiple design cycles, constantly iterating between the building and evaluating of an artifact, facilitated by the relevance and the rigor cycle to incorporate the appropriate environment into the DSR project and simultaneously ensure theoretical grounding through the knowledge base. Accordingly, this research paradigm borrows its overarching structure from the IS research framework. It enables the design of complex, relevant, and rigorous IT artifacts to push the boundaries of organizational and human capabilities (Gregor & Hevner, 2013). These artifacts manifest as different output types: constructs, models, methods, instantiations, or theories (Cleven et al., 2009). Typically, recent DSR-based publications have contributed research artifacts such as taxonomies (Kundisch et al., 2022; Nickerson et al., 2013), design theories (Möller et al., 2020), and reference models (vom Brocke, 2007).

## 1.4 Research Objectives and Thesis Overview

**Research Objectives.** Following the IS research methodology, this thesis combines behavioral science and DSR, as suggested by Meske et al. (2022), in an alternating and sequential fashion to address open research challenges in socio-technical XAI research (cf. Section 1.1). To this end, this thesis aims to shed light on the perception of XAI-based DSSs from an end-user perspective and subsequently addresses the lack of design knowledge for XAI-based DSSs. These research fields consequently serve as overarching ROs. In this regard, RO1 aims to investigate end-user perceptions of XAI-based DSSs and is divided into three sub-ROs, namely a preliminary study on end-user acceptance (RO1.1), a systematization of existing user-centered XAI studies (RO1.2), and investigations on various factors contributing to the perception of (X)AI-based explanations (RO1.3). RO2 subsequently contributes to the design knowledge of XAI-based DSSs (RO2). See Figure 1.3 for an overview.

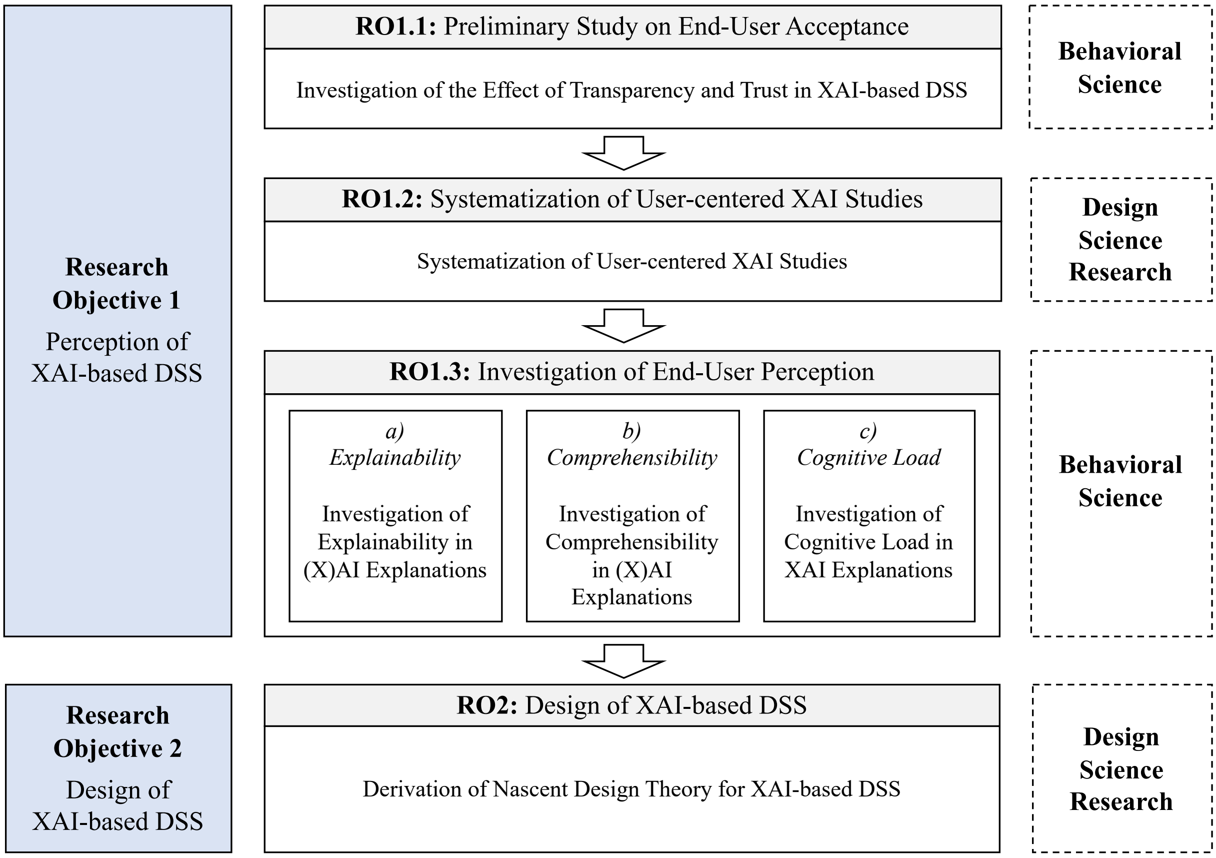

Research Objective 1
Perception of XAI-based DSS
RO1.1: Preliminary Study on End-User Acceptance
Investigation of the Effect of Transparency and Trust in XAI-based DSS
Behavioral Science
RO1.2: Systematization of User-centered XAI Studies
Systematization of User-centered XAI Studies
Design Science Research
RO1.3: Investigation of End-User Perception
a) Explainability
Investigation of Explainability in (X)AI Explanations
b) Comprehensibility
Investigation of Comprehensibility in (X)AI Explanations
c) Cognitive Load
Investigation of Cognitive Load in XAI Explanations
Behavioral Science
Research Objective 2
Design of XAI-based DSS
RO2: Design of XAI-based DSS
Derivation of Nascent Design Theory for XAI-based DSS
Design Science Research

Figure 1.3 Research Design of Thesis

**RO1** aims to research the end-user's perception of XAI-based DSSs. To do so, it is subdivided into three sub-ROs:

First, *RO1.1* investigates the end-user acceptance of XAI-based DSSs by utilizing the behavioral science paradigm. Based on the theoretical assumption of recent XAI-based literature that increased transparency can influence end-users' trust and acceptance of XAI-based DSSs, a preliminary study investigates whether and to what extent this assumption can be empirically proven. This forms the foundation and justification for further examinations of factors influencing end-users' perception of XAI-based DSSs. That is, this sub-RO aims to explore the high-level factors of transparency and trust to determine the need for future in-depth research endeavors that measure end-user perceptions in detail. Here, a theoretical model scrutinizes the end-user acceptance of XAI-based DSSs. As a consequence, this preliminary study utilizes and extends the UTAUT model (Venkatesh et al., 2003) to include the AI-related factors of transparency and trust.

Second, to enable rigorous empirical research on end-user perceptions of explanations for XAI-based DSSs, *RO1.2* conducts an initial systematization of recent XAI-based end-user studies to lay the groundwork for a validated conduction of end-user perceptions in subsequent sub-ROs. Thus, *RO1.2* addresses this lack by relying on the DSR paradigm to develop a taxonomy (Nickerson et al., 2013) that summarizes various aspects of XAI-based end-user studies. In doing so, a structured literature review (vom Brocke et al., 2015) lays the theoretical foundation for deriving a taxonomy for XAI-based studies. In addition, an XAI pattern analysis contributes to the current body of knowledge by providing a starting point to facilitate XAI-based end-user studies.

Third, based on RO1.1 and following Mohseni et al. (2021), a significant effect of transparency on system acceptance, deems it essential to focus on its measurable factors including explainability and comprehensibility to investigate end-user perception. In this context, recent research also calls for the investigation of cognitive load (e.g., Hudon et al., 2021). Thus, using the systematization of RO1.2, *RO1.3* strives to examine relevant factors that contribute to the end-user perception of (X)AI-based DSSs by relying on the behavioral science paradigm. This RO is subdivided into three sub-ROs: RO1.3a, explainability; RO1.3b, comprehensibility; and R1.3c, cognitive load. Here, *RO1.3a* aims to address two research incentives: First, an end-user-focused experiment investigates the explainability of intrinsic explainable ML models and thus examine whether empirical evidence exists for the theoretical performance-explainability trade-off in common ML algorithms. In addition, an end-user investigation compares the perceived level of explainability for various implementation-independent XAI-based explanation types to investigate the explainability of post-hoc models. Following that, *RO1.3b* aims to empirically investigate whether a correlation exists between the perceived level of explainability in common ML algorithms (intrinsic explainability), XAI explanations (post-hoc explainability), and the tested comprehensibility of end-users. *RO1.3c* builds on these findings and aims to examine the perceived cognitive load level of implementation-independent XAI explanations. This also include investigating the cognitive load's impact on task performance and task time. Finally, the results contextualize a metric for mental efficiency that enables its evaluation and thus an assessment of the appropriateness of an explanation for high-stakes use cases.

**RO2** builds on the findings of RO1. Here RO2 aims at deriving design knowledge for the ensured application of XAI-based DSSs, thus facilitating their acceptance. In this context, research applies the DSR paradigm to derive a domain-independent and interdisciplinary nascent design theory (Möller et al., 2020) that includes meta-design requirements, design requirements, design principles, and design features. This also includes integrating the user group developer into the design theory, as developers use XAI explanations to evaluate, improve, and manage the underlying ML models. It is thus essential to include developer-specific design elements to ensure their acceptance and adoption of XAI-based DSSs. Hence, RO2 compromises design considerations from the environment and knowledge base, using DSR evaluation strategies according to Venable et al. (2016)'s FEDS framework to maintain scientific rigor while ensuring relevance for practice.

**Thesis Overview.** To address the proposed ROs and sub-ROs, this thesis utilizes an accumulative research process, compromising six publications published between 2021 and 2023. The following paragraphs describe these publications comprehensively and provide an overview of the thesis structure and scope.

Section 2 addresses *RO1.1* by presenting the publication of Wanner, Herm, et al. (2022a). This contribution seeks to empirically research the factors of trust and transparency within AI-based DSSs in terms of their influence on end-user system acceptance. For this purpose, the widely used UTAUT model (Venkatesh et al., 2003) is extended by constructs on the factors (initial) trust and transparency and evaluated by related scientific literature and practitioners. This extended model is subsequently tested in an industrial maintenance scenario through quantitative studies with domain experts. While the results imply that end-user acceptance depends mainly on performance, transparency and trust have a significant but indirect effect on end-user acceptance. This demonstrates the need for future research on the measurable factors of transparency and trust within XAI research.

Section 3 focuses on the systematization of XAI end-user studies (*RO1.2*) by referring to Herm, Wanner, et al. (2022). Here, a structured literature review (vom Brocke et al., 2015) is conducted to identify a knowledge base of end-user-centered XAI studies, which is systematized into a taxonomy (Nickerson et al., 2013) that includes the meta-characteristics of objectives, participants, methods, and measurements. Drawing from this classification, descriptive, cluster, and archetype analyses are derived for this knowledge base to reveal contemporary research streams and shortcomings. This research contribution also summarizes characteristics to holistically structure future user-centered XAI studies. In doing so, the findings are contextualized in an overarching, cross-disciplinary nomenclature to help researchers address predominant research gaps and precisely position their research incentives.

Using the findings of Section 3, *RO1.3* investigates the end-user perception of XAI-based DSSs in a threefold research incentive:

First, *RO1.3a*, which originates from Herm, Heinrich, et al. (2023) (Section 4), is concerned with investigating the perceived explainability in (X)AI research. In this research contribution, two end-user studies are conducted to examine, first, the lack of empirical evidence on the trade-off between performance and explainability in common ML algorithms and second, the end-user perception of implementation-independent XAI explanation types. To do so, this contribution has deduced and validated hypotheses for both experiments through quantitative end-user studies. By using this research design, the research contributions target the different types of explanation approaches, namely intrinsic explainable ML models and post-hoc explainable models. For the first experiment, the results deviate from the theoretical trade-off assumption and imply a less gradual ratio in terms of end-user perceptions, which fall into three groups: no perceived explainability (e.g., DNNs), mediocre explainability (e.g., support vector machines), and high explainability (e.g., decision trees). The results of the second experiment further indicate a stark differentiation in end-user perception across explanation types, with local types rated as best. Lastly, the research contribution proposes future ROs and implies that evaluating (XAI) explanations should not depend solely on perceived explainability, as this may introduce bias in the end-user's perception. The results suggest that perceived explainability does not automatically translate into comprehensibility and that end-users may prefer local explanation types, surmising a potential preference for explanations demanding less mental effort.

Second, Section 5 addresses *RO1.3b* by reflecting the results of Herm, Wanner, et al. (2021a). Similar to Herm, Heinrich, et al. (2023), this research contribution addresses the lack of empirical evidence in end-user-centered XAI research. Building on the findings of RO1.3a, this contribution examines the theoretical correlation between task-solving ability (degree of comprehensibility) and the explainability of intrinsically explainable ML models and XAI post-hoc explainable models. This contribution thus comprises a quantitative study examining different use cases to measure the perceived degree of explainability and comprehensibility. To this end, the research contribution compares commonly used ML algorithms and the post-hoc-based shapley additive explanation (SHAP) (Lundberg et al., 2020) by benchmarking participants with different task types in real-world scenarios adapted from the cognitive theory of multimedia learning (Mayer & Mayer, 2005). The results indicate a correlation between comprehensibility and explainability, with SHAP scoring highest for both factors.

Third, Herm (2023a) answers *RO1.3c* by researching the end-user's cognitive load, as presented in Section 6. This research contribution emerges from the results of RO1.3a, pointing to the need for an in-depth investigation of the factor of mental effort for XAI explanation types. Following the results of RO1.3a and RO1.3b, wherein post-hoc explanation models perform best regarding end-user perception, this research paper aims to investigate the impact on end-user cognitive load for different XAI explanation types. Here, this contribution presents several hypotheses to investigate the relationship between cognitive load, the resulting task performance, and task time by conducting a quantitative study with prospective physicians within a COVID-19 use case. To this end, the study determines how mental effort, task performance, and task time are affected when using distinct XAI explanation types to solve classification tasks. The results exhibit the lowest mental effort and the highest task performance for local explanation types. Transferring these results into a mental efficiency metric, these explanation types are deemed to be the most efficient. Finally, this yields implications for using XAI in high-stakes use cases and future research incentives.

Lastly, Herm, Steinbach, et al. (2022) addresses *RO2* (Section 7)*,* which centers on the design of XAI-based DSSs. Using the findings from RO1, this research paper develops a nascent design theory (Möller et al., 2020) to facilitate the acceptance of XAI-based DSSs in high-stakes use cases for multiple user groups. In doing so, the nascent design theory aims to provide interdisciplinary design guidelines for both end-users and developers in a domain-independent fashion. To ensure the relevance and rigor of this nascent design theory, the contribution applies a DSR methodology (Vaishnavi & Kuechler, 2007) comprising two design cycles. Furthermore, by conducting a structured literature review (vom Brocke et al., 2015), two qualitative expert studies (Glaser & Strauss, 1967), a quantitative evaluation (Iivari et al., 2021), and a real-world use case instantiation, the contribution incorporates meta-design requirements, design requirements, design principles, and design features. To this end, the design theory targets the topics of global explainability, local explainability, personalized interface design, and psychological/emotional factors.

All contributions are presented as published, with minor corrections (e.g., unification of capitalization, references, and section format; adjustment of section, table, and figure numbering), to standardize the layout of this thesis. See Figure 1.4 for an overview of the thesis structure and a comprehensive description of the sections.

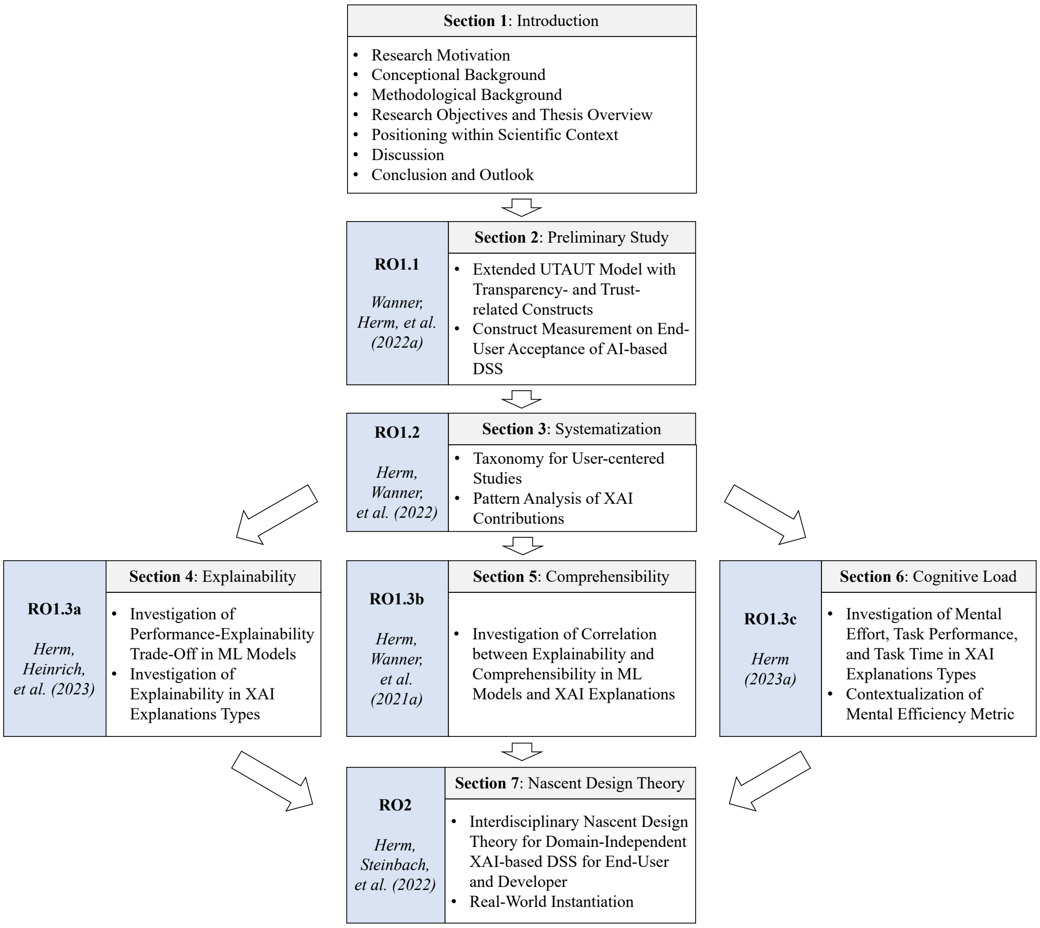

**Section 1**: Introduction
• Research Motivation
• Conceptional Background
• Methodological Background
• Research Objectives and Thesis Overview
• Positioning within Scientific Context
• Discussion
• Conclusion and Outlook
RO1.1
Wanner, Herm, et al. (2022a)
**Section 2**: Preliminary Study
• Extended UTAUT Model with Transparency- and Trust-related Constructs
• Construct Measurement on End-User Acceptance of AI-based DSS
RO1.2
Herm, Wanner, et al. (2022)
**Section 3**: Systematization
• Taxonomy for User-centered Studies
• Pattern Analysis of XAI Contributions
RO1.3a
Herm, Heinrich, et al. (2023)
**Section 4**: Explainability
• Investigation of Performance-Explainability Trade-Off in ML Models
• Investigation of Explainability in XAI Explanations Types
RO1.3b
Herm, Wanner, et al. (2021a)
**Section 5**: Comprehensibility
• Investigation of Correlation between Explainability and Comprehensibility in ML Models and XAI Explanations
RO1.3c
Herm (2023a)
**Section 6**: Cognitive Load
• Investigation of Mental Effort, Task Performance, and Task Time in XAI Explanations Types
• Contextualization of Mental Efficiency Metric
RO2
Herm, Steinbach, et al. (2022)
**Section 7**: Nascent Design Theory
• Interdisciplinary Nascent Design Theory for Domain-Independent XAI-based DSS for End-User and Developer
• Real-World Instantiation

Figure 1.4 Thesis Overview

## 1.5 Positioning within Scientific Context

This section sets the research contributions included in this thesis (cf. Section 1.4) in the holistic context of the author's XAI-related publications. These contributions (*n*=6) form the core of the author's XAI research endeavors and thus the scope of this cumulative thesis. However, the publications not included are relevant because they either serve as input knowledge for the included research contributions or derive research outcomes closely related to the thesis topic. For the sake of clarity, these not-included publications are referred to as research outcomes (*n*=10). In addition, this section demonstrates the connections of these research outcomes and the included research contributions. See Figure 1.5 for an overview. Thus, this section summarizes and contextualizes the author's XAI-related publications. Lastly, Appendix A provides an overview of all authored contributions (*n*=31), including those not related to the research stream of XAI.

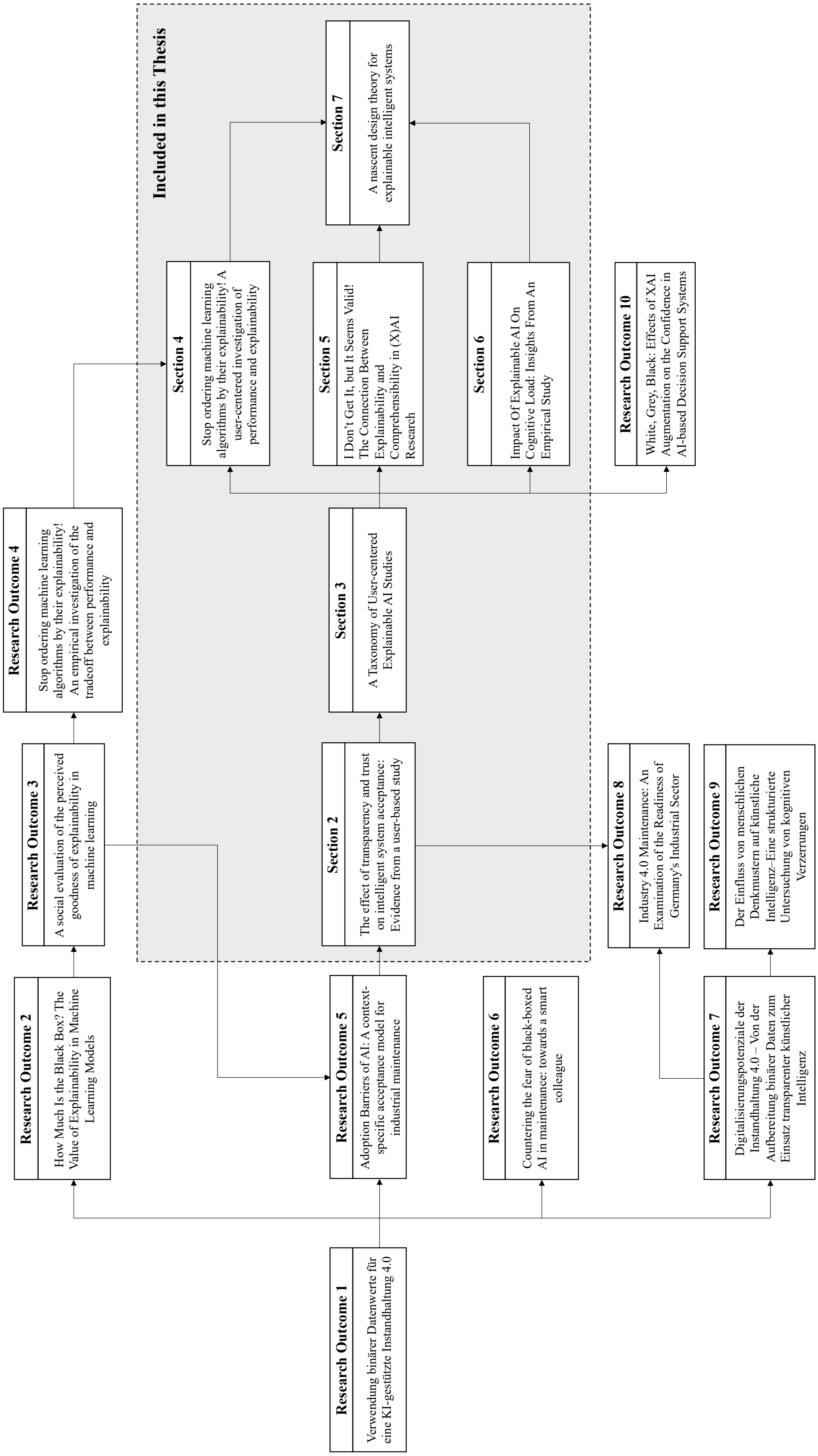

Included in this Thesis
Research Outcome 1
Verwendung binärer Datenwerte für eine KI-gestützte Instandhaltung 4.0
Research Outcome 2
How Much Is the Black Box? The Value of Explainability in Machine Learning Models
Research Outcome 3
A social evaluation of the perceived goodness of explainability in machine learning
Research Outcome 4
Stop ordering machine learning algorithms by their explainability! An empirical investigation of the tradeoff between performance and explainability
Research Outcome 5
Adoption Barriers of AI: A context-specific acceptance model for industrial maintenance
Research Outcome 6
Countering the fear of black-boxed AI in maintenance: towards a smart colleague
Research Outcome 7
Digitalisierungspotenziale der Instandhaltung 4.0 – Von der Aufbereitung binärer Daten zum Einsatz transparenter künstlicher Intelligenz
Research Outcome 8
Industry 4.0 Maintenance: An Examination of the Readiness of Germany's Industrial Sector
Research Outcome 9
Der Einfluss von menschlichen Denkmustern auf künstliche Intelligenz–Eine strukturierte Untersuchung von kognitiven Verzerrungen
Research Outcome 10
White, Grey, Black: Effects of XAI Augmentation on the Confidence in AI-based Decision Support Systems
Section 2
The effect of transparency and trust on intelligent system acceptance: Evidence from a user-based study
Section 3
A Taxonomy of User-centered Explainable AI Studies
Section 4
Stop ordering machine learning algorithms by their explainability! A user-centered investigation of performance and explainability
Section 5
I Don't Get It, but It Seems Valid! The Connection Between Explainability and Comprehensibility in (X)AI Research
Section 6
Impact Of Explainable AI On Cognitive Load: Insights From An Empirical Study
Section 7
A nascent design theory for explainable intelligent systems

Figure 1.5 Author's XAI-related Publications

**Research Outcome 1.** This research outcome (Wanner, Herm, Hartel, et al., 2019) serves as a starting point for the author's XAI-driven research endeavor. It develops an approach to integrate techniques from the field of process mining (PM) and ML into a demonstrator to leverage binary sensor data from today's Industry 4.0 (I4.0) machines. To this end, this research contribution illustrates the applicability of PM and ML to binary data and thus the ability to integrate analytical maintenance strategies into production facilities. Likewise, the evaluation reveals that black-box ML algorithms are unsuitable for high-stakes use cases and that tractable ML models must be used instead, resulting in a research call on perceived explainability in ML models for decision support.

**Research Outcome 2.** Building on the findings of research outcome 1, this publication (Wanner, Herm, & Janiesch, 2020) seeks to investigate the perception of black-box ML models from an end-user perspective. To this end, this paper provides a literature-based classification of technical explanation approaches and intrinsic explainable ML algorithms and proposes a survey design to systematize existing XAI research, laying the foundation for subsequent end-user-centric research.

**Research Outcome 3.** Within this research outcome (Wanner, Herm, et al., 2022b), the survey design developed in research outcome 2 has been extended and applied to measure derived factors that influence the goodness of perceived explainability of ML models. These include the factors of intuitiveness, complexity, trustworthiness, understandability, satisfaction, and sufficiency. This evaluation targeted several ML algorithms, different types of stakes (low and high), and ML problem types (classification and regression). The results indicate that trustworthiness is the most critical factor for perceived goodness of explainability.

**Research Outcome 4.** Using the two-factor XAI survey design of research outcome 3, this publication (Wanner, Herm, Heinrich, et al., 2021) examines the theoretical trade-off between the explainability and performance of common ML algorithms from the end-user perspective. In doing so, it i.a. reveals that tree-based algorithms are perceived to be more explainable than ML algorithms, such as ANNs.

**Research Outcome 5.** As research outcomes 1 and 3 reveal the importance of the factors of trust and explainability, this contribution (Wanner, Popp, et al., 2021) conducts several quantitative and qualitative studies to extend the UTAUT model with constructs related to trust and transparency when dealing with an AI-based DSS. To this end, this contribution lays the foundation for the contribution in Section 2 (Wanner, Herm, et al., 2022a). Thus, the derived UTAUT model has been extended and tested in an industrial maintenance use case (see Section 2). This necessitates systematizing user-centered XAI research (Section 3), as shown in Herm, Wanner, et al. (2022), and provides the requisite to investigate further XAI-related constructs (Sections 4-6), such as perceived explainability (Herm, Heinrich, et al., 2023), comprehensibility (Herm, Wanner, et al., 2021a), and cognitive load (Herm, 2023a) and, lastly, to demonstrate the need to derive a nascent design theory for XAI-based DSSs (Herm, Steinbach, et al., 2022) (Section 7). See Section 1.4 for more information about the contributions included in this thesis.

**Research Outcome 6.** Following the necessity for explainability in AI-based DSSs for I4.0 maintenance (see research outcome 1), this research outcome (Wanner, Herm, & Janiesch, 2019) develops and demonstrates a prototype that leverages trained ML models and expert knowledge to automatically derive intelligible rule sets for real-time anomaly detection within production facilities.

**Research Outcome 7.** Building on research outcome 1, this research outcome (Wanner, Herm, & Janiesch, 2021) extends the developed prototype with a novel technique to translate decision knowledge from more complex ML models, such as random forests, into traceable but accurate decision rules. These decision rules apply for real-time sensor data monitoring in production facilities.

**Research Outcome 8.** This research outcome (Wanner, Herm, Janiesch, et al., 2022) focuses on German manufacturing companies. Its primary objective is to investigate the state of the art in AI-based I4.0 maintenance strategies in German manufacturing companies, and it thus indirectly compares the maturity of the German industry with the developed approach of research outcome 7. This includes tracking, perceiving, and utilizing sensor data for data-based maintenance approaches.

**Research Outcome 9.** While previous work (e.g., research outcome 7) has argued that training ML algorithms with unknown or unsupervised data may introduce bias, this bias can lead to unfair and discriminatory erroneous decisions. Thereon, this research contribution (Herm, Janiesch, et al., 2022) investigates what types of bias may occur. To this end, a structured literature review of AI-based research identifies distinct bias types and provides recommendations for addressing and avoiding bias in AI research and practice.

**Research Outcome 10.** Lastly, the research outcome of Wanner, Herm, Heinrich, et al. (2020) pursues the idea of investigating end-user-related factors for using XAI-based DSSs. Hypotheses and a study procedure have been derived to examine end-user confidence in XAI-based DSSs. Specifically, this aims at the influence of different explanation types for XAI-based DSSs and the use case stake on the end-user's confidence.

## 1.6 Discussion

Recognizing that this cumulative thesis's overall goal is to shed light on the perception and design of XAI-based DSSs, the following section discusses the results in an overarching manner, comprehensively summarizing the findings and providing implications and prospective limitations. As such, this section comprises three subsections, which discuss the perception of XAI-based DSSs (RO1), deliberate on the design of XAI-based DSSs (RO2), and outline the thesis's contribution to research, respectively.

**Perception of XAI-based DSSs.** In recent years, user-centered XAI research has emphasized several under-researched factors to uncover the nature of user's rationale in the decision-making process (Herm, Wanner, et al., 2022). This implies the consideration of complex heuristic human behavior and, accordingly, a multifaceted mental model (Gunning et al., 2019) to facilitate the acceptance of XAI-based DSSs (Wanner, Herm, et al., 2022a). This cumulative thesis consequently aims to address this issue by providing first-hand empirical evidence of end-user perception and subsequent design knowledge for the application of XAI-based DSSs; hence, it represents a starting point for future XAI research to draw upon.

First, a preliminary study establishes the rationale for the research incentive to investigate the end-user's mental model during the utilization of XAI-based DSSs. Here, an extended UTAUT model for AI-based DSSs reveals the importance of the factors of transparency and (initial) trust for end-user acceptance. Although performance expectancy is revealed as the most relevant direct factor, this study indicates the significance of transparency and trust through indirect effects, denoting a multifaceted construct. In this sense, the results imply the need for future research on indirect factors to understand the user's mental model as well.

As part of these research incentives, this thesis focuses on the complex construct of transparency by examining its measurable factors of explainability and comprehensibility (Mohseni et al., 2021). Further following recent research calls (e.g., Hudon et al., 2021), this thesis also emphasizes on the factor of cognitive load to investigate end-user's perception. Using a cumulative research approach, this thesis also examines on related aspects of these factors. That is, the results of the empirical studies indicate that the type of explanation strongly influences the end-user's perception (explainability, mental effort) and performance (comprehensibility, task performance, and task time). It is apparent that end-users prefer local explanation types over global ones. Furthermore, examining mental effort reveals that end-users perceive local explanation types as less demanding. Analogously, perceived mental effort correlates with task time and task performance, which in combination denotes positive mental efficiency only for local explanation types. However, it should be noted that with the preference for a more straightforward (local) explanation type, the amount of information provided decreases compared to global explanation types. That is, relying exclusively on these explanation types can lead to unpredictable side effects. For example, within the investigation of the comprehensibility factor, it became apparent that participants often miscomprehend these gray-box explanations. Recent research (e.g., Rudin, 2019) advocates using comprehensible white-box ML models rather than augmenting black-box models with XAI post-hoc models. Apart from XAI-based explanations, the empirical studies on explainability and comprehensibility demonstrate that end-users prefer intrinsic ML models that represent their decision logic in a tree- or coefficient-based manner, which may be due to prior knowledge or a more accustomed representation format.

Altogether, the results indicate that one must determine the explanation selection individually for different use cases. Similarly, empirical studies address the perceptions of novice end-users without prior knowledge of AI or ML. Thus, these studies do not consider learning effects or additional user groups. Similarly, the research findings may imply an over-reliance on straightforward explanation types, causing unintended downstream consequences. That is, to study these effects, it is necessary to perform long-term studies on different types of end-users. Similarly, these studies have been conducted in a controlled environment where participants must place themselves in a high-stakes scenario without worrying about the consequences for poor or incorrect performance or having to deal with time constraints or interference times. Therefore, one might assume that the pressure placed on participants would lead to new insights and different results regarding the studied factors (Saeed & Omlin, 2023). Similarly, it is currently not apparent how end-users deal with erroneous system recommendations, biases in the ML model's decision logic, or information asymmetries due to human decision biases (Venkatesh, 2022).

The systematization in Section 3 derives several factors that may contribute to heuristic user behavior within current user-centered XAI research. In conclusion, this cumulative thesis addresses some of the stated factors: transparency and trust (Section 2), explainability (Section 4), comprehensibility (Section 5), and cognitive load (Section 6). In turn, open factors remain that must be addressed: accuracy, decision quality, time constraints, and satisfaction. Future XAI research must consequently address this shortcoming and thereby develop a holistic understanding of an end-user's mental model during decision-making (Herm, 2023a; Herm, Heinrich, et al., 2023).

**Design of XAI-based DSSs.** According to the IS research framework (cf. Figure 1.2), the design of XAI-based DSSs is based on preliminary studies such as those conducted in the context of the first RO. Hence, within this cumulative thesis, the derived nascent design theory (RO2) compromises (meta-) design requirements, design principles, and design features that originate from an interdisciplinary research endeavor. To this end, the proposed design theory contains design elements primarily researched within RO1. To facilitate the acceptance of XAI-based DSSs, the design theory addresses transparency and trust factors through meta-design requirements that result in design requirements and design principles to provide global and local explanations. Similarly, the design requirement of cognitive effort is addressed through RO1 by examining the end-user's cognitive load. In turn, the design theory extends this body of knowledge by additionally considering design elements that arise due to developer-specific characteristics and the related research field of human-computer interaction (Herm, Steinbach, et al., 2022).

Overall, the nascent design theory addresses the prevailing lack of design knowledge in IS literature. In particular, the design theory expands the body of knowledge by incorporating technical and socio-technical aspects that can relate to multiple user groups and enables holistic design guidance to facilitate the adoption of XAI-based DSSs in practice. Thus, the design theory provides a starting point for future research and practice to draw upon. To realize this, it aims to provide interdisciplinary design knowledge for domain-independent applications by considering the user groups of end-users and developers. In line with RO1, RO2 aims to provide an initial foundation for the design of XAI-based DSSs. In this sense, this design theory is not intended to be a one-size-fits-all solution; rather, research and practice must tailor the results for specific application domains and extend them for additional user groups (Herm, Steinbach, et al., 2022). However, the design theory reveals that end-users and developers share the same design requirements, three out of four design principles, and half of the design features, indicating a potential overlap between different stakeholders using XAI-based DSSs for diverse types of actions.

So far, research has to delve deeper into distinct design elements, for instance, to derive mechanisms and design knowledge for an ensured trust development and to determine an appropriate explanation selection within use cases (Herm, Heinrich, et al., 2023; Venkatesh, 2022). Hence, it may be critical to incorporate design elements that facilitate the development of interactive XAI-based DSSs by keeping humans in the loop, thus providing a feedback mechanism to improve the decision quality of the system and facilitate user attitudes toward the XAI-based DSS (Meske et al., 2022). Similar to RO1, applicability beyond a real-world initialization must be tested through multiple long field studies. This applicability also includes technical considerations due to the nature of XAI's post-hoc models. This also embraces the definition of computational metrics for evaluation, as post-hoc models approximate the actual black-box model, which may lead to varying inherent ML decision-making rationales (Mohseni et al., 2021). This consequently raises the relevance of legal considerations when addressing liability and legally compliant application in high-risk use cases (Górski & Ramakrishna, 2021; Thiebes et al., 2021).

**Contribution to Research.** During the 1980s and 1990s, research began to investigate the role of explanations within intelligent agents (Gregor & Benbasat, 1999). While these types of systems transitioned into AI-based systems, causing the shift toward "*newer-paradigm intelligent systems*" (Gregor & Yu, 2002, p. 289), the requisite to investigate the impact of explanations leveraged due to the increasing complexity of ML models (Maedche et al., 2021). Unsurprisingly, IS research is calling for examinations of the AI's explainability (Meske et al., 2022). So investigating the explainability of ML models is critical to IS research, as it has a multitude of requirements and implications for the daily tasks of decision-makers (Bauer et al., 2021). Hence, IS research (Bauer et al., 2021; Meske et al., 2022; Venkatesh, 2022) proposes several high-level problem spaces: *Perception*) an examination of how XAI explanations influence users' beliefs and consequently their mental models; *Design*) a concentration on user-centered model explanations, thus addressing the design and implementation of user-centered AI-based systems to accommodate the requirements of multiple user groups; and *Concept shifts*) a determination of how explanations can cause endogenous concept shifts within AI usage.

Building on these problem spaces, this thesis follows the IS research framework (cf. Figure 1.2) to investigate the topic of XAI-based DSSs, from not only a behavioral science perspective but also a DSR one. In particular, this thesis targets on the first and second problem spaces by deriving knowledge about end-users' mentals model based on empirical research (perception), which subsequently transfers into design knowledge for the design of XAI-based DSSs (design). However, the scope of this thesis aims not to address these issues exhaustively. From that, the thesis contributes to an understanding of end-users' mental models during decision-making by focusing on several factors to provide an empirical starting point and subsequently a domain-independent and adaptable design theory. Furthermore, this design theory is extended due to the consideration of developers as an XAI-based DSS user group. This implies that future research must reflect the findings of RO2 from a behavioral science perspective, which can yield additional insights about user perception and thus contribute to a more sophisticated design theory. Also, as this thesis focuses on the user groups of end-users and developers, research has already proposed further user groups (e.g., regulators, managers) (Meske et al., 2022), which in turn implies the relevance of future research endeavors.

However, as this thesis is concerned with the ROs perception and design, it still encountered initial XAI conceptual shifts constituting prospective research endeavors aside from traditional decision support traits. As an example, the results suggest that the application of XAI must be reconsidered, as the conducted expert studies pointed to the potential of using XAI as a toolkit for training novice workers (Herm, Steinbach, et al., 2022). Likewise, recent XAI research proposes a paradigm shift from recommendation-driven toward hypothesis-driven support (Miller, 2023). Given these conceptual shifts, research must refine and rethink the role of AI explanation and thus its perception and design. The application of XAI consequently remains in its infancy, which means that a considerable amount of research must be done to prevent opening Pandora's box.

## 1.7 Conclusion and Outlook

As computational advances drive the integration of AI-based products into any system, these AI-based solutions empower the utilization of new capabilities for decision-making, even in high-stakes scenarios (Berente et al., 2021; Janiesch, Zschech, et al., 2021). To this end, the resulting AI-based DSS includes ML and DL models capable of outperforming domain experts (McKinney et al., 2020). In contrast, performance accompanies a diminishment of these models' inherent explainability. Overall, the resulting information asymmetry can adversely impact the user's mental model and thus reduce their system acceptance (Gunning et al., 2019; Shin, 2021). Hence, the XAI research stream aims to develop techniques that remedy this deficiency by making the model's decision-making process traceable to humans. So far, developers have engineered these techniques primarily for developers, without involving the actual end-users of the system (Meske et al., 2022; van der Waa et al., 2021).

Accordingly, this cumulative thesis aims to address the proposed shortcomings by researching end-user perceptions (RO1) of (X)AI explanations and subsequently deriving design knowledge (RO2) for the rigorous development of an XAI-based DSS. To accomplish this, the thesis embraces six research contributions that contribute to this overarching goal, as follows: First, Wanner, Herm, et al. (2022a) extends and tests a UTAUT model that combines the factors of transparency and (initial) trust and subsequently identifies the indirect influence of both factors on end-user system acceptance. Based on this preliminary study, Herm, Wanner, et al. (2022) systematizes existing user-centered XAI studies and derives a taxonomy that enables future research to construct substantiated studies and address identified research gaps. From that, the included research contributions investigate the end-user's perception of explainability (Herm, Heinrich, et al., 2023), comprehensibility (Herm, Wanner, et al., 2021a), and cognitive load (Herm, 2023a). To summarize, the results indicate that end-user perception varies markedly per explanation type, with primarily local explanation types rated best. In addition, the studies provide empirical evidence for a partially confirmed trade-off between performance and explainability, for the correlation between comprehensibility and explainability, and for the influence of cognitive load on end-user task performance and task time. Drawing on this, the results from the hypothetico-deductive end-user studies were incorporated into a nascent design theory (Herm, Steinbach, et al., 2022). This artifact comprises (meta-) design requirements, design principles, and design features, embodying an initial, interdisciplinary design proposal in a domain-independent fashion, which integrates multiple user groups. Lastly, a real-world instantiation in a predictive maintenance scenario has ensured the applicability of the design theory.

As IS research is driven through the continuous interaction of behavioral science and DSR (Hevner & March, 2003), so is this cumulative thesis, in which design knowledge for XAI-based DSSs is grounded on users' perceptions. This thesis's research results can thus be considered a vehicle for future user-centered XAI research, thereby also contributing to IS research and related research stream endeavors. That is, it provides the rationale for gaining first-hand insights about the perception and design of XAI-based DSSs. In turn, to establish a holistic understanding of an end-user's mental model, it is essential to examine further factors thoroughly (cf. Herm, Wanner, et al., 2022) to obtain insights on behavioral changes that may impact the design of XAI-based DSSs in the long run. Ultimately, this may yield a more nuanced comprehension of end-user perception and thus a sophisticated design theory for XAI-based DSS applications, overcoming the phenomenon of the *"inmates running the asylum"* (Miller, 2019, p. 4) in XAI research.

# 2 The Effect of Transparency and Trust on Intelligent System Acceptance: Evidence from a User-based Study

*Jonas Wanner, Lukas-Valentin Herm, Kai Heinrich, and Christian Janiesch*

**Abstract.** Contemporary decision support systems are increasingly relying on artificial intelligence technology such as machine learning algorithms to form intelligent systems. These systems have human-like decision capacity for selected applications based on a decision rationale which cannot be looked-up conveniently and constitutes a black box. As a consequence, acceptance by end-users remains somewhat hesitant. While lacking transparency has been said to hinder trust and enforce aversion towards these systems, studies that connect user trust to transparency and subsequently acceptance are scarce. In response, our research is concerned with the development of a theoretical model that explains end-user acceptance of intelligent systems. We utilize the unified theory of acceptance and use in information technology as well as explanation theory and related theories on initial trust and user trust in information systems. The proposed model is tested in an industrial maintenance workplace scenario using maintenance experts as participants to represent the user group. Results show that acceptance is performance-driven at first sight. However, transparency plays an important indirect role in regulating trust and the perception of performance.

**Keywords.** User acceptance, intelligent system, artificial intelligence, trust, system transparency.

**Reference.** Wanner, J., Herm, L.-V., Heinrich, K., & Janiesch, C. (2022). The effect of transparency and trust on intelligent system acceptance: evidence from a user-based study. *Electronic Markets*, *32*(4), 2079-2102. https://doi.org/10.1007/s12525-022-00593-5

## 2.1 Introduction

Would you trust a superintelligent computer's recommendation on a critical decision such as turning off crucial machinery if it offered no transparency into the decision-making?

Intelligent systems with human-like cognitive capacity have been a promise of artificial intelligence (AI) research for decades. Due to the rise and sophistication of machine learning (ML) technology, intelligent systems are becoming a reality and can now solve complex cognitive tasks (Benbya et al., 2021). They are being deployed rapidly in practice (Janiesch, Zschech, et al., 2021). More recently, deep learning allows tackling even more compound problems such as playing Go (Silver et al., 2016) or driving autonomously in real traffic (Grigorescu et al., 2020). On the downside, the decision rationale of intelligent systems based on deep learning is not per se interpretable to humans and requires explanations. That is, while the decision is documented, its rationale is complex and essentially intransparent from the point of human perception constituting a perceived black box (Kroll, 2018).

Further, users tend to credit anthropomorphic traits to an intelligent system subconsciously to ascribe the system's AI a sense of efficacy (Epley et al., 2007; Pfeuffer et al., 2019). In this respect, intelligent systems are credited with the trait of agency (Baird & Maruping, 2021), creating a situation comparable to the principal-agent problem as their decision rationale is self-trained (self-interest) and intransparent to the principal. This results in an information asymmetry between the user (principal) and the intelligent system (agent). This information asymmetry constitutes a major barrier for intelligent system acceptance and initial trust in intelligent systems (McKnight et al., 2002; Shin et al., 2020), because the system cannot provide credible, meaningful information about or affective bonds with the agent (Bigley & Pearce, 1998).

Altogether, this lack of transparency and, subsequently, trust can be a hindrance when delegating tasks or decisions to an intelligent system (Shin, 2020a, 2021). More specifically, the acceptance and adoption of AI currently remains rather hesitant (Chui & Malhotra, 2018; Milojevic & Nassah, 2018; Wilkinson et al., 2021). The result is observable user behavior, such as algorithm aversion, where the user will not accept an intelligent system in a professional context even though it outperforms human co-workers (Burton et al., 2020). While this can be attributed at least partially to lack of control and the information asymmetry due to its black-box nature, we also observe the inverse, algorithm appreciation, and, thus, acceptance and use of intelligent systems in other scenarios (Herm, Wanner, et al., 2021a; Logg et al., 2019).

This is a crucial point, as intelligent systems can only be effective if users are willing to engage with them actively and have confidence in their recommendations. Consequently, it is of great importance to understand what the intended users of such systems expect, and which influences have to be considered for mitigation of algorithm aversion and successful acceptance (Mahmud et al., 2022).

While the factors of performance, trust, and transparency have been connected to user perception of technology, a rigorous study to connect them to intended usage behavior of intelligent systems is missing (Venkatesh, 2022). With our research, we expand the body of knowledge on the acceptance of intelligent systems by considering system transparency and trust in combination as pivotal factors (e.g., Adadi & Berrada, 2018; Mohseni et al., 2021; Rudin, 2019). Furthermore, we extend beyond the measurement of direct effects and investigate their mediating, indirect roles regarding the drivers of behavioral intention.

We build a theoretical model by synthesizing explanation theory, user trust theory, and the unified theory of acceptance (UTAUT) to fit the nature of intelligent systems and to understand the human attitude towards them.

Thereby, we offer three key contributions. First, we provide an explanatory model for the context of intelligent systems. It can serve as a starting point for research in distinct fields. Second, by validating established hypotheses, we provide a better understanding of the actual factors that influence the user's acceptance of intelligent systems and explain user behavior towards AI-based systems in general. This allows both the use of this knowledge for the (vendor's) design and implementation of intelligent systems and its use for the (customer's) process of software selection. Third, by establishing new hypotheses that regard the nature of trust and transparency in system acceptance, we take into account the unique attributes of intelligent systems related to the perceived black-box nature of their underlying rationales (Herm, Heinrich, et al., 2023; Mohseni et al., 2021).

Our paper is structured as follows: In Section 2.2, we introduce the theoretical background for our research. In Section 2.3, we describe our research design. In Section 2.4, we describe our research theorizing. This includes the review of existing UTAUT research on trust and system transparency as well as the hypothesis and items of the derived constructs and relationships. In Section 2.5, we describe the empirical testing of the theoretical derivations and their results. Finally, we discuss the implications for theory and practice in Section 2.6, before we summarize and offer an outlook on future research in Section 2.7.

## 2.2 Theoretical Background

### 2.2.1 Artificial Intelligence and Intelligent Systems

AI is an umbrella term for any technique that enables computers to imitate human intelligence and replicate or even surpass human decision-making capacity for complex tasks (Russell & Norvig, 2021). This entails that the meaning and scope of AI is constantly being refined as technology evolves while the reference point of human intelligence remains relatively static (Berente et al., 2021).

In the past, AI focused on handcrafted inference models known as symbolic AI or the knowledge-based approach (Goodfellow et al., 2016). While this approach is inherently transparent and enabled trust in the decision process, it is limited by the human's capability to explicate their tacit knowledge relevant to the task (Brynjolfsson & Mcafee, 2017). More recently, ML and deep learning algorithms have overcome these limitations by automatically building analytical models from training data (Janiesch, Zschech, et al., 2021). However, the resulting advanced analytical models often lack immediate (system) transparency constituting an information asymmetry to the user.

Intelligent systems are software systems that make use of AI technology. They exhibit at least two traits towards end-user that separate them from traditional commercial-off-the-shelf software with decision support such as accounting information systems or enterprise resource planning software. First, intelligent systems enable decision-making with human-like or even super-human cognitive abilities for certain tasks (McKinney et al., 2020). Second, the decision rationale of intelligent systems cannot be looked up conveniently.

That is, intelligent systems do not use handcrafted and thus traceable, deterministic rulesets to make decisions, but intelligent systems exhibit complex probabilistic behavior with superior performance that was learned based on data input rather than explicitly programmed, for example using ML algorithms (Janiesch, Zschech, et al., 2021; Mohseni et al., 2021). While the underlying relations in the analytical models can be analyzed by experts given enough time and resources (and technically constitute white-box decision making), no end-user is capable of extracting explanations on the decision process or individual decisions. Rather, the model constitutes a black box from the perspective of the end-user (Savage, 2022).

This circumstance leads to an increased tension between human agency and machine agency during decision making (Sundar, 2020). In this context, intelligent systems inherit characteristics associated with new, revolutionary technologies, including technology-related anxiety and alienation of labor through a lack of comprehension and a lack of trust (Mokyr et al., 2015). Hence, when facing these properties, due to effecting motivation, the human has a "desire to reduce uncertainty and ambiguity, at least in part with the goal of attaining a sense of predictability and control in one's environment" (Epley et al., 2007).

### 2.2.2 Transparency and Trust in Intelligent Systems

Trust in the context of technology acceptance has widely been studied and derived from organizational trust towards humans. Notably, besides the core construct of the cognition-based trust in the ability of the system, additional affect-based trust aspects like the general propensity to trust technology and the believed goodwill or benevolence of the trustee towards the trustor exist (von Eschenbach, 2021). While it can be argued that the system has no ill will by itself, in the case of black-box systems, we cannot observe whether it acts as intended, possibly hindering initial trust formation (Dam et al., 2018).

Building trust in new technologies is initially hindered by unknown risk factors and thus uncertainty, as well as a lack of total user control (McKnight et al., 2011; Shneiderman, 2020). The main factors in building initial trust are the ability of the system to show possession of the functionalities needed, to convey that they can help the user when needed, and to operate consistently (McKnight et al., 1998; Paravastu & Ramanujan, 2021).

For human intelligence, it is generally an important aspect to be able to explain the rationale behind one's decision, while simultaneously, it can be considered as a prerequisite for establishing a trustworthy relationship (Samek et al., 2017). Thus, observing a system's behavior in terms of transparency plays an important role. In IS research, it has been argued that transparency can increase the cognition-based part of trust towards the system (Shin et al., 2020). In addition, system transparency is assumed to have an indirect influence on IS acceptance via trust in the context of recommending a favorable decision to the user (Wilkinson et al., 2021).

While general performance indicators of ML models can be used to judge the recommendation performance of an intelligent system, the learning process and the inner view of the intelligent system towards the problem can be different from the human understanding, generating a dissonance, suggesting system performance by itself is not sufficient as a criterion (Miller, 2019).

Thus, the ML model underlying an intelligent system cannot address these factors itself. Therefore, it is widely suggested that this issue can be alleviated or resolved by providing an overall system transparency by offering explanations of the decision-making process (i.e., global explanations) as well as explanations of individual recommendations (i.e., local explanations) (Mohseni et al., 2021). That is, in recent AI-based IS literature the perceived explanation quality is defined as the level of explainability (Herm, Wanner, et al., 2021a). The field of explainable AI (XAI) offers augmentations or surrogate models that can explain the behavior of intelligent systems based on black-box ML models (Injadat et al., 2021).

Altogether, the rise of design-based literature on explainable, intelligent systems suggests that the lack of transparency of deep learning algorithms poses a problem for user acceptance, rendering the systems inefficacious (Bentele & Seidenglanz, 2015; Sardianos et al., 2021). It is reasonable to assume that system transparency or its explainability, as well as trust, play a central role when investigating socio-technical aspects of technology acceptance. Furthermore, both seem to be interrelated to one another (Shin, 2021). Nevertheless, it is not evident to what extent an increase in the user's perceived system explainability improves the user's trust factor or how this affects the user's technology acceptance of intelligent systems (Shin et al., 2020; Wang & Benbasat, 2016).

### 2.2.3 Technology Acceptance

Technology acceptance has been widely studied in the context of several theoretical frameworks. In its core idea, a behavioral study is used to draw conclusions regarding the willingness of a target group to accept an investigative object (Jackson et al., 1997; Venkatesh et al., 2012).

Davis (1989) utilized the *Theory of Reasoned Action* to propose the *Technology Acceptance Model* (TAM) that explains the actual use of a system through the perceived usefulness and perceived ease of use of that system. It was later updated to include other factors such as subjective norms (Marangunić & Granić, 2015). An extension of the theory that includes the additional determinant is the *Theory of Planned Behavior* by Taylor and Todd (1995). As a competing perspective of explanation, the *Model of PC Utilization* includes determinants that are less abstract to the technology application environment, such as job-fit, complexity, affect towards use, and facilitating conditions that reflect on the actual objective factors from the application environment, it can differ largely from case to case. The *Innovation Diffusion Theory* by Rogers (2010) is specifically tailored to new technologies and the perception of several determinants like a gained relative advantage, ease of use, visibility, and compatibility. Furthermore, the *Social Cognitive Theory* was extended to explain individual technology acceptance by determinants like outcome expectancy, self-efficacy, affect, and anxiety (Bandura, 2001).

Venkatesh et al. (2003) combined those theories in the *Unified Theory of Acceptance and Use of Technology* (UTAUT). It provides a holistic model that includes adoption theories for new technologies and approaches to computer usage that capture the actual factors of the implementation environment. Compared to ABM, UTAUT is favored due to its ability to explain the variance within the dependent variable more precisely (Demissie et al., 2021). Here, *behavioral intention* (BI) acts as an explanatory factor for the actual user behavior. Determinants of *BI* in the UTAUT model are, for example, *performance expectancy* (PE), *effort expectancy* (EE), or social influence. UTAUT has been used extensively to explain and predict acceptance and use in a multitude of scenarios (Williams et al., 2015).

Despite the fundamental theoretical foundation, it has become a common practice to form the measurement model for a specific use case given by multiple iteration cycles (e.g., Yao & Murphy, 2007). Thus, many authors modify their UTAUT model (e.g., Oliveira et al., 2014; Shahzad et al., 2020; Slade et al., 2015). Typically, an extension is applied in three different ways (Slade et al., 2015; Venkatesh et al., 2012): i) using UTAUT for the evaluation of new technologies or new cultural settings (e.g., Gupta et al., 2008); ii) adding new constructs to expand the investigation scope of UTAUT (e.g., Baishya & Samalia, 2020); and/or iii) to include exogenous predictors for the proposed UTAUT variables (e.g., Neufeld et al., 2007). Furthermore, many contributions such as Esfandiari and Sokhanvar (2016) or Albashrawi and Motiwalla (2017), combine multiple extension methods to construct a new model. Lastly, Blut et al. (2021) introduce four new broad predictors for future technology acceptance and use. However, they do not incorporate the idea of black-box systems common with contemporary AI.

Related work of technology acceptance research is primarily focused on e-commerce, mobile technology, and social media (Rad et al., 2018). The intersection with AI innovations is rather small yet. Despite contributions for autonomous driving (e.g., Hein et al., 2018; Kaur & Rampersad, 2018) or healthcare (e.g., Fan et al., 2018; Portela et al., 2013), only a few studies exist for industrial applications, such as on the acceptance of intelligent robotics in production processes (e.g., Bröhl et al., 2016; Lotz et al., 2019). Also, there is the intention to understand the acceptance of augmented reality (Jetter et al., 2018).

Consequently, knowledge about the technology acceptance of intelligent systems is still limited. In particular, trust and system transparency have not been considered in conjunction as potential factors for technology acceptance of intelligent systems.

## 2.3 Methodological Overview

The focus of our research problem is the acceptance of an intelligent system from an end-user perspective. It is located at the intersection of two fields of interest: technology acceptance and AI, more specifically XAI.

Figure 2.1 presents our methodological frame to develop our UTAUT model for the context of intelligent systems. It corresponds to the procedure presented by Šumak et al. (2010), which we modified to suit our objective. We detail the steps in the respective sections.

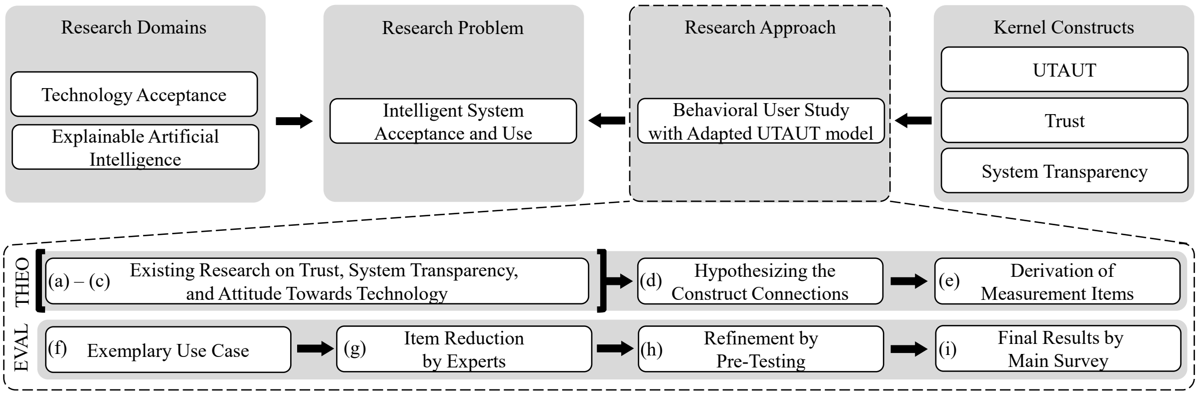

Research Domains
Technology Acceptance
Explainable Artificial Intelligence
Research Problem
Intelligent System Acceptance and Use
Research Approach
Behavioral User Study with Adapted UTAUT model
Kernel Constructs
UTAUT
Trust
System Transparency
THEO
(a) – (c) Existing Research on Trust, System Transparency, and Attitude Towards Technology
(d) Hypothesizing the Construct Connections
(e) Derivation of Measurement Items
EVAL
(f) Exemplary Use Case
(g) Item Reduction by Experts
(h) Refinement by Pre-Testing
(i) Final Results by Main Survey

Figure 2.1 Methodology Overview

The kernel constructs to form our model are derived from the related research on UTAUT, trust, and system explainability. Thus, in the theorizing section (THEO, see Section 2.4), we derive a suitable model from existing UTAUT research on (a-c) system transparency and attitude towards technology. We then (d) hypothesize the derived measurement model constructs and connections based on empirical findings, and we I collect potential measurement items.

In the evaluation section (EVAL, see Section 2.5), we (f) validate and modify our UTAUT model by using an exemplary application case in the field of industrial maintenance. Further, we (g, h) iteratively adapt it in empirical studies, perform the main study, and (i) discuss the results.

As scientific methods, we use empirical surveys (see e.g., Lamnek & Krell, 2010) in combination with a structural equations model (SEM) (see e.g., Weiber & Mühlhaus, 2014). For the analysis of the SEM, we apply the variance-based partial least squares (PLS) regression (see e.g., Chin & Newsted, 1999).

## 2.4 Research Theorizing

### 2.4.1 Trust Extensions in UTAUT

While trust has been widely recognized as an important factor in information system usage in ABM theory, the UTAUT model does not account for trust in its original form (Carter &

Bélanger, 2005). While several extensions of the UTAUT model have been proposed to address this drawback, both the inclusion and definition vary among research contributions (Venkatesh et al., 2016). Table 2.1 depicts a summary of UTAUT extensions regarding the construct of trust.

We can characterize these extensions by inclusion type regarding the dependent variables, which are affected by the trust construct in the respective UTAUT model. Endogenous inclusion refers to a direct connection between trust and BI, while exogenous inclusion refers to an indirect relationship through other variables. Furthermore, we indicate which determinants are included for the trust variable itself.

| Inclusion Type | Dependent Variables | Determinants | Example References |
|---|---|---|---|
| Endogenous | BI | None | Alaiad and Zhou (2013); Carter and Bélanger (2005); Oh and Yoon (2014) |
| | | Personal Propensity to Trust | Oliveira et al. (2014) |
| | | Trust Integrity, Trust Ability | Komiak and Benbasat (2006) |
| | | Trust Property, Satisfaction | Kim (2014) |
| Exogenous | PE | Trust Benevolence, Trust Integrity, Trust Ability | Cheng et al. (2008); Lee and Song (2013) |
| | Perceived Usefulness | Perceived Ease of Use, Consumer Decision Making | Xiao and Benbasat (2007) |
| | Perceived Risk, Perceived Usefulness | System Transparency, Technical Competence, Situation Management | Choi and Ji (2015) |
| Endogenous/ Exogenous | Perceived Risk, BI | None | Slade et al. (2015) |
| | PE, BI | Trust Benevolence, Trust Integrity, Trust Ability | Cody-Allen and Kishore (2006) |

Table 2.1 Trust-based UTAUT Extensions

In terms of trust-based model components, we found i) several theoretical approaches to describe trust itself; ii) multiple determinants of the embedded trust construct (determinants); iii) several different ways of embedding trust into existing technology acceptance models such as ABM or UTAUT (inclusion type/ dependent variable).

While a majority of contributions (e.g., Oh & Yoon, 2014) include trust as a single variable with no determinants in an endogenous manner, other studies (e.g., Cheng et al., 2008) adopted more complex theoretical frameworks. McKnight et al. (2002) present a frequently adopted framework. They define a model to determine the intention to trust a system by building upon trust perception theory by Mayer et al. (1995). Specifically, they determine the users' intention to trust as the willingness of the user to depend on the system. This intention is influenced by three variables: *disposition to trust/trust propensity, institution-based trust*, and *trusting beliefs*. Disposition to trust or trust propensity is the general tendency to trust others, in this case an intelligent system. Institution-based trust refers to the contextual propitiousness that supports trust, indicating an individual's belief in good structural conditions for the success of the system. Trusting beliefs indicate an individual's confidence in the system to fulfill the task as expected (Mayer et al., 1995; Vidotto et al., 2012).

Trusting Beliefs itself is comprised of three determinants: *trust benevolence*, *trust integrity*, and *trust ability/ability beliefs*. *Ability beliefs* (AB) refers to the system's perceived competencies and knowledge base for solving a task, that is trust in the ability of the system. Trust integrity involves the user's perception that the system acts according to a set of rules that are acceptable to him or her. Trust benevolence is indicated to be the belief in the system to do good to the user beyond its own motivation (Cheng et al., 2008; Mayer et al., 1995; McKnight & Chervany, 2000).

Considering the problem of a complex, intelligent system that mimics human functions, we adopted the unified model of McKnight et al. (2002) but made several modifications. First, we adopt *trust propensity towards an AI system* (TP) as the indicator for an aggregation of prior beliefs that potentially allow a professional to become vulnerable to an intelligent system. Moreover, we include this measure as an important determinant as in comparison to other information systems, perception of intelligent systems is different since the belief is also formed by outside media and social influence in more drastic way (e.g., sentient AI) that can increase factors of fear and/or aversion leading to decreased trust. We adopt *AB* as the determinant for *TP*, since it is the component that directly measures trust in the system itself rather than environmental factors and personal factors that are covered by facilitating conditions and moderators of the core UTAUT model already. Following the discussion in the realm of algorithm aversion, we argue that the trust propensity will be changed by seeing the system perform. While this is contrary to related work on trust, we believe that based on findings rooted in the algorithm aversion theory that for intelligent systems, *TP* also encompasses the changeable beliefs regarding the ability of algorithms. As argued above trust propensity also reflects beliefs that reflect external sources like media. This renders its role more important than merely reflecting on a general trusting behavior but rather as an indicator of trusting an intelligent system specifically. Thus, we also used items that express a tendency for trust propensity that can be subject to change. Third, we model *TP* as a direct influence factor of *BI* and as an exogenous factor for *PE*. Following the discussion in Lankton et al. (2015) between human-like and system-like trust, we come to understand that with a system, *AB* reflects the system-like trust properties of reliability, functionality, and helpfulness, as a system is not able to exhibit behaviors on its own that would not adhere to the given rules (trust integrity) or exocentric motive (trust benevolence). Thus, in a bid to limit complexity, we omit the factors trust benevolence and trust integrity and adopt *AB* as the sole determinant for *TP* by assuming that no matter how many functions or tasks are assigned, the system has no hidden intention to extend its tasks beyond its programming and it cannot change its "promise" by itself. For a possible pre-existing perception of ill will towards the system, trust propensity will collect those prior beliefs reflected by the tendency of the user to trust the system prior to use. This is also confirmed by Jensen et al. (2018) who in the case of computer systems attribute the most influence on benevolent beliefs and perception to dispositional characteristics that are already reflected in our model by trust propensity. However, if we extend the definition of the system by including system providers, programmers, and other stakeholders that are involved in the creation and maintenance, this simplification will pose a limitation, since hidden, malicious behavior can then play a pivotal, especially with intelligent systems since they can be subject to manipulation for example via adversarial learning (Heinrich et al., 2020).

### 2.4.2 System Transparency Extensions in UTAUT

Especially in recent years, the transparency of a system, as the backbone of an XAI's system explainability, has been increasingly integrated into studies of technology acceptance of intelligent systems and seems to have a direct influence on the perceived trust of users (e.g., Nilashi et al., 2016; Peters et al., 2020). In the context of intelligence systems, we define *system transparency* (ST) as the ability of the system to explain and reveal its decision rationale to the user by visual means (e.g., a visual panel that shows based on which maintenance-related input variables the suggestion of imminent maintenance was made). Table 2.2 depicts a summary of UTAUT extensions regarding the construct of *ST*.

| **Inclusion Type** | **Dependent Variables** | **Determinants** | **Example References** |
|---|---|---|---|
| Exogenous/ Endogenous | Trust, BI | None | Brunk et al. (2019); Slade et al. (2015); Choi and Ji (2015); Hebrado et al. (2011); Hebrado et al. (2013) |
| | | Explanation | Nilashi et al. (2016) |
| | | Accuracy, Completeness | Peters et al. (2020) |
| | ATT, Trust, BI | None | Shahzad et al. (2020) |
| | Understanding, BI, Users' Privacy Concerns | None | Zhao et al. (2019) |
| | Trusting Beliefs, Understanding, Competence, Acceptance | None | Cramer et al. (2008) |
| | AB, Information Satisfaction | Accuracy, Completeness, Time Information Currency | Cody-Allen and Kishore (2006) |
| | EE, Trusting Beliefs | None | Wang and Benbasat (2016) |

Table 2.2 System-Transparency-based UTAUT Extensions

We can characterize these extensions regarding the dependent variables, which are affected by the *ST* construct in the respective UTAUT model. Again, we found only references for the exogenous/ endogenous inclusion type. Similarly, we indicate which determinants are included for the transparency variable itself.

Among others, Brunk et al. (2019) and Hebrado et al. (2013) define *ST* as a factor to increase the user's understanding of how a system works. It further entails an understanding of the system's inner working mechanisms. That is why specific recommendations were made according to different characteristics and assumptions for a single item (Nilashi et al., 2016; Peters et al., 2020) as well as the system's overall decision rationale. Furthermore, *ST* should be used for required justifications (Shahzad et al., 2020).

Nevertheless, the influence of other factors on *ST* differs in these models. While many contributions, such as Brunk et al. (2019) and Peters et al. (2020), take no further factors into account, Nilashi et al. (2016) consider the type of explanation and the kind of presented information. They measure the factor of explanation through the level of explainability according to the user's perception and, thus, how, and why a recommendation was made and the interaction level within the recommendation process. For Shahzad et al. (2020), it is about characteristics of the information quality, such as for example accuracy and completeness, which influence *ST*.

Further, we noticed *ST* influences many factors: *BI*, *PE, EE,* and trust. As argued above, the factor of trust is modeled as *TP* and *AB*. Here, it is assumed that a highly transparent decision-making process results in an increasing *TP* (Shin, 2020b; Vorm & Combs, 2022), while also increasing transparency results in a better *AB* of the user (Cody-Allen & Kishore, 2006). It is important to know that our assumption reflects a time-dependent use behavior were an introduction of the system takes place and through experiencing performing and through explanation of the decision rationale, the prior trust behavior (TP) changes through a change in the beliefs of the system's ability (AB) after seeing it perform (Dietvorst et al., 2015). *BI* is defined as the degree to how a user's intention changes through the level of *ST* (Peters et al., 2020). Lastly, an increase in *ST* results in a clearer assessment by the user, and thus the user's mental model assumes a higher performance of the system leading to increased *PE*. Likewise, a transparent system can reduce a user's efforts to understand the systems' inner working mechanisms (Wang & Benbasat, 2016).

### 2.4.3 Attitude Towards Technology Extension in UTAUT

Consistent with the theory of planned behavior, an individual's attitude towards technology, in this case *attitude towards AI technology* (ATT), has been found to act as a mediating construct (Dwivedi et al., 2019; Kim et al., 2009; Yang & Yoo, 2004). People are said to be more likely to accept technology when they can form a positive attitude towards it. It is important to note that usually the construct is placed between the endogenous variables in the UTAUT context (e.g., *PE* and *EE*) and intention to use (e.g., *BI*). Furthermore, we believe that *ATT* is influenced the individual's pre-formed opinion about AI technology. The prevailing opinion that forms into attitude is not changed easily and depends on an individual's prior exposure to the technology (Ambady & Rosenthal, 1992). Factors accumulated in *ATT* can be religious beliefs, job security, attitude carried over from popular culture, as well as knowledge and familiarity and privacy, and relational closeness (Persson et al., 2021). Thus, it acts as a place to collect emotional attitude towards a technology, which in the case of AI is reinforced by its anthropomorphic and intransparent nature. While some studies show that not all of these factors are present in an individual's mind, general states of mind like fear towards the technology can influence and form the person's attitude (Dos Santos et al., 2019; Kim, 2019). Thus, we argue to include *ATT* and hypothesize that the mediation strength and thus indirect connections to *ATT* are increasingly present for intelligent systems.

### 2.4.4 Model and Hypotheses

As a result of the above construct derivation, we present our UTAUT model for intelligent systems along with the hypotheses and their respective direction (- or +) in Figure 2.2. The measurement model can be divided into three major parts: i) UTAUT core (*PE*, *EE*, and *BI*), ii) UTAUT AI (*AB*, *TP*, *ST,* and *ATT*), and iii) moderators (*gender*, *age*, *experience*).

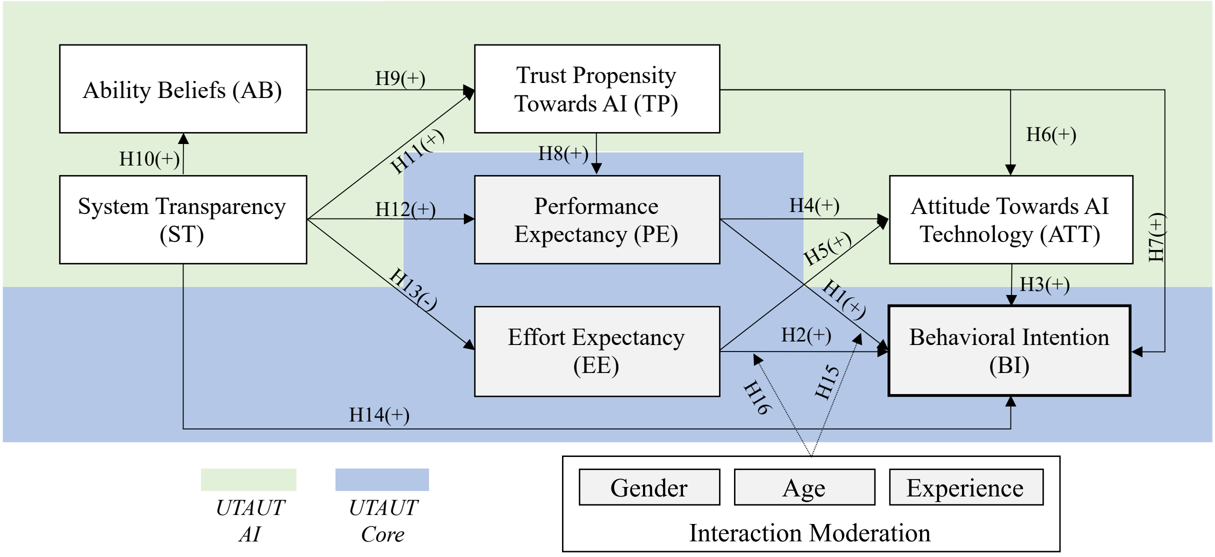


Figure 2.2 Derived Acceptance Model for Intelligent Systems

The derivation of the hypotheses from i) UTAUT core research is primarily based on general research on UTAUT (e.g., Dwivedi et al., 2019; Venkatesh et al., 2003). Nevertheless, these construct interrelations can also be found in UTAUT studies on trust or system transparency (e.g., Lee & Song, 2013; Wang & Benbasat, 2016). Compared to the UTAUT core established by Venkatesh et al. (2003), the constructs of facilitating conditions and social influence are not included due to the results of our multistage reduction process (cf. Section 2.5.2).

The construct *BI* represents our target variable. It measures the strength of a user's intention to perform a specific behavior (Fishbein & Ajzen, 1977). Here, it is about the willingness of a user to adopt an intelligent system or more specifically the willingness of the user to take advice recommended by the intelligent system. This is an important distinction as intelligent system use can be mandated in a professional setting. For example, in the case of intelligent systems for decision support, technically it is not about the intention of the user to adopt the system as such, but about the intention of the user to considers the system's output in his or her work processes.

The construct is influenced endogenously by the two basic UTAUT constructs of *PE* and *EE*. *PE* measures the degree to which an individual believes that using the system can increase their job performance. This includes factors such as perceived usefulness, job-fit, relative advantage, extrinsic motivation, and outcome expectation (Venkatesh et al., 2003). Whereas *EE* measures the degree of individual ease associated with the use of the system, including factors such as perceived ease of use and complexity (Venkatesh et al., 2003). For both constructs, we assume that they have a positive influence on *BI*. This correlation can be seen in UTAUT (e.g., Dwivedi et al., 2019; Venkatesh et al., 2003) as well as in UTAUT model studies on trust and on *ST* (Cheng et al., 2008; Cody-Allen & Kishore, 2006; Lee & Song, 2013; Wang & Benbasat, 2016). Thus, we state:

***H1:** Performance Expectancy positively affects Behavioral Intention.*

***H2:** Effort Expectancy positively affects Behavioral Intention.*

The hypotheses of ii) UTAUT AI, and thus, for *ATT*, *AB*, *TP*, and *ST* are primarily based on the references from Table 1 and Table 2 as well as Venkatesh et al. (2003)'s considerations.

*ATT* is defined as a user's overall affective reaction to using AI technology or an AI system (Venkatesh et al., 2003). While the authors did not include the construct in his final model, it is regularly used in the context of decision support systems. UTAUT research, as well as ABM research on trust, indicate that *ATT* has a positive effect on *BI* (e.g., Chen, 2013; Hwang et al., 2016; Mansouri et al., 2011). That is, people form intentions to engage in behaviors to which they have a positive attitude (Dwivedi et al., 2019). Inversely, it is assumed that both *PE* and *EE* have a positive influence on a user's *ATT*. Suleman et al. (2019) derive this significantly positive influence from the ABM research (Hsu et al., 2013; Indarsin & Ali, 2017) and later confirm it in their own research. Dwivedi et al. (2019) and Thomas et al. (2013) confirm the connection. We summarize these findings by our next hypotheses:

***H3:** Attitude Towards AI Technology positively affects Behavioral Intention.*

***H4:** Performance Expectancy positively affects Attitude Towards AI Technology.*

***H5:** Effort Expectancy positively affects Attitude Towards AI Technology.*

Trust is regarded as a necessary prerequisite to forming an effective intelligent information system (e.g., Dam et al., 2018) and, thus, it is a crucial construct to build our model. In Section 2.4.1, we have explained that *TP* is influenced by AB and thus can be changed by observing the system behavior. *AB* measures the assumed technical competencies of the system to solve a task (Schoorman et al., 2007). *TP* is about the user's general disposition to trust an intelligent system with a task. For *TP*, we expect a positive effect on *ATT* and on *BI*, as this preformed trust is a major influence that has formed from user experienced and external exposure (e.g. through media) which is increasingly critical for intelligent system (Gherheş, 2018). For Suleman et al. (2019), trust in general was the most influential and significant factor affecting a participant's *ATT*. The positive influence of trust on *BI* is well proven by several UTAUT studies (Choi & Ji, 2015; Lee & Song, 2013). In addition, we argue that specifically for intelligent systems, while the assumed ability of the system is quite high, the black-box nature and skepticism or the presence of algorithm aversion in humans can make it increasingly harder to form trust towards a system. As opposed to traditional software systems, intelligent systems make decisions based on a learning process and do not necessarily follow the same reasoning as the human decision-making process. Thus, building trust towards an intelligent system seems more important since the natural state usually assumes a rather critical view of aversion (Mahmud et al., 2022). We summarize this with the next hypotheses:

***H6:*** *Trust Propensity Towards AI positively affects Attitude Towards AI Technology.*

***H7:*** *Trust Propensity Towards AI positively affects Behavioral Intention.*

In turn, we assume that *AB* has a positive influence on a user's *TP*, which in turn has a positive influence on *PE*. However, the direction of the latter influence is disputed in prior research. While Oliveira et al. (2014), Nilashi et al. (2016), and Wang and Benbasat (2016) assume that *PE* has a positive influence on trust, Cody-Allen and Kishore (2006), Lee and Song (2013), and Choi and Ji (2015) think that trust affects *PE*. We additionally argue that the propensity to trust an intelligent system will also result in increased expectance of future performances, while distrust in a system will also lower the expectations towards future high performance. As mentioned earlier with intelligent systems there is a general aversion on the one side, while there is also evidence of performance that exceeds human decision makes. Although intelligent systems can outperform humans, humans are sometimes preferred despite performing worse because of trust issues (Dietvorst et al., 2015). This is partially reflected by a person's trust propensity. Therefore, we assume a positive influence of *TP* on *PE*. The influence of *AB* on *TP* is also relying on the fact that humans change behavior towards algorithms once they observe their behavior, which even can result in switching from a state of aversion to a state of algorithm appreciation (Logg et al., 2019). In fact, we argue that the assumed effect is even stronger with the performance promise that is attributed to intelligent systems compared to a traditional software system. Accordingly, we formulate our hypotheses:

***H8:*** *Trust Propensity Towards AI positively affects Performance Expectancy.*

***H9:*** *Ability Beliefs positively affect Trust Propensity Towards AI.*

Trust - as a multifaced term - is assumed to have a strong correlation with *ST* (e.g., Dam et al., 2018). *ST* measures the user's understanding of the intelligent system's decision rationale (Hebrado et al., 2011). In other words, it represents how openly an intelligent system's inner decision rationale is working as well as how openly characteristics that determine why an intelligent system made a certain decision hare communicated (Mohseni et al., 2021). As the user of such an intelligent system decides whether or not to adopt the system recommendation, *ST* might influence his or her decision-making process. We expect a positive effect of *ST* on *AB* and *TP* based on the findings of Pu and Chen (2007) and Wang and Benbasat (2016) that rely on trust in general. The former found that users assign a recommender system a higher level of competence if the decision-making process is explained in a traceable manner. The latter is supported by the preliminary UTAUT research of Brunk et al. (2019), Hebrado et al. (2013), Nilashi et al. (2016), Peters et al. (2020), and Chen and Sundar (2018). For example, Peters et al. (2020) found that *ST* positively influenced trust in the intelligent system significantly in the context of testing a consumer's willingness to pay for transparency of such black-box systems.

We further argue that the ability to experience the reasoning of the system as a form of self-disclosure will be accepted as some kind of honesty. In addition to the related work that deals with trust as a general construct, we argue that based on an existing state of aversion towards intelligent systems before seeing them perform and experiencing their decision rationale, a method to create transparency not only with regard to the performance but also with regard to the inner logic on a case-to-case basis we increase likelihood of mitigating the aversion (Herm, Heinrich, et al., 2023; Mohseni et al., 2021). While a prior trust (measured through trust propensity) has already been made up, we argue that through the usually missing experience with the design of intelligent systems, this formed opinion can be changed through visual means and demonstrations such as how-to or why-explanations as presented in the XAI literature (Arrieta et al., 2020; Herm, Heinrich, et al., 2023). We conclude with the hypotheses:

***H10:*** *System Transparency positively affects Ability Beliefs.*

***H11:*** *System Transparency positively affects Trust Propensity Towards AI.*

Technology acceptance research supposes that *ST* also influences the residual UTAUT constructs of *PE*, *EE*, and *BI*. We derive the assumed positive effect of *ST* on *PE* from Zhao et al. (2019), who revealed that a higher level of a decision support system supports the user's perception of the performance of that system. If users understand how a system works and how calculations are performed, they will perceive that, in some cases, implementing and using the system requires more effort (Gretzel & Fesenmaier, 2006). This is a special case with black-box intelligent systems since not everything is transparent out-of-the-box and, thus, an associated effort cannot always be clearly derived. However, through ST the effort can be monitored and revealed. Thus, we expect a positive influence of ST onto EE. We also expect a positive influence of *ST* for *BI*. Making the reasoning behind a recommendation transparent allows for an understanding of the recommendation process, significantly increasing acceptance (Bilgic & Mooney, 2005). This significant and strong influence is also reflected in further studies by Venkatesh et al. (2016) and Hebrado et al. (2011). Furthermore, the basic concept of seeing an algorithm perform well can, for some tasks, increase performance expectancy. As a prerequisite of making up one's mind about an algorithm, the ability to experience it is a fundamental necessity (Dietvorst et al., 2015; Logg et al., 2019). We address this through three hypotheses:

***H12:*** *System Transparency positively affects Performance Expectancy.*

***H13:*** *System Transparency positively affects Effort Expectancy.*

***H14:*** *System Transparency positively affects Behavioral Intention.*

The hypothesis for the iii) moderators is also part of the original UTAUT model according to Venkatesh et al. (2003). We assume that *gender*, *age*, and *experience* have a moderating effect on *PE*, *EE*, and *BI constructs*. We derive this assumption from the initial UTAUT model (Venkatesh et al., 2003). It has been confirmed in several other UTAUT studies (e.g., Alharbi, 2014; Esfandiari & Sokhanvar, 2016; Wang & Benbasat, 2016). In contrast, we do not consider voluntariness of use due to the obligatory use of intelligent systems in day-to-day business. From this, we derive the following hypotheses:

***H15:*** *Gender, Age, and Experience moderate the effects of PE on BI.*

***H16:*** *Gender, Age, and Experience moderate the effects of EE on BI.*

## 2.5 Study and Results

### 2.5.1 Study Use Case

In the following, we offer an exploration of the theoretical constructs put forward. For this purpose, we defined a real-world use case and transferred it to the UTAUT model in a step-by-step procedure. In this way, we validate the applicability of our proposed model. Moreover, we gain first insights into the user's willingness to accept intelligent systems at their workplace.

We consider industrial machine maintenance to be a suitable scenario. Its focus is to maintain and restore the operational readiness of machinery to keep opportunity costs as low as possible. In contrast to reactive strategies, anomalous machine behavior can be identified and graded early on using statistical techniques to avoid unnecessary work. Given the technological possibilities to collect large and multifaceted data assets in a simplified manner, intelligent systems based on machine learning are a promising alternative for maintenance decision support (Carvalho et al., 2019).

In this context, rolling bearings are used in many production scenarios of different manufacturers. For example, they are often installed in conveyor belts for transport or within different engines and show signs of wear and tear over time that requires maintenance (Pawellek, 2016).

For our evaluation, we decided to use an automated production process to manufacture window and door handles, as these are common everyday items every respondent can relate to. In our scenario, there shall be several production sections connected by high-speed conveyor belts. Inside these conveyor belts, several bearings are installed. These are monitored by sensors to monitor change (e.g., noise sensor, vibration, and temperature). A newly introduced intelligent system evaluates this data automatically. In case of anomalous data patterns, a dashboard displays warnings and errors with concrete recommendations for action (cf. Appendix B.1).

The respondents of the survey(s) shall be confronted with a decision situation that tests whether or not the user adopts the system recommendation in his or her own decision-making process. That is, they need to decide for or against an active intervention in the production process as recommended by the system. In an extreme case, the optical condition of the conveyor belt bearings is perceived as good. However, the system recommends that the conveyor belt must be switched off immediately. This error does not occur regularly, and the message contradicts the previous experience of the service employee (here, the respondent) with this production section. As additional information, we provide the reliability of the system recommendations and hint at the high follow-up costs in case of a wrong decision.

### 2.5.2 Study Design

Our design and conduct of the survey are based on Šumak et al. (2010), which we modified to our objective. We used five steps to obtain our study results: i) collection of established measurement items; ii) pre-selection by author team; iii) reduction by experts; iv) evaluation and refinement through pre-study; and v) execution of the main study. See Appendix B.2 for the results of each step, as well as the primary and secondary source(s) for all measurement items. A more detailed result table of the validity and reliability measures of the pre-study and main study is available in Appendix B.3.

**Step i).** First, we collected those measurement items that already exist for the respective constructs of interest and are, thus, empirically proven.

As we adopted *PE*, *EE*, and *BI* from Venkatesh et al. (2003), we built on their findings. Venkatesh et al. (2003) chose the measurement items for UTAUT by conducting a study and testing the measurement items for consistency and reliability. For the additional constructs *ATT*, *ST*, *AB*, and *TP*, we examined the source construct measurement items as well as examples of secondary literature and derived constructs. Initially, we used three items to form the construct of ATT - one was adopted from Davis et al. (1992) and two from Higgins and his co-authors (Compeau et al., 1999; Thompson et al., 1991). As we derived *ST* from the perceived local explainability of an intelligent system decision's result visualization as well as the perceived global explainability of the intelligent system's decision process, we initially included five items from Madsen and Gregor (2000) to address the global component and two items from Cramer et al. (2008) to address the local component, as noted in recent XAI-related research (Adadi & Berrada, 2018; Mohseni et al., 2021). The measurement items for *AB* and *TP* were derived from McKnight et al. (2002) (trust competence) and Lee and Turban (2001) (trust propensity). Lastly, the measurement items for facilitating conditions and social influence were adapted from Taylor and Todd (1995), Thompson et al. (1991), Moore and Benbasat (1991), and Davis (1989).

**Step ii).** Next, we discussed the appropriateness of each of the collected measurement items within the team of authors.

The team members merge knowledge in the respective domains of industrial maintenance, technology acceptance, and (X)AI research. Special attention was paid to the duplication of potential item questions and their feasibility for the use case. We reduced the total number of measurement items for the model's constructs from 71 to 24.

**Step iii).** Subsequently, we conducted an expert survey with practitioners from industrial maintenance regarding our intended main study.

The survey with ten experts had two goals: reducing the remaining measurement items and understanding the explainability of intelligent system dashboards. For the former, we briefly explained each of the model measurement constructs to the experts. Thereby, we removed the constructs of facilitating conditions and social influence completely. Subsequently, the experts selected the most appropriate remaining measurement items for the use case per measurement construct. They were given at least one vote and at most votes for half the items. Then, we selected the final measurement items based on a majority vote. For the latter, we presented the experts with four different maintenance dashboards of intelligent systems as snapshots adapted from typical software in the respective field (e.g., Aboulian et al., 2018; Moyne et al., 2013). Here, the experts rated their perceived level of explanation goodness on a seven-point Likert scale. Using the dashboard with the highest overall (median) explanation goodness, ensures that the dashboard for the quantitative survey has inherent explainability to the end-user and thus provides enhanced system transparency (cf. Appendix B.1).

**Step iv).** Then, we conducted a quantitative pilot study to critically examine our questionnaire and research design (Brown et al., 2010). The testing includes checks for internal consistency, convergent reliability, indicator reliability, and discriminant validity.

The study contained 60 valid responses. Here, we ensured representative respondents, that is, maintenance professionals holding a position to use an intelligent system for their job-related tasks (e.g., experience in maintenance). See Appendix B.5 for the demographics of the responses. We provided the participants with a description of the exemplary use case and screenshots of the prototype. We asked them to respond to their perceptions of each of the measurement items on a seven-point Likert scale. See Table 2.3 for the assessment of measurement items and Appendix B.4 for a summary of our decisions on individual items.

| **Construct** | **Assessment Measurement Items** | | | | | |
|---|---|---|---|---|---|---|
| | **CA** | **AVE** | **CR** | **FL Criterion** | **Cross-Loadings** | **Item Loadings** |
| PE | 0.86 | 0.64 | 0.90 | - | - | - |
| EE | 0.74 | 0.57 | 0.84 | - | - | EE1 |
| ST | **0.67** | 0.75 | 0.86 | - | - | - |
| TP | **0.23** | 0.56 | **0.67** | - | TP4 (*BI*) | TP3, 4 |
| AB | 0.74 | 0.66 | 0.85 | - | - | - |
| ATT | **0.67** | 0.59 | 0.81 | - | - | ATT2 |
| BI | 0.83 | 0.74 | 0.89 | - | - | - |

*Internal consistency*: Cronbach's alpha (CA) > 0.7; composite reliability (CR) > 0.7 (Gefen et al., 2000; Hair et al., 2011)

*Convergence reliability*: average variance extracted (AVE) > 0.5 (Hair et al., 2011)

*Indicator reliability*: item loadings 0.7<*x*<1 (Hair et al., 2011)

*Discriminant validity*: cross-loadings; Fornell-Larcker (FL) criterion (Fornell & Larcker, 1981; Hair et al., 2011)

Table 2.3 Validation and Reliability Testing of Pre-Study

**Step v).** Then, we conducted our main quantitative study. Table 2.4 comprises the final set of measurement items. We, again, checked for internal consistency, convergent reliability, indicator reliability, and discriminant validity. Further, while we did not include explicit control variable, we included control questions (CQ) following Meade and Craig (2012) and Oppenheimer et al. (2009) to increase result validity.

| Construct | | Measurement Item | Reference(s) |
|---|---|---|---|
| **PE** | PE1 | Using this system in my job would enable me to accomplish tasks more quickly. | Davis (1989) |
| | PE2 | Using this system would improve my job performance. | |
| | PE3 | Using this system would make it easier to do my job. | |
| | PE4 | I would find this system useful in my job. | |
| | PE5 | Using this system would increase my productivity. | Moore and Benbasat (1991) |
| **EE** | EE1 | Learning to operate this system would be easy for me. | Davis (1989) |
| | EE2 | I would find it easy to get this system to do what I want it to do. | |
| | EE3 | My interactions with this system would be clear and understandable. | |
| | EE4 | I would find this system easy to use. | |
| **ATT** | ATT1 | The actual process of using this system would be pleasant. | Davis et al. (1992) |
| | ATT2 | This system would make work more interesting. | Thompson et al. (1991) |
| | ATT3 | I would like to work with this system. | Compeau et al. (1999) |
| | ATT4 | Using the system would be a bad/good idea. | Peters et al. (2020); Taylor and Todd (1995) |
| | ATT5 | Using the system would be foolish/wise move. | |
| **BI** | BI1 | If this system was available to me, I would intend to use this system in the future. | Venkatesh et al. (2003) |
| | BI2 | If this system was available to me, I predict I would use this system in the future. | |
| | BI3 | If this system was available to me, I would plan to use this system in the future. | |
| **ST** | ST1 | I would understand how this system will assist me with decisions I have to make. | Madsen and Gregor (2000) |
| | ST2 | I would understand why this system provided the decision it did. | Cramer et al. (2008) |
| | ST3 | I would understand what this system bases its provided decision on. | |
| **AB** | AB1 | This system would be competent in providing maintenance decision support. | Cheng et al. (2008); McKnight et al. (2002) |
| | AB2 | This system would perform maintenance decision support very well. | |
| | AB3 | In general, this system would be proficient in providing maintenance decision support. | |
| **TP** | TP1 | It would be easy for me to trust this system. | Cheng et al. (2008); Lee and Turban (2001); Wang and Benbasat (2007) |
| | TP2 | My tendency to trust this system would be high. | |
| | TP3 | I would tend to trust this system, even though I have little or no knowledge of it. | |
| **CQ** | CQ1 | I would not find this system easy to use. | - |
| | CQ2 | Although I may would not know exactly how this system works, I would know how to use it to make decision regarding the quality of its output. Please do not rate this statement and please choose scale point one instead to ensure the data quality of this survey. This only applies to this question. | Meade and Craig (2012) |
| | CQ3 | I have read all questions carefully and answered truthfully. | |
| | CQ4 | Thank you for taking the time to participate in this survey. We end the survey by capturing data about the demographics of the participants. As such, data about gender, age, and experience in the topic of the survey is being collected. In addition, we want to make sure the collected data is reliable. Please select the option "No answer" for the next question that asks about the length of the survey and simply write "I've read the instructions" in the box labeled "Additional remarks". | Oppenheimer et al. (2009) |

Table 2.4 Final Set of Measurement Items for Main Study

We acquired a total of 240 participants who completed the questionnaire via the academic survey platform *prolific.com*. Out of this sample, 240 respondents answered CQ1 correctly. Twenty-three respondents failed CQ2. For CQ3 and CQ4, we decided to add a tolerance of ±1 point. The scale for CQ3 was inverted, and answers compared to PE4, while answers for CQ4 were compared to TP1. The final dataset consists of 160 samples. See the following Table for the demographics of the sample. The demographical data (*gender*, *age*, *experience*) was used for the interaction moderation of the UTAUT model's results presented in Section 2.5.3.

| **Characteristic** | **Attribute** | **Value** | | **Characteristic** | **Attribute** | **Value** | |
|---|---|---|---|---|---|---|---|
| | | **Freq.** | **Percent.** | | | **Freq.** | **Percent.** |
| Gender | Male | 110 | 68.75 | Experience with intelligent systems (EXP2) | None | 87 | 54.38 |
| | Female | 50 | 31.25 | | <1 year | 22 | 13.75 |
| | Others | 0 | 0.00 | | 1-3 years | 24 | 15.00 |
| Age | <=20 | 2 | 1.25 | | 3-5 years | 12 | 7.50 |
| | 21-30 | 62 | 38.75 | | 5-10 years | 10 | 6.25 |
| | 31-40 | 52 | 32.50 | | >10 years | 5 | 5.00 |
| | 41-50 | 31 | 19.38 | Experience with AI (EXP3) | None | 47 | 29.38 |
| | 51-60 | 12 | 0.75 | | <1 year | 39 | 24.38 |
| | >61 | 1 | 0.06 | | 1-3 years | 43 | 26.88 |
| Experience in industrial maintenance (EXP1) | None | 59 | 36.88 | | 3-5 years | 16 | 0.10 |
| | <1 year | 31 | 19.38 | | 5-10 years | 11 | 6.88 |
| | 1-3 years | 30 | 18.75 | | >10 years | 4 | 2.50 |
| | 3-5 years | 17 | 10.06 | Note: Gender, Age, EXP1, EXP2, and EXP3 were used as interaction moderation. | | | |
| | 5-10 years | 15 | 9.38 | | | | |
| | >10 years | 8 | 5.00 | | | | |

Table 2.5 Demographics of Main Study Sample

All constructs achieved reliability and validity across all measurements. Values for Cronbach's alpha, average variance extracted, and composite reliability are well above their respective thresholds. Item ATT2 was below this limit for item loadings (0.59 < 0.7) and was thus excluded from the measurement model, resulting in an overall good reliability of *ATT*. We did not observe any cross-loadings, and none of the constructs failed the Fornell-Larcker criterion (cf. Table 2.6). We additionally checked for collinearity-based indicators of common method bias following the suggestions of Kock (2015). We thus compared the variance inflation factor with the proposed threshold and found that no independent variable exhibits the variance inflation factor threshold of 3.30 and thus no common method bias was detected. See Appendix B.6 for null validation and reliability testing results and Appendix B.7 for the inner and outer variance inflation factor values. Lastly, see Appendix B.8 for the median and standard deviation of the conducted measurement items.

| Assessment Measurement Items | | | | | | | |
|---|---|---|---|---|---|---|---|
| **Construct** | **CA** | **AVE** | **CR** | **Constr.** | **CA** | **AVE** | **CR** |
| PE | 0.91 | 0.74 | 0.93 | AB | 0.89 | 0.82 | 0.93 |
| EE | 0.85 | 0.70 | 0.90 | ATT | 0.86 | 0.71 | 0.91 |
| ST | 0.87 | 0.80 | 0.92 | BI | 0.95 | 0.90 | 0.97 |
| TP | 0.85 | 0.77 | 0.91 | Note: No FL criterion, cross-loadings, and item loadings measured | | | |

*Internal consistency*: Cronbach's alpha (CA) > 0.7; composite reliability (CR) > 0.7 (Gefen et al., 2000; Hair et al., 2011)

*Convergence reliability*: average variance extracted (AVE) > 0.5 (Hair et al., 2011)

*Indicator reliability*: item loadings 0.7<*x*<1 (Hair et al., 2011)

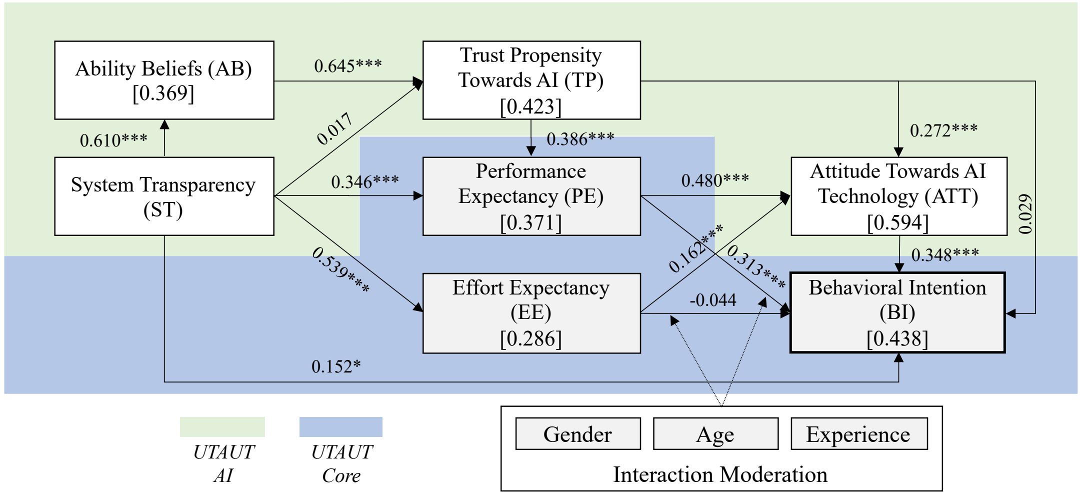
*Discriminant validity*: cross-loadings; Fornell-Larcker (FL) criterion (Fornell & Larcker, 1981; Hair et al., 2011)

Table 2.6 Validation and Reliability Testing of Main Study

### 2.5.3 Study Results

In the following, we present the results from the main study. The estimated model with *direct effect* estimates is depicted in the upper part of Table 2.7, while the lower part contains the observed *indirect effects*.

**UTAUT Results**



| | Construct | Indirectly Effecting | Strength | *p*-Value | Mediation |
|---|---|---|---|---|---|
| **Indirect Effects** | EE | BI | 0.056 | 0.059 | Indirect-only (full mediation) |
| | PE | BI | 0.167 | 0.003 | Complementary (partial mediation) |
| | ST | ATT | 0.441 | 0.000 | Indirect-only (full mediation) |
| | ST | BI | 0.278 | 0.000 | Complementary (partial mediation) |
| | ST | PE | 0.159 | 0.000 | Indirect-only (full mediation) |
| | ST | TP | 0.394 | 0.000 | Indirect-only (full mediation) |
| | AB | ATT | 0.295 | 0.000 | Indirect-only (full mediation) |
| | AB | BI | 0.198 | 0.000 | Indirect-only (full mediation) |
| | AB | PE | 0.249 | 0.000 | Indirect-only (full mediation) |
| | TP | ATT | 0.185 | 0.000 | Complementary (partial mediation) |
| | TP | BI | 0.268 | 0.000 | Indirect-only (full mediation) |

*Confidence levels*: * 0.10, ** 0.05, *** <0.001

Table 2.7 Results of Main Study

In addition, we conducted a mediation analysis based on the indirect and direct effects in our SEM to further investigate the role of system transparency and the two trust constructs following the methodology described in Zhao et al. (2010) and Hair Jr et al. (2021). The type of the mediation effect was derived by comparing direct and indirect effects of the constructs and is subsequently given in Table 2.7. The effects are determined according to the common decision scheme of mediation roles in SEM that was suggested by Hair Jr et al. (2021).

### 2.5.4 The Role of UTAUT Core Constructs

First, we examine the role of the initial exogenous UTAUT constructs *PE* and *EE*. In accordance with Venkatesh et al. (2016) and Dwivedi et al. (2019), *PE* is connected significantly to *BI*. While we observe this effect of *PE* with magnitude 0.313, we cannot confirm a significant effect of *EE* on *BI*. Thus, we can confirm H1 but reject H2. However, we can confirm a significant effect from *ATT* to *BI* in its exogenous role with an effect strength of 0.348. With the established confirmation of H3, we can observe a significant effect of magnitude 0.162 from *EE* to the construct *ATT* in its endogenous role, resulting in an indirect relationship to *BI*. Likewise, with a comparably more substantial effect than its direct connection (0.480), *PE* affects *ATT* significantly. We can therefore confirm H4 and H5, respectively. Comparing the bias-corrected confidence intervals of *EE* to *BI* (width 0.338, from -0.218 to 0.120) and *EE* to *ATT* (width 0.25, from 0.027 to 0.278) further strengthen the notion that *EE* affects *BI* rather indirectly through *ATT* in our context of intelligent systems, confirming results by Dwivedi et al. (2019) and Thomas et al. (2013).

### 2.5.5 The Role of the User's Attitude Towards Intelligent Systems

Since *ATT* is defined as an affective reaction, we conclude that this construct has increased presence in the case of intelligent systems, resulting in its role as a transitory connection of *EE* and *PE* to *BI*. It is reasonable to assume that a user is less affectionate about AI technology when it seems to be complicated to use. However, since intelligent systems are attributed with black-box properties, the ease of use can be difficult to determine beforehand. Hence, a direct connection between *EE* and *BI* seems less likely in the case of intelligent systems. The strong indirect relationship of *PE* to *BI* via *ATT* can be strengthened further by the notion of algorithm appreciation. Logg et al. (2019) found that an algorithmic system that is perceived as complex is expected to have high performance, preferable to that of humans. Thus, the increased *PE* will positively influence their *ATT* before an intention to use is formed. Contrary, the notion of algorithm aversion, as expressed by Castelo et al. (2019) can cause the *PE* to drop if it is observed or expected that the system errs, resulting in a transitory decrease of positive attitude towards the system and making it less likely for the intelligent system to be used. This can be explained by the feeling of missing control over the (partially) autonomous intelligent system (Dietvorst et al., 2015).

### 2.5.6 The Role of Trust Towards Intelligent Systems

Further, we examine the role of the trust-related exogenous constructs. We observe a significant effect from *TP* on *ATT* with a strength of 0.272. Again, we assume an indirect relation to *BI* through *ATT*, since the direct effect of *TP* on *BI* is not significant. This confirms that, especially in the context of intelligent systems, a priori formed trust influences the affection towards technology and, in a transitory fashion, the intention to use said technology. This is further supported by the mediation analysis that found a purely mediating role for TP with regard to BI. Furthermore, we argue that there is a certain order that is important towards forming a decision. While trust is an important catalyst and mediator, it is not the sole determinant and it seems, from the results, inferior to actual performance. Similar findings are confirmed by Wanner, Heinrich, et al. (2020) where in a choice experiment, the performance of an intelligent system played the most pivotal part. For some tasks it has also been found that trust is not a necessary condition for actual use (Logg et al., 2019). Especially for less critical scenarios such as maintenance, one can imagine that while important, pure performance can override trust. However, for tasks that have a more ethical and/or critical nature like healthcare, this might be different. Regarding H6, *TP* and *ATT* are highly affection-based constructs, and thus a connection between them seems highly appropriate. We can therefore confirm H6 and reject H7. We also found that *TP* has an effect on *PE* with a magnitude of 0.345. The confidence interval (width 0.269, from 0.239 to 0.508) confirms a strong effect along with hypothesis H8. The observations are in accordance with the findings of Cody-Allen and Kishore (2006), Lee and Song (2013), Choi and Ji (2015) and thus confirm H8. Drawing from the findings of Logg et al. (2019), we can explain the increase in trust through algorithm appreciation that occurs with increasing algorithm performance. Thus, if a user experiences a well-performing intelligent system, the user is more likely to subsequently change the initial propensity to trust with regard to the system. To no surprise, also the user's trust in the algorithm's ability to perform well, *AB*, has a very strong effect on *TP* with a magnitude of 0.645, confirming H9. This is also expressed by the mediation analysis that shows no sign of a possible omitted mediator and identifying the role as full mediation.

### 2.5.7 The Role of Transparency of Intelligent Systems

Finally, we investigate the role of system transparency. Regarding the trust constructs *AB* and *TP*, we can confirm H10 since we observe a very strong effect of *ST* on *AB* with a magnitude of 0.610. However, we cannot confirm a significant direct connection from *ST* to *TP* and, thus, reject H11. This is not surprising since we expect the user to partially form a pre-existing opinion within trust propensity based on the pre-existing trust in the system's ability that can be better assessed when the user has access to an explanation of the system or the underlying algorithms. This is also supported by the full mediation role of *TP* that was added for this purpose and worked as expected. Furthermore, having a competitive or complementary mediation, while befitting our proposed hypothesis could imply the presence of other trust-related, yet unexplored mediation constructs as suggested in Zhao et al. (2010).

Regarding the initial UTAUT indicators, we find that an understanding of the system also affects the expected performance, as we observe a strong effect of 0.346 of *ST* on *PE*, confirming H12. The effect can be explained by the influences of explanations on perceived performance and decision towards an intelligent system as described by Wanner, Heinrich, et al. (2020) through the means of local and global explanations. Through a global explanation of the intelligent system, the user is made aware of its complexity, leading to increased performance expectancy because intelligent systems based on deep learning models are expected to outperform other systems. Likewise, local explanations that explain a single prediction enable a consensus between the mental model of the user and the system resulting in increased *PE*. An even more potent effect of *ST* was observed regarding *EE* with magnitude 0.539 and confirming H13. Revealing the system's complexity through global explanations also enables the user to realize the effort required to implement an intelligent system, thus increasing *EE*. Besides, we observe a direct effect of *ST* on *BI*, confirming H14 at the 0.10 significance level with a magnitude of 0.152. These results are in line with Wanner, Heinrich, et al. (2020), who indicate that explainability plays a key role when deciding on an intelligent system. The direct effect is rather low compared with the indirect effect via *PE*, which is also in accordance with their findings, where explainability was not as strong a decision factor as performance. While the complementary mediation effect of *ST* regarding *BI* could indicate omitted mediators, we rather suspect the variety of functions of explanations in intelligence systems pose different influences that affect the behavioral intention in an either indirect or direct way. One can imagine the sheer presence of a self-disclosing explanation of how the system forms a decision will positively influence the *BI*, in addition with positively influencing trust in the system. In summary, we find that *ST* poses a strong influential factor concerning the attitude and intention to use an intelligent system either indirectly through previously introduced constructs or as a minor direct effect.

### 2.5.8 The Role of User Characteristics

Lastly, we look at the moderating effects of *age*, *gender,* and *experience* on *PE* and *EE*. We found no significant moderating effects of either variable or construct, contradicting the findings of Alharbi (2014) and Esfandiari and Sokhanvar (2016). We assume that this is because our pre-screening sets boundary conditions that do not allow for a great deal of variance within the participants. Thus, we observed mostly minor experiences and age gaps. In addition, due to the application domain, the sample was skewed towards men (68.75 %), barely allowing for reliable variation.

## 2.6 Discussion

### 2.6.1 Theoretical Implications

**Performance is Crucial (When Looking at Direct Effects).** We extended the modified UTAUT model by Dwivedi et al. (2019), which itself is based on the UTAUT model of Venkatesh et al. (2003), and derived additional constructs and connections in the context of intelligent systems acceptance and use. The direct and indirect effects of *PE* play a major role and are comparable to the findings of Dwivedi et al. (2019). The findings of Wanner, Heinrich, et al. (2020) confirm the dominating role of the expected performance. Contrary, we found that the expected effort is not of major concern when looking at the direct effects since it only delivers impact via indirect connections. We consider this as a first indication of the increased difficulty to build a direct intention to use in the case of intelligent systems, since the intention relies on the affection towards the system more heavily as expressed by the extended UTAUT model of Dwivedi et al. (2019). Thus, while performance is king, it is insufficient to focus only on direct effects when evaluating an intelligent system's acceptance.

**Human Attitude and Trust Steer Acceptance as Latent Indirect Factors**. As mentioned previously, the strength of indirect effects delivered through the more affectionate construct of *ATT* is substantial and shows the necessity for recognizing the deviation from a purely performance- and effort-centered model. Following that thought of increased affection constructs, we found that initial *TP* plays an essential role in determining the *PE* regarding the system. Thus, we revealed a significant indirect influence of *PE* so that we assume it is more likely that a user thinks the system will perform well when he or she trusts the system.

This transitory connection reveals the importance of trust in the context of intelligent system acceptance. The strong effect of *AB* also reveals that a prior belief in the system's problem-solving capability is fundamental. Especially when discussing algorithm appreciation vs. algorithm aversion, this particular construct plays a central role in building up *TP* towards the system. We theorize that the observation of algorithm appreciation or aversion is connected to *AB* and *TP* since they determine what to expect from a system. Trusting a system and expecting super-human performance in the case of algorithm appreciation can turn into mistrust when an aversion is built up due to the individuality of a single task or erratic system behavior (Dietvorst et al., 2015; Logg et al., 2019). However, as argued in XAI literature, an explanation of some sort can help to increase trust in the system (Adadi & Berrada, 2018; Páez, 2019).

**System Transparency Enables Trust Building and Contributes to Performance Expectancy in Both Ways.** Including *ST*, we found that revealing the system's internal decision structure (global explanation) and explaining how it decides in individual cases (local explanation) positively affects almost all constructs. First, we can confirm that an understanding of or at least visibility into the system's decision process has a powerful effect on the user's (initial) trust in the system, confirming the often-postulated connection that motivates much XAI research (Ribeiro et al., 2016b). Second, we find that *ST* also has substantial effects on *PE* and in terms of usability (i.e., *EE*). We expected the strong connection of *ST* to *EE* as a global system explanation is usually required to determine the effort it takes to efficiently train and subsequently use an intelligent system (Wanner, Heinrich, et al., 2020). It is reasonable to assume that the presence of an explanation in a psychological sense reduces uncertainty and thus technological anxiety towards the system (Miller, 2019). Therefore, we theorize that the presence of local and global explanations lets the user shift to a more rational behavior since he or she can make more informed decisions rather than relying on their gut when dealing with black-box intelligent systems.

When comparing the observed effects with related literature such as Wanner, Heinrich, et al. (2020), which deals with determining the decision factors for adopting intelligent systems, we find that the relationship between explanation performance and using a system is a more complex one. While we cannot draw conclusions regarding a trade-off, as stated in Wanner, Heinrich, et al. (2020), we found that the presence of an explanation indirectly influences the expected performance of a system, which is often the dominant influence factor. Therefore, we argue that while performance remains an essential factor for the actual intention to use, *ST* should be attributed a more critical role than current findings suggest since it can significantly increase the *PE* (or lower it depending on the revealed information through the explanation).

Additionally, taking temporal factors into account, we argue that initial trust factors and subsequently expected performance and attitude towards the system are formed by the information that is revealed before the system is used. That is, the availability of *ST* can steer those factors in one direction or another before the user sets his or her *PE*. Thus, we argue that it is less of a situational trade-off and more of a decision process that is repeated with each use and thereby manifesting in the user's attitude toward the system and AI technology in general.

**Intelligent Systems Are a Broad Concept and May Require Contextualization.** There is a plethora of research on trust in and transparency of technology - considering each aspect separately for the most part - as pointed out in the theorizing sections. We have focused our theorizing on UTAUT-related literature, but we have found that the relating constructs have been discussed similarly without relation to UTAUT. Our core contribution is twofold in that we propose to consider the combination of transparency and trust as well as their latent indirect effects to explain a user's intention to use a system. So far, research on AI acceptance has primarily focused on direct effects, where performance stands out (cf. e.g., Wanner, Heinrich, et al., 2020) or considered trust or transparency separately (see Section 2.4).

Our contribution further distinguishes itself from prior art as we focus on intelligent systems as any IT system that can make decisions indistinguishable in performance from or better than a human being based on analytical models that are opaque to the end-user. This definition is independent of the decision task. Consequently, our UTAUT model is designed as a broad model. Hence, its contribution is that it is applicable to multiple types of intelligent systems. As a consequence of this breadth, our model may lack precision for some applications. There may be factors that further affect intention to use in one case but not in another. We do not cover these domain-specific factors. We provide a base model that has merits of its own and can be extended with further constructs such as, for example, facilitating conditions or social influence if the scenario necessitates this.

Much of the extant literature has focused on domain-specific applications such as recommendation systems to support selection processes or human-computer interaction with AI agents that exhibit physical anthropomorphic demeanor. Our model can be used in these contexts but will not measure demarcating aspects such as the effect of physical interaction with AI agents.

### 2.6.2 Practical Implications

**Use Expectation Management to Form Attitude Towards the System**. In order to avoid disappointment and algorithmic aversion, managing the expectations towards performance can increase subsequent intention to use, even if the problem field for application is limited in the process since hesitation is build up through the system's self-signaling of suboptimal performance. In line with Dietvorst et al. (2016), it is important to manage expectations and show the user control opportunities of the system. This can be done with a pre-deployment introductory course involving users in the configuration state while using their knowledge in training the algorithms at the base of the intelligent system (Nadj et al., 2020).

Besides providing support for managing expectations and learning to use the system (Dwivedi et al., 2019), overcoming initial hesitation has a high priority in the case of intelligent systems.

**Control the Level of System Transparency Based on the Target Audience's Capabilities and Requirements**. Global explanations depict the inner functioning and complexity of an intelligent system. They are suitable to manage the expected effort when procuring an intelligent system, specifically through either outsourcing or in-house development. In addition, global explanations can provide a problem/system-fit perspective in that the user can observe whether the complexity of the model is suitable for the task. For example, using a complex deep learning model for an intelligent system to detect simple geometric shapes such as cracks might even decrease performance.

Local explanations can assist with explaining single predictions of intelligent systems, helping the user to compare the decision process by i) visualizing the steps towards the decision (e.g., by creating images of the intermediate layers of the artificial neural network) and by ii) attributing the input data importance regarding the output decision (e.g., by creating a heatmap of input pixels that caused the intelligent system's decision).

Explanations can also prove useful as a communication bridge between developers of the intelligent system who are not domain experts and the domain experts who are AI novices. This helps to diagnose the model and create a common understanding of the decision process from a human point of view enabling all stakeholders to jointly avoid false system behavior that can lead to algorithm aversion, such as learning a wrong input-output relation.

However, disclosing too much information about the principal rationale of the intelligent system can lead to the opposite effect (Hosanagar & Jair, 2018; Kizilcec, 2016). Especially for the stakeholder group of domain experts that are the users of the system, as opposed to developers who are required a global explanation to diagnose system failure.

**Implement Trust Management Independent of Transparency Efforts**. Our results also show that while being influenced by transparency, trust is not solely explained by it. In accordance with Madsen and Gregor (2000), the pre-existing propensity to trust that is reflected by *TP* requires extra treatment that goes beyond simply providing explanations. Thus, trust issues need to be addressed head-on by implementing guidelines for trustworthy AI (Thiebes et al., 2021). Furthermore, companies should think about introducing trust management. For similar reasons, the standard and idea of risk management were introduced decades ago: identify uncertainty roots and trust concerns and create trust policies (Müller et al., 2021).

The uncertainty regarding *PE* and *EE* could be reduced proactively by offering training to the users to experience the intelligent system to form a feeling of beneficence (Thiebes et al., 2021). Using the system in a training session in a non-critical context can support the acceptance of the system and provide a solution to the initial uncertainty about the performance. According to Miller (2019), this could provide partial transparency, in this case, as an indicator of ability and performance.

### 2.6.3 Limitations

In our study, we presented a use case based on a medium-stake scenario. Here, wrong decisions have consequences such as machine breakdown or downtimes within the production plant. This can result in high monetary loss. Nevertheless, wrong decisions do not endanger human lives. We used this scenario for two reasons. First, for the sake of generalization, and second, we tried to replicate a typical industrial medium-stake maintenance use case. However, following Rudin (2019), we need to keep in mind that user behavior may differ in high-stake use cases resulting in bodily harm due to the potential consequences of wrong decisions. Inversely, this also applies to low-stake use cases. Further, using a work system scenario entails that users cannot opt to not use the system. In the consumer space, where consumers can decide to choose a non-intelligent system or use no system at all, the results and necessary constructs may differ.

Further, we focused on user perception. Consequently, we cannot verify if the user's perception corresponds to the actual user behavior. This is especially related to the following: *PE* on whether the system can increase the user's productivity, *EE* on whether the user finds the system easy to use, and *ST* on whether the user understands why the system made the decision it did. The latter is closely related to findings from Herm, Wanner, et al. (2021a), who address the knowledge gap on the perceived explainability of intelligent system explanations and user task solving performance.

Lastly, within our use case, we provided a textual and graphical explanation for intelligent system predictions and did not impose time limits for decision. While many different XAI augmentation techniques have been developed in XAI research, further evaluation of these techniques seems necessary. Similarly, the results may differ when different XAI augmentation techniques are applied. Hereby, inappropriate explanations can cause an overload of the user's cognitive capacity (Grice, 1975). Furthermore, a personalized explanation can increase the behavior intention (Schneider & Handali, 2019).

## 2.7 Conclusion and Outlook

By extending the UTAUT model with factors of attitude, trust, and system transparency, we were able to explain better the factors that influence the willingness to accept intelligent systems in the workplace.

Our extension centers on affection constructs such as *ATT, TP,* and *AB* while simultaneously integrating *ST* as an opportunity to steer both to address the information asymmetry between black-boxed, anthropomorphic agents and their human principal. This combination as well as the consideration of latent indirect factors provides the community with a means to look beyond performance as the dominating decision factor for intelligent system efficacy.

In summary, on the one hand, our model enables researchers to understand the influence of this human factor for intelligent systems and in more general for analytical AI models. On the other hand, our findings can help to create measures to reduce acceptance barriers in practice and thus better leverage AI capabilities. Since our research is based on the UTAUT model and established extensions, we assume that our model is of general nature and generally transferable to or contextualizable in other domains. The results of our model application may be more specific to work system in maintenance as discussed in the limitations.

Since our research results clearly indicate how behavioral intention is influenced by this human factor, we aspire to develop design principles for intelligent systems that contribute to the user's willingness to accept and use these systems in their daily work.

# 3 A Taxonomy of User-centered Explainable AI Studies

*Lukas-Valentin Herm, Jonas Wanner, and Christian Janiesch*

**Abstract.** The progress in the research field of machine learning is fast-paced and it is most noticeable in terms of prediction performance. However, there seems to be a lack of understanding of the explanatory value for the actual user. As only a user-appropriate implementation realizes effective human-machine cooperation, this must be the goal for any intended intelligent system development. Accordingly, some studies have addressed the problem. However, their aims and methods vary, and a meta synthesis of the results is missing. To address these problems, we have developed a taxonomy of user-centered XAI studies. It allows both the conception and the classification of current user-centered XAI studies. Furthermore, through descriptive analytics and a cluster analysis, we identify patterns and archetypes to better conceptualize the field and support future research.

**Keywords.** User-centered XAI, Explainability, Taxonomy, Archetypes.

**Reference.** Herm, L.-V., Wanner, J., & Janiesch, C. (2022). *A Taxonomy of User-centered Explainable AI Studies*. Pacific Asia Conference on Information Systems, Virtual Conference. https://aisel.aisnet.org/pacis2022/9

## 3.1 Introduction

Machine learning (ML) is the state-of-the-art for knowledge discovery and decision support nowadays (Janiesch, Zschech, et al., 2021). Thereby, ML denotes a concept where computer systems learn to solve tasks autonomously using mathematical algorithms (Bishop, 2006). The resulting analytical models offer novel possibilities for automated decision support in complex decision situations and the resulting intelligent systems even outperform humans at certain tasks. Consequently, ML gradually begins to support or replace humans in various areas of business and daily life (Adadi & Berrada, 2018).

Be that as it may, ML technology is not as widely used in practice as its benefits might suggest. While researchers continually try to improve algorithms and tackle robustness problems, they have reduced the traceability of analyses at the same time due to the algorithms increasing complexity (Janiesch, Zschech, et al., 2021). Thus, the underlying decision logic and, hence, the output of these algorithms became inexplicable for humans (Shin, 2021). Particularly for end users, such systems appear as black boxes and the system's lack of transparency leads to a reduction of trust in its predictions (Wanner, Popp, et al., 2021). This could be fatal, especially when ML is applied in sensitive and safety-critical areas such as in medical science or in high-stake maintenance (McKinney et al., 2020).

Explainable artificial intelligence (XAI) addresses this problem and, consequently, is receiving increasing attention in research and practice (Arrieta et al., 2020). By displaying appropriate visualizations, XAI aims at converting or augmenting black-box models into surrogate grey-box or white-box models to enhance user trust and eventually system acceptance, while maintaining high performance at the same time (Yang et al., 2022). Here, explanation refers to the system's ability to provide additional information to address the information asymmetry between the intended end user and the intelligent system in an understandable manner (Shin, 2021). Various frameworks such as SHAP and LIME have been developed to address this (Jesus et al., 2021).

However, XAI technology often only provides data-driven interpretability explicating cause and effect, but additional factors, which influence the decision situation, seem relevant as well (van der Waa et al., 2021). Only predictions that are explainable associate interpretable data with a representation and importance and makes it accessible for its respective target audience. Consequently, the usefulness of a given explanation in an intelligent system crucially depends on the individual receiver's expectations and knowledge base (Wanner, Popp, et al., 2021). Thus, comprehending user priorities and user-specific needs directly impacts the effectivity of the intelligent system. So far, we do not know much about the human evaluation of intelligent systems (Miller, 2019).

To date, several studies already examine the perception of user towards XAI applications in different scenarios (e.g., Khodabandehloo et al., 2021). However, to the best of our knowledge, there is no comprehensive summary and structuring of these user centered XAI studies in academic literature. However, this is necessary to gain a comprehensive overview of the research field and to identify further research needs and potentials. We seek to close this gap by analyzing the state-of-the-art in user-centered XAI studies using a structured literature research as backbone for the development of a holistic taxonomy to analyze the meta-characteristics of XAI as well as their dimensions to ultimately discover patterns and archetypes within.

Thus, our paper's goal can be summarized with the following two research questions (RQ):

*RQ1: How can we systematize user-centered XAI studies in academic literature in a taxonomy?*

*RQ2: Which patterns and archetypes can we derive from the analysis of user-centered XAI studies?*

To address these two research questions, our paper is structured as follows: In Section 3.2, we present the theoretical background of XAI and the related work. Section 3.3 covers our methodology and Section 3.4 outlines the literature search process. In Section 3.5, we develop and present out taxonomy and in Section 3.6, we detail our analysis. Lastly, we present a conclusion, limitation, outlook as well as derived research implications in Section 3.7.

## 3.2 Theoretical Background

### 3.2.1 Machine Learning

ML is a concept belonging to the generic field of artificial intelligence in which machines learn to solve tasks based on mathematical models and algorithms. Without being explicitly instructed to do so, the algorithms learn from observations and improve their analyses automatically (Bishop, 2006). Learning in this context refers to the process of optimizing predefined parameters by using training data (Janiesch, Zschech, et al., 2021). Intelligent systems can provide improved decision support through predictions and help to gain in-depth knowledge into extensive amounts of data (Carvalho et al., 2019). It is therefore a promising technology for applications in various data-driven fields such as the medical diagnostics or industrial maintenance (McKinney et al., 2020).

For several reasons, well-founded decisions in practice are chiefly based on a combination of both, human and machine intelligence also termed hybrid intelligence (Dellermann et al., 2019). Essential prerequisites for hybrid intelligence are the comprehensibility of and the confidence in the system by the user (Das & Rad, 2020). Today, this is often not the case as the algorithm's underlying logic and the resulting outputs are not comprehensible for end users (Herm, Wanner, et al., 2021a). Consequently, humans tend to value those black boxes less, which comprise models that are not transparent, interpretable, or trustworthy (Shin, 2021). As the development of opaque models such as deep neural networks is trending in recent years, the black-box problem intensifies. This finding together with several governmental approaches such as the "right to explanation" by the General Data Protection Regulation (GDPR) of the European Union (Goodman & Flaxman, 2017) or the Explainable AI (XAI) program of the Defense Advanced Research Projects Agency (DARPA) (Gunning, 2017), reinforced the vehement claim of researchers for transparent models and comprehensible ML designs.

### 3.2.2 Explainable AI

XAI is a multidisciplinary field of research that targets the conflict of complexity and effectiveness in ML (Meske et al., 2022). By augmenting black-box models, XAI tries to generate interpretable and comprehensible transparent models that retain a high level of prediction performance (Herm, Wanner, et al., 2021a). Thus, XAI research in a narrower sense focuses on mathematical methods that are used to make the black-box model's internal computational logic interpretable (Yang et al., 2022). At best, this results in a system with a transparent inner reasoning and programming logic (Mohseni et al., 2021). So, an explainable algorithm may provide reasons why a certain result was achieved, or which inputs could be changed to receive a different output (Miller, 2019).

In a wider sense when seeing artificial intelligence as intelligence demonstrated by machines that is not well understood (Makridakis, 2017), XAI is also a general concept aiming at increasing the accessibility of intelligent systems rather than a concrete technological approach (Meske et al., 2022). Consequently, the question of what constitutes a good explanation is not trivial and subject to many research fields (Förster et al., 2020a). Dam et al. (2018) defined explainability as "the degree to which a human observer can understand the reasons behind a decision (e.g. a prediction) made by the model". Cui et al. (2019) described it as the system's capability to provide additional information for filling the information gap between the user and the artificial intelligence in an understandable manner. More precisely, Arrieta et al. (2020) also take the audience into account and state: "Given a certain audience, explainability refers to the details and reasons a model gives to make its functioning clear or easy to understand". As it is in the nature of any research, different approaches and types of explanations have been proposed recently (Mohseni et al., 2021). Overall, these approaches can be differentiated by their focus on a functional (global) and a social (local) focus of explanations (Zhang et al., 2021). While the former term implies the process of adding transparency to the decision model or to individual components such as parameters and algorithms, the latter is characterized by post-hoc explanations that are either textual, visual, or example-based (Wanner, Herm, Heinrich, et al., 2021). Global explanations mainly target explanations among ML experts. Local explanations focus on the communication between experts and intended user (Miller, 2019).

### 3.2.3 User Studies

User studies are a common method in research to investigate the perception and the attitude of a user (group) regarding the subject under examination. Amongst others, methodological approaches are questionnaires, task analyses, behavioral observations, or a combination of several methods. To empirically test the user acceptance of new technologies in the field of IS and to reveal the mediators between system characteristics and user behavior, researchers often build their hypotheses on the technology acceptance model (TAM) (Davis, 1989) or variations thereof. In the TAM, influential factors for the acceptance are perceived usefulness and perceived ease of use that led to a certain attitude and, thus, to a behavioral intention to use the system. Facilitating conditions such as trust or performance expectancy are assumed to have an impact as well (Wanner, Popp, et al., 2021).

Especially in the field of user centered XAI studies, it seems as there is a great diversity of the assessed constructs that influence the perception of the explanation and, thus, the acceptance of the XAI system in practice. While some studies concentrate more on social conditions such as perceived trust or perceived transparency, others emphasize the system performance, and still others differentiate the level of expertise or the user group, respectively. In summary, as Mohseni et al. (2021) and Hoffman et al. (2018) state, most of the constructs are somehow interrelated or even denote the same. A comprehensive analysis to consolidate the available knowledge is necessary.

### 3.2.4 Research Gap

In recent years, numerous contributions dealing with the topic of XAI and the assessment of XAI applications through user-centered studies have been published (Herm, Wanner, et al., 2021a; Wanner, Popp, et al., 2021). Among others, Arrieta et al. (2020) provides a thorough overview of the field of XAI. They explain relevant concepts of XAI, develop a taxonomy of explainability techniques related to different ML models and outline future challenges regarding responsible AI while focusing target audiences. Wang et al. (2019) shed light onto theoretical underpinnings of human decision-making by proposing a framework for building human-centered decision-theory-driven XAI systems. The framework draws on findings from social sciences through an extensive review and is applied in practice. Moreover, Wanner, Herm and Janiesch (2020) conducted a literature review to examine the value of explainability in ML models through XAI model transfers to understand the trade-off in XAI from a user's perspective. Yet, the conduced literature review itself was focused on technical model transfers rather than user-centered XAI studies. Further, Mohseni et al. (2021) surveyed different evaluation methods and measures for interpretable ML systems in computer science literature. By analyzing evaluation measures (e.g., mental model, usefulness, satisfaction) combined with targeted user types or evaluation methods, they build a comprehensive framework for evaluation methods in XAI systems. The framework of Hoffman et al. (2018) also comprises steps and measures for user satisfaction, the understanding of explanations or user trust in XAI. Contrary to our contribution, the emphasis of the latter however was more on the measurement side whereas in our review, we seek to focus the respective user needs in.

Summarizing, to the best of our knowledge, no holistic review of the state-of-the-art in user-centered XAI studies exists. Further, no contribution links the results of the user studies to the needs of special user groups in a comprehensive manner.

## 3.3 Methodology

To address the identified research gaps, we applied a sequence of three major steps: i) literature collection; ii) taxonomy development; and iii) analysis of XAI patterns and archetypes. See Figure 3.1 for an overview.

**Literature Collection.** We have collected user-centered XAI studies as our preliminary work for the structured analysis of the respective research area. For scientific rigor, this was done in accordance with vom Brocke et al. (2009). The recommendation framework comprises five phases to ensure a comprehensive review process of (sub-)areas of research.

**Taxonomy Development.** The research artifact *taxonomy* has established itself in the Information Systems community as a popular method to organize knowledge in a structured way (Kundisch et al., 2022). By enabling a classification of individual objects (or publications), taxonomies allow for analyzing relationships within objects in the respective context and for understanding of complex areas. For the taxonomy development process, we have followed the framework by Nickerson et al. (2013). Lastly, we evaluate our taxonomy according to the guidelines of Rich (1992).

**XAI Pattern Analysis.** To identify research trends and gaps in the field of user-centered XAI studies, we first classified all papers into the developed taxonomy. Then, we applied a descriptive dimension analysis for every characteristic we derived within our taxonomy. Lastly, we made a cluster analysis to find patterns and derive archetypes within these contributions.

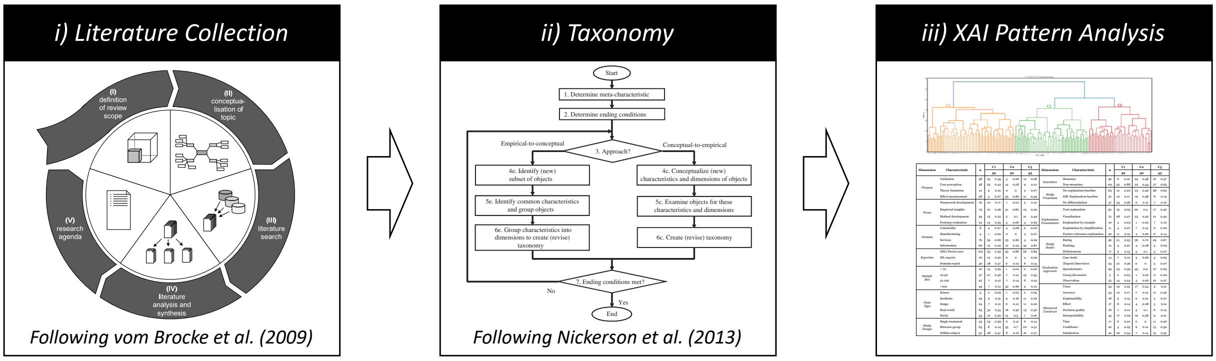

i) Literature Collection
(I) definition of review scope
(II) conceptua-lisation of topic
(III) literature search
(IV) literature analysis and synthesis
(V) research agenda
Following vom Brocke et al. (2009)
ii) Taxonomy
Start
1. Determine meta-characteristic
2. Determine ending conditions
Empirical-to-conceptual
3. Approach?
Conceptual-to-empirical
4e. Identify (new) subset of objects
5e. Identify common characteristics and group objects
6e. Group characteristics into dimensions to create (revise) taxonomy
4c. Conceptualize (new) characteristics and dimensions of objects
5c. Examine objects for these characteristics and dimensions
6c. Create (revise) taxonomy
7. Ending conditions met?
No
Yes
End
Following Nickerson et al. (2013)
iii) XAI Pattern Analysis

Figure 3.1 Overview of Applied Methodology

## 3.4 Literature Collection

vom Brocke et al. (2009)'s framework for structured literature research implies a sequence of five distinct steps: the i) definition of the review scope, the ii) conceptualization of the topic, the actual iii) literature search, the iv) analysis and synthesis of the literature, and the v) identification of research gaps to establish an agenda for future research.

**Definition of Review Scope.** Based on Cooper (1988)'s taxonomy, the focus is on research outcomes of user-centered XAI studies and on XAI applications. Our goal is to derive user concerns connected to XAI to better integrate XAI research and its implementation in practice. Thereby, we take a neutral perspective. We searched the academic literature in an iterative process. However, as we did not include literature from related fields or built upon previous meta-syntheses, the review coverage is neither exhaustive nor central but representational. Results were grouped based on similar concepts, consistent with our research questions. As target audience, we primarily define specialized researchers from the field of XAI or ML, but general Information Systems researchers as well as practitioners that can use the summarized insights from the user-centered XAI studies are also addressees.

**Conceptualization of Topic.** We have conceptualized the relevant research topics previously in the theoretical background (cf. Section 3.2). Therein, we characterized the main concepts and applications of ML and XAI. Furthermore, we clarified key terms such as explainability and presented important aspects associated with user studies. For our search, we define that a pertinent search result must explicitly encompass a user study about an XAI application. A user is thereby any person that uses the application in practice. We separate the user groups (e.g., non-expert, expert) later in the analysis.

**Literature Search.** Following vom Brocke et al. (2009), we successively conducted a journal search, a database search, a keyword search, as well as a forward and backward search. In terms of quality and relevance, the journal search was initially restricted by journals rankings. We removed the restriction later to increase the relevant hits. We conducted the subsequent database search on eight common databases in the field of computer science, business management and information systems. The underlying aim was to open the search for possible business applications and related disciplines rather than restricting it to the technological facets of XAI in the IT-related databases only. For our keyword search, we used a two-part string. The first element consisted of the broader term "explainable AI" and related terminology. The second part comprised specific terms that target the user focus and constructs associated with the model understandability. Our search comprises contributions index by February 2022.

During the first iteration of our search, we retrieved 76.476 possibly relevant publications before a deletion of duplicates. After a closer examination of title, abstract and key words, we reduced the hits to 2.731 unique and potentially relevant publications. After a full-text analysis 118 relevant articles remained. We retrieved them mainly from two databases, SCOPUS and ACM Digital Library. We also conducted a forward search via the citation data using Google Scholar as well as a backward search based on the reference list to identify further 35 articles. The final set comprises 152 articles. To ensure validity of our results, we used a inter-rater reliability test according to Fleiss (1971). See Figure 3.2 for an overview.

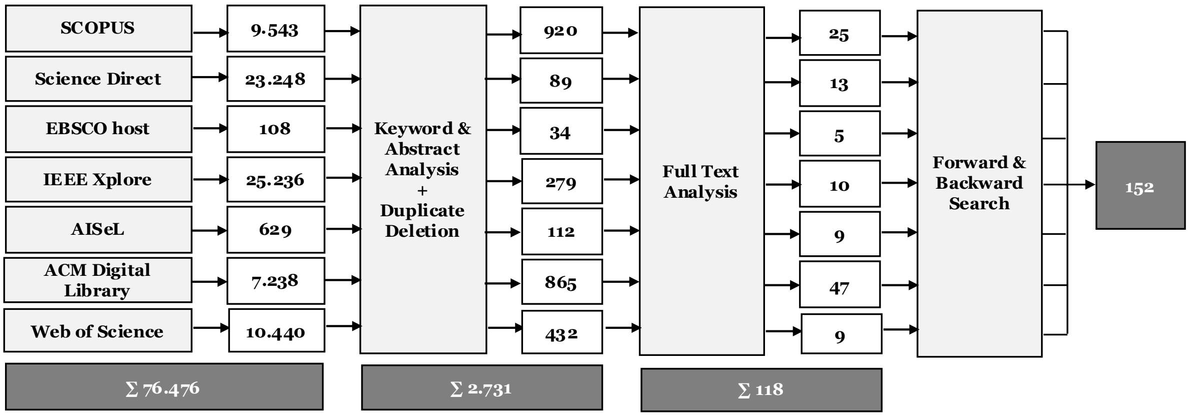


Figure 3.2 Overview of Literature Review Procedure

## 3.5 Taxonomy Development

### 3.5.1 Taxonomy Building

Nickerson et al. (2013)'s framework comprises three main parts: i) determining meta-characteristics, ii) specifying ending conditions, and iii) identifying dimensions and characteristics. Lastly, we validated our results.

**Meta-characteristics.** As we aim at a holistic view of user-centered XAI studies, we have opted for a quadripartite meta-characteristic classification as described below: *objective*, *participant*, *method*, and *measurement*. Nunes and Jannach (2017) define the first dimension that we use, *objective*, as the intention(s) behind the study. In addition, Adadi and Berrada (2018) defined three dimensions of relevance that we use: *participant*, as an ML model has to be understandable by humans, which implies their individual perception shaped by their know-how and experience; *method* refers to the (technical) quest to make an intelligent system explainable; and *measurement* as the approach to address the need to evaluate these methods focusing on the respective users (Adadi & Berrada, 2018).

**Ending Conditions.** Nickerson et al. (2013) defined subjective and objective criterions to ensure a valid status of taxonomy completeness. We adopted their conditions except for the restriction of object exclusivity. This happened in agreement with other authors in the field (e.g., Püschel et al., 2016). Thus, our final taxonomy allows specific objects to be classified in more than one characteristic per dimension. This improves the taxonomy's clarity.

**Dimensions and Characteristics.** We used the defined ending conditions in our iterative process of identification and selection of dimensions and related characteristics. Thus, we performed a total of four iterations. *Iteration I* - In our first iteration, we aimed at an initial conceptualization of our taxonomy. For this purpose, we used publications that also take a holistic view. These were mainly literature reviews and survey articles. Further, we used taxonomies and frameworks. The contained categorization information was used to derive first dimensions and related characteristics. *Iteration II* - In our second iteration, we aimed at an extension and refinement of our first iteration's categorization schema. Thus, we used publications that are topic specific review articles of the dimensions identified beforehand. A precondition for the results is that the respective authors have prepared their data in a taxonomic or categorical way. The new information was used to further improve our developed dimensions and related characteristics. *Iteration III* - Due to the size of the preliminary taxonomy from the second iteration cycle, we decided to move from a conceptualization perspective to an empirical evaluation. For this new purpose, 60 out of our 152 user-centered XAI studies (cf. Chapter 4) were randomly selected to validate the ability of our taxonomy's scheme developed for classifying. Thus, we were able to modify several dimensions and characteristics that were not proofed as suitable for a classification for user centered XAI studies. *Iteration IV* - In the next iteration, we adopted our procedure from the previous iteration. We have randomly selected 40 out of the remaining 92 user centered XAI studies found and further modified dimensions and characteristics until we met our predefined ending conditions. See Table 3.1 for a summary.

| It. | Ap. | Summary | # |
|---|---|---|---|
| I | C2E | Analysis of articles that take a holistic view of the research area to structure it to achieve a first categorization schema. | D=25; C=168; P=13 |
| II | C2E | Analysis of articles that are topic specific with own levels of categorization to further extend and refine our schema. | D=27; C=181; P=15 |
| III | E2C | Classification of study articles with the respective schema to assess its quality and to update detected weaknesses. | D=15; C=65; P=60 |
| IV | E2C | Confirmation of the third iterations' classification scheme and activation of the specified ending conditions. | D=13; C=52; P=40 |

*It.=Iteration; Ap.=Approach; C2E=Conceptual-to-Empirical; E2C=Empirical-to-Conceptual; #=Numbers; D=Dimensions; C=Characteristics; P= Number of Contributions*

Table 3.1 Summary of Taxonomy Development

### 3.5.2 Final Taxonomy

In the following section, we describe our final taxonomy (cf. Table 3.2). We distinguish the four different meta-characteristics as introduced above, which we have expanded into dimensions and characteristics.

| MC | Dimension | Characteristic | | | | |
|---|---|---|---|---|---|---|
| OBJ. | Purpose | Validation | User perception | Theory formation | Effect measurement | |
| | Focus | Framework development | Empirical insights | Method development | Prototype evaluation | |
| PARTICIPANT | Domain | Commodity | Manufacturing | Services | Information | |
| | Expertise | (ML) Novice user | Domain expert | ML expert | | |
| | Incentives | Monetary | Non-monetary | | | |
| | Sample Size | <10 | 10-50 | 51-100 | >100 | |
| MET. | Data Type | Sensor | Image | Social | Synthetic | Real world |
| | Explanation Presentation* | Text | Visualization | Example | Simplification | Feature relevance |
| MEASUREMENT | Study Design | Single treatment | Between-group | Within-subject | | |
| | Study Treatment | No-explanation baseline | Different explanation baseline | No differentiation | | |
| | Study Scales | Rating | Ranking | Dichotomous | | |
| | Eval. Approach | Case study | Interview | Questionnaire | Group discussion | Observation |
| | Measured Construct* | Accuracy | Trust | Explainability | Decision quality | Interpretability |
| | | Satisfaction | Confidence | Effort | Time | |

*MC=Meta-Characteristics; OBJ.=Objective; MET=Method; *Dimensions are non-exclusive*

Table 3.2 Taxonomy of user-centered XAI-Studies

**Objective: Purpose.** The purpose highlights the general intention of the user-centered XAI study. *Validation* describes the evaluation of existing artefacts (Madumal et al., 2019). *User perceptions* is about deriving new insights by the perception of users considering an XAI artifact (Lakkaraju & Bastani, 2020). *Theory formation* deals with the design of new artefacts from summarized knowledge (El Bekri et al., 2019). In contrast to user perception, the *effect measurement* focuses on the overall effects through XAI (Zhang et al., 2020).

**Objective: Focus.** This dimension is about the focus of the study's contribution. The characteristic *framework development* signals the building of a theoretical framework for XAI (Lu et al., 2019). *Empirical insights* focus on gaining new knowledge based on user studies (Zhang et al., 2020). *Method development* focuses on developing new methods for gaining knowledge (Hohman et al., 2019). Lastly the *prototype evaluation* is dealing with the assessment of developed XAI approaches (Schreiber & Bock, 2019).

**Participant: Domain**. The categorization of the domain is oriented on the (extended) sector theory according to Forastié (2020). This clusters the different domains of the respective preliminary works (e.g., Adadi & Berrada, 2018; Nunes & Jannach, 2017). *Commodity* comprises all related topics, such as food or milling, while *manufacturing* summarizes all areas in which end products are physically created or processed such as industrial maintenance. Further, *services* are activities without a physical product exchange such as health care services. *Information* is a new sector covering information-technology-related topics such as ML-based software or social networks.

**Participant: Expertise.** The second dimension of the meta-characteristic participant takes the user-background into consideration. Here *(ML) Novice User* are general end users and laymen without expertise in the field of ML. *Domain Experts* use applications as part of their profession (e.g., medical professionals, data analysts) and have expertise in the application domain itself. However, they lack expertise in the technical aspects of the algorithms behind the applications. *ML experts* (e.g., ML model engineers or data scientists) have a high level of knowledge in the field. They build and improve ML models as part of their daily business (Chromik & Schuessler, 2020; Mohseni et al., 2021).

**Participant: Incentives.** The motivation to participate in a study and the user evaluation itself might be biased when participants receive a monetary reward (Chromik & Schuessler, 2020). Accordingly, we distinguish between *monetary* and *non-monetary* incentives for the participation in the user study.

**Participant: Sample Size.** The sample size considers the number of participants that have been involved in the user study. We used a numerical differentiation of the sample size and split it in *<10*, *10-50*, *51-100,* and *>100* (Nunes & Jannach, 2017).

**Method: Data Type.** Since ML problems are strongly dependent on the respective dataset, there is a subdivision according to the data type (e.g., Wanner, Herm, & Janiesch, 2020). *Sensor* data describes sensor-measured values, such as pressure or temperature. *Image* data represents all kind of images such as thermal images. *Social* data is all data generated by humans such as social media entries. *Synthetic* data does not originate from a real data basis, whereas *real world* is data from business environments such as financial transactions or product ratings.

**Method: Explanation Presentation.** The style and mode of explanations presented to the users are categorized in the dimension explanation presentation (Abdul et al., 2018; Arrieta et al., 2020; Mohseni et al., 2021). *Text* thereby denotes textual explanations such as natural language explanations or word-level feedback. *Visualization* comprises graphical and interactive explanation styles such as saliency maps. Explanation by *example* entails that a certain example is provided to explain the behavior such as specific recommendations. *Simplification* is a presentation mode where the inherent logic of the AI application is explained through simplified rules for example through hierarchical decision trees. *Feature relevance* explanation styles display the main characteristics of items on an instance-level.

**Measurement: Study Design.** The study design specifies the test layout of the user study. In *single treatment* studies, there is only one group of participants. They must solve a certain task or assess an application without being confronted with different explanation conditions (e.g., design studies). *Between-group* studies (randomly) assign the participants to two or more different treatment groups, which only receive one of the treatment conditions. Thus, the analysis is based on a group-comparison. *Within-subject* user studies display different treatment conditions to the same individual until all participants experienced all treatment conditions. The evaluation of within-subject studies can therefore be conducted on an individual level (Nunes & Jannach, 2017).

**Measurement: Study Treatment.** The type of treatment describes the different explanation modes within the user study. Thereby, the characteristic *no-explanation baseline* includes all studies in which a certain explanation type is compared with the original algorithm or model without any explanation (with vs. without explanation). When the investigated explanation type is assessed against another explanation type, it refers to the characteristic *different explanation baseline*. *No differentiation* covers cases when the evaluation comprises only one explanation mode or when there is no further specification (Mohseni et al., 2021).

**Measurement: Study Scales.** Scales measure the constructs and elicit knowledge about the users' attitude. *Rating* scales (e.g., Likert scale) are numeric or descriptive and consist of an ordinal continuum of categories. Users must rate their personal level of agreement or satisfaction. *Ranking* scales imply an order of the objects under assessment where users may choose between a list of options. *Dichotomous* scales comprise questions that are diametrically opposed to each other and only have two choices (e.g., yes / no, true / false) (Albert et al., 2009; Gena et al., 2011).

**Measurement: Evaluation Approach.** Based on how the data was collected methodically, one can distinguish five major evaluations (e.g., Chromik & Schuessler, 2020; Hoffman et al., 2018; Mohseni et al., 2021). While *case studies* provide insights into user studies without collecting the user data directly*,* the other approaches do so. *Interviews* question the users through a conversation or discussion. *Questionnaires* are written surveys where the users must answer or rate specific predefined questions. A *group discussion* is an approach where several participants are questioned together and is often used in design studies. *Observations* comprise the examination of a problem-solving task such as card sorting or prediction tasks, which are often conducted in combination with think-aloud protocols.

**Measurement: Measured Construct.** The independent variables measured and evaluated in the user study are condensed in the dimension measured constructs (Hoffman et al., 2018; Mohseni et al., 2021). *Accuracy* denotes the technical performance of the algorithm itself. Further, *trust* signifies trust in the system and *transparency* respectively transparency of the system. *Decision quality* measures the usefulness of the decisions made by the system.

*Understandability* of the system indicates the compatibility of the explanation with the users' mental model. *Satisfaction* of the user measures the users' overall system perception and rationale. *Confidence* refers to how certain the result is from the user perspective. *Effort* implies measures for the cognitive load of the user. The construct *time* measures the duration to solve a certain user task.

### 3.5.3 Evaluation of Taxonomy

Lastly, to ensure the validity of our findings, we used the evaluation guidelines from Rich (1992). Table 3.3 describes these guidelines and how our taxonomy addresses those. The author defines seven guidelines for the design and evaluation of taxonomies: breadth, meaning, depth, theory, quantitative measurement, completeness, and recognizability.

| Guideline | Description | Application within Taxonomy |
|---|---|---|
| Breadth | Classification system must be typology or taxonomy to shape character selection. | We derived a taxonomy consisting of four meta-characteristics. |
| Meaning | Build upon a philosophical foundation to explain and emerge classification groups. | We build a taxonomy on the existing body of knowledge of user-centered XAI research. This taxonomy enables researcher to uncover lack in XAI research and further address these research gaps. |
| Depth | Basis should be multivariate analysis. | Following Nickerson et al. (2013), we developed through four iterative conceptional and empirical steps a taxonomy, which compromises all characteristics exhaustively. Further, we derived a cluster and archetype analysis. |
| Theory | Use theory to gain qualitative base for determination of units and variables. | We used 152 research contributions to build a qualitative base to detected dimensions and characteristics of user-centered XAI studies. |
| Quantitative Measurement | Placement of taxa into groups through numerical procedures and multivariate data analysis. | We applied empirical steps in our taxonomy building process and a quantitative ex-post analysis to determine the completeness of our result. |
| Completeness and logic | Classification must be thorough, comprehensive, and detailed. | As we adopted and fulfilled five subjective and seven objective ending conditions from Nickerson et al. (2013). |
| Recognizability | Mirror real world with taxonomy. | We derived the taxonomy through a comprehensive literature review. As we focus on XAI research, and we define this area as target. We did not encounter a contribution, that we were not able to classify. |

Table 3.3 Evaluation of Taxonomy According to Rich (1992)

## 3.6 XAI Pattern Analysis

### 3.6.1 Descriptive Dimension Analysis

In the following, we describe the distribution of the identified characteristics and thereby show, from a quantitate point of view, well researched areas as well as topic, that may need further research. For a detailed overview please see Figure 3.3.

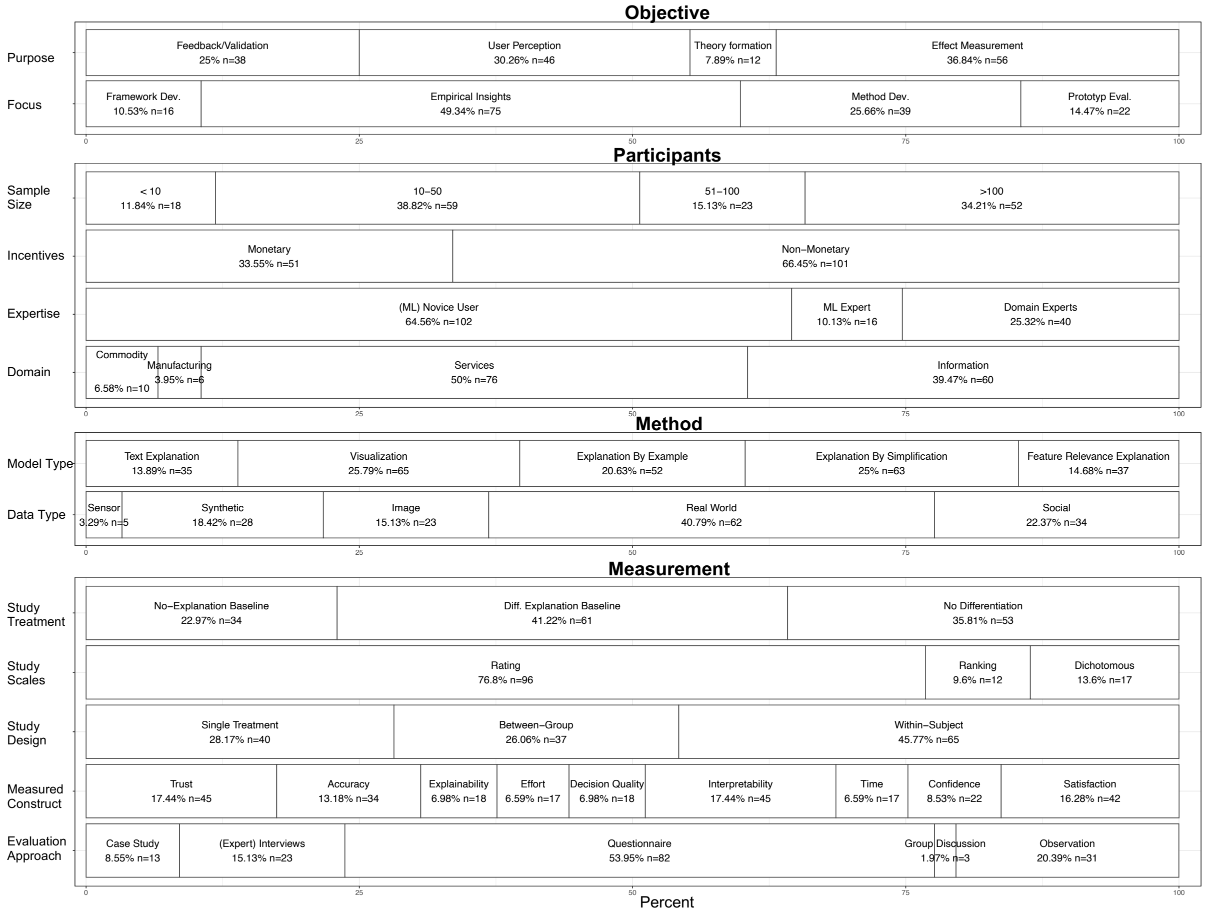


Figure 3.3 Frequency and Occurrence of Characteristics

**Objective.** Looking at the *explanation purpose,* it becomes apparent, that most of the contributions (*n*=58; ≈37 %) deal with the effect measurement of XAI. Also, many (*n*=38; ≈25 %) contributions deal with the validation of XAI approaches or the user perception (*n*=46; ≈30 %). In contrast, only (*n*=12; ≈8 %) studies are forming new theories for the investigation of users in the context of XAI. While the evaluation *focus* is mostly set on gaining empirical insights (*n*=75; ≈49 %) or developing new methods (*n*=39; ≈26 %), the development of frameworks is rarely approached (*n*=16; ≈11 %). The same applies for the evaluation of new XAI prototypes (*n*=22; ≈14 %).

**Participants.** Looking at the meta-characteristic *participants*, most studies are from the field of Service (*n*=76; ≈50 %) or the Information Technology (*n*=60; ≈41 %) *domain*, while the user-centered impact of XAI in the field of manufacturing (*n*=6; ≈4 %) and commodity (*n*=10; ≈7 %) is rarely researched. Regarding *expertise*, most of the contributions focus on ML novice users (*n*=102; ≈65 %) or domain experts (*n*=40; ≈25 %), while only few focus on ML experts (*n*=16; ≈10 %). In that context, user-centered studies primarily use non-monetary *incentives* (*n*=101; ≈69%). Likewise, considering at the *size* of the interviews participants most contributions used a sample size of 10-50 participants (*n*=59; ≈39 %), followed by a sample size of over 100 participants (*n*=52; ≈34 %).

**Methods.** Used ML models are typically applied on real-world *data types* (*n*=62; ≈41 %) or social data types (*n*=34; ≈22 %). However, many contributions are using simulation (*n*=28; ≈19 %) or image data (*n*=23; ≈15 %). Thereby, we noticed a lack of research when it comes to sensor data (*n*=5; ≈3 %). Further, different XAI approaches are used. Most contributions rely on visualization (*n*=65; ≈26 %), explanation by simplification (*n*=63; ≈25 %), or explanation by example (*n*=52; ≈21 %) for their studies. Thereby, it is noticeable that only relatively little research is done with respect to textual explanations (*n*=35; ≈14 %).

**Measurement.** Regarding the dimension *study design,* studies mainly differentiate within subjects (*n*=65; ≈46 %) and between groups (*n*=37; ≈26 %), but many studies also use single treatments (*n*=40; ≈28 %). These studies often employ a differentiation explanation baseline (*n*=61; ≈41 %) or provide no differentiation baseline (n=53; ≈36 %). To measure their constructs, these contributions predominantly use ratings (*n*=96; ≈77 %) for their *study scales*. Consequently, rankings (*n*=12; ≈10 %) and dichotomous (*n*=17; ≈14 %) are less used. As an *evaluation approach*, questionnaires are most frequently used (*n*=82; ≈54%) for the developed XAI artifacts followed by observation (*n*=31; ≈20%) and expert studies (*n*=24; ≈14 %). Lastly, these contributions focus on the satisfaction (*n*=42; ≈16 %), interpretability (*n*=45; ≈17 %), and trust (*n*=45; ≈17 %) as the *measured constructs*. In contrast, relatively little research is conducted to address decision quality (*n*=18; ≈7 %), explainability (*n*=18; ≈7 %), time (*n*=17; ≈7 %), or effort (*n*=17; ≈7 %).

### 3.6.2 Cluster and Archetype Analysis

Further, to uncover hidden patterns or archetypes in current XAI literature, we conducted a cluster analysis and uncovered archetypes these clusters. We show and explain the results in the following.

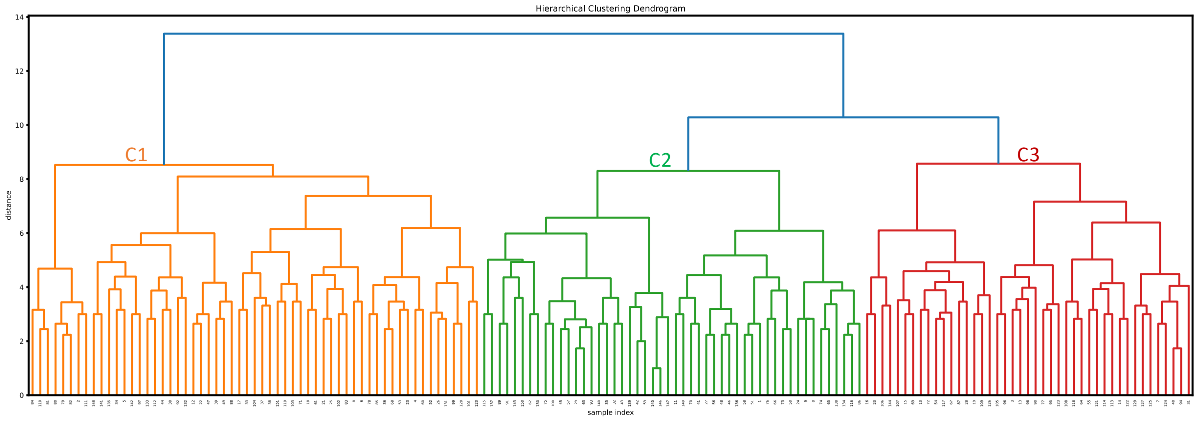


Figure 3.4 Result of Cluster Analysis

| Dimension | Characteristic | *n* | C1 | | C2 | | C3 | | Dimension | Characteristic | *n* | C1 | | C2 | | C3 | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | 59 | | 50 | | 43 | | | | | 59 | | 50 | | 43 | |
| *Purpose* | Validation | 38 | 23 | 0.39 | 3 | 0.06 | 12 | 0.28 | *Incentives* | Monetary | 46 | 6 | 0.10 | 24 | 0.48 | 16 | 0.37 |
| | User perception | 48 | 25 | 0.42 | 14 | 0.28 | 9 | 0.21 | | Non-monetary | 101 | 52 | 0.88 | 22 | 0.44 | 27 | 0.63 |
| | Theory formation | 12 | 9 | 0.15 | 0 | 0 | 3 | 0.07 | *Study Treatment* | No-explanation baseline | 63 | 12 | 0.20 | 23 | 0.46 | 28 | 0.65 |
| | Effect measurement | 58 | 4 | 0.07 | 33 | 0.66 | 21 | 0.49 | | Diff. Explanation baseline | 37 | 10 | 0.17 | 19 | 0.38 | 8 | 0.19 |
| *Focus* | Framework development | 16 | 10 | 0.17 | 1 | 0.02 | 5 | 0.12 | | No differentiation | 47 | 34 | 0.58 | 6 | 0.12 | 7 | 0.16 |
| | Empirical insights | 75 | 21 | 0.36 | 41 | 0.82 | 13 | 0.30 | *Explanation Presentation* | Text explanation | 52 | 15 | 0.25 | 20 | 0.4 | 17 | 0.40 |
| | Method development | 39 | 13 | 0.22 | 5 | 0.1 | 21 | 0.49 | | Visualization | 72 | 28 | 0.47 | 23 | 0.46 | 21 | 0.49 |
| | Prototyp evaluation | 22 | 15 | 0.25 | 3 | 0.06 | 4 | 0.09 | | Explanation by example | 10 | 2 | 0.03 | 1 | 0.02 | 7 | 0.16 |
| *Domain* | Commodity | 8 | 4 | 0.07 | 4 | 0.08 | 0 | 0.00 | | Explanation by simplification | 11 | 4 | 0.07 | 7 | 0.14 | 0 | 0.00 |
| | Manufacturing | 4 | 1 | 0.02 | 0 | 0 | 3 | 0.07 | | Feature relevance explanation | 20 | 11 | 0.19 | 3 | 0.06 | 6 | 0.14 |
| | Services | 76 | 39 | 0.66 | 33 | 0.66 | 4 | 0.09 | *Study Scales* | Rating | 96 | 31 | 0.53 | 36 | 0.72 | 29 | 0.67 |
| | Information | 60 | 13 | 0.22 | 12 | 0.24 | 35 | 0.81 | | Ranking | 12 | 4 | 0.07 | 4 | 0.08 | 4 | 0.09 |
| *Expertise* | (ML) Novice user | 102 | 23 | 0.39 | 43 | 0.86 | 36 | 0.84 | | Dichotomous | 17 | 9 | 0.15 | 5 | 0.1 | 3 | 0.07 |
| | ML experts | 16 | 12 | 0.20 | 0 | 0 | 4 | 0.09 | *Evaluation Approach* | Case study | 14 | 7 | 0.12 | 3 | 0.06 | 4 | 0.09 |
| | Domain expert | 40 | 28 | 0.47 | 6 | 0.12 | 6 | 0.14 | | (Expert) Interviews | 24 | 21 | 0.36 | 0 | 0 | 3 | 0.07 |
| *Sample Size* | < 10 | 16 | 15 | 0.25 | 1 | 0.02 | 0 | 0.00 | | Questionnaire | 95 | 23 | 0.39 | 45 | 0.9 | 27 | 0.63 |
| | 10-50 | 57 | 27 | 0.46 | 7 | 0.14 | 23 | 0.53 | | Group discussion | 3 | 2 | 0.03 | 1 | 0.02 | 0 | 0.00 |
| | 51-100 | 22 | 7 | 0.12 | 7 | 0.14 | 8 | 0.19 | | Observation | 33 | 14 | 0.24 | 3 | 0.06 | 16 | 0.37 |
| | >100 | 49 | 7 | 0.12 | 33 | 0.66 | 9 | 0.21 | *Measured Construct* | Trust | 45 | 19 | 0.32 | 17 | 0.34 | 9 | 0.21 |
| *Data Type* | Sensor | 5 | 2 | 0.03 | 1 | 0.02 | 2 | 0.05 | | Accuracy | 34 | 10 | 0.17 | 7 | 0.14 | 17 | 0.40 |
| | Synthetic | 29 | 9 | 0.15 | 9 | 0.18 | 11 | 0.26 | | Explainability | 18 | 9 | 0.15 | 6 | 0.12 | 3 | 0.07 |
| | Image | 24 | 7 | 0.12 | 6 | 0.12 | 11 | 0.26 | | Effort | 17 | 8 | 0.14 | 4 | 0.08 | 5 | 0.12 |
| | Real world | 63 | 32 | 0.54 | 18 | 0.36 | 13 | 0.30 | | Decision quality | 18 | 7 | 0.12 | 5 | 0.1 | 6 | 0.14 |
| | Social | 34 | 12 | 0.20 | 15 | 0.3 | 7 | 0.16 | | Interpretability | 45 | 17 | 0.29 | 19 | 0.38 | 9 | 0.21 |
| *Study Design* | Single treatment | 35 | 23 | 0.39 | 6 | 0.12 | 6 | 0.14 | | Time | 17 | 6 | 0.10 | 0 | 0 | 11 | 0.26 |
| | Between-group | 65 | 8 | 0.14 | 35 | 0.7 | 22 | 0.51 | | Confidence | 22 | 3 | 0.05 | 6 | 0.12 | 13 | 0.30 |
| | Within-subject | 52 | 28 | 0.47 | 8 | 0.16 | 16 | 0.37 | | Satisfaction | 42 | 20 | 0.34 | 7 | 0.14 | 15 | 0.35 |

Table 3.4 Results of Archetype Analysis

We used the agglomerative hierarchical clustering algorithm from Ward (1963) in combination with the Euclidean metrics to calculate the distance. To find the optimal number of clusters, we analyzed different evaluation metrics (elbow method, silhouette score, and Davies-Bouldin score), resulting in a non-uniform cluster solution. Lastly by following Fischer et al. (2020), we decided on a *n*=3 cluster solution, as we noticed the highest gradient decrease in cluster distance and found it to be a sound foundation for our cluster interpretation. These results are represented as dendrogram in Figure 3.4. In the corresponding Table 3.4 we calculated the absolute and relative occurrence for each characteristic in each cluster. We present the resulting archetypes in the following:

**Cluster C1 - Focused Domain Dialog.** The contributions (*n*=59) in this cluster target ML and domain experts with a small to medium user sample to validate developed frameworks/prototypes or measure user perceptions. Therefore, they use real-world data to evaluate their findings. Since they focus on experts in a more qualitative approach, they merely present their artifact with no differentiation study treatment with a single or within-subject study design, providing no monetary incentives to their participants.

This cluster can be conceptualized as the "focused domain dialog" archetype, which emphasizes the domain-specific nature of XAI applications. It highlights research that is performed in-situ and improves not only the knowledge base of XAI, but it is of direct benefit to the according domains.

In this context, e.g., Ming et al. (2018) and Khodabandehloo et al. (2021) validate their visualization-based prototypes with a small sample size of ML experts. Similarly, Wang, Gou, et al. (2018) and Cabitza et al. (2020) validate their visualization-based framework with a medium size group of domain experts focusing on decision quality. In addition, contributions measure user perception of the constructs of trust (Ghai et al., 2021; Wintersberger et al., 2020) and interpretability (Brennen, 2020; Spinner et al., 2020) with domain or ML experts without any incentives. Lastly, contributions employ a user-centered XAI study in multiple fields of the service domain such as health care (Le et al., 2020; Xie et al., 2020).

**Cluster C2 - Broad Empirical Comparative Study.** Research in this cluster (*n*=50) measures various effects within user-centered studies in a broad audience often surveying more than 100 participants. For their objectives, authors compare different groups of participants using a between-groups study design and often use a differentiation explanation baseline.

Thus, the second cluster can be described as the "broad empirical comparative study" archetype, which focuses on gaining empirical insights from (ML) novice users. It relies on survey instruments to tap into a large audience and receive representative albeit not contextualized results.

Contributions measure the construct of trust (Cai et al., 2019; Dominguez et al., 2020) or interpretability (Fürnkranz et al., 2020) through studies with more than 100 participants within the service domain (Quijano-Sanchez et al., 2017) by providing different types of ML visualizations. As these types of studies often use survey platforms such as Amazon Mechanical

Turk (Poursabzi-Sangdeh et al., 2021) or Prolific (Herm, Wanner, et al., 2021a), they typically provide monetary incentives.

**Cluster C3 - Medium Group Intervention Testing.** In addition to clusters C1 and C2, the contributions (*n*=43) within this cluster examine technological interventions such as new methods in the field of XAI. In these contributions, observations, or questionnaires with a medium to large number of users are used to measure the effects of constructs such as accuracy, satisfaction, confidence, or time primarily in the domain of information, often using a no explanation baseline.

The third cluster can be described as the "medium group intervention testing" archetype, which stresses the significance of XAI applications as interventions that affect user perception and behavior. Although this is the smallest group in our analysis, we noted several types of user-centered XAI studies, as these papers address less researched user-centered XAI topics.

For example, Kenny et al. (2021) test a rarely studied explanation type in the context of image classification through a qualitative analysis. Similarly, Das and Rad (2020) focus on user performance from the application of an XAI-based recommender system. In contrast, Wiegand et al. (2020) investigate initial user needs for decision explanations for autonomous driving systems, while Alipour et al. (2020) measure the correlation between system accuracy and user perception.

## 3.7 Discussion and Conclusion

With our taxonomy, we present a comprehensive framework that can be used to evaluate and classify existing user studies and plan new XAI research. While there is growing body of knowledge in this area, its focus is often on ML performance rather than its explainability, which is important in a socio-technical context. In answering our research questions, we brought structure to the discussion to overcome these shortcomings and we identified initial relations within the dimensions. To do so, we investigated the literature in a descriptive approach as well as a cluster analysis. Thereby, we were able to derive three archetypes in contemporary XAI research. We can use the gained knowledge to shed light on some apparent deficiencies in current research and point to potential future research directions.

In analyzing the different research contributions derived during the literature collection, we noticed several shortcomings of current literature and identified research streams future research should address:

Looking at Table 3.4 it become obvious that XAI research is still in its infancy, as current contributions focus mainly on the measurement constructs of explainability and interpretability. Cognitive effort and decision quality have hardly been investigated. Consequently, it remains vague whether proposed XAI approaches are suitable in real-world conditions rather than in a laboratory experiment. Further, while many studies examine similar constructs, there is no general definition of measurement constructs. For example, articles empirically test the interpretability of XAI frameworks. However, in research, this measurement construct is often also referred to as understandability (Wang, Gou, et al., 2018) or comprehensibility (Herm, Wanner, et al., 2021a). The same is true for explainability (Wanner, Herm, Heinrich, et al., 2021) and transparency Peters et al. (2020), which are often used as synonyms.

Throughout our research, we identified many contributions that aim at breaking down the black-box behavior of ML models in healthcare or autonomous driving. For example, while many researchers describe their research as promising, the lack of consistency with jurisdiction is evident (e.g., Wiegand et al., 2020). Therefore, this highly interdisciplinary topic should aim at involving legal experts to validate these results. This becomes even more critical as many different jurisdictions across various countries need to be addressed (Górski & Ramakrishna, 2021).

As mentioned above, many contributions describe different XAI approaches. Here, it is noticeable that most of them do not follow any design patterns or guidelines (e.g., Khodabandehloo et al., 2021). Since these approaches can be used in challenging scenarios, it is essential that users are not overwhelmed by the complexity of the explanation. Future research should therefore propose validated design patterns that can serve as a starting point for further XAI research (Herm, Wanner, et al., 2021a). Relying on the XAI design jungle, the identified XAI contributions use specific XAI argumentations such as visualizations for their use cases. Often, it is not apparent, why this type of explanation was used (e.g., Kenny et al., 2021). While Mohseni et al. (2021) has introduced an overview of different XAI argumentation types, the holistic comparison and evaluation is still missing.

In summary, these deficiencies emphasize the emerging nature of the field and the need for an overarching discipline-spanning nomenclature and understanding to enable interdisciplinary research. Our taxonomy can be seen as a vehicle of structuring and our cluster analysis provides a first step in this direction. That is, researcher can use our findings, to address different lacks as well as position their research more precisely according to our taxonomy.

In terms of limitations, a taxonomy is never complete and should always be considered as a starting point for further contextualization. We could not delve into each dimension in very detail but focused on interesting relations and connections. Our analysis shows more potentially interesting patterns that need to be explored. In the future, our taxonomy could be contextualized for specific applications or to derive novel best practices for XAI implementations and their evaluations.

# 4 Stop Ordering Machine Learning Algorithms by their Explainability! A User-Centered Investigation of Performance and Explainability

*Lukas-Valentin Herm, Kai Heinrich, Jonas Wanner, and Christian Janiesch*

**Abstract.** Machine learning algorithms enable advanced decision making in contemporary intelligent systems. Research indicates that there is a tradeoff between their model performance and explainability. Machine learning models with higher performance are often based on more complex algorithms and therefore lack explainability and vice versa. However, there is little to no empirical evidence of this tradeoff from an end user perspective. We aim to provide empirical evidence by conducting two user experiments. Using two distinct datasets, we first measure the tradeoff for five common classes of machine learning algorithms. Second, we address the problem of end user perceptions of explainable artificial intelligence augmentations aimed at increasing the understanding of the decision logic of high-performing complex models. Our results diverge from the widespread assumption of a tradeoff curve and indicate that the tradeoff between model performance and explainability is much less gradual in the end user's perception. This is a stark contrast to assumed inherent model interpretability. Further, we found the tradeoff to be situational for example due to data complexity. Results of our second experiment show that while explainable artificial intelligence augmentations can be used to increase explainability, the type of explanation plays an essential role in end user perception.

**Keywords.** Machine Learning, Explainability, Performance, Tradeoff, XAI

**Reference.** Herm, L.-V., Heinrich, K., Wanner, J., Janiesch, C. (2023). Stop Ordering Machine Learning Algorithms by their Explainability! A User-Centered Investigation of Performance and Explainability. *International Journal of Information Management*, *69*, 102538. https://doi.org/10.1016/j.ijinfomgt.2022.102538

## 4.1 Introduction

Artificial intelligence (AI) technology enables human-like cognitive capacity for advanced decision making in contemporary intelligent systems (Dwivedi et al., 2021; Hradecky et al., 2022). Despite these advances, research shows that human perceptions can trigger behavioral responses that affect an organization's capability to leverage such systems when predictions are presented to end users (Berger et al., 2021; Chiu et al., 2021). As certain algorithms exhibit a higher degree of explainability through their inherent interpretability, end users may perceive them more benevolently than others (Shin, 2021; Wanner, Herm, et al., 2022b). In contrast, much of the current AI research focuses solely on the statistical performance measures of machine learning (ML) models (Collins et al., 2021; La Cava et al., 2019), and data competitions are dominated by deep neural network algorithms outperforming shallow ML algorithms (e.g., Hyndman, 2020; Rudin & Radin, 2019).

The processing of these deep neural network algorithms is based on complex calculation logic, which is practically untraceable. While the decision-making is documented in the learned model and could be traced, its inner complexity renders it unfeasible for humans to do so and interpret its decision-making process or even actual prediction results. It practically renders the model a black box as it does not provide any explanations for its predictions (Dwivedi et al., 2021; Loyola-Gonzalez, 2019). This results in a tradeoff between performance and explainability, which is not yet sufficiently understood from a user-centered perspective.

The performance of an algorithm can be measured by indicators such as accuracy, precision, recall, or the F-score. Yet, it remains unclear which ML algorithm's inherent interpretability is perceived as more explainable by end users. Even more so, explanations can be provided in several ways. Currently, it is unclear which types of explanation humans perceive more benevolently (Shin et al., 2020; von Eschenbach, 2021).

This is crucial as the perceived explainability of a prediction determines the effectiveness of an intelligent system: if human decision-makers can interpret the behavior of an underlying ML model, they are more willing to act based on it, especially in cases where the predictions do not conform to their own expectations (Berger et al., 2021; Ribeiro et al., 2016b). Even more so, intelligent systems without sufficient explainability may even be inefficacious if end users disregard their advice (Shin et al., 2020; Wanner, Popp, et al., 2021).

In scholarly literature, several theoretical considerations on the tradeoff of performance and explainability exist (e.g., Angelov & Soares, 2019; Arrieta et al., 2020; Dam et al., 2018; Gunning, 2019; James et al., 2013; Nanayakkara et al., 2018; Yang & Bang, 2019). Yet, they are of theoretic nature, and an empirical investigation of model explainability is missing. We intend to close this gap with our first research question:

***RQ1:*** *How do common classes of machine learning algorithms compare empirically in the tradeoff between their performance as measured by model accuracy and their explainability as perceived by end users?*

While we cannot increase the performance of individual ML models without modification of the actual analytics process (e.g., data preparation, hyperparameter tuning, etc.), any ML model's predictions can be augmented with external explanations. This is especially important for high-performing algorithms based on deep learning as they offer no inherent explainability to end users (Arrieta et al., 2020; Sharma et al., 2021). In response, several types of explainable AI (XAI) augmentations have been developed. Their visualizations can be grouped into several common types of explanations (Guidotti et al., 2018; Mohseni et al., 2021). There is little to no empirical evidence on actual end user perception when considering their role in intelligent system use (Hoffman et al., 2013; Hoffman et al., 2018). We formulate our second research question accordingly:

***RQ2:*** *How do common types of explanations compare empirically in their explainability as perceived by end users?*

Our insights have a high potential to explain better AI adoption of different classes of ML algorithms, contributing to a better understanding of AI decision-making and the future of work. On the one hand, the results can help us to understand to what extent various classes of ML algorithms differ in their perceived explainability from an end user perspective. It allows us to draw conclusions about their future improvement and their suitability for a given situation in practice. On the other hand, the results can help us understand how much performance end users may be willing to forfeit in favor of explainability. Ultimately, Rudin (2019)'s call to avoid explaining black-box models in favor of using inherently interpretable white-box models could be better approached if the tradeoff was sufficiently understood from a social-technical, end user perspective (Arrieta et al., 2020; Herm, Wanner, et al., 2021a).

The remainder of the paper is structured as follows: Section 4.2 introduces related work on ML algorithms and model explainability. In Section 4.3, we discuss the hypothesis development for our research questions, and in Section 4.4, we introduce our methodology as well as outline datasets and algorithms, implementation details, and survey design. In Section 4.5, we present the results of the empirical study before we discuss them in Section 4.6 and present implications. The paper closes with a brief summary.

## 4.2 Literature Review

### 4.2.1 Machine Learning Algorithms

ML focuses on algorithms that improve their performance through experience. They are able to find non-linear relationships and patterns in datasets without being explicitly programmed to do so (Bishop, 2006; Russell & Norvig, 2021). The process of analytical modeling building to turn ML algorithms into concrete ML models for the use in intelligent systems is a four-step process comprising data input, feature extraction, model building, and model assessment (Goodfellow et al., 2016; Janiesch, Zschech, et al., 2021).

ML algorithms are commonly grouped into shallow and deep ML algorithms. Each ML algorithm has different strengths and weaknesses regarding its ability to process data. Shallow

ML algorithms generally require the feature selection of relevant attributes for model training. This task can be non-trivial and time-consuming if the dataset is high-dimensional or if the context is not well-known to the model engineer. Common classes of shallow ML algorithms are linear regressions, decision trees, and support vector machines (SVM). Deep neural networks with multiple hidden layers and advanced neurons for automatic representation learning provide a computation- and data-intensive alternative (Janiesch, Zschech, et al., 2021; Mahesh, 2020). They master feature selection on increasingly complex data by themselves (LeCun et al., 2015; Schmidhuber, 2015). The performance of these deep-learning-based models surpasses shallow ML models and even exhibits super-human performance in specific applications such as data-driven maintenance (e.g., Wang, Ma, et al., 2018) or medical image classification (McKinney et al., 2020). On the downside, the resulting models have a nested, non-linear structure, which is not per se interpretable for humans, and thus their predictions are difficult to retrace. In summary, many shallow ML algorithms are considered interpretable and, thus, white boxes, but deep learning algorithms tend to perform better but are non-transparent and, thus, black boxes (e.g., Adadi & Berrada, 2018; Wanner, Herm, et al., 2022b).

### 4.2.2 Interpretability and Explainability in Machine Learning

In this context, interpretability signifies how accurately an ML model can associate cause and effect. It is an inherent, data-driven property that is related to the ML model's ability to provide meaning in understandable terms to a human by itself (Fürnkranz et al., 2020; Rudin, 2019).

In turn, explanations have the ability to fill the information gap between the intelligent system and its end user similar to the situation in the principal-agent problem (Arrieta et al., 2020; Baird & Maruping, 2021) whenever the ML model is non-transparent and therefore not sufficiently interpretable. Explanations are decisive for the efficacy of the intelligent system as the end user decides based on the given information whether he or she integrates the prediction into his or her decision-making or not (Shin, 2021; Thiebes et al., 2021).

The question of what constitutes explainability and how explanations should be presented to be of value to human users fuels an interdisciplinary research field, that consists of various disciplines, including philosophy, social science, psychology, computer science, and information systems (Collins et al., 2021; Miller, 2019). From a socio-technical perspective, explainability can be considered as the perceived quality of an explanation by an individual or user group (Adadi & Berrada, 2018; van der Waa et al., 2021). While the perceived quality can be circumstantial, we assume that there is a shared perception across user responses that can be used to explain at least part of the judgement.

From a technical point of view, explainability in intelligent systems is about two questions: the "how" question and the "why" question. The former is about global explainability, which provides answers to the ML algorithm's internal processing (Dam et al., 2018; Rudin, 2019). The latter is about local explainability, which answers the ex-post reasoning about a concrete prediction by an ML model (Arrieta et al., 2020; Dam et al., 2018). In this context, as noted above, many shallow ML models are considered to be white boxes that are interpretable per se

(Arrieta et al., 2020; Janiesch, Zschech, et al., 2021). In contrast, a black-box ML model is either too complex for humans to understand or opaque for a reason and, therefore, equally hard to understand (Dwivedi et al., 2021; Rudin, 2019). Consequently, we consider an ML model's innate interpretability as its explainability towards end users not using any XAI augmentations (Adadi & Berrada, 2018; Kenny et al., 2021).

Theoretical contributions typically assume a continuous decrease in explainability with increasing performance of ML models. While it is generally depicted as a linear or cubic curve, it is not apparent whether this relation of explainability and performance is consistent across different ML models in a socio-technical evaluation with end users.

Likewise, there are numerous ways explanations can be presented with XAI. These augmentations can be summarized in six common explanation types (Mohseni et al., 2021).

*How* explanations represent a global view of the ML algorithm; common XAI augmentations display decision boundaries or model graphs. *Why e*xplanations represent a local view and describe why a prediction was made based on a singular input, demonstrating the importance of input variables for the decision of the model. Contrastive visualizations can be used to produce *Why-Not* explanations that outline the difference between an actual and the expected prediction. Furthermore, the algorithms' reaction to change in data or algorithmic hyperparameters can be outlined by *What-If* explanations. In a similar way, *How-To* or counterfactual explanations provide an interactive user experience, where the input of the model is changed in a way so that the output changes. Lastly, *What-Else* explanations offer explanations by example in providing training data that generate similar outputs from the model.

Aside from the type of explanation, it is essential to distinguish the target audience of the explanation. Research distinguishes four different groups: developers, theorists, ethicists, and users (Mohseni et al., 2021; Preece et al., 2018). As empirical findings differ for the stakeholder groups, we solely aim our study at the (end) user (Herm, Wanner, et al., 2021a; Meske et al., 2022). In our case, end users are domain experts that use an intelligent system in their work routines to obtain predictions that assist their decision making. They do not participate directly in the system's planning, engineering, maintenance, or support and typically do not possess technical knowledge about its analytical model.

## 4.3 Theoretical Background and Hypotheses Development

### 4.3.1 Machine Learning Tradeoffs

Considerations about the (hypothesized) tradeoff between model performance and model explainability have been the subject of discussion for some time. Originating from theoretical statistics, a distinction for different ML algorithms was first made between model interpretability and flexibility (James et al., 2013). More recently, this changed towards a comparison between model accuracy and interpretability (e.g., Arrieta et al., 2020; Yang & Bang, 2019) or algorithmic accuracy and explainability (Dam et al., 2018; Rudin, 2019). All tradeoffs address the same compromise of an algorithm's performance versus the algorithm's degree of result traceability.

Many subjective classifications of this tradeoff exist (Angelov & Soares, 2019; Arrieta et al., 2020; Dam et al., 2018; Gunning, 2019; Nanayakkara et al., 2018; Vempala & Russo, 2018; Yang & Bang, 2019). Overall, there is high conformity between the subjective classifications of the different authors. We synthesized these schemes into a generalized classification scheme.[1] The resulting Cartesian coordinate system shows five common classes of ML algorithms ordered by their assumed performance (*y*-axis) and explainability (*x*-axis).

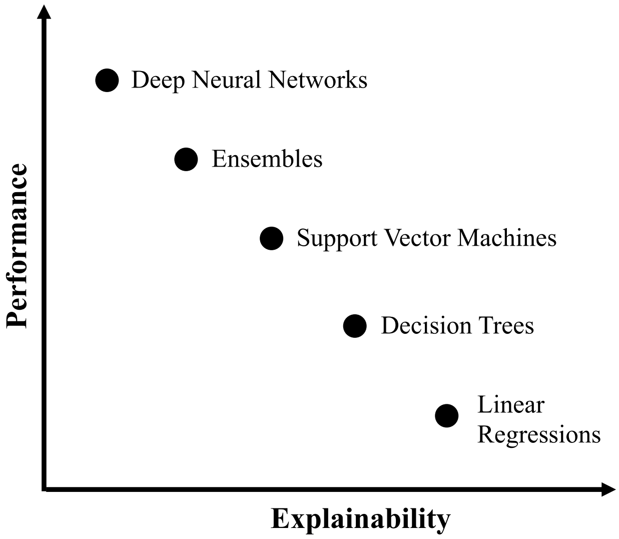


Figure 4.1 Synthesis of Common ML Algorithm Classification Schemes

There is a general agreement on key classes of ML algorithms, but there are some differences in their placement and the granularity of representation. The general notion is that with a loss of performance, algorithms provide better explainability so that algorithms can be ordered on some curve. Hence, deep neural networks are categorized as the most powerful algorithms with the least degree of explainability, followed by ensemble algorithms, which consist of multiple ML models. SVMs serve as a large margin classifier based on data point vectors and come third in performance, superior to decision trees that use sorted, aligned trees for the development of decision rules. Finally, linear regressions (or linear models in general) are considered least in performance yet straightforward to interpret (Goodfellow et al., 2016; James et al., 2013). Note certain classes of ML algorithms are thought to perform closer to each other than the conceptual equidistant visualization of Figure 4.1 (e.g., Dam et al., 2018; Gunning, 2019; Guo et al., 2019).

In essence, these theoretical classification schemes represent a hypothetical and data-centered view on the tradeoff of model accuracy vs. model interpretability. They have neither yet been validated for specific applications based on real data nor confirmed by including end users to

[1] See Appendix C.1 for an enlarged version of the figure, including the different classification schemes from literature.

unearth their true pertinency towards said tradeoff between performance vs. explainability. An empirical quantification of end user explainability is necessary to provide first-hand knowledge to the engineers of intelligent systems for the development of intelligent systems (Jauernig et al., 2022; Meske & Bunde, 2022). Despite this apparent deficiency, they are commonly referenced as a motivation for a user- or organization-centered XAI research or intelligent system deployment (e.g., Asatiani et al., 2021; Guo et al., 2019; Rudin, 2019).

Augmented models (i.e., ex-post explainers) were the subject of several of those studies (e.g., Angelov & Soares, 2019; Nanayakkara et al., 2018). We have not included them in this synthesis as our focus was on the ML algorithms' inherent explainability. Yet, as noted above, XAI augmentations aim to provide more transparent ML models with both high performance and high explanatory power to improve the acceptance of predictions by end users (Arrieta et al., 2020; Gunning, 2019). Hence, it is self-evident that we need to consider how XAI augmentations can improve the explainability of the highest-performing - supposedly least explainable - ML algorithm.

There are human-centered evaluations of XAI algorithms. However, most of the evaluations revolve around testing certain novel algorithms. They often focus on variable importance and semi-automated evaluations by perturbating features to identify the most influential features and compare the result set with the XAI algorithms' explanations (Doshi-Velez & Kim, 2017; Nguyen, 2018). Thus, they measure the algorithms' explanatory performance towards a truth value rather than their explanatory quality towards end users.

In summary, it remains unclear how end users perceive explainability and how this is in line with the considerations presented above. We propose to approach this as follows: First, we focus on the tradeoff between performance and an ML model's inherent explainability to avoid biases introduced by model transfer techniques from the field of XAI. Second, we focus on five of Mohseni et al. (2021)'s six common types of explanation to augment the best-performing ML model to uncover which types of XAI explanations end users prefer independently of their potential to correctly explain an ML model or its predictions.

### 4.3.2 Hypotheses

**Performance vs. Explainability Tradeoff.** Our research design uses a simple group structure, where the independent variable is the *choice of algorithm,* and the dependent variables are *performance* and *perceived goodness of explanation*. The independent variable reflects the nature of an algorithm as it is applied to practical problems. The dependent variable performance measures the objective performance of the algorithm. The perceived goodness of explanation is more subjective and we base the choice of this second dependent variable on the proposed tradeoff that requires a quantification of explanation as it is perceived by users and knowingly can influence the user's mindset towards algorithms (Berger et al., 2021; Jauernig et al., 2022). The moderating group variable *data complexity* is expressed through different cases using different datasets reflecting *low complexity* and *high complexity*. We choose to introduce this variable to also reflect on more complex practical problems that involve large, non-tabular datasets like image and video data. These complex datasets are massive in size, high-dimensional, possibly biased, and not straightforward to explore by the human user. This also combats the predisposition that it is always viable to choose a simple algorithm that is explainable when clearly in those complex cases these types of algorithms are not interpretable as the data is neither (Castiglioni et al., 2021; Wang et al., 2021). When it comes to using algorithms for complex cases, post hoc XAI explanations can be used to provide insights into the decision-making process (Meske & Bunde, 2020; van der Waa et al., 2021) as they have an increased need for explainability (Lebovitz et al., 2021; Liu et al., 2007).

In terms of performance, we hypothesize that for less complex cases, which use tabular data, the performance of the ML models will be very close and not significantly distinguishable (Rudin & Radin, 2019; Zhang & Ling, 2018). Recent replication studies show that prediction scenarios using tabular data can be solved with small-scale models and will hold similar performance or even outperform the more advanced models for those low complexity cases (DeVries et al., 2018; Mignan & Broccardo, 2019). Thus, we expect no significant difference in performance following the ordering of Figure 4.1.

*H1a: The choice of algorithm has no significant impact on the performance for cases with low complexity.*

Contrary, since shallow learning algorithms are limited in their way of extracting higher-level features for complex data, we expect the performance to deviate for complex cases (Janiesch, Zschech, et al., 2021; LeCun et al., 2015). Related research shows a decline in error rates for deep neural networks when applied to image datasets that even outperform human judgment (Heinrich et al., 2019; McKinney et al., 2020). Hence, we theorize that the performance of the shallow ML algorithms will be sub-par to deep neural networks and less grouped since they will fall off at different paces.

*H1b: The choice of algorithm has a significant impact on the performance for cases with high complexity.*

As a next step, we introduce the hypotheses regarding the goodness of explanations of common classes of ML algorithms. While we expect the black-box deep learning model to be the poorest in explainability (Meske et al., 2022; Rudin, 2019), we can only offer some thoughts on the ordering of shallow ML algorithms. First, we believe that the design of the explanation plays an important role in conveying the intended level of transparency (Miller, 2019; Shin, 2021). Shallow ML algorithm classes that have intrinsic means of local interpretability, such as SVM and linear regression in terms of their input feature weights, still have no natural way of visually presenting local variable importance out-of-the-box. The only exception are decision trees that present a logical structure, which is in line with the human thought process (Herm, Wanner, et al., 2021a; Subramanian et al., 1992). Thus, we believe that contrary to the existing theories, only decision tree explainability will be distinguishable from the rest of the ML algorithms.

*H2a: The choice of algorithm has a significant influence on the perceived goodness of explanation for cases with low complexity.*

Following our argumentation, we believe that post hoc analysis will reveal that the significant overall difference can only be attributed to the difference between the decision tree and the other groups of algorithms for less complex cases.

Further, we believe this will not hold for the less complex case. Complex data structures like images, even when referred to by a tree structure, have no convincing explanatory value since there are so many input variables (in the image case: pixels) to choose from (Chandra & Bedi, 2021; Heinrich et al., 2019). For this type of data, an additional step is required to produce high-level features that humans can relate to, such as specific geometric forms.

*H2b: The choice of algorithm has no significant impact on the perceived goodness of explanation for cases with high complexity.*

**Perception of Explanation Types.** For our second research question, we follow up on the first experiment to investigate the question of how the high-performing but black-box class of deep learning algorithms should be augmented with XAI explanation types. We aim to find out which type of explanation augmentation will help to elevate the transparency of deep learning models in the case of complex data. Thus, as independent treatment variables, we use the *type of explanation* (Mohseni et al., 2021). As a dependent variable, we again use the *perceived goodness of explanation*. Since deep learning algorithms come with no explanation at all, we theorize that any explanation will be favorable or equal to no explanation (Adadi & Berrada, 2018; Miller, 2019). Furthermore, we suspect that explanation types with straightforward, non-complex visualizations (e.g., *Why*) will be perceived more benevolently by end users. Thus, we formulate the following hypothesis:

*H3: The type of explanation has a significant impact on the perceived goodness of explanation.*

We believe that the distinction between more local-oriented (*Why*, *Why-Not*, and *What-Else*) and global-oriented types (*How* and *How-To*) will be notable. We grounded this expectation in our focus on end users. Developers and theorists need to understand the nature of the algorithm. End users are satisfied with simpler and more targeted example-based explanations of predictions that rationalize their belief in the system and assist in solving the task at hand (Miller, 2019; Preece et al., 2018). Figure 4.2 summarizes the research models for RQ1 and RQ2. Note that both experiments stand alone, and that experiment 2 has a set choice of algorithm (deep neural network) and dataset (high complexity) and uses post hoc XAI augmentations rather than presenting the non-augmented, inherent explanations as in RQ1.

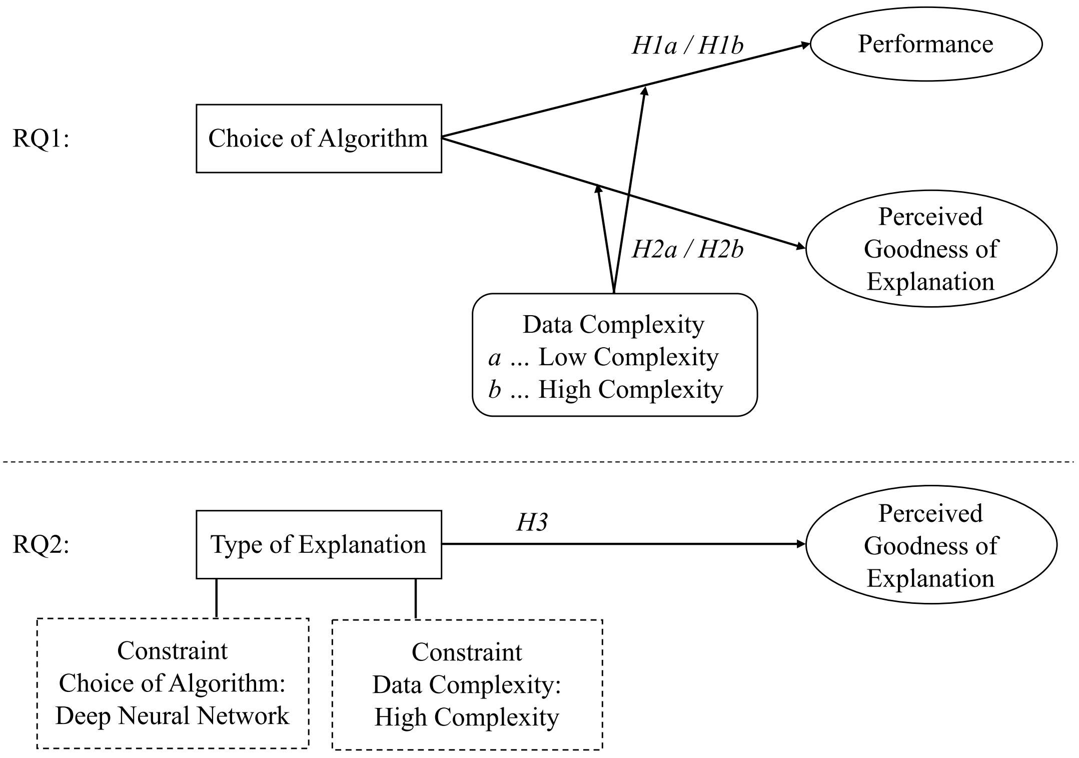

RQ1:
Choice of Algorithm
H1a / H1b
Performance
H2a / H2b
Perceived Goodness of Explanation
Data Complexity
a ... Low Complexity
b ... High Complexity
RQ2:
Type of Explanation
H3
Perceived Goodness of Explanation
Constraint Choice of Algorithm: Deep Neural Network
Constraint Data Complexity: High Complexity

Figure 4.2 Proposed Research Models for RQ1 and RQ2

## 4.4 Methodology

**Performance vs. Explainability Tradeoff.** To execute our research design for RQ1, we use a standard ML analysis process and subsequently conduct an empirical analysis (Müller et al., 2017). In our experiment, low data complexity is represented by a standard tabular dataset with a moderate number of observations and low dimensionality. High data complexity is represented by a large image dataset that exhibits high dimensionality and initial non-tabular form (i.e., pixel-tensors). See Table 4.1 for details on the datasets. Both datasets represent classification tasks.

| Dataset | Moderator Effect | Description |
|---|---|---|
| HEART (Janosi et al., 1988) | Low complexity | The heart disease dataset (HEART) is a low complexity dataset, which is used to classify the presence of heart disease based on medical patient features. It contains 303 observations of 13 different features and a binary target variable. |
| BRAIN (Bohaju, 2020) | High complexity | The brain tumor dataset (BRAIN) contains images of brain MRIs of which 2.079 depict no brain tumor and 1.683 depict a brain tumor. The images have dimensions of 224x224 pixels each. A binary label indicates the tumor status. |

Table 4.1 Overview of Datasets and Moderating Effects

For each dataset, we apply several classes of ML algorithms that represent the levels of the independent treatment variable *choice of algorithm* (see Table 4.2). We applied data preprocessing and grid search optimization for every ML algorithm. We ensured that each algorithm provides an acceptable answer for each case so that the class of algorithms can - in principle - be considered fully interchangeable from an end user perspective except for their performance and explanation.

| **Class of ML Algorithm** | **Implementation** |
|---|---|
| Linear Regression | Due to data preprocessing, we skipped default normalization and used the default settings. For the non-centered datasets, we included the intercept of the model. |
| Decision Tree | We did not restrict the models by regulations such as the minimum sample split numbers of the estimators. The resulting trees have a depth of five or six, depending on the treatment. |
| SVM | For all datasets, we applied an SVM using a radial basis function as kernels. |
| Random Forest (Ensemble) | We used the bagging algorithm random forest as a proxy for ensembles. Random forests consist of 100 estimators each, and their complexity was not restricted (see decision tree). |
| Deep Neural Network | For HEART, we used a multi-layer-perceptron with eight hidden layers, including dropout layers. For BRAIN, we used a convolutional neural network consisting of 13 hidden layers also including maxpooling and dense layers. |

Table 4.2 Overview of ML Algorithm Implementations

After implementation and execution, we measure the dependent variable *performance* by measuring the accuracy of the classification by applying 15-fold cross-validation for each ML algorithm to ensure algorithm behavior in terms of reliability and possible variations. The accuracy of the model refers to the system’s ability to correctly predict an outcome and is given by the ratio of correctly classified entities to all entities. The measure allows us to objectively estimate the performance without including perceived advice quality that can be biased by user perception (Janiesch, Zschech, et al., 2021; Mahmud et al., 2022). In addition, we reviewed recall and F-score to ensure that a single performance metric did not produce outliers.

While a model’s performance can be evaluated independently of the user, its explainability depends on the perceptions of its users (Miller, 2019; Shin, 2021). Therefore, we measured the dependent variable, *perceived goodness of explanation,* in a survey. We used the platform prolific.co providing a monetary incentive. Since both our cases are healthcare cases and we aim our analysis at end users, we assume that the users of such systems would be healthcare professionals. Therefore, we used the filtering functionality of the platform to narrow down subjects to this group. Furthermore, to ensure a basic acceptance of AI among the group, we selected novice healthcare professionals (i.e., enrolled medical students) since experienced healthcare professionals can have a substantial bias to medical AI applications (Logg et al., 2019; Strohm et al., 2020). This also ensures the continuance of the results as those novices constitute the core of the future workforce. We find the notion of catering to as many groups as possible intriguing, but it is out of the scope of this research to consider technology acceptance as a factor (Straub & Burton-Jones, 2007).

For reasons of duration and repetitiveness, we designed two studies, one for each case. The cases were assigned at random. The procedure within each variant was identical. To ensure the validity and reliability of our study, we first asked a senior researcher for a review of our study. Then, we conducted a pre-study to check whether participants had understood the research design and the intended focus of the questions correctly. Furthermore, we asked about any difficulties encountered in completing the survey.

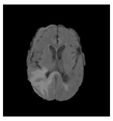
In the survey, we first collected demographics, prior experience with AI, as well as the participant's willingness to take risks. In the second part, we provided them with an introduction to the respective case and the task that the system is carrying out in that context. Third, we informed the participants that they had to put themselves in the situation of a physician who could not delegate the case. Then, we evaluated the participant's perceived goodness of explanation based on the propositions of Hoffman et al. (2018): We provided the participants with a graphical visualization of specific predictions dependent on the algorithm's inherent means to produce such explanations. Thereby, we account for each algorithm's natural way of explanation without adding further augmentations. For each ML model, the participants had to rate their overall perceived goodness of explanation of the model on a seven-point Likert scale. After yielding results for both dependent variables, to check on our hypotheses, we conducted an analysis of variance based on the design in Figure 4.2.

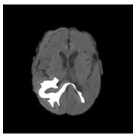
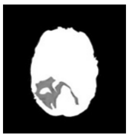
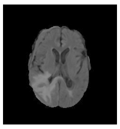
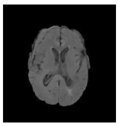
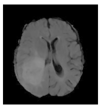
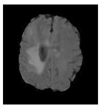
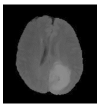
To reduce participant bias, we applied different mechanisms and design elements. First, we randomized the order of treatments within every study to avoid any learning effects or sequence bias. Second, we did not use any colors or explanations from common ML implementation packages, as the participants could be biased through the presentation type (confirmation bias). Third, we only provide input information, an explanation, and a comprehensive description to each explanation, to not force anchoring biases. As an example, the participants did not receive any information about the performance of the ML models to avoid performance bias. Fourth, we assume no focus effects took place, as novice end users (generally) do not have actual prior experience in ML model explanations. Lastly, we applied a validation question as an awareness check. See online Appendix B of (Herm, Heinrich, et al., 2023) for the used study designs to answer RQ1.

**Perception of Explanation Types**. For RQ2, we adopted a similar approach to measure the quality of explanation. As a dataset, we used BRAIN as it represents the more complex dataset and applied only the deep neural network algorithm. Then, we implemented different XAI algorithms that reflect the levels of the independent treatment variable *types of explanation* levels as derived from Mohseni et al. (2021). We measured the *perceived goodness of explanation* analog to RQ1 by conducting a survey. We presented the case and the different explanations as a treatment to the participants. The treatment consists of three boxes: the original input image, the ML algorithm's decision (tumor/no tumor), a form of explanation, and a short textual description of the explanation to ensure a basic understanding. See Appendix C.4 for the study design to answer RQ2.

Table 4.3 comprises the images presented as treatments to the participants for all explanation types. Please note that for our study, we combined Mohseni et al. (2021)'s explanation types *How-To* and *What-If* (in the following: *How-To*) as both focus on hypothetical adjustments to the input to generate what-if scenarios and counterfactual explanations. Due to our focus on end users, the wider scope of *What-If* explanations to include the changing model parameters is not applicable to our analysis. Furthermore, *What-If* explanations are not well suited for high-dimensional data and deep neural networks (Mohseni et al., 2021).

**Input Image**



| Explanation Type | Description of Explanation Types | Treatment Visualization |
|---|---|---|
| *How* | Explanation of which input areas are relevant to the trained model, i.e., the ML algorithm's inner logic.[1] | <br>Result: Tumor |
| *Why* | Explanation of which areas of the given input are relevant to the outcome of the prediction.[1] | <br>Result: Tumor |
| *Why-Not* | Explanation of which areas of the given input are not relevant to the outcome of the prediction.[1] | <br>Result: Tumor |
| *How-To* | Explanation of how hypothetical adjustments of the given input (e.g., the bright shades in the MRI) would result in a (different) model prediction. | <br>Result: Tumor<br>*Input Adjustment:* ⟹<br><br>Result: No Tumor |
| *What-Else* | Explanation by example; showing similar inputs and their respective predictions. | <br>Sample 1 - Result: Tumor<br><br>Sample 2 - Result: Tumor<br><br>Sample 3 - Result: Tumor |

[1] The described XAI augmentations are highlighted as a white area in the treatment visualization.

Table 4.3 Overview of Explanation Type Treatments for RQ2

## 4.5 Results

### 4.5.1 Result Experiment I: Performance vs. Explainability Tradeoff (RQ1)

**Participant Demographics.** In total, we received feedback from $n$=223 subjects (HEART $n$=111; BRAIN $n$=112). To ensure the data quality of our findings, we excluded feedback by applying various preprocessing techniques, such as using control questions, detecting lazy patterns, deleting randomly filled questionnaires, and considering time constraints. This results in a final sample of $n$=100 for HEART and $n$=101 for BRAIN. Table 4.4 shows the demographics for both surveys.

| **Characteristics** | **Attributes** | **HEART**[1] | | **BRAIN**[2] | |
|---|---|---|---|---|---|
| | | *Freq.* | *Percent.* | *Freq.* | *Percent.* |
| Gender | Male | 46 | 46.00 | 49 | 48.85 |
| | Female | 53 | 53.00 | 52 | 51.15 |
| | Others | 1 | 1.00 | - | - |
| Age (years) | <20 | 13 | 13.00 | 20 | 19.80 |
| | 20-30 | 84 | 84.00 | 78 | 77.23 |
| | 31-40 | 2 | 2.00 | 3 | 2.97 |
| | 41-50 | 1 | 1.00 | - | - |
| Location | Africa | 19 | 19.00 | 19 | 18.81 |
| | Europe | 38 | 38.00 | 39 | 38.61 |
| | North America | 32 | 32.00 | 37 | 36.63 |
| | South America | 11 | 11.00 | 6 | 5.94 |
| Experience with AI (years) | None | 44 | 44.00 | 41 | 40.59 |
| | <2 | 28 | 28.00 | 33 | 32.67 |
| | 2-5 | 19 | 19.00 | 19 | 18.81 |
| | 6-10 | 4 | 4.00 | 5 | 4.95 |
| | >10 | 5 | 5.00 | 3 | 2.97 |

[1] $n$=100; [2] $n$=101

Table 4.4 Descriptive Statistics of Subjects for Surveys from Experiment I (RQ1).

**Performance.** In general, the performance results support the theoretical ordering in Figure 4.1 ($y$-axis). Nevertheless, the relative performance and thus the interval of the ordering differs. Especially, the difference between ensemble and SVM is more negligible than assumed. In our case, this may be due to the datasets and the ensemble algorithm. It reveals that the ordering of algorithms by their performance is as assumed in theory, but hardly deterministic.

Further, the performance difference between shallow ML algorithms and deep learning algorithms can be almost neglectable in scenarios with low complexity, such as HEART. Still, linear regression constantly performed worst while the deep neural network performed best. For the more complex case BRAIN, we encounter a strong decline in performance for all models except for the deep neural network. Table 4.5 illustrates the results of our performance evaluation derived through a mean calculation.

| Choice of Algorithm | HEART<br>Mean Accuracy[1, 2] | <br>StdDev Accuracy[2] | BRAIN<br>Mean Accuracy[1, 2] | <br>StdDev Accuracy[2] |
|---|---|---|---|---|
| Linear Regression | 63.43 | 0.09 | 44.47 | 0.03 |
| Decision Tree | 73.86 | **0.06** | 57.98 | 0.03 |
| SVM | 82.34 | 0.09 | 65.66 | 0.08 |
| Random Forest (Ensemble) | 77.52 | 0.08 | 66.04 | **0.02** |
| Deep Neural Network | **84.42** | 0.11 | **89.45** | 0.04 |

[1] higher = better, in %; [2] calculations based on results from 15-fold cross validation

Table 4.5 Descriptive Statistics of Performance for Choice of Algorithm

Using the folds from cross-validation, we executed a one-way analysis of variance (ANOVA) to check hypotheses H1a and H1b, respectively. Table 4.6 shows the results of the ANOVA. The full table with post hoc test results can be found in Appendix C.2.

| Dataset | Variable | Df[1] | Sum Sq[2] | Mean Sq[3] | *F*-value[4] | Pr(>F)[5] |
|---|---|---|---|---|---|---|
| Low complexity (HEART) | Choice of Algorithm | 4 | 0.4074 | 0.10186 | 16.85 | <0.00001 |
| | Residuals | 70 | 0.4230 | 0.00604 | - | - |
| High complexity (BRAIN) | Choice of Algorithm | 4 | 1.4974 | 0.3743 | 165 | <0.00001 |
| | Residuals | 65 | 0.1475 | 0.0023 | - | - |

[1] Degree of freedom; [2] sum squares; [3] mean squares; [4] result *F*-test; [5] result *p*-value

Table 4.6 ANOVA Results for Choice of Algorithm and Performance

Following the ANOVA, we cannot support H1a, but we can support H1b. The post hoc testing (see 0) revealed the hypothesized differences between the two scenarios. In the case of low complexity, we found that linear regression significantly diverted from all other algorithms at the $p<0.01$ level. We found no significant distinction for the other models.

Observing the post hoc result for the complex case BRAIN yields another picture: While we can show the worst-performing role of linear regression in this scenario as well, the decision tree falls off as well and shows significant differences to all other models. Random forest and SVM exhibit nearly similar performance that does not distinguish significantly. In the top end, we find that deep neural network performance is a group of its own with significant distances to all other ML algorithms. In summary, we find that in the complex case BRAIN, the performance differences are more discernible and all ML algorithms except deep neural networks perform notably worse.

**Explainability.** We present the perceived goodness of explanation from the conducted survey for each choice of algorithm in Table 4.7. We followed Boone and Boone (2012) and applied a median calculation for the Likert-type data. As the standard deviations appear normal with no natural anomalies, we applied an ANOVA for the results.

| Choice of Algorithm | HEART<br>Median Explainability* | <br>StdDev Explainability** | BRAIN<br>Median Explainability* | <br>StdDev Explainability** |
|---|---|---|---|---|
| Linear Regression | 4.00 | 1.31 | 3.00 | 1.14 |
| Decision Tree | **6.00** | 1.59 | **4.00** | 1.25 |
| SVM | 3.50 | **1.29** | 3.00 | 1.09 |
| Random Forest (Ensemble) | 5.00 | 1.59 | **4.00** | 1.21 |
| Deep Neural Network | 2.00 | 1.32 | 2.00 | **0.89** |

* Median of seven-point Likert scale: 1,00 = very low; 7,00 = very high; ** standard deviation of seven-point Likert scale

Table 4.7 Descriptive Statistics of Perceived Goodness of Explanation for Choice of Algorithm

Across all treatments, random forests and decision trees achieved the highest or second-highest ratings. We show the results of the ANOVAs for perceived goodness of explanation in Table 4.8, and we can support H2a, but we cannot not H2b.

| Dataset | Variable | Df[1] | Sum Sq[2] | Mean Sq[3] | *F*-value[4] | Pr(>F)[5] |
|---|---|---|---|---|---|---|
| Low complexity (HEART) | Choice of Algorithm | 4 | 486.1 | 121.52 | 59.78 | <0.00001 |
| | Residuals | 70 | 1006.3 | 2.03 | - | - |
| High complexity (BRAIN) | Choice of Algorithm | 4 | 343.7 | 85.92 | 68.28 | <0.00001 |
| | Residuals | 65 | 629.1 | 1.26 | - | - |

[1] Degree of freedom; [2] sum squares; [3] mean squares; [4] result *F*-test; [5] result *p*-value

Table 4.8 ANOVA results for Choice of Algorithm and Perceived Goodness of Explanation

For the low complexity case, we find the expected distribution of ML algorithms with interpretable models being superior in terms of explainability. Hence, the perceived explanation quality of the ML algorithms is significantly distinguishable with some notable exceptions: decision tree and random forest are perceived as similar, presumably due to both being based on tree algorithms and providing tree-structure visualization. In addition, we found that SVM and linear regression are perceived as equal when it comes to explanation goodness.

Surprisingly, for the complex case we find a similar picture. Although the perceived goodness of the decision tree has declined significantly, the perceived goodness between the groups SVM/linear regression and decision tree/random forest is still significantly distinguishable. The post hoc test indicates a strong deviation of the deep neural network from any other algorithm. The reason is straightforward as a deep neural network offers no inherent interpretability. It also shows that decision trees are still perceived as valuable explanations in complex cases.

### 4.5.2 Result Experiment II: Perception of Explanation Types (RQ2)

**Demographics.** For the second experiment, we obtained *n*=109 responses using the high-complexity dataset BRAIN for the perception of different XAI explanation types (in the following, we refer to the sample as BRAIN-XAI) for the deep neural network. To ensure the data quality of our findings, we applied the same preprocessing techniques as in the first survey. The final sample consists of *n*=98. The following table describes the demographics of the survey.

| **Characteristics** | **Attributes** | **BRAIN-XAI**[1] | |
|---|---|---|---|
| | | *Freq.* | *Percent.* |
| Gender | Male | 50 | 51.02 |
| | Female | 48 | 48.98 |
| Age (years) | <20 | 17 | 17.35 |
| | 20-30 | 77 | 78.57 |
| | 31-40 | 4 | 4.08 |
| Location | Africa | 17 | 17.35 |
| | Europe | 41 | 41.84 |
| | North America | 35 | 35.71 |
| | South America | 5 | 5.20 |
| Experience with AI (years) | None | 40 | 40.82 |
| | <2 | 33 | 33.67 |
| | 2-5 | 17 | 17.35 |
| | 6-10 | 4 | 4.08 |
| | >10 | 3 | 3.06 |

[1] $n = 98$

Table 4.9 Descriptive Statistics of Subjects for Survey from Experiment II (RQ2)

**Explainability.** We present the perceived goodness of explanation from the survey for each type of explanation in Table 4.10. Since we used the same survey design as in the first experiment, we also applied a median calculation. Likewise, the standard deviations appear normal with no discernible anomalies.

| **Explanation Type**[1] | **Median Explainability**[2] | **StdDev Explainability**[3] |
|---|---|---|
| Baseline (Black Box) | 2.00 | **1.13** |
| How | 3.00 | 1.94 |
| Why | **6.00** | 1.43 |
| Why-Not | 5.00 | 1.87 |
| How-To | 4.00 | 1.19 |
| What-Else | **6.00** | 1.35 |

[1] All augmentations are based on same CNN model (acc: 89.45 %); [2] median of seven-point Likert scale: 1,00 = very low; 7,00 = very high; [3] standard deviation of seven-point Likert scale

Table 4.10 Descriptive Statistics for Perceived Goodness of Explanation for Type of Explanation

Conducting the respective ANOVA, we can support H3 (see Table 4.11). Looking at the post hoc results (see Appendix C.4) in alignment with the descriptive statistics from Table 4.9, we find that there is a significant difference between no explanation and some sort of explanation, no matter what the type. The local explanations of *Why*, *What-Else*, *Why-Not* scored best, while the global explanation of *How* scored worst aside from the *baseline* of no explanation. Furthermore, we only find *Why-Not* and *How-To* explanations not significantly distinguishable.

| Dataset | Variable | Df[1] | Sum Sq[2] | Mean Sq[3] | *F*-value[4] | Pr(>F)[5] |
|---|---|---|---|---|---|---|
| High complexity (BRAIN-XAI) | Explanation Type | 5 | 960.5 | 192.11 | 82.94 | <0.00001 |
| | Residuals | 582 | 01348.0 | 2.32 | - | - |

[1] Degree of freedom; [2] sum squares; [3] mean squares; [4] result *F*-test; [5] result *p*-value

Table 4.11 ANOVA results for Type of Explanation and Perceived Goodness of Explanation

## 4.6 Discussion and Implications

### 4.6.1 Discussion

As the baseline for the discussion and the generalization of our findings to analyze the tradeoff between performance and explainability (RQ1), we have merged all data from the first experiment. We normalized the data to the range of 0 to 1 to allow for a relative comparison of the ML algorithms regarding the different use cases. We transferred our findings into a Cartesian coordinate system as in Figure 4.1 to visualize our results next to the theoretical assumption. We used mean values to yield a position for each algorithm. Figure 4.3 shows the resulting averaged scheme calculated from the data in Table 4.5 and Table 4.7. It mirrors the results of (Wanner, Herm, Heinrich, et al., 2021).

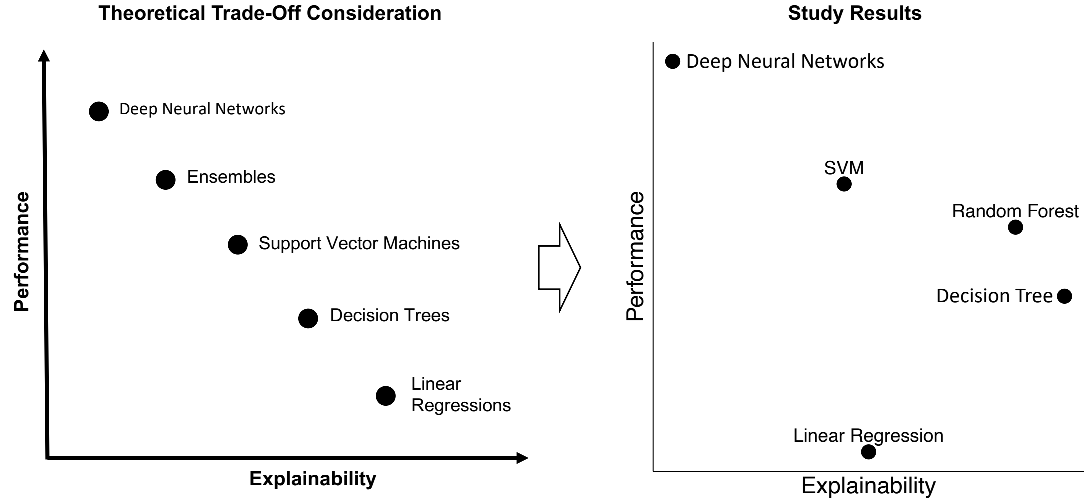


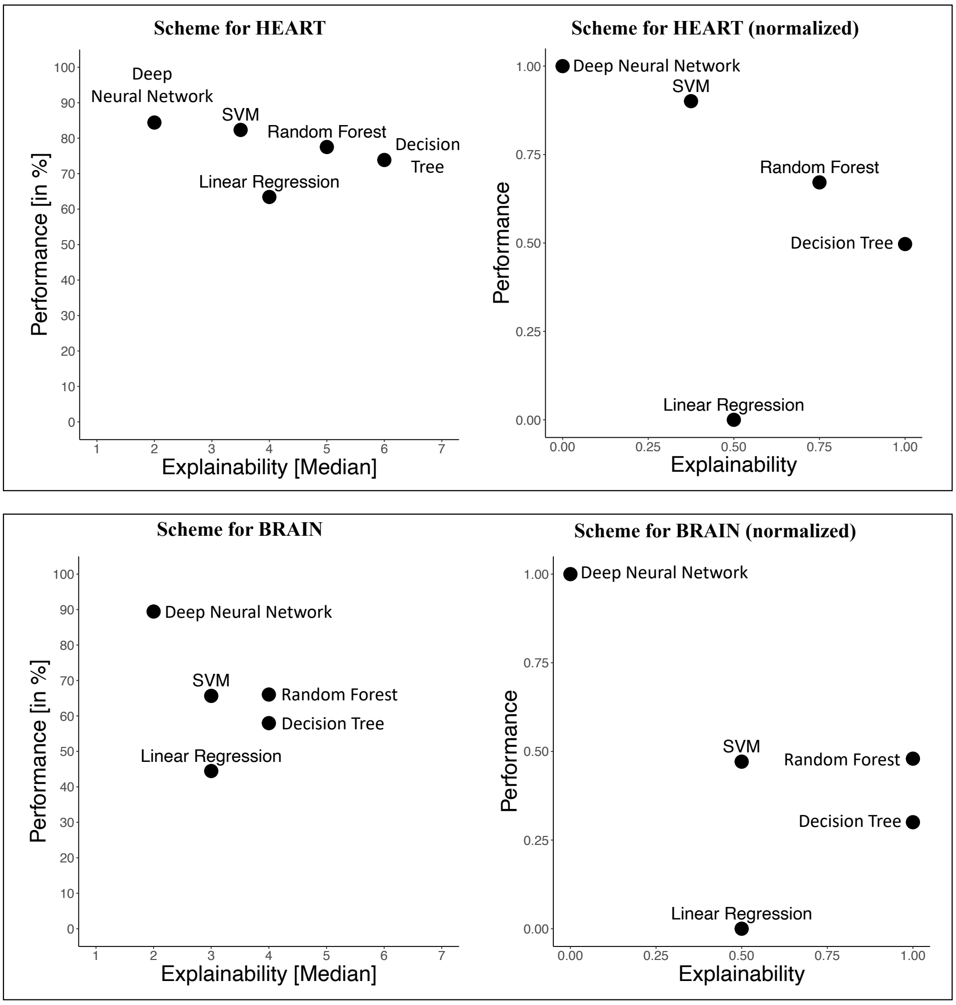
Figure 4.3 Theoretical vs. Empirical Scheme for the Tradeoff of Performance vs. Perceived Explainability in Machine Learning

We can support some tendencies mostly concerning ML model performance as reflected by accuracy. A few things are notably different from the theoretical proposition. In particular, the hypothetical curve between ML model performance and ML model explainability assumed by prior research (left) does not hold in our user-centered treatments (right).

As a result, our empirical evidence shows a grouped structure and challenges the assumption of a tradeoff curve. It is visible in both treatments in terms of explainability. We find that the tree-based models, such as decision trees and random forests, are perceived to provide the best explainability of the five ML models from an end user's perspective. While random forests fall into the ensemble class, the base class model for the ensemble is the decision tree. This explains the random forest's comparably high scores despite being an ensemble algorithm. Contrary to our expectations, we could not substantiate that a single decision tree is perceived as substantially more explainable than a random forest with many unbalanced decision trees. We assume that this may be the case since we did not present all resulting trees of the random forest to the participants for review.

This observation provides new knowledge about the perceived explainability of ML algorithms and renders a more realistic picture of the performance vs. explainability tradeoff than the predominantly theoretical discussion considered the state-of-the-art (e.g., Arrieta et al., 2020; Gunning, 2019; Rudin, 2019).

Figure 4.4 shows the non-normalized and normalized results for the two cases.



Scheme for HEART

Scheme for HEART (normalized)

Scheme for BRAIN

Scheme for BRAIN (normalized)

Figure 4.4 Non-normalized and Normalized Empirical Schemes for the Tradeoff Between Performance and Explainability for Both Cases (HEART and BRAIN)

Continuing with the non-normalized schemes, we can also see that the shallow ML algorithms are positioned relatively close together in terms of their performance. For low complexity datasets, performance gains of deep learning over shallow ML are neglectable. Performance only becomes a factor for high complexity datasets. Hence, as expected, the performance distance between all algorithms widens with increasing dataset complexity and, thus, the choice of algorithm has a larger impact on an intelligent system's performance. Notably, we can also see that the absolute perceived explainability of tree-based algorithms wanes in the non-normalized schema with increasing dataset complexity, and the overall explainability distance decreases.

Consolidating both axes, we find the tradeoff to be dependent on the complexity of the case as well as on real-world performance and explainability requirements. As performance behaves differently than explainability, the tradeoff is non-trivial and consequently a multi-criteria decision (Gunning, 2019; Meske et al., 2022; Wanner, Heinrich, et al., 2020). We provide further evidence of this cause.

Since XAI augmentations can be used to provide post hoc explanations of predictions, the tradeoff becomes even more complex as it becomes evident that - at least for certain applications - the use of high-performing deep learning algorithms may become an option despite their lack of explainability. To investigate how one can best augment these algorithms for end users, we implemented five common types of explanation in a subsequent survey (RQ2).

In Figure 4.5, we summarize the perceived goodness of explanation regarding the explanation types on the employed Likert scale.

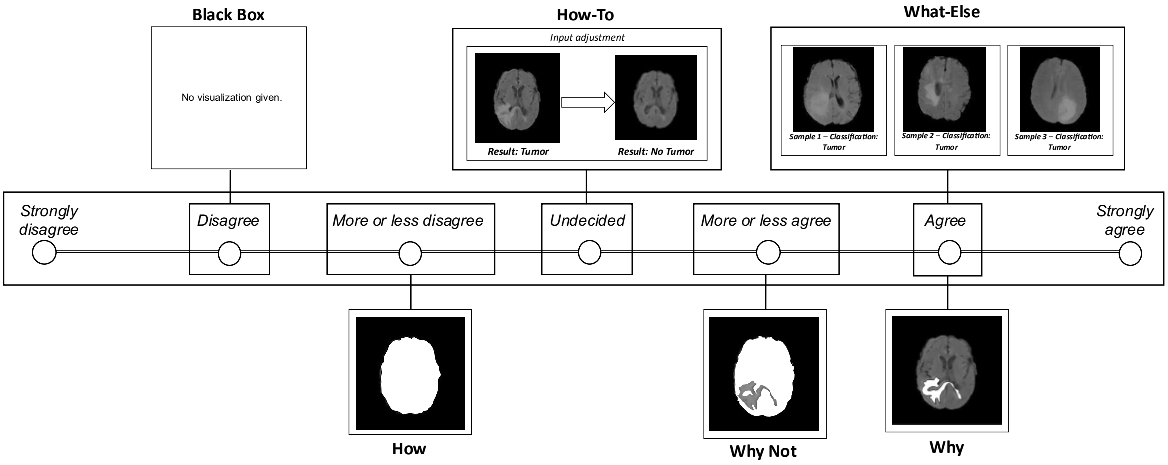


Figure 4.5 Average End User Ratings of Explanation Type Visualized on Likert Scale

We can clearly see the local explanations such as *Why* and *Why-Not* are superior in comparison to the global explanation type *How*. Pointing out the tumor in a direct manner seems to resonate most with the end user as it supports their perception of the tumor's size and location. Even though similar, *Why-Not* explanations that require higher cognitive effort received lower scores than *Why* explanations. A similar argument can be made for *How-to* explanations as those explanations reference a decision to certain input changes to convey an understanding of the decision behavior. Surprisingly, the design of the *How-To* explanation seems to be perceived more benevolently than the *How* explanation. Lastly, *What-Else* explanations are well-received and show that it is beneficial to provide examples for the user to check whether the decision behavior of the algorithm is in line with his or her personal expectancy.

Across all types, we found that end users prefer local explanations that explain the result of a prediction either by pointing to a reason (*Why*) or giving examples (*What-Else*). Inversely, end users tend to disdain global explanations that merely visualize where the algorithms look (*How*) or show the mechanism behind a prediction (*How-To*). The results for contrastive explanations (*Why-Not*) were ambiguous, indicating that there may be applications for this kind of explanation, but while end users prefer them over global explanation, they generally favor non-negated reasoning.

With our socio-technical analysis, we provide first-hand evidence that the design of the explanation (i.e., how it is presented) plays an important role besides the content that is displayed (Das & Rad, 2020; Mualla et al., 2022). We also provide indicative knowledge of which types of explanation are more suitable for end users.

### 4.6.2 Theoretical Implication

Our in-depth discussion focused on three major observations: the generalizability of results across the treatments, the relation between assumed model interpretability and perceived explainability, and end user preferences for explanation types. This allows us to summarize the key points of our theoretical findings as follows:

**Tackling the tradeoff between performance and explainability is non-trivial.** We showed that the tradeoff curve assumption between performance and explainability does not always hold. While we cannot prove that the relationship always exhibits a grouped structure, other evidence points to the fact that the tradeoff can be characterized as a group decision-making process where explanation and performance cannot be approached in isolation but in alignment with organizational policy and external factors such as laws (Ebers, 2020; Goodman & Flaxman, 2017). The decision process is also influenced by the perception of the system by the decision maker. Especially in intelligent systems that use ML algorithms as a basis for decision making, a plethora of individual factors like self-efficacy, general distrust, or neuroticism can influence the view of the tradeoff (Mahmud et al., 2022; Zhou et al., 2021). It requires a weighted multi-criteria decision process to represent and quantify elements for both dimensions, performance, and explainability (Meske et al., 2022; Wanner, Heinrich, et al., 2020), which is complicated by increasing and decreasing distances between the respective algorithms. Furthermore, identifying decision elements is a challenge as some effects may only be detectable as latent indirect factors (Wanner, Popp, et al., 2021). Lastly, it is important to note if enhanced explainability does not translate into firm productivity, investing in XAI may be in vain for businesses. Further research using mixed-method approaches, including qualitative studies, can provide more detailed insights into end users valuations and avoid an explainability paradox.

**Model-inherent interpretability does not entail explainability.** The discrepancy between what is assumed in theory and our empirical findings can be explained at least in parts by the nature of our observations. Theoretical contributions look at the algorithmic and mathematical description of objects (data-centered perspective). We have employed a socio-technical and thus user-centered perspective. In our study, we targeted the naturally biased perception of end users of an ML algorithm directly and found that the difference between performance and explainability is not constantly increasing. Instead, we found that there are three groups of perceived explainability: none (deep neural networks), mediocre (coefficient-based algorithms such as linear regressions and SVM), and high (tree-based algorithms such as decision trees and random forests). The former group represents the concept of deep learning. The latter two represent shallow ML.

Decision trees are considered highly interpretable by humans in terms of their global and local explainability since it is possible to retrace a path of variables from the root node to a leaf node containing the final decision (Arrieta et al., 2020; Herm, Wanner, et al., 2021a). This explainability by design makes the model itself (global) as well as every prediction (local) intuitively accessible. The similarity of coefficient-based models can be explained by the fact that both offer variable weights in the form of coefficients and, hence, any visualization that can be done for SVM would be valid for visualizing linear regression coefficients as well. This suggests that the goodness of the explanation can be attributed largely to its design and, thus, the basic type of the explanation rather than the actual content of the explanation. Lastly, deep neural networks are considered to be black boxes to the end users that are not interpretable by humans. They need to be augmented with XAI to offer any explainability to end users.

**Explanations and XAI augmentations that require low cognitive effort fare better.** End users clearly indicated that they perceive local explanations as more explainable. From this, we infer that people prefer explanations that require less cognitive effort to process and translate into their mental model. Tree-based algorithms that offer intuitive text-based access are perceived as more explainable than the competition. They combine local post hoc explanations with global decision process knowledge. This is in line with the observations of algorithm aversion theory, where it is required to see how the algorithm behaves to form a proper judgment (Berger et al., 2021; Jussupow et al., 2020). In line with Miller (2019), we also observed that end users prefer straightforward XAI augmentations, which require low cognitive effort such as local *Why* and *Why-Not* explanations, over complex global explanations such as *How* or *How-To*. Explanations that show decision examples, such as *What-Else* explanations, have a positive influence on the perception of explanations. It is important to note that the experiment only showed reactions to initial exposure and did not consider learning that occurs through either teaching or experience. The learning effects might reduce cognitive effort and make other means of explanation more accessible.

### 4.6.3 Practical Implications

Our analysis and theoretical findings bear relevance for business practice as we can derive several practical implications from observing how people perceive algorithmic advice and explanations. Below, we summarize our findings by providing three guidelines that should be considered when employing an intelligent system for a specific task:

**Start with the performance threshold.** If an analytical model does deliver the required performance, it is not fit for the task. An explainable model that cannot provide the requested minimum quality will have no value in practice. Hence, all candidate algorithms must fulfill this requirement. The threshold will usually be determined by the overarching goal of the system involving business goals (e.g., cost savings) that can be realized by using the system.

**Consider organizational or project context beyond performance.** Other constraints typically influence the choice of algorithm. They largely depend on environmental factors such as cost (training and inference), time constraints, end user abilities, and laws. For example, models with high inference times cannot be used in real-time settings (e.g., defect detection in production). Tree-based algorithms are particularly accessible for ML laypeople, while coefficient-based algorithms may provide better performance and still be explainable enough for a trained workforce. Furthermore, laws such as the GDPR could require you to implement either a per se interpretable algorithm or a post hoc XAI augmentation. Therefore, the candidate pool must deliver acceptable explanations not only to end users but potentially also to the authorities. In that regard, consider Rudin's call for using inherently interpretable models whenever possible and keep note that the performance gain through deep learning can be neglectable for low complexity datasets (Rudin, 2019).

**Consider the degree of explanation that end users need.** Do not confuse model interpretability (required by experts to analyze the decision-making process) and prediction explainability (necessary for end users to make decisions in their work processes). After deploying an intelligent system, end users will use the system to fulfill their work tasks and not to analyze the model's decision process. These end users should be included in the explanation design process (explanation type, but also colors, visuals, etc.). This ensures that the explanations are appropriate for end users to assess the quality of the system's predictions and consider its advice appropriately. For novice users of the system, local *Why* and *What-Else* explanations promise the best user acceptance. In contrast, more global *How* and *How-To* explanations require more cognitive effort but may help them to better understand the decision process and gather expertise faster.

### 4.6.4 Limitations and Future Research Directions

As with any empirical research, our study faces some limitations.

First, our study uses online surveys with benchmarking datasets. While we only allowed for participants with a certain background, participants may have been exposed to the scenarios and several of the ML algorithms for the first time. Hence, we measured an *initial* explainability. Both datasets stem from the healthcare sector. This may introduce bias. In response, we have piloted similar surveys in different domains with comparable results (Wanner, Herm, Heinrich, et al., 2021). Furthermore, there was no time constraint for viewing and assessing an explanation. We expect results to differ in a high-velocity treatment where faster inference time becomes more valuable. Moreover, we compared inherently interpretable shallow ML algorithms and deep neural networks without further augmentations. We assume that XAI augmentations will affect explainability positively and initial evidence points to the fact (Herm, Wanner, et al., 2021a), but we refrained from including it in the first experiment due to the diversity of explanation types and visualization options that we only began to explore in the second experiment. Nevertheless, a comprehensive evaluation of the explainability of XAI augmentations is necessary to gain a better understanding. This would include assessing whether single, isolated explanations work best or if users should be presented with explanations in pairs, triplets, or even more explanation types at the same time.

Second, choosing an ensemble model (in our case, random forest) always yields bias toward the interpretability of the base algorithm class of the ensemble model. In our case, the choice of random forest caused an overestimation of ensemble explainability due to the high degree of explainability of decision trees. Consequently, we expect other ensemble models to perform consistently according to their respective algorithm base classes. To evaluate this, we suggest testing multiple ensemble models that use a variety of base class permutations to give a more objective overview of ensemble explainability. Performing such an analysis was out of scope for our research.

Third, it is possible that participants were biased in their judgment by the perceived capability or promise of algorithms and therefore assumed a higher value (Hilton, 1996; Mehrabi et al., 2021). That is, shallow ML algorithms such as SVM and linear regression offer a form of internal explainability. Hence, they were supposed to result in a better-perceived explainability than black-box models with no internal explainability, such as deep neural networks. However, we found that difference to be smaller than expected. This may be due to participants who were not able to understand the presentation of SVM and linear regression as they lacked prior knowledge (Amershi et al., 2019; Arrieta et al., 2020), which may be a practical problem in real-life cases as well. Due to high expectancy in one category (performance), end users may attribute higher scores in another category (explainability), resulting in a halo effect. Furthermore, the perceived overall impression of an algorithm can be attributed to other factors that were omitted from the study in a controlled manner. Lastly, we did not measure whether the use of certain ML algorithms or XAI augmentations improved end user task performance and thus productivity. We assume a correlation between perception, understanding, and task performance as Herm, Wanner, et al. (2021a) report. However, we did not directly measure this.

## 4.7 Conclusion

Despite its fundamental importance for human decision-makers, empirical evidence regarding the tradeoff between ML model performance and explainability is scarce. In response, we conducted an empirical study to develop a more realistic understanding of this relationship (RQ1) and explore the effect of various explanation types on end users (RQ2).

We underscore that existing theoretical propositions on the tradeoff are data-centered and misleading oversimplifications in terms of end user explainability. You cannot exchange performance for explainability and vice versa in a continuous fashion. Rather than a tradeoff curve assumption, we found a grouped structure of no, mediocre, and high explainability, where the explanation quality of decision trees and random forests constantly dominates other ML models. Further, we found that explanations fare better when they require less cognitive effort such as local explanations.

In our research, we measured the naturally biased perception of explanations by end users and not their understanding, learning effects, or task performance. Research into the usefulness of AI and human biases in ML is still in its infancy and requires substantial advances to pinpoint the effects of the various factors in play.

# 5 I Don’t Get It, But It Seems Valid! The Connection Between Explainability And Comprehensibility In (X)AI Research

*Lukas-Valentin Herm, Jonas Wanner, Franz Seubert, and Christian Janiesch*

**Abstract.** In explainable artificial intelligence (XAI), researchers try to alleviate the intransparency of high-performing but incomprehensible machine learning models. This should improve their adoption in practice. While many XAI techniques have been developed, the impact of their possibilities on the user is rarely being investigated. Hence, it is neither apparent whether an XAI-based model is perceived as more explainable than existing alternative machine learning models nor is it known whether the explanations actually increase the user’s comprehension of the problem, and thus, their problem-solving ability. In an empirical study, we asked 165 participants about the perceived explainability of different machine learning models and an XAI augmentation. We further tasked them to answer retention, transfer, and recall questions in three scenarios with different stake. The results reveal high comprehensibility and problem-solving performance of XAI augmentation compared to the tested machine learning models.

**Keywords.** Explainable Artificial Intelligence, Empirical Study, Explainability, Comprehensibility, Problem-Solving Performance

**Reference.** Herm, L.-V., Wanner, J., Seubert, F., Janiesch, C. (2021). *I Don’t Get It, But It Seems Valid! The Connection Between Explainability and Comprehensibility in (X)AI Research*. European Conference on Information Systems, Virtual Conference. https://aisel.aisnet.org/ecis2021_rp/82

## 5.1 Introduction

Decision support systems (DSS) based on Artificial intelligence (AI) are increasingly being used in research and practice to support humans in various areas of daily life and business (Zhang et al., 2018). Thereby, AI describes a concept of data-driven problem solving using multiple mathematical algorithms, often related to area of machine learning (ML) (Goodfellow et al., 2016). During the decades, several types of ML algorithms have been developed, with different kinds of calculation logic (Bishop, 2006). In practice, it is noticeable that lower complexity algorithms, that is algorithms that are transparent from a user's perspective, are often preferred over higher complexity algorithms that lack traceability even though they may outperform their counterparts (Adadi & Berrada, 2018). Hence, users prefer *white-box ML algorithms* over *black-box ML algorithms*. The assumed reason is that AI-based DSS users would have to trust the recommendation of a black-box without understanding its rationale (Rudin, 2019). Such a circumstance holds several dangers for decision-makers and also brings up legal issues regarding general data protection regulation (GDPR) (Goodman & Flaxman, 2017).

The research domain of explainable AI (XAI) deals with this issue by developing solutions to overcome the intransparency of black-box ML algorithms while maintaining their high model performance (Gunning, 2017). There are already some XAI transfer techniques for transferring black-box models into a more comprehensible form (Wanner, Herm, & Janiesch, 2020). However, there is criticism as two independent models with their related complexity are trained instead of using a white-box ML model from the very beginning and improve it iteratively to achieve a comparable performance (Rudin, 2019). However, since the performance of black-box models is indispensable, there is increasing research on ex-post explanatory approaches. These are referred to as *grey-box ML models* (Gunning et al., 2019). Through XAI augmentation techniques, methods are applied to the trained model to make its internal logic or predictions transparent (Slack et al., 2020) and there are already encouraging results (Lundberg et al., 2020). Nevertheless, on the one hand, scientists say that these methods are only an approximation and, therefore, inherently inaccurate (Rudin, 2019). On the other hand, little is known about how users perceive (X)AI explanations (Adadi & Berrada, 2018).

XAI research should therefore address both points of criticism to resolve this trade-off (Doran et al., 2017; Gilpin et al., 2018; Guidotti et al., 2018). A particular problem seems to be that an AI-based DSS's high performance is still associated with a high decision quality. However, the decision quality only becomes efficacious, if the user of the system includes the recommendation of the algorithm in his or her decision process, which requires the perception of credibility (Nawratil, 2013). Existing XAI research shows that this depends heavily on the extent to which a person understands the behavior of a model (Ribeiro et al., 2016b). Therefore, the given information gap between the AI-based DSS and its user must be closed by appropriate explanations (Cui et al., 2019; Dam et al., 2018; Gilpin et al., 2018).

However, a high perceived explainability should not be considered as the final objective of XAI. Further, research demonstrates that a hybrid intelligence consisting of humans (here: system users) and machines (here: AI-based DSS) can be considered the (future) state-of-the-art to accomplish tasks (Dellermann et al., 2019). What seems to be problematic is that precise explanations are often not easy to interpret for humans, and conversely, understandable explanations often lack predictive power (Doran et al., 2017; Gilpin et al., 2018; Guidotti et al., 2018). In addition to an explanation that is perceived as interpretable, the question arises, to what extent the user really comprehends what the system explains to be able to act as a validator and form a functional hybrid intelligence (Futia & Vetrò, 2020).

As current XAI research focuses primarily on solving the trade-off between model performance and model explainability from a feature perspective, we are trying to understand the correlation between perceived explainability and subsequent comprehensibility. To do so, we ask to what extent this circumstance already exists in today's ML algorithms as the backbone of an AI-based DSS and to what extent a popular XAI augmentation (feature influence method) can compete with those or even surpass them. Thus, we first try to determine if and to what extent an (X)AI explanation is perceived as explainable by system users. Following that, we examine if the perceived explainability improves the comprehension for problem-solving. Thus, we formulate the following research question:

***RQ:*** *What is the relationship between the perceived explainability and comprehensibility of predictions for the user of AI-based DSS and how do XAI augmentations influence this relationship?*

To answer the RQ, we proceed as follows: In Section 5.2, we present the theoretical background and the related work based on (XAI) dimension and interrelation, as well as a structured literature review on the research gap. Section 5.3 describes our research design, including the methodology, the theoretical derivation of the research model, the scenarios, technical realization, and the survey design. In Section 5.4 we describe the survey results. We critically discuss these in Section 5.5, including the own implications. Concluding in Section 5.6, we describe limitations, and provide an outlook for future research.

## 5.2 Theoretical Background and Related Work

### 5.2.1 Artificial Intelligence

AI in the Information Systems (IS) discipline research is a generic term for *intelligent agents*. Thereby, through data-based observations, these agents generate decision knowledge and further use this knowledge to solve related tasks with high accuracy (Poole et al., 1998). To do so, they need the cognitive abilities of pattern detection and problem solving resembling the intelligent abilities of a human being (Nilsson, 2014). Recently, ML as a major class of AI algorithms has gained a lot of interest, especially for real-life applications (Janiesch, Zschech, et al., 2021). Thereby, ML is the science of using mathematical models and algorithms that improve their performance through experience (Goodfellow et al., 2016). Hereby, they learn iteratively from empirical data, enabling them to find non-linear relationships and complex patterns without being explicitly programmed to do so (Bishop, 2006).

The current focus of ML research is on the optimization of the model performance. More specifically, deep learning (DL) models regularly outperform other types of ML models. Thereby, DL models, represent a specific type of ML algorithms, by using (deep) artificial neural networks (ANN). These models are especially good at analyzing highly complex datasets (Zhang et al., 2018). However, due to their complex structure, they are intransparent. Thus, a user often can trace neither the inner model logic nor specific decision making (Ribeiro et al., 2016b). Therefore, these models are black boxes that face the problem of a lack of trust, which reduces the willingness of users to accept the recommendations of such a system (Adadi & Berrada, 2018).

### 5.2.2 Explainable Artificial Intelligence

Since complex deep learning models tend to outperform lower complexity models, they are considered to have the greatest potential for further optimization (Rudin, 2019). The research area of XAI tries to develop methods to explain these black-box models by converting them into comprehensible *grey-box models* (Gunning et al., 2019), while preserving their high model performance (Lundberg et al., 2020). Here, comprehension refers to the ability to understand a decision logic within a model and therefore the ability to use this knowledge in practice (Futia & Vetrò, 2020). Therefore, grey-box models should enable users to understand two different components of the model (Dam et al., 2018; Lipton, 2018): the inner logic (global explainability) and the reasoning for a specific prediction (local explainability).

Multiple XAI techniques have been developed. On the one hand, there is the option of XAI model transfers. Here, a second, white-box model, that is a model that is perceived as per-se explainable (global), is used to explain the black-box model (Adadi & Berrada, 2018). On the other hand, there are XAI augmentations calculated on top of the black-box model (local). Multiple augmentation techniques can be used such as explanation by simplification, visualization, knowledge extraction, or influence methods. In terms of influence methods, Shapley additive explanations (SHAP) is a commonly used XAI tool (Lundberg et al., 2020). XAI toolsets estimate the influence of a single feature on a specific prediction post-hoc. By iterative manipulation of the feature values, the tools analyze how these features truly influence the prediction or the overall model's decision behavior (Ibrahim et al., 2019).

However, many researchers, such as Rudin (2019) claim that these explanations are only a mathematical approximation to the actual values and thus inferior as the techniques are insufficiently detailed to enable users to use the AI as a DSS (Hoffman et al., 2018). Aggravating this issue, there is no shared understanding of how a proper explanation should look like to ensure explainability and also comprehensibility (Miller, 2019).

### 5.2.3 (X)AI Dimensions and Interrelations

The trade-off that XAI research tries to solve in the best possible way is between model performance and model explainability (Adadi & Berrada, 2018; Angelov & Soares, 2020; Arrieta et al., 2020; Dam et al., 2018). It can be assumed that these two dimensions are related to the comprehensibility of the explanation for the AI-based DSS user. Thereby, the user acts as a validator (cf. Table 5.1).

| **Dimension** | **Description** | **Reference(s)** |
|---|---|---|
| Performance | Accuracy of an AI model regarding its predictions. | Arrieta et al. (2020) |
| Explainability | Perceived quality of a given explanation by the user. | Adadi and Berrada (2018) |
| Comprehensibility | Degree of user understanding of the explanation enabling the user to apply the information for new tasks. | Fürnkranz et al. (2020) |

Table 5.1 Dimensions of XAI Research

Many authors have tried to classify different types of ML models according to the trade-off between model performance and model explainability. Typically, a two-dimensional grid is used for this purpose. Commonly classified ML models here are support vector machines (SVM), linear regressions, rule set algorithms, decision trees, ensemble learning, and ANNs (Arrieta et al., 2020; Dam et al., 2018; Luo et al., 2019; Morocho-Cayamcela et al., 2019). SVMs are a margin-based classifier for datapoint vectors. Decision trees are sorted decision rules, which are aligned in a structured tree hierarchy. Ensemble learning models are a combination of different ML models combined with a majority voting. An ANN consists of many (hidden) computational layers and perceptrons. Input is processed through these layers and their perceptrons using mathematical operations. Linear regression is a popular statistical technique included in these comparisons, aiming to find a linear function to describe a dependent variable according to one or more independent ones (Goodfellow et al., 2016). We did not include rulesets in our analysis as they are rarely used in practice nowadays (Nosratabadi et al., 2020) and they were already examined by Fürnkranz et al. (2020).

The left side of Figure 5.1 illustrates the trade-off of performance vs. explainability by a cross-section of the authors' classification. We further integrated the classification advances by Angelov and Soares (2020) and Nanayakkara et al. (2018) who also consider XANN. The *y*-axis represents the accuracy-based performance metric, for every model. The *x*-axis represents the relative explainability scoring. The authors' classification entails that complex models achieve higher performance compared to less complex models, but at the cost of explainability (Arrieta et al., 2020; Dam et al., 2018; Duval, 2019; Gunning, 2017; Luo et al., 2019; Morocho-Cayamcela et al., 2019; Salleh et al., 2017; Yang & Bang, 2019).

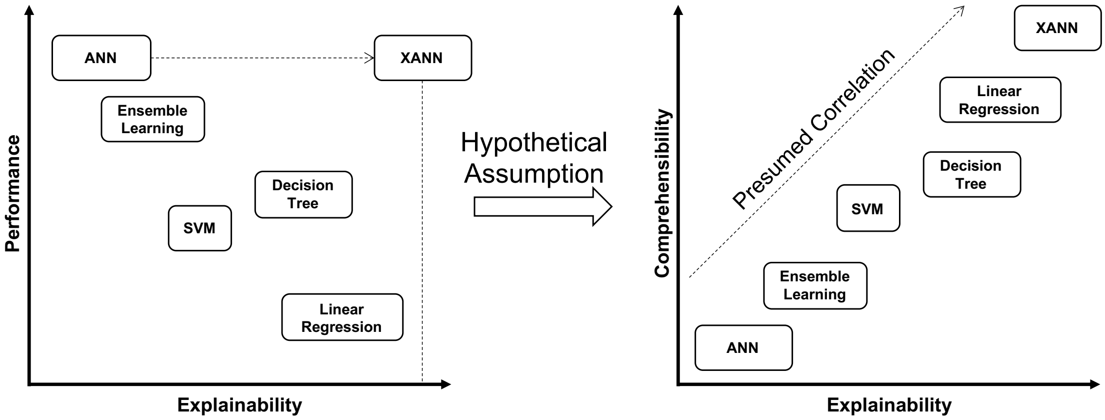


Figure 5.1 Relation of Performance, Explainability, and the Presumed Comprehensibility

Despite a general agreement on the trade-off classification of existing ML algorithms, the suitability of the users' explanations has not been evaluated yet. Hence, we need to verify if a higher perceived model explainability also results in a higher comprehensibility, and thus, problem-solving performance. Several contributions theoretically assume that there is a linear correlation between explainability and comprehensibility (Blanco-Justicia & Domingo-Ferrer, 2019; Došilović et al., 2018; Futia & Vetrò, 2020; Holzinger et al., 2019; Páez, 2019). This leads to the hypothetical assumption of certain comprehensibility levels for the different ML models (cf. Figure 5.1, right side). Here, the *y*-axis represents the achieved comprehensibility scoring, while the *x*-axis represents the explainability scoring.

### 5.2.4 Preliminaries and Research Gap

To investigate the state-of-the-art on empirical user-based studies about the correlation between perceived model explainability and user comprehensibility of (X)AI models, we conducted a structured literature review according to Webster and Watson (2002). We focused on the Computer Science related databases IEEE Xplore and ACM Digital Library. Further, we queried relevant Information Systems databases: AIS eLibrary, Science Direct, and Web of Science. Due to the subject's novelty, we did not restrict our search to (journal) rankings. We used the following pseudocode for our search term: "*((Expla* | Interpreta* | Comprehensib* | Decision Quality | Black box | Blackbox) AND (Machine Learning | Artificial Intelligence | AI | Deep Learning | Neural Net* | ANN) AND (XAI) | Explainable Artificial Intelligence)*". Through the extension of a forward and backward search, we identified 12,321 publications. After an abstract and keyword analysis, and full-text analysis, we considered *n*=42 publications to be relevant.

**Theoretical Contributions.** Most preliminary work (*n*=26) is about the theoretical evaluation of the usefulness criterion of (X)AI explanations. In particular, authors try to theoretically assess factors that affect the perceived model explainability, such as explanation fidelity (e.g., Guidotti et al., 2018), trust (e.g., Guo, 2020), effort (e.g., Calegari et al., 2020), privacy (e.g., Ras et al., 2018), and interpretability (e.g. Tjoa & Guan, 2019). The factor of user comprehension is only examined to a limited extent so far. Research has attempted to derive measurements and influences for AI explanations by using literature from related topics such as Cognitive Science (Arrieta et al., 2020; Schneider & Handali, 2019). Often, the term interpretability is used instead of comprehensibility (e.g., Freitas, 2014).

**Empirical Contributions.** Several contributions (*n*=16) already evaluated their findings on (X)AI empirically. Here, different contributions examined the influence of fidelity and interpretability (Lakkaraju et al., 2019), trust (Weitz et al., 2019), effort (Wang et al., 2019) as well as privacy (Pereira & de Carvalho, 2019). Furthermore, authors such as Förster et al. (2020a) compare different XAI methods to reveal key criteria for XAI augmentations' adaption. Likewise, a large stream of research investigates the willingness to adopt different (X)AI explanation in practice, such as, for example, in medicine (Gale et al., 2019), industrial maintenance (Wanner, Heinrich, et al., 2020), or education (Putnam et al., 2019). Two contributions investigate the influence of comprehensibility within a rule-based system (Fürnkranz et al., 2020) and decision trees (Huysmans et al., 2011). Contrary, comprehensibility within (X)AI-based systems is only proposed to be examined by Kuhl et al. (2019) who plan to do an exploratory study to analyze the task-solving performance of AI models through the influence of the compliance with (X)AI learning algorithms and explanations.

**Summary and Research Gap.** Research on the perceived explainability of X(AI) models and the resulting user comprehension has so far been theoretical rather than practical in investigation. Further, we did not find any contributions dealing with the comparison of XAI comprehensibility with other AI models by using augmentation techniques.

## 5.3 Research Design

### 5.3.1 Methodology Overview

To ensure the quality of our research, we follow the methodology according to Müller et al. (2017). This methodology is divided into four steps: *(1) Research Questions, (2) Data Collection, (3) Data Analysis,* and *(4) Results Interpretation*. We explain the steps in Figure 5.2 and briefly below.

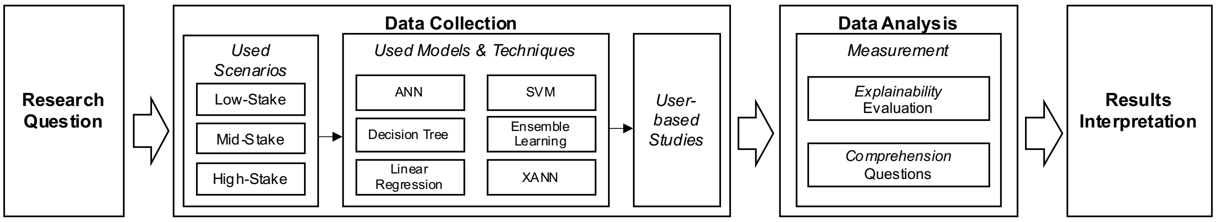


Figure 5.2 Research Methodology According to Müller et al. (2017)

*(1) Research Question.* We identified a research gap through a structured literature review: First, we want to investigate how users perceive the level of explanation of different types of ML models and an XAI-transferred ANN (XANN) augmentation. Second, we want to check whether higher perceived explainability leads to better problem-solving performance, requiring explanation comprehension.

*(2) Data Collection.* We perform an empirical study to answer the questions of interest. Further, we use different scenarios with different stakes to enable result generalization. *(3) Data Analysis.* After steps of data preprocessing, we apply various quantitative as well as qualitative analyses for knowledge discovery. *(4) Results Interpretation.* In the last step, we analyze our data and clarify the relation between the perceived level of explainability and users' comprehension regarding the different ML models and scenarios. Finally, we discuss our findings in comparison to existing research and theories.

### 5.3.2 Measurement Model

We developed a corresponding measurement model through theoretical research before we conducted the survey to investigate the assumed connection between the two dimensions.

**Variables and Dependency.** The theoretical relationship between the (perceived) model explainability and the related user comprehension can be found in explanation and ML theory (cf. Section 3.2). So, explanations that are perceived as more complex are assumed to decrease user comprehension (Futia & Vetrò, 2020). Especially for users that are inexperienced with AI systems, this might negatively impact their acceptance of AI-based DSS as they do not comprehend the results (Došilović et al., 2018). Nevertheless, Blanco-Justicia and Domingo-Ferrer (2019) recommend further investigation since an XAI surrogate model, and thus theoretically explainable but not comprehensible model, can confuse users since many XAI researchers build XAI augmentations for their purposes rather than for the intended system user (Miller et al., 2017). Therefore, we assume a linear relationship between both variables, whereby a misapprehension may exist.

**Stake of Scenario.** It has been shown that people act differently in terms of their decision-making behavior, depending on the criticality of the scenario they are confronted with (Arnott & Pervan, 2005). We therefore assume that criticality (i.e., *stake of scenario*) has a moderating effect on the connection between the perceived explainability and subsequent user comprehension. Here, low-stake scenarios describe user decisions that have only minor (cost) effects. In contrast, high-stake scenarios are associated with user decisions that may even potentially cost human lives (Kunreuther et al., 2002).

**Measurement Method.** The *explainability* of a model is a sociological measure and must therefore be approximated, and thus objectivated, by user perceptions toward the presented explanations of an ML model (Hoffman et al., 2018; Miller, 2019). The *comprehensibility* of ML model explanations can be measured by asking the user related to the given scenario and results (Lage et al., 2018; Poursabzi-Sangdeh et al., 2018). Based on the Cognitive Theory of Multimedia Learning (Mayer & Mayer, 2005), our examination uses three types of tasks: retention, transfer, and recall. Retention is about understanding the model's prediction, and thus what the AI model presents to the user. Transfer is about the user's ability to use the gained knowledge, for example to process further tasks based on the AI model's decision. Finally, recall tests the ability to reproduce the knowledge. This tests whether participants have difficulties remembering the information due to limitations of the user's cognitive abilities regarding the given explanations. We used a group interaction calculation to measure the influence of the *stake of scenario* on the relationship between explainability and comprehensibility. We divide the test sample into groups per scenario so that the answers can be calculated separately. This allows post-hoc comparability of the group results (e.g., Tausch et al., 2007). We present the final measurement model in Figure 5.3.

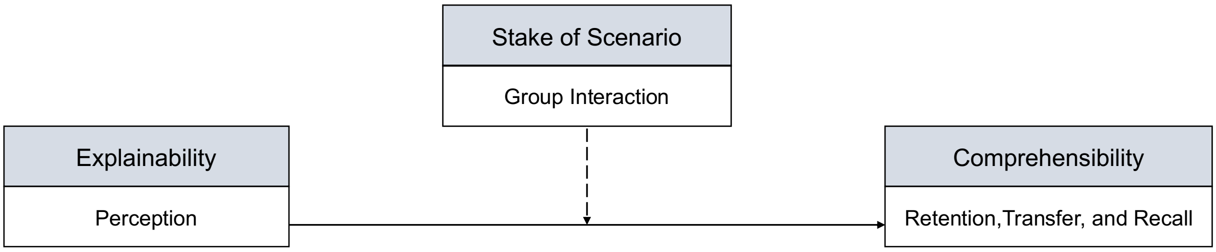


Figure 5.3 Measurement Model

### 5.3.3 Scenario Selection and Technical Realization for Survey

To enable a generalization of our findings, we use three different regression scenarios. These three scenarios differ in their stake (low, medium, and high) and are described in the following. Further we describe the technical realization of our model implementations.

**Dataset WINE.** We used the dataset on WINE quality from the UCI machine learning library for our low-stake scenario (Cortez, 2009). We consider it low stake since a wrong prediction only results in falsely predicted wine quality. The dataset consists of 11 different features describing different “vinho verde” wines from Portugal. For our approach, we have used the red wine dataset only. This dataset includes 1599 wines, which are ranked numerically between 0 to 12 in their quality level.

**Dataset Bakery.** We cooperated with a local German bakery retailer to obtain sales data from 40 stores over the last 3.5 years. In total, we used 11 different features, such as weather data, past sales, or school holidays, to predict the sales quantity for the next day. Since wrong order decisions reduce the company’s profit, we used this dataset as our medium-stake scenario.

**Dataset C-MAPSS.** We use the regression dataset modular aero propulsion system simulation (C-MAPSS) from the NASA Prognostic Center of Excellence as our high-stake scenario (Saxena & Goebel, 2008). The dataset contains simulation data from different turbofan engines. The simulation of each turbofan is tracked by 25 sensors and contains over 93 turbines on 50 simulation cycles each. After each cycle, the remaining useful lifetime (RUL) is verified. A wrong decision and, thus, a turbine failure can lead to the loss of human life. Hence, we consider the scenario to be of high stake.

**Technical Realization.** Starting with data processing, we stick to the recommendation of García et al. (2016) and deleted any incomplete observations and outliers as well as applied a feature selection and normalization. Subsequently, we have implemented the different common ML models (ANN, XANN, random forest, decision tree, SVM, and linear regression) for each scenario, we described in Figure 5.2. For the implementation of the models, we use the python package scikit-learn and keras as well as the package SHAP for the XAI augmentation. We selected the parameters through hyperparameter tuning using scikit-learn's GridSearchCV. Further, each result presentation is set in the same color scheme. For further information about the technical realization see Herm, Wanner, et al. (2021b).

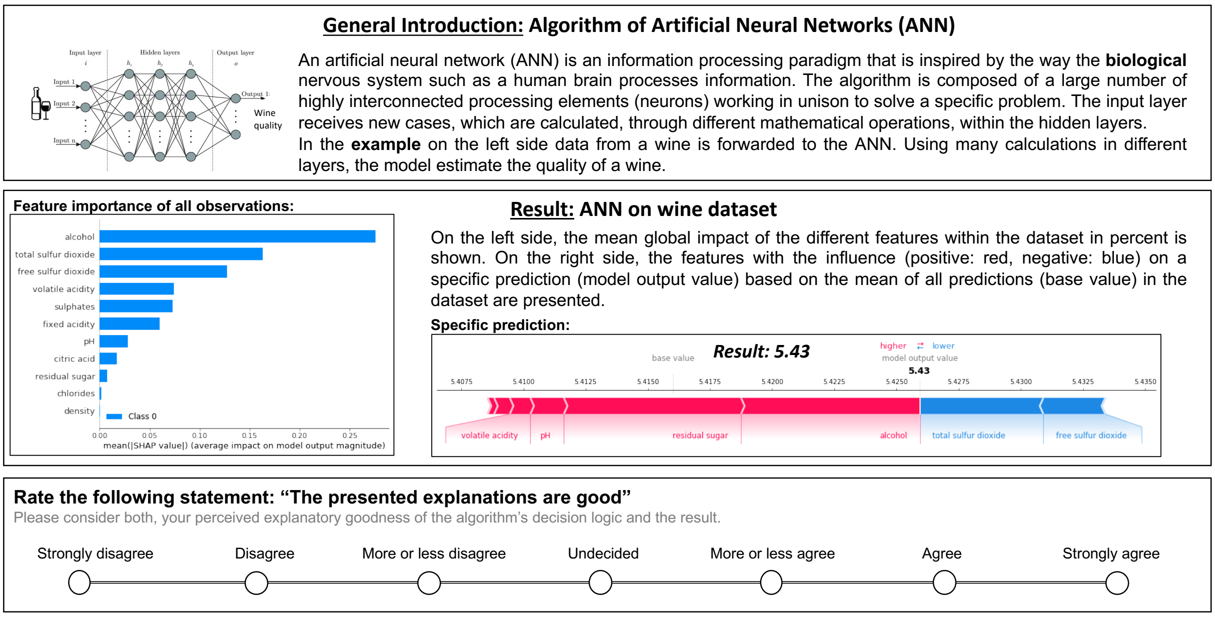
### 5.3.4 Survey Design

To evaluate the three use scenarios (cf. Section 5.4.1), we set up three separate but identical surveys (cf. Section 5.3.3). Further, each survey is divided into three parts: i) demographics and introduction, ii) perceived model explainability, and iii) examination of the user comprehensibility.

**Demographics and Introduction.** First, we examined the participants' demographics. In addition to gender, age, and location, we asked them whether they already have had experience with AI systems and whether they were willing to adopt them. Subsequently, we presented the procedure of the survey. We also introduced AI to ensure a shared understanding of the necessary knowledge.

**Perceived Model Explainability.** Second, we examined the users' perceived explainability of the implemented ML models (cf. Section 5.4.1). We started with a description of the respective scenario. This includes information about the scenario, the dataset, the task objective, and the criticality of wrong decisions. Afterwards, we presented a prediction of a particular observation for each of our five ML models to the participants. To avoid sequence bias, we randomized the order of the models. Also, we instructed user to assume similar model performance to avoid performance bias. First, we explained the algorithm themselves theoretically to ensure an understanding of their general global explainability. Also, where technically applicable, we include an average feature importance or feature impact calculated by the trained model (partial dependence plot) as well as the local explanation of the calculated result by the ML model (visualization). Figure 5.4 shows a SHAP-based XANN for the scenario of WINE as an example. Based on this information, we asked the user to rate the statement "*The presented explanations are good*" on a seven-point Likert-scale (*strongly disagree* to *strongly agree*) (Joshi et al., 2015). This kind of question is based on the recommendation for XAI-based studies by Hoffman et al. (2018) and Luo et al. (2019). The full questionnaire is available at Herm, Wanner, et al. (2021b).

**Examination of User Comprehensibility.** Lastly, we performed a review of the user's comprehension based on the given explanations to examine the effect of the AI model support (Miller, 2019). Therefore, we asked the participants to choose their preferred explanation from the five ML models. The remainder of the survey is conducted based on the comprehensibility of the selected explanation. Based on Mayer's Cognitive Theory of Multimedia Learning (Mayer & Mayer, 2005), we examined user comprehension by asking three types of questions for retention, transfer, and recall. An example from the WINE dataset for retention is "*Does the pH level have a significant influence on the quality of the wine relative to the other features?*" We provided single choice options as answers. To examine the ability to transfer, we asked for example “*Explain the influence of the sulfate level on the quality of the specific wine in comparison to the other wines*”. These questions had to be answered as open text. To test the recall of users by using cloze questions, we asked for example “*The explanations of the models show how the different [features] influence the [quality] of the wine*”. Here, the complete sentence with the missing words was presented at the beginning of the survey and can be reconstructed from the tasks conducted during the investigation.



**General Introduction:** Algorithm of Artificial Neural Networks (ANN)

An artificial neural network (ANN) is an information processing paradigm that is inspired by the way the **biological** nervous system such as a human brain processes information. The algorithm is composed of a large number of highly interconnected processing elements (neurons) working in unison to solve a specific problem. The input layer receives new cases, which are calculated, through different mathematical operations, within the hidden layers.
In the **example** on the left side data from a wine is forwarded to the ANN. Using many calculations in different layers, the model estimate the quality of a wine.

**Feature importance of all observations:**

**Result:** ANN on wine dataset

On the left side, the mean global impact of the different features within the dataset in percent is shown. On the right side, the features with the influence (positive: red, negative: blue) on a specific prediction (model output value) based on the mean of all predictions (base value) in the dataset are presented.

**Specific prediction:**

**Rate the following statement: “The presented explanations are good”**
Please consider both, your perceived explanatory goodness of the algorithm’s decision logic and the result.

Strongly disagree ○ Disagree ○ More or less disagree ○ Undecided ○ More or less agree ○ Agree ○ Strongly agree ○

Figure 5.4 Example: Introduction to the XANN Model and Prediction

## 5.4 Data Analysis

### 5.4.1 Survey and Demographics Overview

We used the platform *Prolific.co* to recruit our participants, granting them a monetary incentive of £10 per hour. The platform allows specifying one’s target group by characteristics and abilities to achieve valid results within research tasks (Peer et al., 2017). In this way, we have ensured that we survey appropriate experts for each use cases. We received feedback from *n*=175 participants. To ensure the data quality of the answers, we further used several validating checks, looking for randomly filled questionnaires, lazy patterns, failure in answering control questions, and time constraints. Subsequently, we used the feedback from *n*=165 participants (WINE *n*=55, Bakery *n*=53, C-MAPSS *n*=57). The raw survey data is available at Herm, Wanner, et al. (2021b). Out of those *n*=165 participants, *n*=98 were male, while *n*=66 were female and *n*=1 answered diverse. Most of the participants were between 20 and 30 (≈58%) or between 31 and 40 (≈24%) years old. Most answers came from Europe (≈89%). Overall, on a five-point Likert-Scale, the participants shared a medium (median) willingness to accept AI at their workplace and a medium (median) trust in AI.

### 5.4.2 Result Analysis

In the following, we present the survey results according to the study structure. First, we detail the results of the participants' perceived explainability per model. This is followed by the selection of the interviewees' preferred ML model to solve problems. Finally, we discuss the comprehension tasks (retention, transfer, and recall). The calculated results can be found in Table 5.2. They are presented per scenario, ML model, and question, subdivided into perceived explainability, retention, transfer, and recall. Further, the table includes standard deviations, aggregations, mean, and min-max-normalized results to support generic insights according to the theoretical assumption (Boone & Boone, 2012).

| Case | Model | Perceived Explainability / Standard Dev. 1) | Comprehensibility / Accuracy of Answers 2) | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | Retention | | Transfer | | Recall | | Overall / Standard Dev. 3) |
| | | | RT1 | RT2 | T1 | T2 | RC1 | RC2 | |
| WINE | *ANN* | 4.00 / 1.63 | 0.13 | 0.25 | 0.00 | 0.38 | 0.63 | 0.63 | 0.34 / 0.19 |
| | *XANN* | 6.00 / 0.97 | 0.64 | 0.71 | 0.86 | 0.86 | 0.93 | 0.86 | **0.81** / 0.19 |
| | *Ensemble Learning* | 5.00 / 1.15 | 1.00 | 1.00 | 0.00 | 0.67 | 0.67 | 0.67 | 0.67 / 0.16 |
| | *Decision Tree* | 5.00 / 1.46 | 0.58 | 0.83 | 0.83 | 0.67 | 0.83 | 1.00 | 0.79 / 0.29 |
| | *SVM* | 5.00 / 1.37 | 0.25 | 0.33 | 0.33 | 0.75 | 0.83 | 0.83 | 0.55 / 0.20 |
| | *Linear Regression* | 5.00 / 1.14 | 0.50 | 0.67 | 0.50 | 0.67 | 0.67 | 1.00 | 0.67 / 0.14 |
| **Bakery** | *ANN* | 3.00 / 1.38 | 0.71 | 0.43 | 0.29 | 0.86 | 0.43 | 0.86 | 0.60 / 0.19 |
| | *XANN* | 6.00 / 1.41 | 0.93 | 0.93 | 0.79 | 1.00 | 0.79 | 1.00 | **0.91** / 0.24 |
| | *Ensemble Learning* | 6.00 / 1.45 | 0.88 | 0.63 | 0.50 | 0.75 | 0.50 | 1.00 | 0.71 / 0.34 |
| | *Decision Tree* | 5.00 / 1.46 | 1.00 | 0.82 | 0.45 | 0.91 | 0.55 | 0.91 | 0.73 / 0.29 |
| | *SVM* | 5.00 / 1.48 | 1.00 | 0.50 | 0.50 | 1.00 | 0.50 | 1.00 | 0.38 / 0.23 |
| | *Linear Regression* | 5.00 / 1.47 | 0.36 | 0.55 | 0.73 | 0.91 | 0.73 | 1.00 | 0.71 / 0.20 |
| **C-MAPSS** | *ANN* | 3.00 / 1.72 | 0.50 | 1.00 | 0.00 | 0.00 | 0.50 | 0.50 | 0.42 / 0.11 |
| | *XANN* | 5.00 / 1.23 | 0.95 | 0.95 | 0.95 | 0.82 | 0.91 | 0.82 | **0.90** / 0.17 |
| | *Ensemble Learning* | 5.00 / 1.47 | 1.00 | 0.89 | 0.56 | 0.78 | 0.67 | 0.89 | 0.80 / 0.20 |
| | *Decision Tree* | 4.00 / 1.40 | 1.00 | 0.00 | 0.50 | 1.00 | 1.00 | 0.00 | 0.58 / 0.35 |
| | *SVM* | 5.00 / 1.36 | 0.78 | 0.33 | 0.33 | 0.78 | 1.00 | 0.89 | 0.69 / 0.16 |
| | *Linear Regression* | 5.00 / 1.28 | 0.92 | 0.38 | 0.46 | 0.62 | 0.92 | 0.92 | 0.70 / 0.16 |
| **Overall 4)** | *ANN* | 0.00 | 0.45 | 0.56 | 0.10 | 0.69 | 0.52 | 0.66 | 0.00 |
| | *XANN* | **1.00** | 0.84 | **0.86** | **0.87** | **0.89** | **0.88** | 0.89 | **1.00** |
| | *Ensemble Learning* | 0.86 | **0.96** | 0.84 | 0.35 | 0.73 | 0.61 | 0.85 | 0.60 |
| | *Decision Tree* | 0.57 | 0.86 | 0.55 | 0.59 | 0.86 | 0.79 | 0.64 | 0.60 |
| | *SVM* | 0.71 | 0.68 | 0.39 | 0.39 | 0.84 | 0.75 | 0.91 | 0.43 |
| | *Linear Regression* | 0.71 | 0.59 | 0.53 | 0.56 | 0.73 | 0.77 | **0.97** | 0.51 |

Legend: *1) Perceived Explainability* by median / *Standard Dev.; 2) Comprehensibility / Accuracy of Answers* by relative number (number correct answers / total number of answers); *3) Overall* as average of Comprehensibility / Accuracy of Answers per model and Standard Dev. of Overall; *4) Overall* as normalized average for explainability and comprehensibility per scenario and tasks types

Table 5.2 Results of Explainability and Comprehensibility Questionnaire

**Explainability.** First and as expected, the ANN model is perceived worst across all scenarios and in relative comparison to all other ML models (0.00). However, if an XAI augmentation (here via SHAP) is used, the user's perception changes profoundly. Across all scenarios that we tested, the XANN was perceived to be highly explainable. In relative comparison, XANN even scored best (1.00). However, the perception seems to decrease with an increasing stake. The relative positioning of decision tree, SVM, and ensemble learning between 0.57 and 0.86 is generally consistent with the theoretical assumption across the different stakes and complexities. However, our results contradict theory regarding ensemble learning's positioning within this group. The assumption was that a single decision tree is better explainable than a complex ensemble learning model. In our case, participants preferred ensemble learning to a decision tree in terms of explainability.

**Choice of Best Model.** Following the presentation of the scenario and ML models, the participants had to choose their preferred model for solving different comprehension evaluation questions. We present the selection results per scenario in Figure 5.5. The ratings of the participants' perceived explainability per ML model and scenario are given in Table 5.2. As expected, there is a strong correlation between evaluating the participants' perceived explainability and their best model choice for solving the problem. In each scenario, XANN was rated best in explainability and is accordingly also most frequently selected (*n*=50). On the contrary, we could confirm that the ANN was selected least as the model with the worst perceived explainability (*n*=17). Furthermore, decision trees were chosen primarily for scenarios with a lower stake (*n*=12, *n*=11), whereas they do not seem to be an alternative for high-stake scenarios (*n*=2). This is also in line with their perceived explainability scoring. The results for the decision tree are contrary to the choice of linear regression. Here, users seem to prefer its option with an increasing stake (*n*=6, *n*=11, *n*=13). Similarly, there is a similar tendency in the choice of random forest as the best model (*n*=3, *n*=8, *n*=9). However, the selection of SVM seems to depend strongly on the respective scenario (*n*=12, *n*=2, *n*=9).

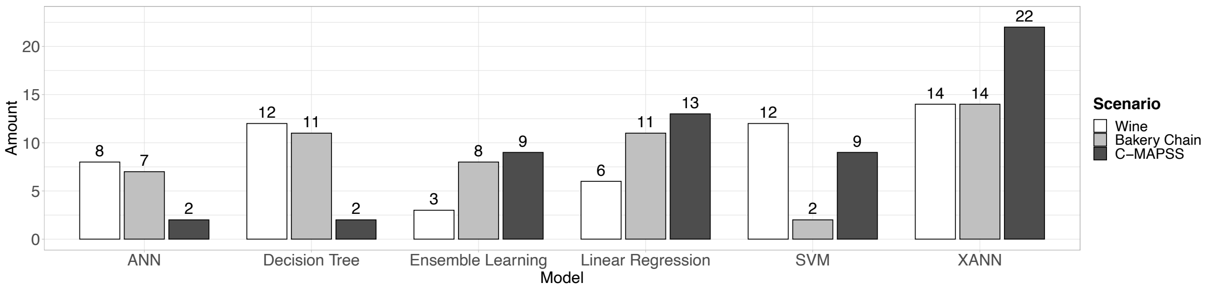


Figure 5.5 Choice of Preferred ML Model for Problem-Solving

**Comprehensibility.** Examining the users' comprehension of the ML model explanations shows similar results regarding the perceived explainability. Again, the results of the ANN are worst (0.00) and those of the XANN are best (1.00). The understanding even seems to increase with an increased stake. XANN seems to support users well, especially for transfer tasks in comparison to other ML models. On the other hand, ensemble learning seemed particularly useful in supporting users for retention tasks, but fall behind in transfer tasks. In a relative comparison, decision trees show equally good user comprehension (0.60). These, on the other hand, are characterized by an excellent balanced rating across the comprehension evaluation categories, but do not stand out in any of them. SVM (0.43) and linear regressions (0.51) perform worst (except for ANN) for comprehensibility. Yet, they scored higher in perceived explainability than decision trees. Nevertheless, both were helpful in answering recall questions.

Figure 5.6 provides an overview of the distributions based on the relative number of correct answers per model across all scenarios as the number of correct answers relative to the six questions. Overall, as expected ANN has the lowest median. In contrast, XANN has the highest median compared to the other models. Decision tree, linear regression, ensemble learning, and

SVM share roughly the same median. Nonetheless, the top quartile of the decision tree is much higher in comparison to these models.

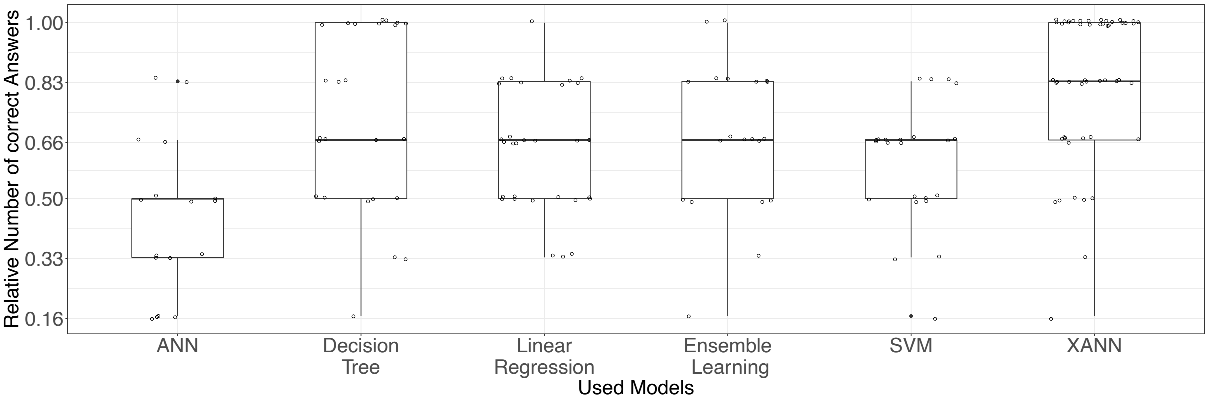


Figure 5.6 Boxplot on the Relative Comprehensibility per Model across all Scenarios

## 5.5 Discussion and Implications of Findings

### 5.5.1 Discussion of Findings

To investigate the correlation between the perceived ML model's explainability and ML model's user comprehension, we plotted the results into a two-dimensional grid presented in Figure 5.7.

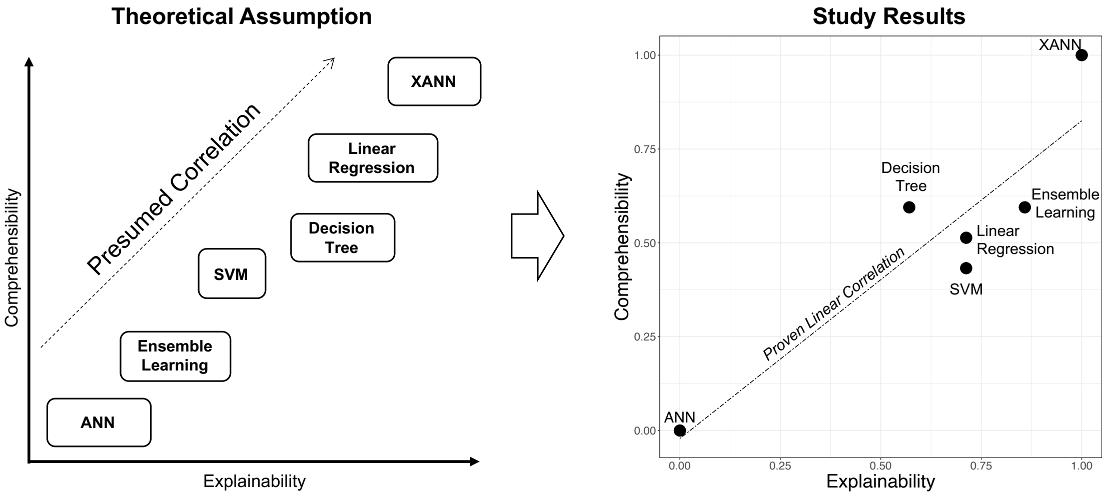


Figure 5.7 Theoretical Correlation of Explainability and Comprehensibility vs. Study Results

The left part of Figure 5.7 shows the hypothetical assumptions (see Section 5.3.2). The right part of the figure shows the empirical results of our survey. For the sake of generalization, we merged the results of the different scenarios. By doing so, we enable more general claims across criticalities as described in literature. Furthermore, we used the normalized results of the two dimensions to compare the theoretical assumption within Figure 5.2 and our results (cf. Table 5.2). To prove the presumed linear correlation, we followed Meng et al. (1992) and applied a linear correlation between both dimensions using Pearson's *r* (*p*-value: 0.01; corr: 0.91).

**General Correlation.** The results, especially considering Pearson's linear correlation, show a high agreement with the theoretical assumptions from the preliminary work of various authors such as Blanco-Justicia and Domingo-Ferrer (2019) and Holzinger et al. (2019). We also have to agree with Páez (2019) describing that explanations focusing on the post-hoc interpretability (XANN) lead to a better task solving performance comparing a model's *inherent transparency* (e.g., linear regression). Further, we can also validate the assumptions from Freitas et al. (2008) and Verbeke et al. (2011) that models such as decision trees are more comprehensible in comparison to non-linear-models such as ANN or SVM. Likewise, Ribeiro et al. (2016a) assume black-box ANNs to be the worst explainable models. Nonetheless, we also found discrepancies in our results: Futia and Vetrò (2020) indicate that comprehensibility for users should correlate with the focus on user-centered explanations. They claim that interactive explanations are necessary for problem-solving. This contradicts our results, since our participants were able to solve the different tasks (retention and transfer). Therefore, we link this assumption instead to the willingness to accept (X)AI-based DSS (Burton et al., 2020). We see that the user's perceived model explainability does not perfectly transfer to the actual user understanding, but both dimensions share a high correlation, and thus, a strong dependency. Since, we conducted a study with experts, the results may differ compared to non-experts (Castelo et al., 2019).

**Model Details.** We show that especially the decision tree ML model offers above-average comprehensibility. Ensemble learning also scored above the presumed expectations in our study. We concluded that the choice of using a random forest algorithm might have influenced this, as it allows for similar interpretability as the decision tree model. These findings are discussed in Subramanian et al. (1992). Following them, these representations have the ability to show decision patterns for the data more clearly than other ML models. Nonetheless, we also recognized different task performances levels and thus a high standard deviation within the group of participants who chose decision tree and random forest. Referring to Huysmans et al. (2011), we assume differences through users' preexisting knowledge in AI and, thus, their understanding of the explanation representation (Amershi et al., 2019). Further, it is noticeable that users perceive the ensemble behavior as more explainable than the individual tree. One possible explanation is that users often expect that a model perceived as more complex should provide a more precise prediction (Nawratil, 2013; Pratt & Zeckhauser, 1985). Likewise, the influence of the factor trust concerning the model's complexity through the applied majority or average vote can lead to those user rankings (Guo, 2020; Tintarev & Masthoff, 2012). However, looking at the transfer questions, it appears that participants using decision trees perform better in reusing the given information compared to ensemble learning. A further indication of this can be seen in comparing the SVM model and the linear regression model. Both models are perceived as equally good in their explainability. However, linear regression models had better user comprehension. A possible explanation of the result is that the intuitiveness of those models for a human decision maker may be increased due to his or her preexisting knowledge (Narayanan et al., 2018; Weld & Bansal, 2019). Nonetheless, keeping the clarity and theoretical considerations (cf. Figure 5.2) of linear regression models in mind, they perform worse (transfer). We assume that this is due to the resulting complexity of many features in real-life data. Further we assume that due to the good recall performance, linear regression does not overload the cognitive abilities of the participants.

**XANN Details.** XANN showed the best performance in both dimensions (explainability=1; comprehensibility=1). Thus, the XAI augmentation improved the worst perceived and problem-solving ANN model (explainability=0; comprehensibility=0) substantially (cf. Figure 5.6). The high appropriateness of the augmentations is also reflected in the users' frequent choice of XANN as the best model to support comprehension tasks (cf. Figure 5.5). This corresponds with the theory regarding the overload of user's cognitive capacity having inappropriate explanations (Grice, 2019). Fürnkranz et al. (2020) argue that there must be an appropriate way to explain the user's prediction. Due to relatively low standard deviation of comprehensibility of XANN (compared to the other models), we assume that most participants were able to use these explanations. Further, we noticed, that the color scheme of SHAP and therefore the presentation style can cause misunderstandings since the participants were not familiar with the explanation type and a uniform and standard visualization seems therefore necessary to support adoption (Förster et al., 2020a; Schneider & Handali, 2019). Looking at the transfer questions of the high-stake scenario, we noticed a strong primary use of local explanations by the participants instead of the global explanation. This is in line with Wolf and Ringland (2020) stating that the importance of local explanations helps solve tasks correctly in real-life scenarios and understanding the overall decision logic becomes less relevant for the respective decision instance. This goes in line with our findings, where XANN also performs good at retention questions. The high potential of XANN for problem-solving seems to be particularly evident in high-stake scenarios (cf. Table 5.2) and is also recognized by the users (cf. Figure 5.5). Both findings stand in contrast to the recommendations from Rudin (2019). However, in the low-stake scenario, the XANN shows weaknesses in retention. Also, its average overall result shows further potential for increasing the comprehensibility dimension. Nonetheless, it showed that XANN scored best overall compared to the other common ML algorithms (cf. Table 5.2). We assume that a personalized explanation, as Schneider and Handali (2019) suggested, can further increase the XANN's comprehensibility.

### 5.5.2 Implications of Findings

**Theoretical Implications.** We noticed a lack of knowledge regarding the tested comprehensibility at (X)AI models. While contributions such as Fürnkranz et al. (2020) already describe the importance of comprehensibility in theory and Kuhl et al. (2019) intend to investigate this in future research, we aimed at closing this gap. Due to the participants ability to choose their model for task solving on their own, the sample size for specific models is relatively low. Thus, using statistical tests indicate misleading results. Nonetheless, our results provide a first overview of participants' problem-solving performance on common (X)AI models and clearly highlight user preference based on scenario stake. Following that, future research can use our findings to concentrate on promising models and test their significance in more detail. Further, our findings provide a first insight, where models explanations are lacking in terms of user's comprehensibility and scenario stake. Nonetheless, further investigations are necessary regarding different dataset types and XAI augmentation techniques, as we only used SHAP and regression datasets. Similarly, we assume that the differences between perceived explainability and tested comprehensibility often result from different factors such as trust. Thus, further research methods such as technology acceptance models may be necessary to understand perceived explainability better.

**Practical Implications.** Likewise, due to the proven correlation, we can assume that the commonly used SHAP-based XAI augmentation technique is suitable to support problem-solving. However, this recommendation must be taken with a grant of salt, since this approach consists of a post-hoc trained grey-box. Therefore, in many cases, using white-box models such as decision trees is necessary due to governmental regulations (Rudin, 2019). While literature stated that linear regression is the most explainable model, we noticed a lack of comprehensibility. In contrast, in our study the decision tree performed well. Nonetheless, the use of decision tree requires skilled employees due to the data-centered and thus complex representation of decision logic (Subramanian et al., 1992).

## 5.6 Conclusion, Limitation, and Outlook

A high perceived explainability does not necessarily require user understanding and vice versa (e.g., Gilpin et al., 2018 ). While XAI research focuses on explainability, the comprehensibility of models and their predictions is of great relevance to form an effective hybrid intelligence that outperforms man and machine individually (Dellermann et al., 2019). Therefore, the goal of our research was to understand the connection between explainability and comprehensibility as well as the extent to which XAI augmentations can compete with existing ML models in user comprehension in real-life scenarios.

In our approach, we performed an empirical study with different stakes as moderator with common ML models and the XAI augmentation SHAP. Our results indicate that grey-box XAI explanations achieve the best results and are perceived to be even superior to inherently interpretable white-box ML models. One explanation seems to be that local explanations are more helpful in solving tasks correctly, while understanding the overall decision logic becomes less relevant for concrete decision situations. This entails that for example the need to explain the decision logic within a black-box ANN seems to be less critical than representing the approximated impact of the features on a decision. The results for decision trees showed that the importance of user-centered rather than data-centered explanations are especially related to good user comprehension. Likewise, our results reveal that XANN models perform best in the users' perceived explainability. We also show that there is a good correlation between the perceived explainability and the associated user comprehension across all other ML models, and thus, problem-solving performance. Subsequently, we have shown that there is a linear correlation between perceived explainability and comprehensibility of the models, with decision trees and XANNs being most consistent. However, while XANN's perceived explainability excelled in low- and medium-stake scenarios, it decreased with high-stake scenarios, which underlines Rudin (2019)'s call for the use of (novel) white-box models rather than developing new XAI augmentations.

There are some limitations to our contribution. We assume a correlation between scenario complexity and scenario stake for our datasets. A further isolated observation with more observations may help to differentiate this better. In addition, our literature review revealed additional factors for the usefulness of (X)AI explanations such as explanation fidelity, that we did not examine here. Lastly, we also must consider further XAI augmentation techniques to examine influences such as the cognitive load within these augmentations. Likewise, using XAI augmentations for example for image classification can produce different results that with numerical data.

Looking forward, we also intend to investigate the perceived level of comprehensibility within different XAI augmentation techniques as well as to overcome the lack of design principles for (X)AI in practical use. Further research also needs to extend our study and give user-centered, socio-technical recommendations for the development and sophistication of XAI frameworks to overcome the issue of "inmates running the asylum" in XAI research (Miller et al., 2017).

# 6 Impact Of Explainable AI On Cognitive Load: Insights From An Empirical Study

*Lukas-Valentin Herm*

**Abstract.** While the emerging research field of explainable artificial intelligence (XAI) claims to address the lack of explainability in high-performance machine learning models, in practice, XAI targets developers rather than actual end-users. Unsurprisingly, end-users are often unwilling to use XAI-based decision support systems. Similarly, there is limited interdisciplinary research on end-users' behavior during XAI explanations usage, rendering it unknown how explanations may impact cognitive load and further affect end-user performance. Therefore, we[2] conducted an empirical study with 271 prospective physicians, measuring their cognitive load, task performance, and task time for distinct implementation-independent XAI explanation types using a COVID-19 use case. We found that these explanation types strongly influence end-users' cognitive load, task performance, and task time. Further, we contextualized a mental efficiency metric, ranking local XAI explanation types best, to provide recommendations for future applications and implications for sociotechnical XAI research.

**Keywords.** Explainable Artificial Intelligence, Cognitive Load, Empirical Study.

**Reference.** Herm, L.-V. (2023). *Impact Of Explainable AI On Cognitive Load: Insights From An Empirical Study*. European Conference on Information Systems, Kristiansand, Norway. https://aisel.aisnet.org/ecis2023_rp/269/

[2] Although this research contribution is single-authored, "we" is used for consistency.

## 6.1 Introduction

Due to recent advances in computing, the spectrum of potential use cases for the application of artificial intelligence (AI) is constantly expanding, enabling end-users to rely almost solely on data-driven decision support systems (DSS) (Berente et al., 2021; Janiesch, Zschech, et al., 2021). That is, integrating AI into information systems forms intelligent systems to enhance end-users' and organizations' effectiveness (Gregor & Benbasat, 1999; Herm, Steinbach, et al., 2022). In this context, AI refers to an abstract concept mimicking human cognitive abilities through the application of mathematical and statistical algorithms, to generate (i.a.) machine learning (ML) models capable of automatically finding nonlinear relationships within data. So, decision knowledge is derived without the need for explicit programming (Goodfellow et al., 2016; Russell & Norvig, 2021). Research has focused on overcoming algorithmic constraints by increasing the decision complexity of ML algorithms, resulting in ML applications capable of outperforming domain experts even in complex and high-stakes use cases (Janiesch, Zschech, et al., 2021; McKinney et al., 2020). Furthermore, a subclass of ML algorithms, called deep learning (DL), uses deep neural network architectures to achieve unsurpassed performance. In turn, the inner decision logic of these models is no longer traceable by humans, which reduces end-users' willingness to use these AI-based DSSs; thus, their overall acceptance is decreased, potentially leading to algorithm aversion (Berger et al., 2021; Wanner, Herm, et al., 2022a).

To address this issue, the research stream of explainable AI (XAI) has developed approaches to overcome the lack of traceability while maintaining the performance of these black-box models (Meske et al., 2022). However, as all that glitters is not gold, these approaches are mostly mathematically driven models that provide technical explanations, as opposed to addressing the actual end-users of the system with a sound explanatory scope. That is, recent XAI approaches have mainly been designed by developers for developers (Arrieta et al., 2020; van der Waa et al., 2021). Consequently, first research endeavors emerged proposing research agendas (Laato et al., 2022), first-hand end-user evaluations (Herm, Heinrich, et al., 2023; Shin, 2021), and design knowledge (Herm, Steinbach, et al., 2022; Meske et al., 2022). Yet, it is not completely apparent how an end-user's heuristic mental model behaves, in terms of different perception factors, when operating within a use-case (Laato et al., 2022). Unsurprisingly, IS research calls for further examinations of AI-based explanations from a sociotechnical perspective (Gregor & Benbasat, 1999; Herm, Heinrich, et al., 2023), which also interfere with the research streams of human-computer interaction and cognitive science (Langer et al., 2021; Liao & Varshney, 2022).

Crucially, an explanation is a social and cognitive process of knowledge transfer from an XAI-based DSS to the end-user (Miller, 2019). It is unclear how end-users perceive these explanations, as  increased cognitive load may be imposed when end-users rely on them to solve real-world tasks (Hudon et al., 2021). Furthermore, it is unknown whether this increased cognitive load affects end-user performance or the time required to solve a task (Hemmer et al., 2021). Consequently, XAI explanations should be perceived as mentally efficient to prevent end-users from feeling overwhelmed, stressed, and unable to perform well (Buçinca et al., 2020; Paas et al., 2016).Complicating matters further, due to the increasing attention to XAI in research and practice, numerous XAI applications are being developed, creating an XAI jungle from which to select an appropriate XAI approach (Das & Rad, 2020; Dwivedi et al., 2023). That is, organizations must determine how XAI explanations affect end-user behavior and what type of explanation should be used to form a sound DSS application within intelligent systems (Gregor & Benbasat, 1999). Hence, Mohseni et al. (2021) proposed an initial systematization that groups XAI explanation types in an implementation-independent manner, providing a foundation for future research and further facilitating generalizable findings that can be transferred to any type of XAI application in practice.

Following Mohseni et al. (2021)'s systematization, we contribute to IS research (Gregor & Benbasat, 1999; Meske et al., 2022) by comparing these explanation types in an end-user-centered manner. Therefore, we conduct an empirical study in the field of medicine using COVID-19 X-ray images to measure end-users' cognitive load. Similarly, we benchmark end-users' task performance and time required to solve the task. Lastly, we combine these findings to put the mental efficiency metric of Paas et al. (2016) into the context of sociotechnical XAI research. To summarize our research intent, we propose the following research question (RQ):

> ***RQ:*** *Do XAI explanation types affect end-users' cognitive load and what are the ramifications for task performance and task time?*

The remainder of this paper is organized as follows: Section 6.2 presents the theoretical foundations, related work, and our measurement model. Section 6.3 describes the research methodology by following Müller et al. (2017) and the applied study design. Section 6.4 presents the data analysis, including demographic data, descriptive statistics, and hypotheses testing. Then, Section 6.5 discusses the findings to answer our RQ, derives implications for research and practice, and describes the study's limitations and recommendations for future research. Finally, Section 6.6 summarizes our research findings by drawing conclusions.

## 6.2 Theoretical Foundation

### 6.2.1 (Explainable) Artificial Intelligence

**Artificial Intelligence.** Following Berente et al. (2021), AI can be envisioned as an arbitrary frontier of computational advancements that mimics human-like or superhuman intelligence, enabling DSSs to assist end-users in accomplishing any task. A DSS employs ML models to enable these artificial cognitive capabilities. Here, ML is an umbrella term that encompasses mathematical and statistical algorithms used to automatically infer decision knowledge using historical data (Goodfellow et al., 2016). To this end, recent research has developed increasingly complex algorithms with high predictive power, making the models' rationale less tractable (Janiesch, Zschech, et al., 2021). Unsurprisingly, research has derived the performance-explainability trade-off, where inherently understandable models have been proposed to have the lowest performance and - conversely - DL models to have the highest performance (Herm, Heinrich, et al., 2023). Here, DL is subsumed under the umbrella term ML and refers to a deep neural network architecture with decision logic that is no longer comprehensible to humans (Janiesch, Zschech, et al., 2021). DL applications can generate promising outcomes, even in high-stakes use cases (e.g., medicine) where a wrong decision could cost human lives  (Dwivedi et al., 2023; McKinney et al., 2020). However, this may reduce end-users' willingness to use the system as non-traceability could lead to ambiguity and uncertainty in task solving (Epley et al., 2007). Alternatively, end-users may not be allowed to use the system due to regulations, such as the General Data Protection Regulation (GDPR) (Goodman & Flaxman, 2017).

**Explainable Artificial Intelligence.** In response, the multidisciplinary research stream of XAI has emerged. Its objective is to develop transfer techniques that make these opaque black-boxes comprehensible to users while preserving the predictive power of the underlying DL model (Arrieta et al., 2020; Meske et al., 2022). Thus, post-hoc explainability methods have been developed for specific types of ML models (model-specific) or a subset of them (model-agnostic); for different dataset formats (e.g., images, text, or tabular); and for different task types (e.g., classification or regression). They can also be distinguished by the nature of their explanatory scope - either explaining predictions for individual observations (local) or explaining the ML model's inner decision logic (global) (Das & Rad, 2020; Speith, 2022). This results in a plethora of explanation possibilities for depicting a rationale. In conjunction, countless distinct XAI applications have been developed in practice, creating an XAI jungle from which to choose and thus complicating the development process. Therefore, Mohseni et al. (2021) systematized these explanation types in an application-independent fashion. Table 6.1 summarizes these explanation types, their respective descriptions, and an exemplary excerpt of XAI's implementation jungle for each explanation type:

| **Type**[1] | **Description**[1] | **Exemplary Implementations**[2] |
|---|---|---|
| *How* | Holistic representation of how the ML model's inner decision logic operates - global explanation type. | ProfWeight, SHAP, DALEX, Saliency |
| *How-To* | Hypothetical adjustment of the ML model's input yielding a different output (counterfactual explanation) - local explanation type. | DiCE, KNIME, PDP |
| *What-Else* | Representation of similar instances of inputs that result in similar ML model outputs (explanation by example) - global explanation type. | SMILY, Alibi |
| *Why* | Description of why a prediction was made by informing which input features are relevant to the ML model - local explanation type. | SHAP, LIME, ELI5, Anchor |
| *Why-Not* | Description of why an input was not predicted to be a specific output (contrastive explanations) - local explanation type. | CEM, ProtoDash |

Legend: *1)* Types and definitions adapted from Mohseni et al. (2021); *2)* exemplary classification of frequently mentioned XAI implementation packages based on Das and Rad (2020), Dwivedi et al. (2023), Liao and Varshney (2022), and Mohseni et al. (2021).

Table 6.1 Description and Implementation of XAI Explanation Types

In practice, this XAI jungle is exacerbated by developers primarily designing these XAI implementations for developers without prioritizing the actual end-users (van der Waa et al., 2021). As first research endeavors target the end-user of an XAI-based DSS, these interdisciplinary research outcomes must be incorporated into practical applications to design valuable explanations for end-user (Arrieta et al., 2020). Following the explanation theory of Miller (2019), a useful explanation is defined as a social and cognitive process of knowledge transfer from an XAI-based DSS to the end-user. Thus, if an explanation is perceived as inadequate, contradicts an end-user's mental model, or does not appeal to their emotions or beliefs, trust issues can occur and user acceptance may be reduced, leading to algorithm aversion (Berger et al., 2021; Shin, 2021). Following recent IS research, a mental model defines any type of mental representation used to encode beliefs, facts, and knowledge when conceptualizing cognitive processes (Bauer et al., 2023). In this sense, the extent to which these explanation types affect end-users' cognitive load is unknown, which is an essential factor in the design and development of appropriate XAI implementations (Herm, Heinrich, et al., 2023).

### 6.2.2 Cognitive Load Measurement

Although the human cognitive system can be considered an information-processing engine, its capacity is limited when using information systems. Providing too much or distracting information in an instructional design can lead to a high cognitive load for the end-user (Bahari, 2022). The cognitive fit theory (CFT) (Vessey, 1991) posits the relationship between a task and the required information presentation (i.e., the type of XAI explanation), where an inappropriate explanation type leads to poor end-user task performance. Moreover, end-users are unlikely to have a solid understanding of the instructional design or to build a representative mental model of the task problem (Simon, 1955). This leads to them feeling overburdened, stressed, and incapable of performing sound decision-making (Anderson et al., 2020; Paas et al., 2004). Therefore, cognitive research developed a computational approach that combines mental effort, task performance, and task time into a quantitative variable called mental efficiency to classify the goodness of instructional design with respect to end-users' information processing to prevent excessive mental workload in complex cognitive tasks (Paas et al., 2016). Accordingly, XAI explanations should require an appropriate level of cognitive load to represent the model's decision and facilitate seamless knowledge transfer to the end-user. Paired with an appropriate level of task performance and task time, the high mental efficiency of an XAI explanation constitutes a well-designed XAI-based DSS (Herm, Heinrich, et al., 2023; Hudon et al., 2021).

While cognitive load is a multifaceted construct comprising various components, cognitive science research has developed several approaches for measuring it. Objective measures exist, such as eye activity, along with subjective measures, such as self-reported mental effort (Schmeck et al., 2015). While the former focuses on the identification of unconscious factors among participants, the latter targets conscious factors. Accordingly, both approaches behave complementary (Tams et al., 2014).

### 6.2.3 Preliminaries and Research Gap

To investigate the extent to which cognitive load from the perspective of XAI explanations has already been researched, we conducted a structured literature review according to Webster and Watson (2002). We focused on the information systems-related databases ScienceDirect, AIS eLibrary, and Web of Science, as well as the computer science-related databases ACM Digital Library and IEEE Xplore. Specifically, we used the following search term: *"((expla* | interpreta*) AND (explainable artificial intelligence | artificial intelligence | deep learning | machine learning | AI | XAI)) AND (cognitive load | mental load | mental effort | mental workload | cognitive capacity)".* Without restricting our search in terms of (journal) rankings, we identified *n* = 2,814 publications as potentially relevant. Hence, we consider publications that examine or discuss the effects of XAI explanations (packages) on end-users' cognitive load as relevant. This results in *n* = 17 publications after performing an abstract, keyword, and full-text analysis.

**Theoretical Considerations.** Most publications (*n* = 12) have merely centered the theoretical relevance of cognitive load (e.g., Herm, Wanner, et al., 2021a) and assumed that reduced cognitive load positively affects end-user performance (e.g., Hemmer et al., 2021) and assists end-users to solve the task faster (Bertrand et al., 2022). It is also hypothesized that increased problem complexity might be perceived as cognitively demanding (Cai et al., 2019). Similarly, research suggests that increased cognitive load reduces end-user trust in the system (e.g., Sultana & Nemati, 2021). In this context, research has derived tentative design principles (Fahse, Blohm, Hruby, et al., 2022) or design frameworks that assume reduced information granularity diminishes cognitive load (Barda et al., 2020).

**Empirical Research.** Only scarce research (*n* = 5) has focused on testing XAI's cognitive load. These contributions have mainly compared a single XAI implementation package or single XAI explanation type with a black-box implementation (e.g., Abdul et al., 2020) under simplified conditions, such as proxy tasks (Buçinca et al., 2020). From that, these contributions provide first evidence, that increased explainability will reduce end-user's cognitive load (Kulesza et al., 2013). In addition, research has focused on the connection between the end-user's cognitive load and their confidence or trust (Davis et al., 2020; Karran et al., 2022), implying that increased cognitive load slightly negatively affects perceived confidence and trust.

**Research Gap.** In summary, this sparse stream of research contains merely a handful of theoretical and empirical contributions. Regarding the former, theoretical considerations already hypothesize that use case complexity may affect end-users' cognitive load, which in turn affects task performance, task time, and trust. Concerning the latter, previous empirical contributions have mainly examined the cognitive load of end-users on a single XAI explanation package or type. Most strikingly, there is currently no research contribution that examines multiple implementation-dependent XAI explanation types simultaneously to provide conceptual guidance for a domain-independent XAI-based DSS application. Furthermore, while research has emphasized the potential impact of cognitive load on end-user task performance and time to solve a particular task, empirical evidence is lacking. As a result, to the best of our knowledge, we are the first to use these preliminary results to perform a holistic empirical cognitive load evaluation of these implementation-independent XAI explanation types and their impact on task performance and task time.

### 6.2.4 Measurement Model

To conduct our research, we derive and test hypotheses to investigate the cognitive load of the aforementioned explanation types and their effects on *task performance* and *task time*. Beyond this hypothesis testing, the findings are then used as input for the mental efficiency metric of Paas et al. (2016) to enable a summative evaluation (cf. Table 6.2). In the following, we describe the derivation of the hypotheses for our RQ and provide an overview of the measurement model.

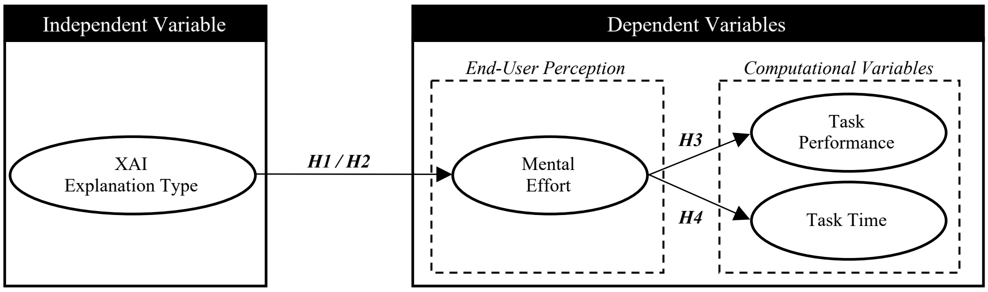


Figure 6.1 Measurement Model

In line with the CFT, we derive a group structure consisting of one independent variable and three dependent variables. The independent variable is the type of *XAI explanation*, while the dependent variables are *mental effort, task performance,* and *task time*. Here, the independent variable represents the choice of *XAI explanation* types to provide reasoning for the DL model's decision logic. The dependent variable of *mental effort*, defined as the total sum of cognitive processing that a human is engaged in, indicates the perceived level of cognitive load required to comprehend the provided *XAI explanation* for task solving (Leppink & Pérez-Fuster, 2019; Paas & Van Merriënboer, 1993). Similarly, the dependent variable of *task performance* results from the end-user's ability to use the provided *XAI explanation* to solve a task within a use case. Finally, the dependent variable of *task time* results from the time required by an end-user to solve a task when using an *XAI explanation*.

First, we assume that assisting an end-user with any type of XAI explanation would reduce the mental effort required to comprehend an ML model's reasoning for a classification (Mohseni et al., 2021). This is because these explanations pinpoint towards relevant parts of the observation for the model's classification, compared with end-users who have to figure this out for themselves (Meske et al., 2022). Therefore, we propose the following hypothesis:

*H1: Any type of XAI explanation reduces mental effort compared with no explanation.*

Second, while research suggests that XAI explanations differ in terms of their perceived explainability (Herm, Heinrich, et al., 2023), we assume that this degree of explainability is in line with the perceived *mental effort* required to comprehend the reasoning of an ML model.

That is, while explanation types such as *Why* and *Why-Not* explanations are local explanations - and therefore have a more straightforward explanatory fashion and scope than global explanation types (e.g., *How*) (Buçinca et al., 2020; Speith, 2022) - we hypothesize that variations in explanatory scope and style would result in a different level of required mental effort for each *XAI explanation* type. Thus, we formulate the following hypothesis:

*H2: Each type of XAI explanation differs in terms of mental effort.*

Third, providing information that requires a high cognitive load may overwhelm people during task solving, resulting in weak *task performance* (Hemmer et al., 2021). This may be the case when too much information is presented in a complex scenario, wherein humans are either incapable of comprehending all of it or deliberating among the levels of relevance within it (Hudon et al., 2021). Bringing this into an XAI perspective, we hypothesize that *XAI explanation* types that require less mental effort would improve end-user *task performance*. Therefore, we propose the following hypothesis:

*H3: Less mental effort when using XAI explanations leads to improved end-user task performance.*

Fourth, in research, cognitive load is considered as the number of items processed within a limited time period (Leppink et al., 2014); thus, it impedes any other cognitive tasks or activities (Barrouillet et al., 2007). That is, tasks that require a relatively significant amount of time to solve are perceived as requiring increased *mental effort*, resulting in a linear relation (Leppink & Pérez-Fuster, 2019; Otto & Daw, 2019). Hence, we hypothesize that *XAI explanations* that require less *mental effort* would help end-users to solve tasks faster than explanations that demand more *mental effort*. Therefore, we formulate the following hypothesis:

*H4: Less mental effort when using XAI explanations leads to reduced end-user task time.*

## 6.3 Research Design

### 6.3.1 Methodology Overview

We follow the methodology of Müller et al. (2017) to ensure the rigor of our research. This involves a four-step process, namely *1)* RQ, *2)* data collection, *3)* data analysis, and *4)* results interpretation. Figure 6.2 presents an overview of the research design, followed by descriptions of the four research steps.

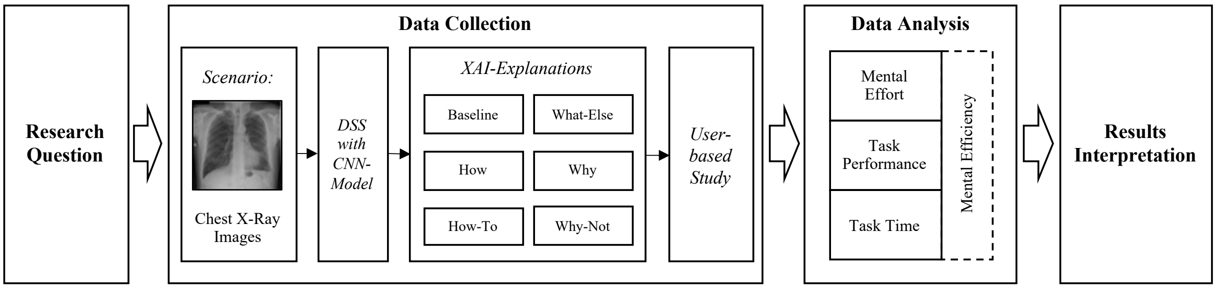


Figure 6.2 Overview of the Research Design

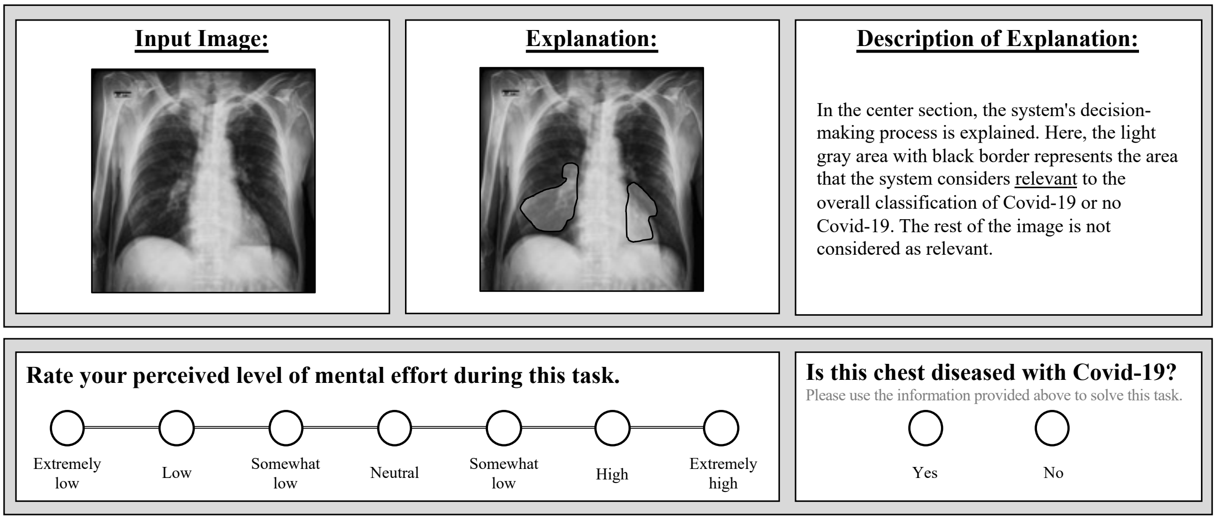
*1) RQ*: Based on the structure literature review, we found that only a handful of contributions have investigated various XAI explanation types in a holistic and implementation-independent manner. Moreover, research assumes that these explanations differ in their cognitive load. Building on this knowledge gap, we derived hypotheses to investigate this assumption and further demonstrate whether this also affects task performance and task time. We use these findings to calculate the mental efficiency of these XAI explanation types. *2) Data collection*: We use a publicly available dataset of COVID-19 chest X-ray images, a DL model, and the aforementioned XAI explanation types to test our hypotheses through a user-based study. *3) Data analysis*: From the analysis, we derive a knowledge base for our research. 4) *Results interpretation*: Ultimately, we answer our RQ and derive implications for research and practice.

### 6.3.2 Survey Design

To answer our RQ, we focused on a high-stakes use case from medicine. Specifically, we used chest X-ray images of COVID-19-infected and healthy humans (Tawsifur et al., 2022) to train a DL model - a convolutional neural network (CNN) - consisting of 11 layers, which yields a classification accuracy of 96.43% on the validation set. Then, we had the CNN classify several images of infected and healthy humans and enriched the images with the explanation style of the aforementioned XAI explanation types from Section 6.2.1.

Following the study design of Herm, Heinrich, et al. (2023), we chose a within-subjects design for our study. First, we asked participants about their demographics, introduced the high-stakes use case, and described how the XAI-based DSS operates, enabling them to put themselves in the position of a physician deciding on a patient's well-being. Subsequently, we asked each participant to perform one assignment for every explanation type: Within each assignment, they received an input image of a chest X-ray, the corresponding XAI augmented image (XAI explanation), and a comprehensive description of the XAI explanation. For each explanation type, we designed two variants, one image with an infected chest and one for healthy patients. Only one variant is shown at a time (evenly and randomly distributed). Using the provided XAI explanation, each participant was asked to classify whether the depicted chest is infected with COVID-19 or not. Then, they were asked to rate the mental effort required for this classification task on a seven-point Likert scale (extremely low to extremely high).

Using their classification, we measured their (task) performance (correct or incorrect) and clocked the required time to complete the task (task time). Both measurements were performed for every assignment and every participant. An example of the study design for the explanation type Why is presented in Figure 6.3. See Herm (2023b) for the complete questionnaire.



**Input Image:**

**Explanation:**

**Description of Explanation:**

In the center section, the system's decision-making process is explained. Here, the light gray area with black border represents the area that the system considers relevant to the overall classification of Covid-19 or no Covid-19. The rest of the image is not considered as relevant.

**Rate your perceived level of mental effort during this task.**

○ Extremely low ○ Low ○ Somewhat low ○ Neutral ○ Somewhat low ○ High ○ Extremely high

**Is this chest diseased with Covid-19?**

Please use the information provided above to solve this task.

○ Yes ○ No

Figure 6.3 Example of the Study Design

To avoid bias, we did not present the performance metrics of the used CNN (performance bias); did not use colors nor representations of XAI implementation packages (e.g., SHAP) to avoid confirmation bias; avoided learning effects through randomization; and only provided a comprehensive description for the XAI explanation to avoid forcing anchoring bias. Additionally, we incorporated several mechanisms, including attention checks, to ensure the validity of responses. Furthermore, we asked an XAI researcher to appraise our study design and a physician to review whether the classification tasks are equally difficult. Also, we conducted a preliminary study to test its validity. As we focused on the actual end-user of an XAI-based DSS, we targeted novice users in terms of AI experience. That is, we focused on prospective physicians currently enrolled as medical students, since experienced physicians might exhibit bias toward XAI-based DSS, and moreover, we wanted to focus on the future healthcare workforce (Herm, Heinrich, et al., 2023; Logg et al., 2019).

## 6.4 Data Analysis

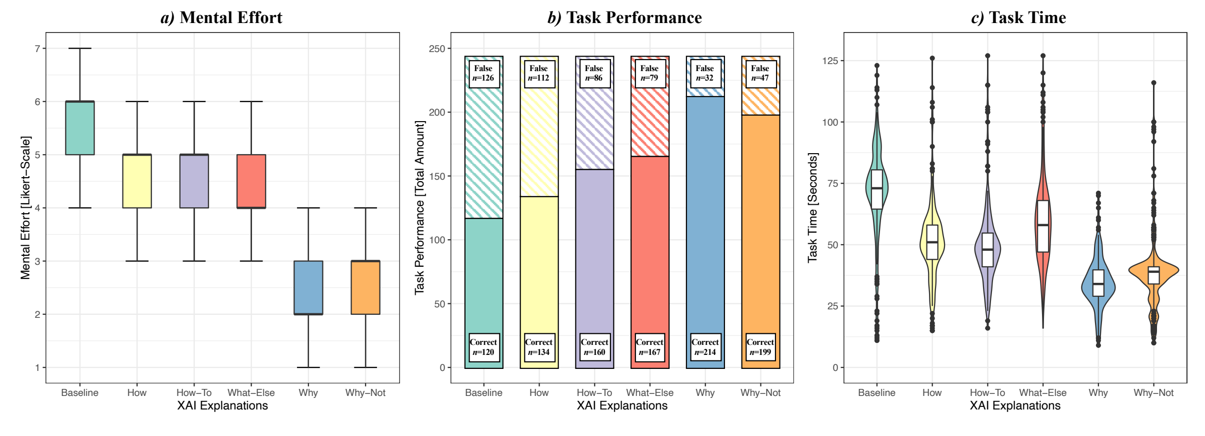
### 6.4.1 Survey Overview and Demographics

To recruit our participants, we used the Prolific.co platform, where we offered a monetary incentive of £10 per hour. Using this platform, we were able to specify and address our target group of prospective physicians (Peer et al., 2017). For this purpose, we gathered feedback from *n* = 271 participants. Since we performed several validation checks, such as randomly completed questionnaires, time-based outliers, lazy patterns, and control questions, we considered feedback from *n* = 246 participants to be optimal for our study. Among these, *n* = 130 participants were female, *n* = 115 were male, and *n* = 1 was diverse. Since we targeted enrolled students, *n* = 12 participants were younger than 20 years, *n* = 193 were 20-30 years old, and *n* = 41 were older than 31 years. They were located in Europe (*n* = 125), North America (*n* = 64), or Africa (*n* = 50). Regarding AI experience, *n* = 95 had no experience, while *n* = 112 had fewer than 2 years, and only *n* = 39 had more than 2 years. Further, *n* = 84 participants had less than 2 years of experience in medicine, $n$ = 101 had 2-5 years, and $n$ = 61 had more than 6 years.

### 6.4.2 Data Results

First, we provide an overview of the results and their distribution for the dependent variables for each explanation type (cf. Table 6.2). Subsequently, we utilize the findings to test our hypotheses in Table 6.3.

**Descriptive Statistics.** Table 6.2 highlights the results of the dependent variables: First, the mental effort findings, including medians and deviations, are plotted in Figure *a)*. Second, the total numbers of correct and incorrect answers are presented in Figure *b)*. Here, the average task performance is calculated by the ratio between correct and incorrect answers per type. Third, an overview of the distribution and kernel density of the time required per task is provided in Figure *c)*. These results are also summarized in tabular form. Thereon, we calculate the mental efficiency of the explanation types (Paas et al., 2016).



| Type | Median Mental Effort[1] | SD Mental Effort[2] | AVG Task Performance[3] | AVG Task Time[3] | SD Task Time[2] | AVG Mental Efficiency[3,4] |
|---|---|---|---|---|---|---|
| *Baseline* | **6** | **1.34** | 0.49 | **72.59** | **26.15** | -0.34 |
| *How* | 5 | 1.15 | 0.55 | 51.68 | 17.49 | -0.15 |
| *How-To* | 5 | 1.05 | 0.65 | 49.84 | 16.71 | -0.11 |
| *What-Else* | 4 | 1.20 | 0.68 | 60.10 | 18.49 | -0.08 |
| *Why* | 2 | 0.92 | **0.87** | 34.50 | 10.25 | **0.34** |
| *Why-Not* | 3 | 0.90 | 0.81 | 38.92 | 15.40 | 0.23 |

Legend: *1)* Median on a 7-point Likert scale [1,7] according to Boone and Boone (2012); participant-fixed model (LSDV): RSE: 1.047, multiple $R^2$ = 0.632, adjusted $R^2$ = 0.563, $F$=7.741, $p$ < 2.2e-16; *2)* standard deviation of mental effort/ task time; *3)* average of task performance [0,1]/ task time (in seconds)/ mental efficiency {-1..1}; *4)* mental efficiency as $ME = \frac{z_{task\,performance} \times z_{task\,time} - z_{mental\,effort}}{\sqrt{2}}$ adapted from Paas et al. (2016), mental effort and task performance standardized and task time standardized and reversed scale applied for computation.

Table 6.2 Descriptive Results of Cognitive Load Questionnaire

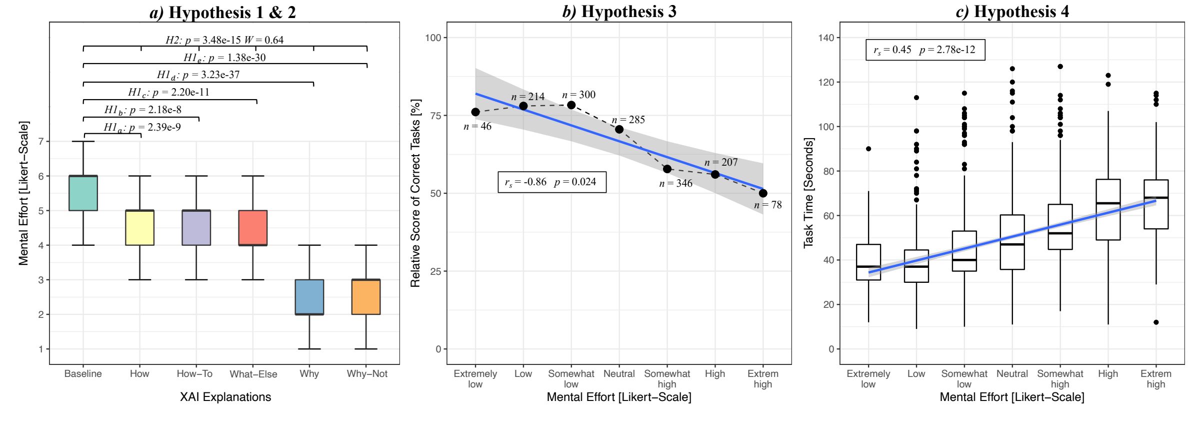
*Mental Effort.* The absence of an XAI explanation (*Baseline*) led to the highest required mental effort in this study (median = 6). The local explanations *Why* (median = 2) and *Why-Not* (median = 3) required the least mental effort to solve the task. By contrast, the global explanation *How* (median = 5) and the *How-To* explanation (median = 5) required the most mental effort across all XAI explanation types. Within this range, providing multiple images for a task to indicate similar examples (*What-Else*, median = 4) was rated as requiring moderate mental effort.

*Task Performance.* Regarding task-solving performance, without any XAI explanation (*Baseline*), the participants solved approximately 49% of the tasks correctly. Consistent with the mental effort results, using the explanations *Why* (87%) and *Why-Not* (81%) led to the highest task performance. When participants were supplied with a global explanation (*How*), their task performance increased slightly (55%) compared with no XAI explanation. Finally, the explanation types *How-To* (65%) and *What-Else* (68%) were in the middle of this comparison.

*Task Time.* When participants did not use XAI explanations (*Baseline*), they took the longest time on average (72.59 sec) to solve a task. By contrast, the explanations *Why* (34.50 sec) and *Why-Not* (38.92 sec) almost halved the elapsed time. We noticed that these explanation types exhibited a high density around this meantime compared with explanations the *What-Else* or *How-To*. In this respect, the mean task times of the *How-To* (49.84 sec), *How* (51.68 sec), and *What-Else* (60.10 sec) explanations were much closer to the baseline than those of the *Why* and *Why-Not* explanations.

*Mental Efficiency.* Since an ME above null would indicate that the end-users' performance was higher than expected compared with the mental effort invested (Paas et al., 2016), the explanations *Why* (0.34) and *Why-Not* (0.23) can be considered highly efficient. By contrast, the most mental effort was required to solve a task when no XAI explanation (*Baseline*, -0.34) or *How* explanation (-0.15) was presented. The explanations *How-To* (-0.11) and *What-Else* (-0.08) also performed slightly better in this calculation but still yielded negative values.

**Hypotheses Testing.** To test our hypotheses (cf. Section 6.2.4), we follow Motulsky (2014) and apply different testing methods for H1-H4 depending on the type of test case, as demonstrated in Table 6.3. For each hypothesis, the results are plotted and then the statistical method, resulting *p*-value, and corresponding decision of acceptance or rejection are provided below.



| **H.** | **Description** | **Test** | ***p*-Value$^{1,2}$** | **Dec.$^3$** |
|---|---|---|---|---|
| *H1* | Mental effort of every XAI explanation is lower than baseline. | Kruskal-Wallis | Cf. Figure *a)**** | Acc. |
| *H2* | Mental effort of every XAI explanation differs. | Friedman | 3.48e-15*** | Acc. |
| *H3* | Decreased mental effort results in increased task performance. | Spearman | 0.024** | Acc. |
| *H4* | Decreased mental effort results in decreased task time. | Spearman | 2.78e-12*** | Acc. |

Legend: *1)* * <0.10, ** <0.05, *** <0.001; *2)* for H1, each test yielded high significance; *3)* decision of acceptance (acc.) or rejection (rej.) of the hypothesis.

Table 6.3 Results of Hypotheses Testing

Using the results in Table 6.3, we decide whether to accept or reject our hypotheses as follows: First, to test whether providing an XAI explanation reduces the mental effort required to solve a task compared with no explanation (*Baseline*) (**H1**), we performed five Kruskal-Wallis tests that compared each XAI explanation with our baseline individually. This yielded highly significant results for each comparison, which confirm H1. Second, to investigate whether, due to the different explanatory scopes, each XAI explanation led to different levels of perceived mental effort, we performed a Friedman test and compared all types. Since this procedure revealed highly significant results, we accept **H2**. Third, to test whether using XAI explanations perceived as less demanding in terms of mental effort led to higher task performance (**H3**), we performed a Spearman correlation test to find an association between these two dependent variables. We found evidence of a significant correlation and thus accept H3. Finally, to test whether lower mental effort also correlates with lower task time (**H4**), we performed a Spearman correlation test to detect an association between mental effort and task time. We obtained a highly significant correlation, confirming H4.

## 6.5 Results Interpretation

### 6.5.1 Discussion of Results

To address our RQ, we interpret the results presented in Section 6.4.2 compromising end-users' cognitive load, task performance, and task time. Subsequently, we discuss the computed metric mental efficiency (cf. Table 6.2) to combine the findings of these dependent variables.

**Impact of XAI Explanation Types on Cognitive Load.** Recent research (Karran et al., 2022) already assumed that any type of XAI explanation assists the end-user in solving tasks, thereby reducing the required cognitive effort, as any type of explanation helps to render the end-user's mental model more congruent with the task problem compared with no explanation. We support this assumption through H1. Conversely, unstable explanations can influence the mental model. Thus, explanations are likely to impact end-user trust, especially when abductive reasoning is engaged (e.g., in complex or high-stake use cases) (Lakkaraju & Bastani, 2020). Hence, providing explanations to end-users encourages them to simplify their mental model based on the information supplied and potentially rely solely on the ML model's rationale, which could theoretically lead to mispredictions (Janssen et al., 2022).

As previous work (e.g., Buçinca et al., 2020) has already assumed that end-users perceive explanations to be individually demanding due to variations in the amount of information available and the style of explanation, we found that each XAI explanation type differs in terms of mental effort (H2). These results are reinforced, as end-users reasoning can be distinguished into a rational or an intuitive cognitive process, emphasizing a salience features evaluation or a systematic evaluation (Hamilton et al., 2016). In this regard, local explanations, namely the explanation types *Why* and *Why-Not*, yielded the lowest median and standard deviation among our results concerning mental effort. While this is consistent with Weerts et al. (2019)'s empirical study, which tested a local explanation using the SHAP package, Herm, Wanner, et al. (2021a) also found that using this package can lead to misinterpretation due to confusing color palettes or additional information. Combining our research and related studies, we expected the use of color-free *Why* or *Why-Not* explanations to impose the least mental effort on end-users. By contrast, the XAI explanation *How* returned the highest mental effort score of our study and ranked close to the baseline. In research, this type of explanation is highly debated as it provides the most information compared with other types; hence, it can be presumed to have the highest explanatory scope (Hudon et al., 2021). Still, it could possibly also overwhelm non-ML experts (Fürnkranz et al., 2020).

Comparing our findings with Herm, Heinrich, et al. (2023)'s explainability evaluation and their assumption that explainability is concomitant with cognitive load, we observe tendencies indicating a correlation between the two factors. That is, when comparing explainability and mental effort, comparative results emerged for the *Baseline*, *How*, *How-To*, and *Why* explanation types. In turn, we identified differences for *Why-Not* and *What-Else* explanation types. Here, end-users perceive What-Else explanations as more explainable (presumably) due to their information scope, but requiring increased mental effort to comprehend, which is congruent with the assumption of Miller (2019). Still, in this clinical context, distinct requirements for the explanatory scope and domain-specific regulations necessitate a detailed level of granularity (Ghanvatkar & Rajan, 2022). Also, contrary to the research of Herm, Heinrich, et al. (2023), the local *How-To* explanation demanded an increased mental effort compared to the global *What-Else* explanation. Reflecting Sultana and Nemati (2021), we surmised that this was due to the complexity of our task, which might be different with fewer features or image segments. Given the broad distribution of task time when using the *What-Else* explanation, we assumed that mental effort also depends on whether end-users grasp or struggle with this type of explanation. Still, researchers argued that this type is relatively facile to realize and its application merits prior training of end-users (Kim et al., 2016). Lastly, considering local explanation types (*Why*, *Why-Not*) perceived best, these explanation types might cause difficulties, as end-users tend to rely on features that are highlighted by the explanations (Bauer et al., 2023). Hence, guidelines are required to ensure the application of XAI in high-stakes use cases (Kloker et al., 2022).

**Impact of Cognitive Load on Task Performance.** As we obtained significant results for a linear correlation between perceived mental effort and task performance (H3), this denotes a general surplus in end-user task performance. Yet, our results are consistent with Fahse, Blohm and van Giffen (2022)'s and Hemmer et al. (2021)’s assumptions that cognitive load and task performance are diametrically related. However, we recognize some relative outliers. In particular, one might expect the responses rated “extremely low” in terms of mental effort to have produced the best task performance results; however, we found moderate to high relative task performance. This could be due to the relatively small sample size, and an outlier could skew the results. In addition, a related study already found that participants become negligent when a task is not mentally demanding (so-called “cognitive underload”), and thus, errors accumulate (Lavie, 2010). Nevertheless, when explanations require less cognitive load, the end-user’s mental model is more capable of retrieving information and recognizing new circumstances more quickly (Abdul et al., 2020). In this regard, these results should be taken with a grain of salt as we targeted novice medical end-users who were unlikely to have actively used an XAI-based DSS before. These results may change once end-users are taught how to use these types of systems or use them more frequently due to the iterative learning process (Engström et al., 2017). Thus, the explanations *What-Else* or *How* may be favored due to their increased information scope but cease to overwhelm eligible participants.

**Impact of Cognitive Load on Task Time.** Given that research has previously assumed a linear relationship between perceived mental effort and task time (Bertrand et al., 2022; Leppink & Pérez-Fuster, 2019), highly significant results also emerged for H4. Here, the task time per level of mental effort was consistent with the results and the corresponding mental effort medians. The high density within the *Why* and *Why-Not* explanation types indicates general straightforward intelligibility for novice end-users. Surprisingly, considering this linear relationship, the global *What-Else* explanation was perceived as less demanding, yet participants were able to solve our tasks faster with the local *How-To* explanation. We attribute this to the nature of the explanation, as participants might not have used such support before. Further, Buçinca et al. (2020) argue that increased task time results from end-users’ commitment to comprehend the provided explanation, as they may not trust the AI's recommendations. Conversely, a comparatively low task time could indicate over-reliance on explanations. In research, the task time factor is highly controversial: Liao and Varshney (2022, pp. author-year) stated that in the absence of time pressure, more complex explanations should be preferred as an end-user is able to iteratively discover new relationships within the explanations. Contrary, in real-world applications, a thorough evaluation process is temporarily infeasible (Shaft & Vessey, 2006). However, research has already discussed that increased task time may impact end-user satisfaction (Hsiao et al., 2021).

**Mental Efficiency Ramifications.** Local explanation types perform best regarding mental efficiency, resulting in a positive value. Therefore, end-users employing *Why* and *Why-Not* explanations exceed the performance-mental effort ratio, leading to a higher-than-expected result (Paas et al., 2016). However, compared to the cognitive load results of *What-Else* explanations, these explanations are about the participants' expected value. That is, while our findings imply relatively high mental effort, this mental efficiency result hints at relatively high end-user commitment levels, which could be consistent with the perceived explainability results of Herm, Heinrich, et al. (2023). Comparatively, there is limited research evaluating XAI explanation metrics that incorporate end-user understanding (Gentile et al., 2021). Merely Ghanvatkar and Rajan (2022) and Fahse, Blohm and van Giffen (2022) derived metrics to measure a person's effectiveness. Since we target cognitive load, we also transfer the cognitive load into the context of XAI. Accordingly, we distinguish as follows: Ghanvatkar and Rajan (2022) consider layer-wise relevance propagation (global explanation) to be the most effective as it provides the utmost information. However, this can also be critical as end-users may be unable to complete a task when overloaded. Conversely, for domain-specific requirements (e.g., in a clinical context), XAI explanations mandate a certain level of information, raising the importance to focus on effectiveness rather than efficiency. With this in mind, while one should not rely solely on effectiveness or efficiency, the trade-off should be determined based on the use case at hand (Forsythe et al., 2014).

### 6.5.2 Implications, Limitations, and Future Research

**Theoretical Implications.** While research (e.g., Buçinca et al., 2020; Hudon et al., 2021) has only partly investigated the cognitive load of singular XAI implementations, holistic comparisons of distinct implementation-independent XAI explanation types are lacking. This is especially critical when considering potential bias, which may confuse end-users or even force them to make erroneous decisions (Nourani et al., 2022). From a theoretical perspective, we have contributed to the existing body of human-technology interaction knowledge, one of the cores in IS research (Riefle & Benz, 2021), by researching XAI's cognitive load and related effects on task performance and task time to ultimately derive a mental efficiency metric for the evaluation of XAI explanations. To best of our knowledge, we are the first to place this type of metric into the context of XAI and thus also take the end-user's mental model into account. Likewise, by directly comparing task performance and task time to cognitive load, we contribute to this relatively sparse body of knowledge in XAI research.

Here, we demonstrate that XAI explanations are essential for recommendation-based decision support because they reduce cognitive load, increase task performance, and reduce task time. Consequently, local explanations perform best in terms of mental efficiency. Following the ongoing (IS research) debate on the selection of explanation types (Gregor & Benbasat, 1999; Herm, Heinrich, et al., 2023; Meske et al., 2022), we therefore provide initial insights on cognitive load for implementation-independent XAI explanation types. Drawing on this, this jigsaw piece contributes to the overall puzzle of the end-user's heuristic mental model. Although we did not measure trust and reliance during the experiment, we can identify some tendencies that could indicate end-user over-reliance especially on more straightforward explanations (e.g., *Why*). Thus, while Miller (2019) posits four requirements for the goodness of an explanation, our research indicates that focusing on causal reasoning and selective representation likely facilitates misclassification when the AI's recommendation is inaccurate. This may also be related to the type of end-users, as we focus on young professionals who tend to use the XAI-based DSS for support and pattern learning. In contrast, experts are prone to focus on using these explanations for verification (Gregor & Benbasat, 1999) and AI-experienced individuals have more reservations about AI explanations (Herm, Steinbach, et al., 2022). In this regard, our results may differ as we focus on additional end-user groups, which means that the role of explanations may vary (Bauer et al., 2023). As recent research has shown that providing an explanation has a positive impact on trust and attitudes toward an AI-enabled DSS (Wanner, Herm, et al., 2022a), this end-user over-reliance can lead to unwarranted trust that results in automation bias, even in high-risk use cases (Jacovi et al., 2021). Therefore, cognitive forcing strategies should accompany the utilization of XAI-based explanations.

Although recent IS research calls for a paradigm shift in XAI, proposing the application of hypothesis-driven support instead of recommendation-driven support to accommodate the end-user's cognitive process (Miller, 2023); the evaluation of explanations remains critical to ensure appropriate knowledge transfer of inferred evidence for an end-user action. Moreover, this approach forces end-users to be more committed, which increases their cognitive load and consequently emphasizes the need for mentally efficient explanations. To this end, we further advance the theoretical debate through the provision of a sociotechnical metric to evaluate XAI explanation types. As our results can be considered as a cognitive load-centered starting point for the discussion on the role of explanations in IS research, it currently lacks longitudinal analysis to determine additional aspects such as learning effects. This includes a combined study of other factors such as trust, acceptance, and satisfaction, which appear to be essential to understand the end-users' heuristic behavior. Ultimately, the benefits of providing XAI-based explanations in DSS will facilitate the integration of ML algorithms into organizational information systems, thus embedding the potentials of AI into intelligent systems (Gregor & Benbasat, 1999; Wanner, Herm, et al., 2022a).

**Practical Implications.** In the early days of XAI research, XAI was seen as the silver bullet for end-user adoption of AI in any use case (Goebel et al., 2018); however, we found significant differences in perceived cognitive load, task performance, and required task time among the XAI explanation types. Thus, several considerations must be made: First, our research identified that developers of recent XAI implementations (cf. Table 6.1) must reconsider their applications, building upon our results, with respect to sociotechnical factors (e.g., cognitive load) and redesign them to match end-users' mental model. Second, we demonstrated that not every explanation type is appropriate for every situation; thus, practitioners must determine an appropriate explanation based on various factors, such as performance constraints, time constraints, or use case requirements. Third, it should be considered that explanations are usually a simplification of the ML model's rationale, and therefore, they are unlikely to contain the entire decision logic, which may include bias, adversarial attacks, or open Pandora's box due to the non-applicability of the GDPR (Slack et al., 2021). That is, relying on inherently explainable ML models could be essential once a defined performance threshold is fulfilled (Rudin, 2019).

**Limitations and Future Research.** Like any empirical study, ours has its limitations. We focused our fundamental XAI research on implementation-independent explanation types to derive unbiased insights for further XAI development, which must be translated into concrete and use case-specific applications. Specifically, researchers and practitioners must integrate these insights into their XAI explanations and then re-evaluate their improved artifacts. Our research could also be expanded as follows: Following the triangulation approach of Tams et al. (2014), future research should validate our findings by using further complementary measurement approaches such as eye-tracking studies and electro-encephalograms to identify how end-users behave when using these explanation types. Second, our study should be expanded by examining additional factors to determine a holistic understanding of an end-user's behavior. Third, given the assumption of a tendency toward over-reliance within our study, future research should carry out dedicated research to examine potential trust miscalibrations. This includes investigating whether end-users are able to detect erroneous recommendations from the XAI-based DSS. Fourth, while we focused on a representative use case from the medical field, future research should leverage our findings to conduct further studies in other high-stakes use cases and with different types of end-user groups. However, our results lay the foundation for the end-user-centered design of XAI explanations and the derivation of design principles for XAI-based DSSs (Herm, Steinbach, et al., 2022).

## 6.6 Conclusion

AI is emerging as a frontier of computational advances for mimicking or surpassing human intelligence. However, in high-stake decision-making use cases, the models' internal decision logic hinders the use of DL-based applications due to being incomprehensible to end-users and thus reducing their willingness (Berente et al., 2021; Wanner, Herm, et al., 2022a). XAI has gained momentum by making these black-boxes understandable while maintaining the predictive power of the underlying model (Janiesch, Zschech, et al., 2021). Despite the proliferation of XAI applications, actual end-users are not sufficiently addressed (van der Waa et al., 2021). Unsurprisingly, using these systems for high-stakes use cases will likely result in overwhelmed and stressed end-users, which might not perform well due to high cognitive load (Hudon et al., 2021). In this regard, actual user-centered XAI research is relatively limited (Laato et al., 2022). To address this knowledge gap on end-users' cognitive behavior, we used COVID-19 X-ray images to conduct an empirical study, thereby investigating how distinct implementation-independent XAI explanation types affect end-users' cognitive load, task performance, and the time required to solve a task. Combining our results, we calculate the mental efficiency of these explanation types. This facilitates an in-depth empirical study and thus, the derivation of implications for future research and practice. In doing so, we contributed to the current body of XAI knowledge to surmount the *"inmates running the asylum"* situation (Miller, 2019) in sociotechnical XAI research.

# 7 A Nascent Design Theory for Explainable Intelligent Systems

*Lukas-Valentin Herm, Theresa Steinbach, Jonas Wanner, and Christian Janiesch*

**Abstract.** Due to computational advances in the past decades, so-called intelligent systems can learn from increasingly complex data, analyze situations, and support users in their decision-making to address them. However, in practice, the complexity of these intelligent systems renders the user hardly able to comprehend the inherent decision logic of the underlying machine learning model. As a result, the adoption of this technology, especially for high-stake scenarios, is hampered. In this context, explainable artificial intelligence offers numerous starting points for making the inherent logic explainable to people. While research manifests the necessity for incorporating explainable artificial intelligence into intelligent systems, there is still a lack of knowledge about how to socio-technically design these systems to address acceptance barriers among different user groups. In response, we have derived and evaluated a nascent design theory for explainable intelligent systems based on a structured literature review, two qualitative expert studies, a real-world use case application, and quantitative research. Our design theory includes design requirements, design principles, and design features covering the topics of global explainability, local explainability, personalized interface design, as well as psychological/emotional factors.

**Keywords.** Artificial Intelligence, Explainable artificial intelligence, XAI, design science research, design theory, intelligent systems.

**Reference.** Herm, L.-V., Steinbach, T., Wanner, J., & Janiesch, C. (2022). A nascent design theory for explainable intelligent systems. *Electronic Markets*, *32*(4), 2185-2205. https://doi.org/10.1007/s12525-022-00606-3

## 7.1 Introduction

As the frontier of computational advancements, artificial intelligence (AI) is currently pushing the boundaries of what is feasible in data-driven problem-solving (Berente et al., 2021). In this context, AI can be considered as an abstract concept for solving data-driven problems by using mathematical and statistical algorithms to build machine learning (ML) models that do not require explicit programming (Hutson, 2017; Janiesch, Zschech, et al., 2021). Unsurprisingly many kinds of systems are using AI today to achieve or surpass human intelligence for selected tasks (Berente et al., 2021). AI-based decision support systems (DSS) are a particular type of such systems capable of supporting human decision-making in many situations (Herm, Heinrich, et al., 2023; Mohseni et al., 2021) such as evaluating heat-flux sensor data to track plastic welding processes and ensure the durability of the welding seam (see Section 5).

As past research has primarily focused on solving mathematical constraints and thereby improving the performance of ML models, their inherent algorithmic complexities steadily increased (Arrieta et al., 2020; Meske et al., 2022). Lately, a class of ML algorithms called deep learning (DL) algorithms is employed increasingly as their deep ML models regularly outperform shallow ML models (Janiesch, Zschech, et al., 2021). In turn, these models are particularly opaque to the user, making them de facto black boxes for human users. Hence, these models cause difficulties in interpreting or even understanding the model's inherent processing logic or even their predictions in complex real-world use cases (Herm, Wanner, et al., 2021a; Sharma et al., 2021). This lack of explainability of the decision-making process leads to reduced trust and lowers the acceptance of intelligent systems, especially in high-stake use cases (Shin, 2021; Thiebes et al., 2021). Hence, their overall adaptation in practice is still hesitant (Hradecky et al., 2022; Kelly et al., 2019). In response, multiple studies have shown that explainability can directly contribute to adopting these models for decision support in practice (Sardianos et al., 2021; Wanner, Popp, et al., 2021).

The research domain of explainable AI (XAI) addresses this issue by developing diverse techniques to maintain the high level of performance of black-box algorithms while increasing the level of explainability at the same time (Mohseni et al., 2021). Consequently, the integration of such XAI techniques in intelligent systems and the development of explainable intelligent systems (EIS) for decision support is considered a key factor for intelligent system acceptance (Gunning et al., 2019; Mohseni et al., 2021). Due to the novelty of the research domain, there are several unsolved problems (Abedin et al., 2022; Meske et al., 2022). Despite numerous applications and developments of XAI techniques, there is still a lack of a holistic reappraisal of design factors to enable the integration of XAI techniques into intelligent systems (Abedin et al., 2022; Herm, Wanner, et al., 2021a; Meske et al., 2022; Mohseni et al., 2021). Complicating matters further, recent XAI techniques are predominantly developed by ML experts for ML experts leading to a situation where the desired explainability of the models only becomes accessible to experts but is barely accepted by end-users in practice. In this context, ML experts are developers with in-depth knowledge of ML algorithms to build and evaluate ML models. In contrast, end-users are users who are skilled in their application domain and thus use EIS in support of decision making without having any profound ML background (Arrieta et al., 2020; Herm, Wanner, et al., 2022). As intelligent systems rapidly emerge as a core assistance for daily work, in our research we predominantly address the future workforce that will be affected by such systems (Berente et al., 2021; McKinney et al., 2020). Users come with various age and experience profiles. We focus on educated people with some work experience as well as little (for end-users) to pre-existing (for developers) AI background. We do not consider in-training or late-career specificities. In this respect, through our requirements analysis and evaluations we focus on work systems and professional work situations and do not consider EIS for private uses such as entertainment.

In our research, we address this lack of system development guidelines and the consideration of both user groups to foster the acceptance of EIS. Employing design science research (DSR), we investigate which design requirements, design principles, and design features, cumulated as a nascent information systems design theory, are relevant for EIS in theory and practice. The following research questions (RQ) summarize our socio-technical research intent:

*RQ1) What are design requirements, design principles, and design features of a nascent design theory for EIS?*

*RQ2) How do the results vary for end-users and developers?*

To answer our research questions, we applied a two-cycled DSR methodology according to Vaishnavi and Kuechler (2007). In the first design cycle, we conducted a structured literature review to derive an initial theory-based design theory, which we then adjusted and validated through expert interviews. In the second design cycle, we refined our design theory and evaluated it against a real-world use case application. Ultimately, we propose a nascent design theory crafted for domain-independent development of EIS comprising multiple user groups. Due to its multidisciplinary nature, our design theory takes the diverse facets of XAI's human-agent interaction (Miller, 2019) into account and can be considered as a starting point for adaptations for all types of use cases, including electronic market scenarios that require decision support such as e-business, supply chain, or service management.

Our paper structures as follows: In the second section, we present the theoretical background and related research of EIS. Section 7.3 describes the used DSR methodology, including a comprehensive description of the two design cycles. Section 7.4 introduces the final nascent design theory and Section 7.5 presents an EIS real-world use case application and evaluation. We discuss the results in Section 7.6, before we conclude with a summary.

## 7.2 Research Background

### 7.2.1 From Decision Support Systems to Intelligent Systems

While DSS gained significant momentum in information systems research in the 1970s and 1980s, their application is still essential today (Liu et al., 2008). In this context, DSS are interactive and computer-based software systems that use decision rules and models to aid decision makers in solving unstructured problems (Turban & Watkins, 1986). Since this is a broad definition, any system that contributes to a decision-making process can be defined as a DSS (Sprague, 1980). Unlike expert systems, DSSs do not replace users but rather provide them with decision recommendations (Turban & Watkins, 1986). In the early days of the DSS era, software engineers handcrafted decision rules and decision models underlying the DSS. That is, knowledge workers had to transfer their skills into DSS's logic explicitly (Sprague, 1980). Since then, computational breakthroughs due to advances in ML technology have enabled the use of DSS in highly complex and critical situations (Janiesch, Zschech, et al., 2021). Recent examples can be found in all kind of application fields, such as medicine (McKinney et al., 2020), manufacturing (Nor et al., 2022), or social media (Meske & Bunde, 2022). For the following, we align with Herm, Heinrich, et al. (2023) and Mohseni et al. (2021) by referring to these types of AI-based DSS or intelligent DSS as *intelligent systems*.

### 7.2.2 Artificial Intelligence and Intelligent Systems

According to definition of Berente et al. (2021, p. 4), AI is the *"frontier of computational advancements that references human intelligence in addressing ever more complex decision-making problems"*, which is pushed further by intelligent systems to provide decision-making with human-like or even superhuman cognitive abilities (Herm, Heinrich, et al., 2023; Janiesch, Zschech, et al., 2021). To enable these decision-making abilities for decision support, intelligent systems use ML to allow for the autonomous generation of decision knowledge based on observations (Nilsson, 2014; Poole et al., 1998). The field of ML has gained increasing attention due to groundbreaking computational advances (Thiebes et al., 2021). Here mathematical and statistical algorithms are used to iteratively learn nonlinear relationships and complex patterns from empirical data to train ML models (Goodfellow et al., 2016; Janiesch, Zschech, et al., 2021). This includes models from DL, which are based on (deep) artificial neural network (DNN) (LeCun et al., 2015). Nowadays, the predictive performance of DNNs exceed that of domain experts (McKinney et al., 2020). On the downside, while their architectural structure is becoming more complex, the user's ability to comprehend the inner decision logic decreases (Ribeiro et al., 2016b). In practice, this results in a complex tradeoff between the performance and the explainability of these models (Herm, Heinrich, et al., 2023). That is, models with high predictive accuracy also tend to be more challenging to comprehend and vice versa (Herm, Wanner, et al., 2021a). Since we do not make a distinction between shallow ML and DL in this article, as we focus on any non-white-box model, in the following we subsume DL under the larger umbrella term ML.

When integrating ML models into intelligent systems, this results in an increased tension between a user and the intelligent system during a decision-making process (Sundar, 2020), as a user may not be able to understand the underlying rationale of the ML model. Consequently, the user's willingness to adopt this system diminishes as humans desire to reduce uncertainty and ambiguity in their environment (Epley et al., 2007). Ultimately, the overall goal should be to implement intelligent systems, which can describe their rationale with sufficient explanations to aid in decision making (Mohseni et al., 2021; Rudin, 2019). We define those systems as EIS.

### 7.2.3 Explainable Artificial Intelligence in Explainable Intelligent Systems

According to Miller (2019), explanations as the product of explanation theory are about the assignment of causal responsibility derived through a cognitive and social process of knowledge transfer. Hence, he outlines that explanation theory for AI must account for multiple dimensions ranging from information requirements, information access, functional capacities to pragmatic goals of the explainer and explanatory tool to address cognitive aspects as well as beliefs, desires, intentions, emotions, and thoughts derived from the theory of mind to address social aspects.

Correspondingly, we define explainability as the ability to use information to comprehend an event by formalizing logic-based causal chains (Arrieta et al., 2020; Lewis, 1986). In this regard, missing explainability can cause trust issues and reduce the acceptance of those systems (Shin et al., 2020; Zerilli et al., 2022), resulting in so-called algorithmic aversion (Berger et al., 2021). As an explanation includes both the product of cognitive reasoning and the social process, an explanation may be inappropriate if it is not correctly understood by the receiver or perceived as irrelevant (Hilton, 1996). Accordingly, recent research has demonstrated the importance of considering a plethora of factors to provide the receiver with an adequate explanation (Mahmud et al., 2022; Shin et al., 2020).

Explaining ML decisions is of paramount importance as misclassified training data can have devastating consequences when human lives are at stake (Lebovitz et al., 2021). To achieve explainability in intelligent systems, the system must either apply inherently explainable shallow ML models (e.g., decision trees), that is white-box models, and thus potentially forfeit predictive power or consider more complex models (e.g., DNNs) that are black boxes if considered in isolation and require explanation augmentations (Arrieta et al., 2020; Rudin, 2019).

The multidisciplinary research field of XAI addresses this objective by developing transfer techniques that provide users with comprehensible explanations of an intransparent model's decision logic or insights from the utilized data of a decision (Das & Rad, 2020; Meske et al., 2022). XAI is gaining momentum due to policy initiatives and regulations such as the "right to explanation" in the wake of the General Data Protection Regulation (GDPR) (Goodman & Flaxman, 2017). In addition, the integration of XAI into intelligent systems for decision support is motivated by the need to manage, control, and improve intelligent systems (Arrieta et al., 2020; Mohseni et al., 2021), establishing the need of EIS (Herm, Heinrich, et al., 2023).

Hence, various techniques have been developed for DNNs (Adadi & Berrada, 2018), showing a promising suitability for resolving the tradeoff between performance and explainability (Arrieta et al., 2020; Herm, Heinrich, et al., 2023). In this context, using model-agnostic techniques enable the transformation of opaque black-box models into transparent white-box models, with the coincident goal of maintaining their predictive power (Mohseni et al., 2021). They can be distinguished in two different post-hoc explanation types (Gunning et al., 2019): global explanations and local explanations. Global explanations allow a deeper traceability of the model's behavior, making the holistic decision-making process of models transparent (Lundberg et al., 2020). In theory, these types of explanations are mainly used by developers to validate trained models (Miller, 2019). In contrast, local explanations, primarily aimed at end-users, provide explanations for specific predictions presented in the form of visual, textual, or example-based explanations (Arrieta et al., 2020; Herm, Wanner, et al., 2021a; Lipton, 2018). However, literature claim the lack of user-centered evaluation of existing XAI techniques, which may lead to inadequate XAI explanations and thus hinder successful human-agent interaction (Miller, 2019; van der Waa et al., 2021).

### 7.2.4 Related Work

Apart IS-related contributions such as Förster et al. (2020b) who provide a design process for user-centric XAI systems and Herm, Wanner, et al. (2022) who introduce a taxonomy to assist user-centered XAI research, we were only able to identify a handful of DSR-based contributions that focus on user-based studies for EIS (Bunde, 2021; Cirqueira et al., 2021; Landwehr et al., 2022; Meske & Bunde, 2022; Schemmer et al., 2022). Meske and Bunde (2022) and Bunde (2021) provide design principles for explainable DSS limited to detecting hate speech. Landwehr et al. (2022) derive design knowledge for image-based DSS. Further, Cirqueira et al. (2021) stated design principles for XAI-based systems in fraud detection and Schemmer et al. (2022) propose design principles for an XAI-based DSS at real estate appraisals.

Related to this, we identified further XAI design studies in the field of human-computer interaction (HCI) relevant to our cause. Here, Amershi et al. (2019) and Mohseni et al. (2021) provide some generic design recommendations for XAI research. Moreover, Sokol and Flach (2020) and Liao et al. (2020) primarily focus on design needs for EIS. Similarly, current research in the field of HCI-based XAI investigates how users perceive user interfaces (UI) and thereby their expectations towards the use of intelligent systems (e.g., Mualla et al., 2022; Stumpf et al., 2009). This research aims to reveal the influence of HCI in the field of XAI research (e.g., Abdul et al., 2018; Bove et al., 2022). Lastly, research addresses the impact of interactive UI elements within intelligent systems (e.g., Evans et al., 2022; Khanna et al., 2022).

In addition, we identified XAI-related research, which implicitly derives challenges and thus requirements for the use of EIS. This includes human-in-the-loop for EIS development (Chou et al., 2022), identifying the degree of EIS's decision explainability (Herm, Heinrich, et al., 2023), or defining new responsibilities to handle EIS's outcome (Storey et al., 2022).

While preliminary research has already derived a first theoretical foundation for the derivation of a design theory, it is apparent that this research has not been synthesized to design knowledge as starting point for the derivation of use case dependent design theories yet. In contrast, recent research primarily focuses on specialized use cases. To this end, this manifests the deficit and thus the need for first-hand and use case independent design knowledge to enhance and ensure future EIS design theory development.

## 7.3 Research Methodology

### 7.3.1 Design Science Research Methodology

**Design Science Research.** DSR is a problem-solving-oriented research approach to generate IT artifacts (e.g., design theories) for a more effective and efficient use, implementation, and management of information systems or to solve a specific organizational problem. The goal is to transform a defined problem state into a solution state by intervening with a defined IT artifact (Hevner et al., 2004; Möller et al., 2020). In this context, the role of DSR is twofold. First, a kernel theory initiates the search progress for an appropriate solution state. As elaborated above, explanation theory (Miller, 2019) serves as a kernel theory with XAI as its instantiation to enable AI-based applications in DSS resulting in EIS. Second, the application of DSR aims at providing prescriptions for how to solve a defined problem state. These prescriptions can be provided by a *design theory* (Vaishnavi & Kuechler, 2007). Design theories contain certain classes of (meta-) design requirements, practices for IT artifact development (e.g., design principles), and IT artifacts themselves or distinctive design features that contribute to design knowledge (Meth et al., 2015). Gregor and Hevner (2013) distinguish situated implementation from nascent design theories from well-developed design theories. While the former deals with instantiations and the latter encompasses mid-range to grand theories, nascent design theories focus on knowledge as operational principles expressed through design principles. Design principles are precepts that are inductively or deductively derived from experience or empirical evidence to support achieving a prosperous solution state. Finally, the concrete problem is solved by visualizing the design principles into concrete design features (Fu et al., 2015; Möller et al., 2020).

**Application of Design Science Research.** The aim of our research is to develop a nascent design theory. To ensure the quality of the IT artifact, we applied the DSR methodology according to Vaishnavi and Kuechler (2007) and extended it by including multiple theory-building elements (Glaser & Strauss, 1967; vom Brocke et al., 2015). This combination of qualitative and quantitative research is also recommended by Mohseni et al. (2021). The resulting methodology divides into five phases: problem awareness, suggestions, design & development, evaluation, and conclusion. For our research, we applied two of these design cycles (see Figure 7.1).

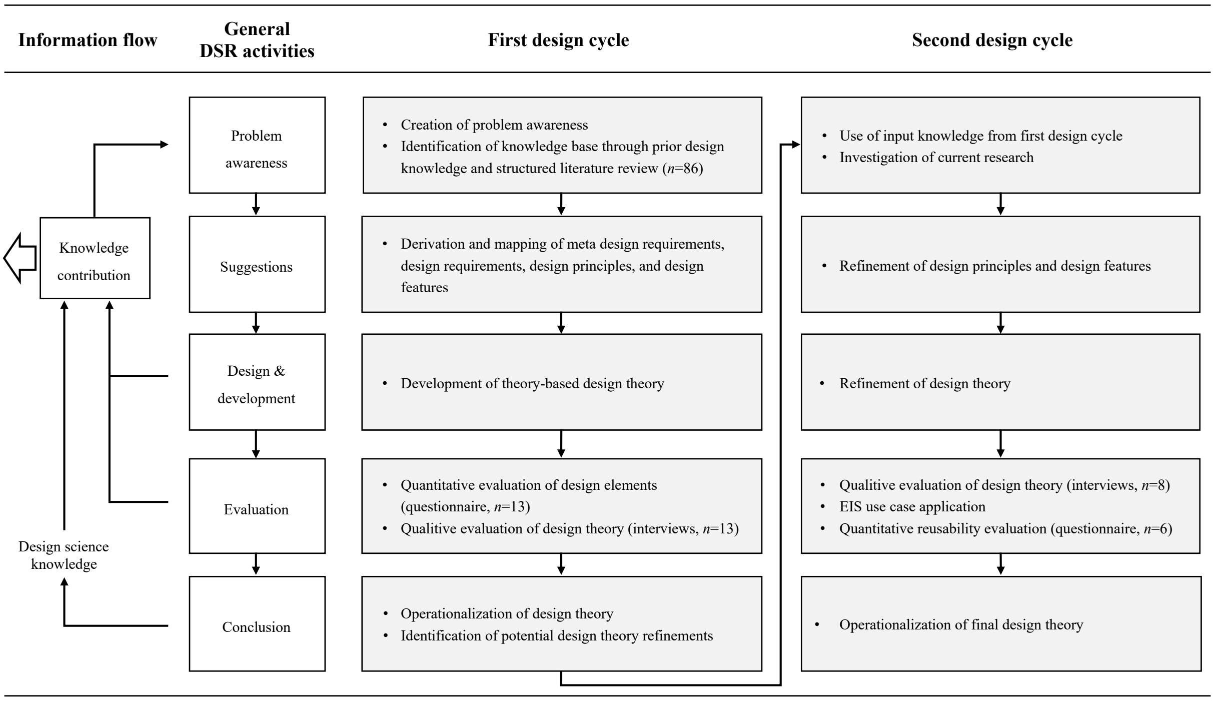

Information flow
General DSR activities
First design cycle
Second design cycle
Problem awareness
• Creation of problem awareness
• Identification of knowledge base through prior design knowledge and structured literature review (n=86)
• Use of input knowledge from first design cycle
• Investigation of current research
Knowledge contribution
Suggestions
• Derivation and mapping of meta design requirements, design requirements, design principles, and design features
• Refinement of design principles and design features
Design & development
• Development of theory-based design theory
• Refinement of design theory
Evaluation
• Quantitative evaluation of design elements (questionnaire, n=13)
• Qualitive evaluation of design theory (interviews, n=13)
• Qualitive evaluation of design theory (interviews, n=8)
• EIS use case application
• Quantitative reusability evaluation (questionnaire, n=6)
Design science knowledge
Conclusion
• Operationalization of design theory
• Identification of potential design theory refinements
• Operationalization of final design theory

Figure 7.1 Application of DSR according to Vaishnavi and Kuechler (2007)

**Overview of First Design Cycle.** Initially, the design cycle began with the phase of *problem awareness* where we identified the lack of design knowledge and built the knowledge foundation. Here, we identified that information systems research currently lacks design knowledge for the derivation of use-case-independent design theories for EIS (cf. Section 7.2.4). To address this lack, we used prior design knowledge as input for the derivation of three meta design requirement proposals (vom Brocke et al., 2020). In order to do so, we conducted a structured literature review according to vom Brocke et al. (2015), including design studies, case studies, scenarios, and reviews. During the *suggestions* phase, we extracted goals, design requirements, design principles, and design features from the structured literature review to address our meta design requirements (Möller et al., 2020). Extending this, we follow the guidelines of Gregor et al. (2020) to propose an initial design theory. In the subsequent *design & development* phase, we specified design principles using the development process of Möller et al. (2020) to materialize the theory-based design theory. In the *evaluation* phase, we enriched the theory-based design theory and demonstrated as well as validated it with practitioners and researchers in qualitative semi-structured interviews according to Kaiser (2014). This preliminary nascent design theory constitutes the result of the *conclusion* phase of the first design cycle and as input for the second design cycle.

**Overview of Second Design Cycle.** As we observed improvement potential during the evaluation of the first design cycle, we conducted a second design cycle, including findings from recent XAI publications and input from the evaluation phase of the first design cycle in the *awareness of problem* phase. Then, we refined the design principles and features in the *suggestions* phase and, consequently, the overall design theory in the *design & development* phase. Subsequently, we performed a threefold evaluation in the *evaluation* phase with experts from a German predictive maintenance project to prove the rigor of our design theory (Hevner et al., 2004; Mohseni et al., 2021). This includes a qualitative study to ensure the validity our design theory and reveal possible improvement potentials, an instantiation of the design theory through the implementation and evaluation of an EIS through a real-world use case within the maintenance project, and lastly a quantitative evaluation against Iivari et al. (2021)'s reusability criteria. Lastly, we operationalized the final design theory and thereby contribute to theory and practice by revealing novel design knowledge (Vaishnavi & Kuechler, 2007). Section 7.4 introduces and details our final nascent design theory, while Section 7.5 comprises the design theory instantiation and the quantitative evaluation.

### 7.3.2 Results of First Design Cycle

**Awareness of Problem, Suggestions, Design, and Development.** To obtain the theoretical foundation for the derivation of the design theory, we applied a structured literature review according to vom Brocke et al. (2015). Due to the interdisciplinary nature of the topic, we considered databases from economics (Emerald Insight, EBSCOhost), computer science (IEEE Xplore, ACM Digital Library), and from information systems (AISeL, ScienceDirect). We queried contributions focusing on the topics of XAI, HCI, explainability, and (design) requirements. Please see Appendix D.1 for a comprehensive overview of the search strings, the used terms, and synonyms. Further, due to the novelty of the subject, we did not restrict search in terms of rankings. This resulted in 1.426 potential hits, which we then screened and analyzed using reduction criteria consisting of title, keyword, abstract analysis, as well as duplication and language checking. This leads to 114 remaining contributions, of which we classified 86 as relevant using full-text and forward/backward search analysis. As inclusion criteria, we considered contributions from the XAI domain, focusing on requirements, guidelines, best-practices, and different explanatory concepts from a (non-)technical perspective. Figure 7.2 summarizes the process of the literature review.

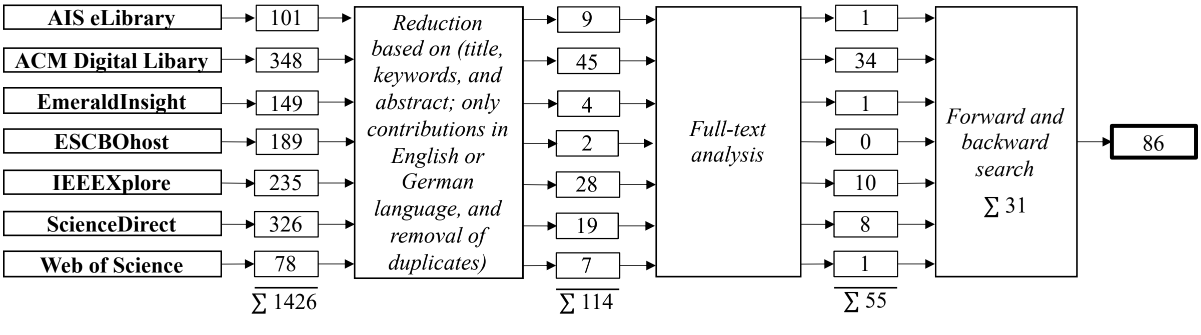


Figure 7.2 Process of Structured Literature Review according to vom Brocke et al. (2015)

We iteratively developed a concept matrix using these 86 contributions by following Möller et al. (2020), including three iterations to develop a theory-based design theory. Please note that to improve readability, we will only provide details on the evaluated design theory of the first design cycle within the following subsection. See Appendix D.2 for a full overview of the iterations of the first design cycle and a visualization of the initial theory-based design theory.

**Adjustment and Evaluation of Theory-based Design Theory.** Following the FEDS framework from Venable et al. (2016), we conducted an artificial summative evaluation to "demonstrate the utility, quality, and efficacy" (Venable et al., 2016, p. 77) of our design theory. First, we conducted two preliminary expert test interviews (TI) to make initial adjustments to the design theory (cf. Appendix D.3). Then, we conducted eleven additional semi-structured expert interviews to evaluate the design theory (Kaiser, 2014). Here, we define an expert as a person who has theoretical and practical knowledge in the field of AI and XAI. In this context, we interviewed German-speaking researchers and practitioners who classified themselves in the role of an end-user (*n*=5) or a developer (*n*=6). All interviews were in the age group of late 20s to mid-40s. See Table 7.1 for more information also on their demographics such as experience with AI.

| I# | Group[1] | Role | Duration[2] | Demographics | | |
|---|---|---|---|---|---|---|
| TI1 | R | Postdoctoral researcher | 32 | | End-user | Developer |
| TI2 | R | Professor | 53 | Experience with AI[3] | 2.4 | 3.8 |
| I1 | P | Head of innovation | 39 | | | |
| I2 | R | Research associate | 40 | | | |
| I3 | R | Research associate | 39 | | | |
| I4 | R | Professor | 53 | Acceptance in AI[4] | 5.0 | 4.0 |
| I5 | R | Research associate | 32 | | | |
| I6 | R | Postdoctoral researcher | 37 | | | |
| I7 | R | Postdoctoral researcher | 34 | | | |
| I8 | P | Head of digitalization | 42 | Trust in AI[4] | 4.0 | 4.0 |
| I9 | P | Process engineer | 49 | | | |
| I10 | P | Data scientist | 61 | | | |
| I11 | P | Data scientist | 48 | | | |

[1] Group: R: Researcher, P: Practitioner; [2] In minutes; [3] Mean in years; [4] Median scale 0-5

Table 7.1 Overview Interviewees and Demographics (First Design Cycle)

We divided the interviews into four phases: *1)* At the beginning, we asked the experts about their demographics and their knowledge and experience in the field of XAI, including their estimation about potential barriers for the adoption of intelligent systems to carry out an initial completeness check of our meta design requirements. *2)* Furthermore, we asked them to classify themselves as either end-users or developers. *3)* We then evaluated our nascent design theory with these experts by presenting the theory-based design theory and openly discussing it with them. Here, we assessed appropriateness and completeness by asking them if they would add, change, or replace any elements. As additional support, we used hypothetical use cases to empower the participants to put themselves in a corresponding situation. *4)* Lastly, we asked them to rate the perceived relevance of the design requirements, design principles, and design features based on a seven-point Likert-scale.

In line with Glaser and Strauss (1967), we transcribed and classified the results by creating inductive and deductive codes. Likewise, according to Flick (2020), we made a qualitative analysis. As a single coder primarily coded the data, we obtained intercoder reliability according to O'Connor and Joffe (2020) through coding a sample of data by an additional coder. Altogether, the interviews comprise 559 minutes of audio material, which is equivalent to 126 pages of transcripts (Herm et al. 2022).

**Initial Design Theory.** Using the relevance rating of the experts, we categorized the design requirements, principles, and features into a user group if the median of the perceived relevance is "slightly important" or above. The following table illustrates the derived and evaluated design requirements, design principles, and design features, as well as the related rating from the experts of the first design cycle. See Appendix D.4 for a graphical overview of the detailed description of the applied steps and the corresponding design theory, in Section 7.4 we will provide a comprehensive explanation of each element of the design theory except DF11[3].

| **Type**[1] | **Description** | **Relevance rating**[2] | |
|---|---|---|---|
| | | End-user | Developer |
| ***DR1*** | Improve explainability | 6.0 | 7.0 |
| ***DR2*** | Support human in own decision-making | 6.5 | 6.5 |
| ***DR3*** | Increase user motivation | 5.0 | 5.0 |
| ***DR4*** | Reduce cognitive effort | 5.5 | 5.0 |
| ***DP1*** | Provide global explanations | 3.5 | 6.5 |
| ***DP2*** | Provide local explanations | 7.0 | 7.0 |
| ***DP3*** | Provide personalized interface design (preference, needs) | 4.0 | 6.0 |
| ***DP4*** | Provide ability to address psychological/emotional factors (intrinsic barriers) | 5.0 | 5.0 |
| ***DF1*** | Provide (technical) information | 5.0 | 6.0 |
| ***DF2*** | Provide (performance) metrics | 6.0 | 7.0 |
| ***DF3*** | Provide input information | 6.0 | 7.0 |
| ***DF4*** | Provide archive of historical decisions | 7.0 | 4.0 |
| ***DF5*** | Provide associative information | 6.0 | 5.0 |
| ***DF6*** | Provide information about decision alternatives | 7.0 | 5.5 |
| ***DF7*** | Provide hypothetical scenarios | 7.0 | 3.5 |
| ***DF8*** | Use visualization techniques | 6.0 | 5.0 |
| ***DF9*** | Incorporate granularity and navigability | 4.5 | 6.0 |
| ***DF10*** | Group and prioritize explanations | 4.0 | 6.0 |
| ***DF11*** | Use anthropomorphic content and designs | 2.0 | 1.5 |

[1] DR = Design requirement; DP = Design principle; DF = Design feature; [2] Median of "*How do you perceive the relevance of [DRx; DPx; DFx]?*" on seven-point Likert scale from 1 - "very unimportant" to 7- "very important".

Table 7.2 Design Requirements, Design Principles, and Design Features of First Design Cycle Including their Relevance

During the expert study, we found that there was improvement potential for our design theory. We used this as input knowledge for the second design cycle.

[3] DF11 characterizes design considerations that represent human-like behaviors such as emojis or chatbots. We discarded DF11 in the second design cycle.

### 7.3.3 Results of Second Design Cycle

**Awareness of Problem, Suggestions, and Design & Development.** In the second design cycle, we refined the nascent design theory. Thereby, we included the input from the expert study of the first design cycle and revisited current XAI and HCI research. That is, we adapted DR1 to "improve intelligibility of system's decision" to emphasize that users must have some access to the logic of ML models for decision support rather than explanations per se. Explanations represent one means to do so as introduced by the subsequent design principles. With this change, we acknowledge that the solution space may actually be larger than only considering explanations. In addition, we assigned DP3 to end-user relevance because a personalized interface design decreases the perceived cognitive effort and increases end-users' motivation to use the EIS for decision-support (Arrieta et al., 2020; Conati et al., 2021). Likewise, we made DF1 only applicable for developers as end-users are often overwhelmed by (technical) details about the used ML model and are not able to comprehend the provided information (Evans et al., 2022; Holzinger et al., 2022). Further, we added the need for visualization technique explanation into DF8, which results from the fact that XAI visualizations are often difficult to understand for non-technical users and thus may hamper decision support (Herm, Wanner, et al., 2021a; Mualla et al., 2022; van der Waa et al., 2021). Lastly, following the first evaluation we discarded DF11, since "*users are used receiving abstract information from different systems, so [they] don't need these anthropomorphic stories*" (I8) and the experts rated the relevance of this design feature as overall unimportant. We could not identify any further aspects through the inclusion of recent XAI-related literature.

**Expert Study, Use Case Application, and Reusability Evaluation.** The evaluation phase in the second design cycle consists of a threefold naturalistic summative evaluation (Venable et al., 2016). First, we conducted a semi-structured expert study, consisting of a pre-test (TU1-2) and the main expert study (U1-6), with four end-users and four developers (Kaiser, 2014) that are part of an AI project in the field of predictive maintenance involving two German companies. Since we observed theoretical saturation, we did not include further expert interviews in our evaluation (Strauss & Corbin, 1994). In line with the first semi-structured expert interview study, we asked the participants about their demographics. Subsequently, we showed the adjusted design theory to them and asked them about their perception and if they would modify, add, or remove any elements within the design theory. Again, all interviews were in the age group of late 20s to mid-40s. See Table 7.3 for more information also on their demographics such as experience with AI. To minimize group bias, we conducted the interviews with each expert individually. Altogether, the interviews comprise 271 minutes of audio.

| U# | Group[1] | Role | Duration[2] | Demographics | | |
|---|---|---|---|---|---|---|
| TU1 | D | Full stack developer | 19 | | End-user | Developer |
| TU2 | E | Process owner | 22 | | | |
| U1 | D | Lead ML developer | 31 | Experience with AI[3] | 1.7 | 6.7 |
| U2 | E | Team lead | 32 | | | |
| U3 | D | ML developer | 46 | Acceptance in AI[4] | 2.0 | 5.0 |
| U4 | D | Head of research | 30 | | | |
| U5 | E | Process engineer | 43 | Trust in AI[4] | 3.0 | 3.0 |
| U6 | E | Process engineer | 48 | | | |

[1] Group: E: End-user, D: Developer; [2] In minutes; [3] Mean in years; [4] Median scale 0-5

Table 7.3 Overview Interviewees and Demographics (Second Design Cycle)

In the second step, we presented the implemented EIS following our design theory to them. We provided them with the opportunity to use this system and think about the corresponding design theory once again. Lastly, we asked them to rate the design principles according to the reusability evaluation criteria of Iivari et al. (2021). We illustrate the use case application of the design theory as well as the results from the evaluation according to Iivari et al. (2021) in Section 7.5.

## 7.4 Final Nascent Design Theory

While contemporary intelligent systems can support users with precise recommendations for decision support, their application is hampered especially in high-stake scenarios due to their lack of explainability (Shin, 2021), which highlights the need for EIS (Herm, Heinrich, et al., 2023). However, due to the novelty of the subject, there is only scarce research on EIS design theories, which are predominantly developed for domain-dependent tasks (e.g., Landwehr et al., 2022). To this end, we propose a broad and domain-independent nascent design theory for EIS, that facilitates the adaptation to different types of use cases (RQ1). Moreover, since XAI research has primarily focused on developers as target group and not the actual end-user of an EIS (van der Waa et al., 2021), we extend this body of knowledge through the differentiated consideration of end-users and developers within the design theory (RQ2). In Figure 7.3, we comprehensively visualize the results of the derived design theory for EIS and its dependencies. We present meta design requirements that form the basis for our design requirements and subsequently for the design principles and design features. In addition, we present the user group relevance for each element. When both user groups deemed an aspect necessary, we marked it as "end-user and developer relevance".

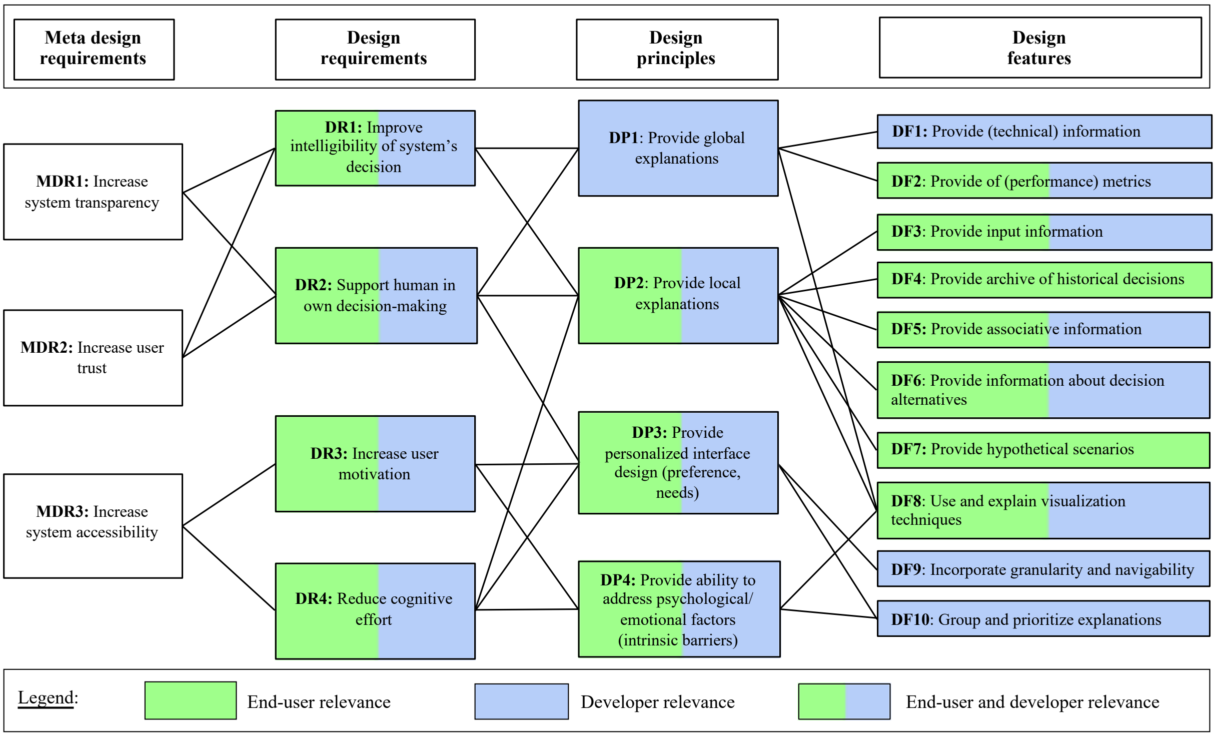

Meta design requirements
Design requirements
Design principles
Design features
MDR1: Increase system transparency
MDR2: Increase user trust
MDR3: Increase system accessibility
DR1: Improve intelligibility of system's decision
DR2: Support human in own decision-making
DR3: Increase user motivation
DR4: Reduce cognitive effort
DP1: Provide global explanations
DP2: Provide local explanations
DP3: Provide personalized interface design (preference, needs)
DP4: Provide ability to address psychological/ emotional factors (intrinsic barriers)
DF1: Provide (technical) information
DF2: Provide of (performance) metrics
DF3: Provide input information
DF4: Provide archive of historical decisions
DF5: Provide associative information
DF6: Provide information about decision alternatives
DF7: Provide hypothetical scenarios
DF8: Use and explain visualization techniques
DF9: Incorporate granularity and navigability
DF10: Group and prioritize explanations
Legend:
End-user relevance
Developer relevance
End-user and developer relevance

Figure 7.3 Visualization of Final Nascent Design Theory

### 7.4.1 Meta Design Requirements and Design Requirements

**Meta Design Requirements.** Baskerville and Pries-Heje (2019) state that DSR-based research must be projectable to propagate design knowledge. Following the argument of Zschech et al. (2020), we used prior design research as input knowledge for our IT artifact (vom Brocke et al., 2020) to gather meta design requirements (Chandra Kruse et al., 2022; Lee & Baskerville, 2003). To this end, we derived the three meta design requirements: system transparency, user trust, and system accessibility, as described below.

*MDR1: Increase System Transparency*. The lack of transparency of the system is a significant barrier to the adoption of AI in practice (Wanner, Herm, et al., 2022a), as users are incapable of comprehending a models' internal logic or the reasoning behind a models' recommendation, rendering EIS for decision support inefficacious (Arrieta et al., 2020; Sardianos et al., 2021). Consequently, system transparency can be seen as a prerequisite for enabling a trustworthy user interaction with the EIS (Landwehr et al., 2022; Samek et al., 2017; Shin et al., 2020). Increasing system transparency also results in a shift in user perception making decisions more conscious (Chazette & Schneider, 2020). Simultaneously, system transparency increases the acceptance of using an EIS in work environments (Arrieta et al., 2020; Bhatt et al., 2020).

*MDR2: Increase User Trust*. The acceptance of EIS and, consequently, their adoption depends on trust in the results a system provides (Carvalho et al., 2019; Thiebes et al., 2021; Wanner, Herm, et al., 2022a). Especially for critical decisions, users have to rely on these results to make an informed decision (Choi & Ji, 2015; Herm, Wanner, et al., 2021a). Consequently, it is only possible to establish initial trust in a (new) intelligent system if there are no unknown risk factors present or users are not afraid of losing control due to a lack of information about the results (McKnight et al., 2011; Slade et al., 2015). However, while this may lead to the perception that trust is influenced by system transparency (e.g., Schmidt et al., 2020), empirical research has proven that there is no significant direct effect of system transparency on the perceived level of trust (Cramer et al., 2008; Wanner, Herm, et al., 2022a). Lastly, the EIS must take into account several influencing factors, such as keeping humans in the loop during system development, to ensure that users perceive the EIS as a competent decision support system for their use case, leading to increased user trust and thus acceptance of EIS (Mualla et al., 2022; Shin, 2021).

*MDR3: Enhance System Accessibility*. Crucial in using EIS is the transfer of knowledge towards the user (Berger et al., 2021). Here, a fluent and non-restrictive interaction must be ensured if recommendations differ from user expectations due to the user's reservations or domain knowledge (Chander et al., 2018; Meth et al., 2015). The use of XAI transfer techniques to ensure an interaction enables the increase of acceptance and the improvement of the intrinsic attitude towards the systems (Sokol & Flach, 2020). This also includes the adaptation of the system's recommendation (Ferreira & Monteiro, 2020) as well as the ability to generate causalities for following actions (Liao et al., 2020).

**Design Requirements.** Design requirements describe how general meta design requirements from related fields of the IT artifact's topic should be addressed in a way that allows for an evaluation of a developed design solution (Baskerville & Pries-Heje, 2019; vom Brocke et al., 2020). During our structured literature review, we scrutinized the meta requirements unearthed initially and operationalized them into more output-related design requirements. We ensure their validity and completeness through the expert interviews in the first and second design cycle (see Section 7.3.2 and 7.3.3). We describe them in the following.

*DR1: Improve Intelligibility of System's Decision.* The use of EIS empowers end-users and developers to compare their intrinsic mental model and consequently their expectations with the recommendation of an EIS. So, when user's expectations conform with the recommendation explanations, their willingness to use the system in practice increase (Carvalho et al., 2019; Malhi et al., 2020). In doing so, EIS must provide recommendations with associated accounts in a way that adequately supports users during the decision process (Longo et al., 2020).

*DR2: Support Human in own Decision-Making.* To support and improve a human's own decision-making by providing accounts for predictions, those need to be enriched with domain knowledge and situation-specific context (Dikmen & Burns, 2022). Providing such accounts increases the user's confidence during the decision-making process (Evans et al., 2022). Once end-users can understand the recommendation, they are skilled in making sound decisions. This is also true for developers when they intent to understand the internal processing logic of the model (Malhi et al., 2020).

*DR3: Increase User Motivation.* In case users are extrinsically or intrinsically motivated to use the EIS, the degree of motivation increases, and consequently their system acceptance will increase as well (Stumpf et al., 2009). EIS should therefore incorporate features that rise the motivation of the end-users using an EIS for decision support (Ferreira & Monteiro, 2020). This could include different paradigms, as they are directly related to user expectations, leading to a well-perceived user experience (Nunes & Jannach, 2017).

*DR4: Reduce Cognitive Effort*. If users require a long time to understand recommendation and their accounts, for example if they are counterintuitive or complex, it may be perceived as cognitively demanding and lead to frustration and rejection (Fürnkranz et al., 2020). It is worth noting that the perceived cognitive load may vary by an individual due to context-specific circumstances (Oviatt, 2006). Hence, EIS must provide accounts in a manner that reduces the cognitive effort of users (Zschech et al., 2020).

### 7.4.2 Design Principles and Corresponding Design Features

Design principles and design features are intended to explain how derived design requirements can be addressed in a design theory (Baskerville & Pries-Heje, 2019; vom Brocke et al., 2020). In the following, we present the final and validated design principles and design features of our nascent design theory. For each design principle, we first provide a comprehensive rationale, followed by a tabular formulation of the design principle using the design principle schema established by Gregor et al. (2020) (see Table 7.4 - Table 7.7). Lastly, we present corresponding design features to illustrate how the design principles can be implemented into an associated instantiation (Gregor et al., 2020; Seidel et al., 2018).

**DP1: Principle of Global Explanations.** With an EIS, users can understand the general behavior of an intelligent system within the decision-making process and thereby comprehend the inner logic of the model to a certain level. For this purpose, the internal logic of the system must be represented in a user-friendly manner in order for the developer to understand the ML model (Das & Rad, 2020). It is essential to grasp the capabilities of the model beforehand because *"it is pointless using an ML model that makes completely insufficient predictions"* (I5). Furthermore, Rudin (2019) calls for per-se interpretable but performance-wise appropriate ML models, when deploying intelligent systems in highly critical environments as this may be necessary due to regulatory constraints (Vale et al., 2022).

| Design principle title | Provide global explanations |
|---|---|
| Aim, implementer, and users | For the EIS (enactor) to provide global explanations for developers (implementors) enabling them to understand the general behavior of the EIS's ML model for debugging and optimization purposes (aim) |
| Context | During implementation and during usage of EIS |
| Mechanism | Ensures that developers comprehend the inner decision logic of the EIS's ML model |
| Rationale | Inner decision logic of ML model must be transparent for evaluation purposes or due to regulatory constraints |

Table 7.4 Principle of Global Explanations

On the one hand, (technical) information (DF1), such as system capabilities of the ML model, (hyper-) parameters, and information about the training data and training history, must be provided to ensure lawfulness and fairness of the training process (Hepenstal & McNeish, 2020; Kaur et al., 2022) (U3; U4). This is primarily relevant to developers, since if the logic of an ML model *"is far above the level of knowledge, then it's all magic [for them]"* (U5). Furthermore, (performance) metrics must be provided (DF2) to quantitatively evaluate the decision support capability of an EIS (e.g., accuracy, *F1*-score, decision certainty) (Glomsrud et al., 2019; Sun et al., 2022).

**DP2: Principle of Local Explanations.** To render the recommendation of individual observations explicable, an EIS must provide local explanations. This allows (end-)users to validate or adjust their own expectations if certain recommendations *"fit somewhere in [their] expectations"* (I8). This internal process can assist in resolving cognitive restrictions (Hepenstal & McNeish, 2020). Local explanations complement global explanations and make recommendations easier to understand. Consequently, they are necessary, especially for end-users and novices (Hohman et al., 2019; Mohseni et al., 2021). Moreover, our research shows that this representation is also relevant for developers, since they *"[..] can use local explanations to analyze the pre-trained models for reliability by manipulating data and seeing how the model's outputs change"* (U1). This becomes specially important if transfer-learned models are used.

| Design principle title | Provide local explanations |
|---|---|
| Aim, implementer, and users | For the EIS (enactor) to provide local explanations for end-users (users) and developers (implementors) to understand the reason for a concrete EIS recommendation (aim) |
| Context | During usage of EIS |
| Mechanism | Ensures that developers and end-users comprehend the reasoning of an EIS's recommendation |
| Rationale | Users can only make an appropriate decision if they can trace the reasoning process by comparing their expectations for a particular recommendation with those of the EIS |

Table 7.5 Principle of Local Explanations

The EIS must display related input data to enable end-users and developers to trace the specific data input used (DF3) for the recommendations and the resulting data output (Liao & Varshney, 2022; Nunes & Jannach, 2017). This is also true for associative information (DF5) to understand causal decision chains of the EIS in a user-friendly way (Haynes et al., 2009; Nunes & Jannach, 2017). This also includes process diagrams, graphical explanations (e.g., correlation matrixes) (U4), and look-up glossaries to understand complex issues in time-constrained situations (U1; U3). Similarly, filterable historical information about past decisions (DF4), including the used visualizations, must be displayed (Atkinson et al., 2020) (U3) as users can form their decision based on previous data and receive information about the decision-making process when legal issues arise (e.g., in high-risk cases) (U1). Moreover, additional information about possible decision alternatives (DF6) must be presented especially in cases of low decision certainty (Nor et al., 2022). In addition, providing input options to customize the input data allows developers to validate and debug an ML model according to (regulatory) unit tests (U3). Lastly, providing hypothetical scenarios (DF7), for example simulations to end-users, would reveal the potential impacts of the provided recommendations (Amershi et al., 2019).

**DP3: Principle of Personalized Interface Design.** When using EIS, different user groups have varying preferences and needs for information presentation (Arrieta et al., 2020; Bhatt et al., 2020). Only flexible customization of system components can ensure user comprehension and consequently increase adoption of an EIS (Conati et al., 2021; Mualla et al., 2022). In addition, it is essential to pay attention to reducing the cognitive effort for the user when designing individual EIS components (Carvalho et al., 2019; Cheng et al., 2019). That is, established UI design guidelines (e.g., Shneiderman & Plaisant, 2016), and best practices from numerous application domains must be consulted (Amershi et al., 2019) to avoid being *"a confusing system with a thousand numbers and variables and layers"* (I8). While developers primarily identified this requirement, it is apparent that this is meant to support end-users.

| **Design principle title** | **Provide personalized interface design (preference, needs)** |
|---|---|
| Aim, implementer, and users | For the EIS (enactor) to provide the end-users (users) and developers (implementors) with a personalized interface design that meets their preferences and needs (aim) |
| Context | During usage of EIS |
| Mechanism | Ensures that users are not cognitively overwhelmed when using the EIS |
| Rationale | A personalized interface design reduces perceived cognitive effort and consequently increases the system's accessibility |

Table 7.6 Principle of Personalized Interface Design (Preference, Needs)

To enable personalized adaptation, several visualization techniques, for example XAI-based argumentations, should be used (DF8) (Jesus et al., 2021), including justifications for why these types of visualizations are used to gain the trust of end-users and developers (U1). Therein, these visualizations should offer different levels of granularity in information presentation (DF9) and should be independently adjustable by users (Amershi et al., 2019). An example would be zooming into an explanation *"so [it] can be successively traced further and further in detail"* (I2). Similarly, it is necessary to group and prioritize (DF10) individual explanation components for specific user groups to enable adequate presentation and consequently not overwhelm users cognitively (Schneider & Handali, 2019).

**DP4: Principle of Ability to Address Psychological/Emotional Factors.** For successful interaction with end-users and developers, the EIS should address their emotions, beliefs, and expectations to achieve the intended goals (Arrieta et al., 2020). This includes situational representations to support the user emotionally and psychologically (Kocielnik et al., 2019), thus addressing their *"[..] personal idiosyncrasies and preferences so that they are satisfied with the results"* (I1). This improved interaction increases the perceived ease of use, leading to higher adoption of the EIS (Ferreira & Monteiro, 2020).

| Design principle title | Provide ability to address psychological/emotional factors (intrinsic barriers) |
|---|---|
| Aim, implementer, and users | For the EIS (enactor) provides the ability to address psychological and emotional factors (aim) of end-users (user) and developers (developers) |
| Context | During usage of EIS |
| Mechanism | Increase the perceived ease of use for the EIS |
| Rationale | Addressing psychological and emotional factors to reduce users' intrinsic barriers leads to greater user motivation and system accessibility resulting in an improved EIS adoption |

Table 7.7 Principle of Ability to Address Psychological/Emotional Factors (Intrinsic barriers)

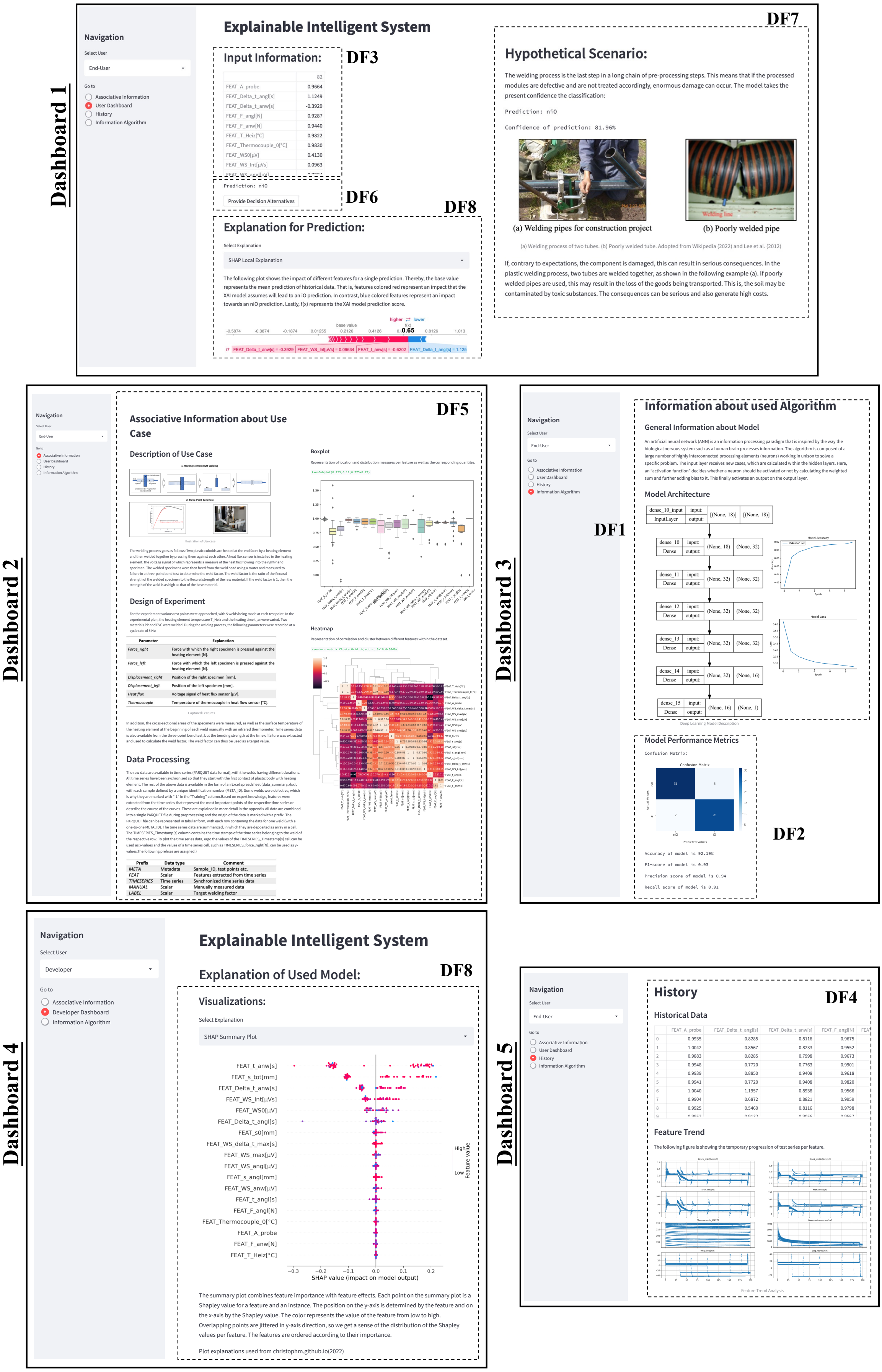
The incorporation of multiple visualization techniques (DF8) enables users to handle individual emotions, such as stress, when faced with time-critical decisions by allowing them to customize the UI to their individual preferences (Chromik & Butz, 2021). In addition, end-users must be able to reexamine textual explanations to the corresponding visualizations, in case of interpretational uncertainties during process execution. Besides, end-users require training prior to using EIS to reduce the cognitive effort required (U1; U2).

## 7.5 Evaluation of the Final Nascent Design Theory

Overall, the naturalistic summative evaluation in the last design cycle consists of a threefold evaluation following the FEDS framework of Venable et al. (2016). While we demonstrate the qualitive expert study and their findings in Section 7.3.3 and 7.4, in this subsection, we describe the instantiation of the nascent design theory using an EIS prototype implemented in a production-ready environment, including a subsequent reusability evaluation (Iivari et al., 2021) through use-case-related employees.

The use case is part of an AI-based predictive maintenance project performed by the two German companies ROBOUR Automation GmbH and SKZ - German Plastics Centre. In this project, heat-flux sensors track plastic welding processes of polypropylene homopolymer pipes (Lambers & Balzer, 2022). This welding process is used when setting up infrastructural underground pipes for freshwater or wastewater supply. The application of poorly welded pipes can lead to the loss of the transported goods and, consequently to the contamination of the soil with potential toxic substances.

According to tracked senor data, a multi-layer DNN predicts the ratio between the flexural strength of the welded specimen and the raw materials, whereby a ratio lower than 0.7 indicates an insufficient welding process. Taking the DNN's ability to outperform experts and the relatively low acceptance of DNNs in this high-risk scenario into account, the application of an EIS that supports the decision-making process of experts is promising for evaluating our nascent design theory. As an in-depth pre-test with one developer and one end-user during EIS development revealed, splitting the EIS into multiple dashboards reduces the cognitive load of end-users and developers. As illustrated in Figure 7.4, the implemented EIS consists of five different dashboards. Following the proposed nascent design theory, the user specific dashboards are only accessible to the certain user groups.



## Explainable Intelligent System

### Input Information:

| | 82 |
|---|---|
| FEAT_A_probe | 0.9664 |
| FEAT_Delta_t_angl[s] | 1.1249 |
| FEAT_Delta_t_anw[s] | -0.3929 |
| FEAT_F_angl[N] | 0.9287 |
| FEAT_F_anw[N] | 0.9440 |
| FEAT_T_Heiz[°C] | 0.9822 |
| FEAT_Thermocouple_0[°C] | 0.9830 |
| FEAT_WS0[µV] | 0.4130 |
| FEAT_WS_Int[µVs] | 0.0963 |
| FEAT_WS_angl[µV] | [illegible] |

```
Prediction: niO
```

Provide Decision Alternatives

### Explanation for Prediction:

Select Explanation: SHAP Local Explanation

The following plot shows the impact of different features for a single prediction. Thereby, the base value represents the mean prediction of historical data. That is, features colored red represent an impact that the XAI model assumes will lead to an iO prediction. In contrast, blue colored features represent an impact towards an niO prediction. Lastly, f(x) represents the XAI model prediction score.

### Hypothetical Scenario:

The welding process is the last step in a long chain of pre-processing steps. This means that if the processed modules are defective and are not treated accordingly, enormous damage can occur. The model takes the present confidence the classification:

```
Prediction: niO

Confidence of prediction: 81.96%
```

(a) Welding pipes for construction project

(b) Poorly welded pipe

(a) Welding process of two tubes. (b) Poorly welded tube. Adopted from Wikipedia (2022) and Lee et al. (2012)

If, contrary to expectations, the component is damaged, this can result in serious consequences. In the plastic welding process, two tubes are welded together, as shown in the following example (a). If poorly welded pipes are used, this may result in the loss of the goods being transported. This is, the soil may be contaminated by toxic substances. The consequences can be serious and also generate high costs.

## Associative Information about Use Case

### Description of Use Case

Illustration of Use case

The welding process goes as follows: Two plastic cuboids are heated at the end faces by a heating element and then welded together by pressing them against each other. A heat flux sensor is installed in the heating element, the voltage signal of which represents a measure of the heat flux flowing into the right-hand specimen. The welded specimens were then freed from the weld bead using a router and measured to failure in a three-point bend test to determine the weld factor. The weld factor is the ratio of the flexural strength of the welded specimen to the flexural strength of the raw material. If the weld factor is 1, then the strength of the weld is as high as that of the base material.

### Design of Experiment

For the experiment various test points were approached, with 5 welds being made at each test point. In the experimental plan, the heating element temperature T_Heiz and the heating time t_anwere varied. Two materials PP and PVC were welded. During the welding process, the following parameters were recorded at a cycle rate of 5 Hz:

| Parameter | Explanation |
|---|---|
| *Force_right* | Force with which the right specimen is pressed against the heating element [N]. |
| *Force_left* | Force with which the left specimen is pressed against the heating element [N]. |
| *Displacement_right* | Position of the right specimen [mm]. |
| *Displacement_left* | Position of the left specimen [mm]. |
| *Heat flux* | Voltage signal of heat flux sensor [µV]. |
| *Thermocouple* | Temperature of thermocouple in heat flow sensor [°C]. |

Captured Features

In addition, the cross-sectional areas of the specimens were measured, as well as the surface temperature of the heating element at the beginning of each weld manually with an infrared thermometer. Time series data is also available from the three-point bend test, but the bending strength at the time of failure was extracted and used to calculate the weld factor. The weld factor can thus be used as a target value.

### Data Processing

The raw data are available in time series (PARQUET data format), with the welds having different durations. All time series have been sychronized so that they start with the first contact of plastic body with heating element. The rest of the above data is available in the form of an Excel spreadsheet (data_summary.xlsx), with each sample defined by a unique identification number (META_ID). Some welds were defective, which is why they are marked with "-1" in the "Training" column.Based on expert knowledge, features were extracted from the time series that represent the most important points of the respective time series or describe the course of the curves. These are explained in more detail in the appendix.All data are combined into a single PARQUET file during preprocessing and the origin of the data is marked with a prefix. The PARQUET file can be represented in tabular form, with each row containing the data for one weld (with a one-to-one META_ID). The time series data are summarized, in which they are deposited as array in a cell. The TIMESERIES_Timestamp[s] column contains the time stamps of the time series belonging to the weld of the respective row. To plot the time series data, ergo the values of the TIMESERIES_Timestamp[s] cell can be used as x-values and the values of a time series cell, such as TIMESERIES_force_right[N], can be used as y-values.The following prefixes are assigned:)

| Prefix | Data type | Comment |
|---|---|---|
| *META* | Metadata | Sample_ID, test points etc. |
| *FEAT* | Scalar | Features extracted from time series |
| *TIMESERIES* | Time series | Synchronized time series data |
| *MANUAL* | Scalar | Manually measured data |
| *LABEL* | Scalar | Target welding factor |

### Boxplot

Representation of location and distribution measures per feature as well as the corresponding quantiles.

```
AxesSubplot(0.125,0.11;0.775x0.77)
```

### Heatmap

Representation of correlation and cluster between different features within the dataset.

```
<seaborn.matrix.ClusterGrid object at 0x16c0c56d0>
```

## Information about used Algorithm

### General Information about Model

An artificial neural network (ANN) is an information processing paradigm that is inspired by the way the biological nervous system such as a human brain processes information. The algorithm is composed of a large number of highly interconnected processing elements (neurons) working in unison to solve a specific problem. The input layer receives new cases, which are calculated within the hidden layers. Here, an "activation function" decides whether a neuron should be activated or not by calculating the weighted sum and further adding bias to it. This finally activates an output on the output layer.

### Model Architecture

dense_10_input | InputLayer | input: | output: | [(None, 18)] | [(None, 18)]
dense_10 | Dense | input: | output: | (None, 18) | (None, 32)
dense_11 | Dense | input: | output: | (None, 32) | (None, 32)
dense_12 | Dense | input: | output: | (None, 32) | (None, 32)
dense_13 | Dense | input: | output: | (None, 32) | (None, 32)
dense_14 | Dense | input: | output: | (None, 32) | (None, 16)
dense_15 | Dense | input: | output: | (None, 16) | (None, 1)
Model Accuracy
Validation Set
Accuracy
Epoch
Model Loss
Loss
Epoch

Deep Learning Model Description

### Model Performance Metrics

```
Confusion Matrix:
```

```
Accuracy of model is 92.19%

F1-score of model is 0.93

Precision score of model is 0.94

Recall score of model is 0.91
```

## Explainable Intelligent System

### Explanation of Used Model:

#### Visualizations:

Select Explanation: SHAP Summary Plot

The summary plot combines feature importance with feature effects. Each point on the summary plot is a Shapley value for a feature and an instance. The position on the y-axis is determined by the feature and on the x-axis by the Shapley value. The color represents the value of the feature from low to high. Overlapping points are jittered in y-axis direction, so we get a sense of the distribution of the Shapley values per feature. The features are ordered according to their importance.

Plot explanations used from christophm.github.io(2022)

## History

### Historical Data

| | FEAT_A_probe | FEAT_Delta_t_angl[s] | FEAT_Delta_t_anw[s] | FEAT_F_angl[N] |
|---|---|---|---|---|
| 0 | 0.9935 | 0.8285 | 0.8116 | 0.9675 |
| 1 | 1.0042 | 0.8567 | 0.8233 | 0.9552 |
| 2 | 0.9883 | 0.8285 | 0.7998 | 0.9673 |
| 3 | 0.9948 | 0.7720 | 0.7763 | 0.9901 |
| 4 | 0.9939 | 0.8850 | 0.9408 | 0.9618 |
| 5 | 0.9941 | 0.7720 | 0.9408 | 0.9820 |
| 6 | 1.0040 | 1.1957 | 0.8938 | 0.9566 |
| 7 | 0.9904 | 0.6872 | 0.8821 | 0.9959 |
| 8 | 0.9925 | 0.5460 | 0.8116 | 0.9798 |
| 9 | 0.9952 | 0.9132 | 0.9056 | 0.9567 |

### Feature Trend

The following figure is showing the temporary progression of test series per feature.

Feature Trend Analysis

Figure 7.4 Overview of the Different Dashboards of the EIS Instantiation

These five dashboards comprise the different views for the end-users and the developers of the EIS and consequently postulate a meaningful representation of the derived nascent design theory. The first dashboard provides an overview of the input information (DF3) from the tracked sensors, the corresponding prediction from the ML model, and a (local) explanation of this prediction and thus the resulting decision recommendation (DF8). By clicking on a button below the shown prediction (DF6), the dashboard highlights decision alternatives. In conjunction with the prediction, a hypothetical scenario is presented to the end-user (DF7). The second dashboard contains the associative information for end-users and developers, including (graphical) information about the related sensors, process execution, and data processing steps (DF5). The third dashboard provides (technical) information about the EIS, including a comprehensive description, the applied ML model architecture, information about ML model training (DF1), and the corresponding performance metrics (DF2). Comparable to the first dashboard, the fourth dashboard addresses DF8 by providing (global) explanations of the ML model for the developer. The last dashboard contains an archive of historical decisions including the associated sensor data and its history (DF4). By dividing the EIS into multiple dashboards, we ensure granularity and navigability throughout the EIS (DF9). Similarly, within the first and fourth dashboards, we provide drop-down menus that allow end-users and developers to group and prioritize explanations concurring to their own preferences (DF10).

We asked the experts using the system to speak unreservedly about their impressions and whether they would change, add, or remove any elements. In doing so, we qualitatively analyzed their feedback to identify if this would affect the proposed design theory. In this regard, we noticed that our experts, except for occasional comments, are satisfied with this EIS instantiation. Here a developer stated, that *"The system is well designed and offers all necessary functions to assist me during my work"* (U3) or "*I would like to use the system in our production. As a minor improvement, more technical information about data gathering and preprocessing would be appreciated, at least for our use case*" (U1). Likewise, an end-user concluded *"The system seems to offer a solid and comprehensible approach to support end-users."* (U5), while another one claimed that *"At first, I perceived the dashboard as complex, which is why I believe that a short introduction is necessary, especially for new end-users. Afterwards, the system appears complete and well designed."* (U6).

Lastly, we evaluate the derived design principles by following the reusability evaluation propositions for DSR-based design principles of Iivari et al. (2021). We performed this quantitative evaluation at the end to verify that users are aware of the implemented EIS and thus of our nascent design theory, as real-world use of an EIS may reveal additional changes to the proposed design theory. To do so, we asked the participants to rate the constructs of accessibility, importance, novelty & insightfulness, actability & guidance, as well as effectiveness through multiple questions constructs on a 5-Point Likert scale (1 = strongly disagree, 5 = strongly agree). We conducted the evaluation anonymously via an online survey, to not force biases. The following Figure illustrate the corresponding results. Please see Appendix D.5 for the questionnaire.

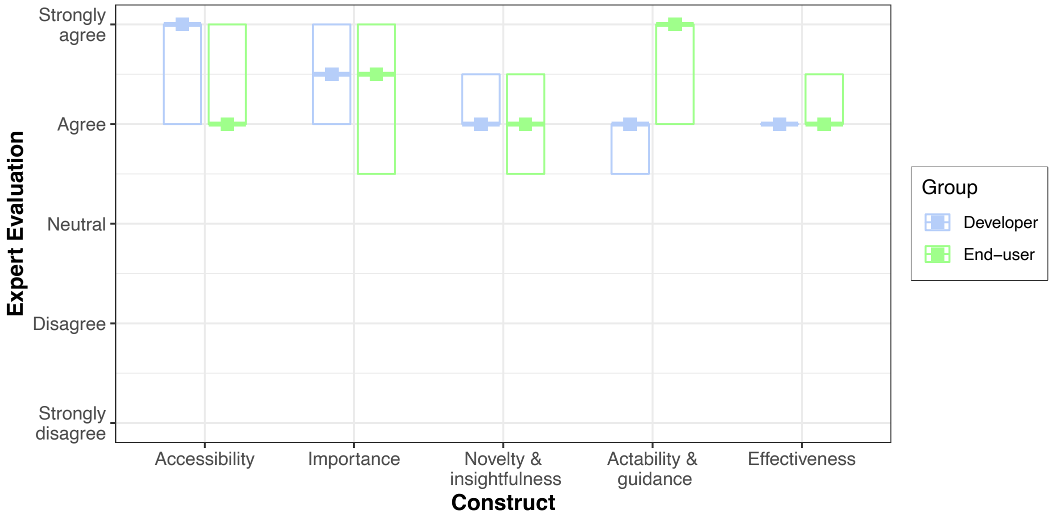


Figure 7.5 Results of reusability evaluation according to Iivari et al. (2021)

Since we used multiple questions per construct, we calculated the median for each construct and expert group. Then, we used the median, minimum, and maximum of this data for the overall construct evaluation per user group (Boone & Boone, 2012).

This results in overall positive expert feedback. As, the experts considered no further changes within our design theory, as *"the design theory seems complete"* (U4) and had a positive perception of the design principles (cf. Figure 7.5), we consider our nascent design theory ready-to-use.

## 7.6 Discussion of Findings

### 7.6.1 Discussion and Implications

**Discussion.** There are several contributions dealing with design approaches for EIS (e.g., Bunde, 2021; Landwehr et al., 2022; Meske & Bunde, 2022; Schemmer et al., 2022) to create a hybrid intelligence as Dellermann et al. (2019) have called it.

While we conclude, that intelligibility (DR1) expressed through global and local explanation is both important, Meske and Bunde (2022) and Landwehr et al. (2022) are limited to local explanations; only Schemmer et al. (2022) describe the need for providing an overall explainability. Further, recent DSR-based XAI contributions (e.g., Landwehr et al., 2022; Meske & Bunde, 2022) do not include the support of own decision-making (DR2) within their design theory. In contrast, these research findings are primary derived from the HCI field (e.g., Dikmen & Burns, 2022) and demonstrate the need for an interdisciplinary design theory. The same applies for increasing the user motivation (DR3) (e.g., Ferreira & Monteiro, 2020) and reducing cognitive effort (DR4) (e.g., Oviatt, 2006). Moreover, while we observed the need for increasing user motivation and reducing cognitive effort within recent literature, end-users and developers did barely envision this need, when talking about both design requirements on a theoretical basis. Nonetheless, we were able to uncover, during EIS application, that users still require design principles related to DR3 and DR4.

In terms of the derived design principles, our study also extends the current body of design knowledge. That is, while recent research targets end-users and is thus limited to addressing local explainability (e.g., Bunde, 2021; Landwehr et al., 2022; Meske & Bunde, 2022), our nascent design theory does not only include local explainability (DP2) but also incorporates global explainability (DP1) for developers. In addition, while theoretical contributions (e.g., Mohseni et al., 2021) are mainly assigning DP2 to end-users, our research indicate, that developers also benefit from using local explanations. This extension of design science knowledge based on our research applies for DP3 and DP4 as well. While personalized interface design (DP3) is considered important (Conati et al., 2021), during our first design cycle only developers confirmed this finding. Nonetheless, during the second design cycle, end-users also confirmed the importance of DP3. Regarding the consideration of psychological/emotional factors (DP4) for end-users and developers our findings are in line with recent research (Arrieta et al., 2020).

Lastly, matching theoretical foundations with our research findings also reveals differences. Comparing our findings with related design theories (Bunde, 2021; Landwehr et al., 2022; Meske & Bunde, 2022; Schemmer et al., 2022) shows that only four out of our ten design features have been mentioned earlier. This includes design features such as providing input information (DF3) and historical information (DF4) as well as using explanation techniques (DF8) and incorporating granularity and navigability (DF9). Six out of our ten design features were derived from interdisciplinary contributions. Comparing the targeted user groups from theory with our findings uncovers further distinctions: while the six design features DF1 (Hepenstal & McNeish, 2020), DF2 (Sun et al., 2022), DF3 (Nunes & Jannach, 2017), DF4 (Atkinson et al., 2020), DF6 (Nor et al., 2022), and DF7 (Amershi et al., 2019) are in line with recent interdisciplinary research, four design features are not. Although previous research consider DF5 (Haynes et al., 2009), DF8 (Jesus et al., 2021), DF9 (Amershi et al., 2019), and DF10 (Schneider & Handali, 2019) for both user groups, our evaluations reveal, that DF5 and DF8 have a purely unilateral preference towards end-users and DF9 and DF10 towards developers. While our theory-based initial design theory, drawing on scholarly literature, included the need for anthropomorphic design language, as in chatbots, to reduce adaptation barriers (Weitz et al., 2019), we did not include this design principle in our final nascent design theory because our experts rejected this, as non-novice users are accustomed working with abstract information, which leads to undesirable complexity within the EIS. We could not find evidence with the EIS instantiation either. We acknowledge though that DF11 may be relevant in situation where end-users possess no technical skills at all (e.g., private use of intelligent assistance services, chatbots, etc.).

**Theoretical Implications.** DSR seeks to develop prescriptive design knowledge by developing and evaluating novel IT artifacts to solve practical problems (Hevner et al., 2004). Corresponding to mode 3B of Drechsler and Hevner (2018)'s design theorizing modes, we derived a nascent design theory that provides explicit prescriptions for entity realization for a class of explainable AI-based DSS, so-called EIS. Further, following Gregor and Hevner

(2013)'s DSR knowledge contribution framework, we contribute with a nascent design theory including (meta) design requirements, design principles, and design features (level 2 contribution) and a situated implementation of the IT artifact (level 1 contribution). Since we applied two design cycles, the design theory can be considered rigorous and consequently can serve as input for future research (Hevner, 2021).

Looking at previous design science research reveal that the integration of AI in DSS leads to intelligent systems that are capable of supporting users in their decision-making process (Janiesch, Zschech, et al., 2021). However, due to their focus on user performance, these systems are primarily developed for low-stake use cases wherein users do not rely on comprehending the reasoning of a ML model (e.g., Zschech et al., 2020) as an incorrect recommendation has no significant impact on humans or the environment (Rudin, 2019). In contrast, utilizing these systems in high-stake use cases, wherein incorrect decisions may endanger human lives or may have vast consequences, designing intelligent systems require the explicit consideration of techniques such as XAI to make the ML model's behavior traceable (Mohseni et al., 2021), resulting in the need of EIS applications (Herm, Heinrich, et al., 2023). Hence, recent research has already developed first design principles for domain-dependent EIS development (e.g., Landwehr et al., 2022). To extend this sparse research, we position our research as a broad design theory for EIS development (Chandra Kruse et al., 2022), that distinguishes itself from recent research:

First, to best of our knowledge, there is no other scholarly contribution providing a nascent design theory for a domain-independent EIS including an instantiation. That is, compared to current research contributions that develop DSR-based design principles for specific use cases (e.g., Bunde, 2021; Landwehr et al., 2022; Meske & Bunde, 2022), our research provides a first-hand design knowledge as a starting point for adoption and refinement for all types of decision support use cases. As an example, applying our design theory to a healthcare use case may lead to the consideration of additional factors to assist physicians in high-stake cases when human lives could depend on a decision.

Second, in our design theory we consider recent findings from design-based XAI, interdisciplinary XAI, and HCI research. To this end, our design theory compromises not only technical XAI aspects but also socio-technical aspects that origin from the field of HCI and psychology. In doing so, we take into account the diverse facets of human-agent interaction that unfold due to XAI's nature (Miller, 2019).

Third, our design theory also includes the consideration of different user groups. Since previous XAI research has not sufficiently addressed the integration of end-users, we have focused our design theory not only on the developer and ML expert, but also on the end-users. However, we recognize that there is no one-size-fits all EIS. That is, during the interview studies, we mostly rely on end-users that are domain-expert but mostly unskilled in terms of ML. During our qualitative research, we identified this type of end-user as widely spread. Hence, we take our design theory as a starting-point for the consideration of end-users, with the potential need of design theory adjustment, when it comes to specific use cases, for instance, when novice users perform tasks.

**Practical Implications**. During our research, we found that XAI is not a silver bullet. That is, in practice the use of XAI does not automatically ensure utilization of EIS. Even when using XAI-based transfer techniques, novice users need to be empowered to use these EIS and thereby develop a widespread understanding. This is especially true for high-stake scenarios, where recommendations and explanations must be comprehensible to users at all times. In addition, this can (psychologically) support users, when they compare explanations with their own expertise and expectations.

Besides, companies should discuss the required cognitive effort with their end-users. Surprisingly, as we particularly focused on reducing this effort, end-users told us, that using this EIS seemed quite complicated for them at first. Consequently, conducting training before using an EIS guides these novice users and similarly reduces the required cognitive effort, as they become familiar with the system.

Nevertheless, we revealed that some end-users do not only want to comprehend the recommendation but also want to determine the quality of the ML model based on metrics such as accuracy, $F1$-score, or decision certainty to critically evaluate the provided recommendation. In contrast, these users are not interested in understanding how the models operate. Instead, we have found that end-users trust the model development and selection by the EIS designers. Conversely, talking to the experts shed light on the correlation between AI knowledge and trust in AI. This means that AI experts tend to have more reservations about AI because they are aware of potential difficulties during selecting, training, and developing ML models.

Finally, in the second evaluation phase of the design cycle, we found that experts not only view the implemented EIS as an opportunity to deploy AI into practice in an explainable fashion but also to use the data-driven generated knowledge to train end-users for use case execution. In doing so, we noticed that the utilization of an EIS fosters the acceptance of AI and allows experts to view AI as trustworthy.

### 7.6.2 Limitation and Future Research

Although we ensured scientific rigor by applying established DSR guidelines (Gregor & Hevner, 2013; Iivari et al., 2021; Vaishnavi & Kuechler, 2007), we noticed certain limitations in our research. This includes the two expert studies we conducted while adjusting and evaluating the proposed nascent design theory, where experts already had several years of experience in the field of AI. Hence, we must assume that the results could differ for novice users. Further, all interviewees were early to mid-career employees. Hence, our results are more likely to apply for this age group than for mid-50s and older. We conducted the last evaluation phase based on an exemplary and thus context-dependent scenario, which is why the results could vary in other scenarios. Also, end-users did not have to make time-critical decisions in the use case application. With this in mind, we assume that the design of EIS systems may differ, when there are additional technical, privacy, or cognitive constraints to consider. Lastly, we did not test all 15 possible design principles configuration to ensure design principle expressiveness (Janiesch et al., 2020). Our design theory represents a nascent design theory, it is not yet a fully developed grand theory.

During our research, we noticed several shortcomings in current XAI literature and XAI applications in practice leading to novel research opportunities. As part of a DSR-based research project, we provide research prospects that future research projects can use as a starting point and thus as meta design requirements for their work (Peffers et al., 2007).

Contrary to existing theoretical assumptions (e.g., Liao et al., 2020), global explanations are not necessarily suitable for developers, as they as well may be cognitively overwhelmed. For future research, it is therefore necessary not only to investigate interactive XAI-based explanations with different levels of granularity for end-users but also to consider developers as a relevant user group. This is especially true since the algorithmic output of common XAI tools can be challenging for these user group (Herm, Wanner, et al., 2021a; van der Waa et al., 2021), as not all developers have a data science related background.

Connected to this, we found that all experts emphasized the importance of adequate XAI-based explanations during the evaluation of the use case. However, none of these experts were able to provide dedicated requirements for such an explanation. Consequently, research should target the derivation of frameworks and guidelines for selecting context specific and appropriate XAI explanation types to assist decision-making. This includes evaluation metrics and standards to define the quality of an explanation. This evaluation may also differ due to different use case scenarios. While previous research has already endeavored to define criteria such as clarity, fairness, bias, completeness, and soundness (e.g., Zhou et al., 2021), it is not evident how these can be objectively measured and whether they are sufficient in constrained scenarios. In addition, the use of EIS requires interdisciplinary research to define guidelines and norms that ensure legally compliant utilization of EIS across different application domains, transitioning EIS into trustworthy AI (Thiebes et al., 2021).

Lastly, we found divergent results for the relevance of user motivation (Ferreira & Monteiro, 2020). Here, we assume that the inclusion of components to increase user motivation is primarily necessary for novice users, since experienced users have already internalized the benefits provided by an EIS. Although our experts have mentioned the potential of using gamification concepts to reduce EIS acceptance barriers through play, recent research has not yet focused on this approach. While research has already shown how students can learn and perform new content through an interactive, game-based learning platform (Xinogalos & Satratzemi, 2022), a gamified approach with a leaderboard could provide employees with necessary EIS knowledge and potentially increase adaptation or reduce learning barriers when it comes to using yet unknown technologies. However, our experts were unable to define how such a learning platform should be designed to support their employees without overwhelming them.

## 7.7 Conclusion and Outlook

The lack of explainability of intelligent systems inhibits their acceptance. XAI offers a potential path out of this dilemma. In response, we have developed a rigorous nascent design theory for EIS that includes four design principles and ten design features to foster the acceptance of AI-assisted decision-making focusing on local and global explanation, personalization as well as addressing intrinsic barriers. In doing so, we incorporate both technical and socio-technical aspects of XAI to address the needs of different user groups, including end-users and developers to develop a broad, domain-independent design theory also considering human-agent interaction. In summary, our nascent design theory provides novel knowledge design knowledge for a symbiosis of expert and system and can further foster the integration of AI into operational practice.

# Appendix

## Appendix A. List of Authored Contributions

| Year | Title | Outlet | Impact-Factor (2021) | VHB-Jourqual 3 (2015) | Included in Thesis | Related Work of Thesis |
|---|---|---|---|---|---|---|
| 2023 | Impact of Explainable AI on Cognitive Load: Insights From an Empirical Study | European Conference on Information Systems | - | B | yes | - |
| 2023 | Stop Ordering Machine Learning Algorithms by their Explainability! A User-Centered Investigation of Performance and Explainability | International Journal of Information Management | 18.958 | C | yes | - |
| 2023 | A Framework for Implementing Robotic Process Automation Projects | Information Systems and e-Business Management | 2.775 | C | - | - |
| 2022 | A Nascent Design Theory for Explainable Intelligent Systems | Electronic Markets | 6.017 | B | yes | - |
| 2022 | The Effect of Transparency and Trust on Intelligent System Acceptance: Evidence from a User-based Study | Electronic Markets | 6.017 | B | yes | - |
| 2022 | A Taxonomy of User-centered Explainable AI Studies* | Pacific Asia Conference on Information Systems | - | C | yes | - |
| 2022 | Industry 4.0 Maintenance: An Examination Of The Readiness Of Germany’s Industrial Sector | International Conference on Business Informatics | - | - | - | yes |
| 2022 | Applications and Challenges of Task Mining: A Literature Review | European Conference on Information Systems | - | B | - | - |
| 2022 | Der Einfluss von menschlichen Denkmustern auf künstliche Intelligenz - Eine strukturierte Untersuchung von kognitiven Verzerrungen | HMD Praxis der Wirtschaftsinformatik | - | D | - | yes |
| 2022 | A Social Evaluation of the Perceived Goodness of Explainability in Machine Learning | Journal of Business Analytics | - | - | - | yes |

| | | | | | | |
|---|---|---|---|---|---|---|
| 2022 | Plattform für das integrierte Management von Kollaborationen in Wertschöpfungsnetzwerken (PIMKoWe) | Working Paper Series of the Institute of Business Management | - | - | - | - |
| 2021 | From Symbolic RPA to Intelligent RPA: Challenges for Developing and Operating Intelligent Software Robots | International Conference on Business Process Management | - | C | - | - |
| 2021 | A Framework of Cost Drivers for Robotic Process Automation Projects* | RPA Forum of the International Conference on Business Process Management | - | C | - | - |
| 2021 | Stop Ordering Machine Learning Algorithms by their Explainability! An Empirical Investigation of the Tradeoff between Performance and Explainability | IFIP Conference e-Business, e-Services, and e-Society | - | - | - | yes |
| 2021 | Design Principles for Shared Maintenance Analytics in Fleet Management | International Conference on Design Science Research in Information Systems and Technology | - | C | - | - |
| 2021 | I Don't Get It, But It Seems Valid! The Connection Between Explainability and Comprehensibility in (X)AI Research | European Conference on Information Systems | - | B | yes | - |
| 2021 | Adoption Barriers of AI: A Context-Specific Acceptance Model For Industrial Maintenance | European Conference on Information Systems | - | B | - | yes |
| 2021 | Managing RPA implementation projects | C. Czarnecki & P. Fettke (Ed.), Robotic Process Automation (Book) | - | - | - | - |
| 2021 | Digitalisierungspotenziale der Instandhaltung 4.0 - Von der Aufbereitung binärer Daten zum Einsatz transparenter künstlicher Intelligenz | Meinhardt, S. & Wortmann, F. (Ed.), IoT - Best Practices (Book) | - | - | - | yes |
| 2021 | Critical Success Factors for Process Modeling Projects-Analysis of Empirical Evidence | Pacific Asia Conference on Information Systems | - | C | - | - |
| 2021 | Towards an Implementation of Blockchain-based Collaboration Platforms in Supply Chain Networks: A Requirements Analysis | Hawaii International Conference on System Sciences | - | C | - | - |

| | | | | | | |
|---|---|---|---|---|---|---|
| 2021 | Towards a Reference Architecture for Female-Sensitive Drug Management | Hawaii International Conference on System Sciences | - | C | - | - |
| 2021 | Entscheidungsunterstützung in KI - Eine Analyse technischer und sozialer Faktoren für die industrielle Instandhaltung in Deutschland | Industrie 4.0 Management | - | - | - | - |
| 2020 | Bridging the Architectural Gap in Smart Homes Between User Control and Digital Automation | International Conference on Design Science Research in Information Systems and Technology | - | C | - | - |
| 2020 | A consolidated framework for implementing robotic process automation projects* | International Conference on Business Process Management | - | C | - | - |
| 2020 | White, Grey, Black: Effects of XAI Augmentation on the Confidence in AI-based Decision Support Systems | International Conference on Information Systems | - | A | - | yes |
| 2020 | How much is the black box? The value of explainability in machine learning models | European Conference on Information Systems | - | B | - | yes |
| 2020 | A Moral Consensus Mechanism for Autonomous Driving: Towards a Law-compliant Basis of Logic Programming | International Conference on Wirtschaftsinformatik | - | C | - | - |
| 2019 | Verwendung binärer Datenwerte für eine KI-gestützte Instandhaltung 4.0 | HMD Praxis der Wirtschaftsinformatik | - | D | - | yes |
| 2019 | Countering the fear of black-boxed ai in maintenance: Towards a smart colleague | Pre-ICIS SIGDSA Symposium | - | - | - | yes |
| 2019 | Anforderungsanalyse für eine Kollaborationsplattform in Blockchain-basierten Wertschöpfungsnetzwerken | Working Paper Series of the Institute of Business Management | - | - | - | - |

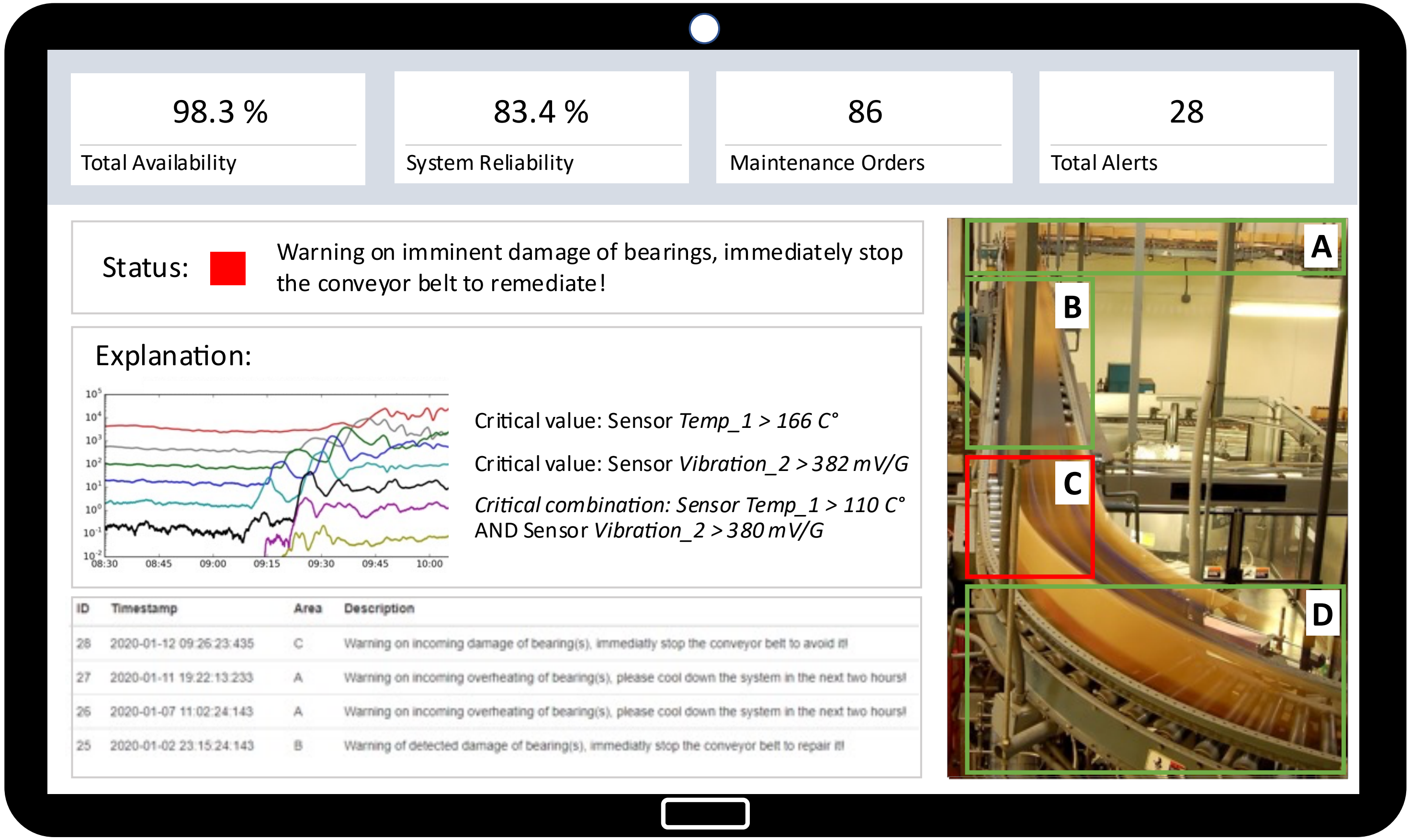
Legend: * - Recognized with the "Best Paper Award" of the respective conference.

Note: Most of the publications are authored by at least one co-author. Please see the Section References for detailed references.
Impact-Factor derived from respective journal homepages (accessed: 08.01.2023).
VHB-Jourqual 3 derived from vhbonline.org/vhb4you/vhb-jourqual/vhb-jourqual-3 (accessed: 08.01.2023).

Table A.1 List of Authored Contributions

# Appendix B. The Effect of Transparency and Trust on Intelligent System Acceptance: Evidence from a User-based Study

## Appendix B.1. Dashboard Design



98.3 %
Total Availability

83.4 %
System Reliability

86
Maintenance Orders

28
Total Alerts

Status: 🟥 Warning on imminent damage of bearings, immediately stop the conveyor belt to remediate!

Explanation:

Critical value: Sensor *Temp_1 > 166 C°*

Critical value: Sensor *Vibration_2 > 382 mV/G*

*Critical combination: Sensor Temp_1 > 110 C°* AND Sensor *Vibration_2 > 380 mV/G*

| ID | Timestamp | Area | Description |
|---|---|---|---|
| 28 | 2020-01-12 09:26:23:435 | C | Warning on incoming damage of bearing(s), immediatly stop the conveyor belt to avoid it! |
| 27 | 2020-01-11 19:22:13:233 | A | Warning on incoming overheating of bearing(s), please cool down the system in the next two hours! |
| 26 | 2020-01-07 11:02:24:143 | A | Warning on incoming overheating of bearing(s), please cool down the system in the next two hours! |
| 25 | 2020-01-02 23:15:24:143 | B | Warning of detected damage of bearing(s), immediatly stop the conveyor belt to repair it! |

Figure B.1 Dashboard Design

## Appendix B.2. Measurement Item Collection Procedure

| Construct | | Measurement Item | Abbrev. | Primary Source | Secondary Source | Sequential Reduction | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | *1* | *2* | *3* | *4* | *5* |
| Performance Expectancy (PE) | Perceived Usefulness | Using the system in my job would enable me to accomplish tasks more quickly. | PE1 | Davis (1989) | Venkatesh et al. (2003) | • | • | • | • | • |
| | | Using the system would improve my job performance. | PE2 | | | • | • | • | • | • |
| | | Using the system would enhance my effectiveness on the job. | - | | | • | • | • | | |
| | | Using the system would make it easier to do my job. | PE3 | | | • | • | • | • | • |
| | | I would find the system useful in my job. | PE4 | | | • | • | • | • | • |
| | Job-fit | Use of the system will have no effect on the performance of my job (reverse scored). | - | Venkatesh et al. (2003) | Thompson et al. (1991) | • | | | | |
| | | Use of the system can decrease the time needed for my important job responsibilities. | | | | • | | | | |
| | | Use of the system can significantly increase the quality of output on my job. | | | | • | • | • | | |
| | | Using this system can significantly increase the quantity of output for the same amount of effort in my job | | | | • | • | • | | |
| | | Use of the system can increase the effectiveness of performing job tasks. | | | | • | | | | |
| | | Use can increase the quantity of output for the same amount of effort. | | | | • | • | • | | |
| | | Considering all tasks. the general extent to which use of the system could assist on the job. | | | | • | • | • | | |
| | Relative Advantage | Using the system enables me to accomplish tasks more quickly. | - | Moore and Benbasat (1991) | Venkatesh et al. (2003) | • | • | | | |
| | | Using the system improves the quality of the work I do. | | | | • | • | • | | |
| | | Using the system makes it easier to do my job. | | | | • | | | | |
| | | Using the system enhances my effectiveness on the job. | | | | • | | | | |
| | | Using the system in my job would increase my productivity. | PE5 | | | • | • | • | • | • |
| | Outcome Expectations - Performance | If I use the system, I will increase my effectiveness on the job. | - | Venkatesh et al. (2003) | Compeau and Higgins (1995) | • | | | | |
| | | If I use the system, I will spend less time on routine job tasks. | | | | • | • | • | | |
| | | If I use the system, I will increase the quality of output of my job. | | | | • | | | | |
| | | If I use the system, I will increase the quantity of output for the same amount of effort. | | | | • | • | • | | |
| | | If I use the system, my coworkers will perceive me as competent. | | | | • | • | • | | |
| | | If I use the system, I will increase my chances of obtaining a promotion. | | | | • | • | • | | |
| | | If I use the system, I will increase my chances of getting a raise. | | | | • | • | • | | |
| Effort Expectancy (EE) | Perceived Ease of Use | Learning to operate the system would be easy for me. | EE1 | Venkatesh et al. (2003) | Davis (1989) | • | • | • | • | • |
| | | I would find it easy to get the system to do what I want it to do. | EE2 | | | • | • | • | • | • |
| | | My interaction with the system would be clear and understandable. | EE3 | | | • | • | • | • | • |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | I would find the system easy to use. | EE4 | | | • | • | • | • | • |
| | | I would find the system to be flexible to interact with. | - | | | • | • | • | | |
| | | It would be easy for me to become skillful at using the system. | - | | | • | | | | |
| | Complexity | Using the system takes too much time from my normal duties. | - | Venkatesh et al. (2003) | Thompson et al. (1991) | • | • | • | | |
| | | Working with the system is so complicated. it is difficult to understand what is going on. | | | | • | • | • | | |
| | | Using the system involves too much time doing mechanical operations (e.g.. data input). | | | | • | • | • | | |
| | | It takes too long to learn how to use the system to make it worth the effort. | | | | • | • | • | | |
| | Ease of Use | My interaction with the system is clear and understandable. | - | Venkatesh et al. (2003) | Moore and Benbasat (1991) | • | | | | |
| | | I believe that it is easy to get the system to do what I want it to do. | | | | • | | | | |
| | | Overall. I believe that the system is easy to use. | | | | • | | | | |
| | | Learning to operate the system is easy for me. | | | | • | | | | |
| Attitude Towards AI Technology (ATT) | Intrinsic Motivation | I find using the system to be enjoyable | - | Davis et al. (1992) | - | • | | | | |
| | | The actual process of using the system would be pleasant. | ATT1 | | Venkatesh et al. (2003) | • | • | • | • | • |
| | | I have fun using the system. | - | | - | • | | | | |
| | Affect Toward Use | This system would make work more interesting. | ATT2 | (Thompson et al., 1991) | Venkatesh et al. (2003) | • | • | • | • | • |
| | | Working with the system is fun. | - | Venkatesh et al. (2003) | - | • | | | | |
| | | The system is okay for some jobs. but not the kind of job I want. (R) | | (Thompson et al., 1991) | | • | • | • | | |
| | Affect | I would like to work with the system. | ATT3 | Compeau et al. (1999) | Venkatesh et al. (2003) | • | • | • | • | • |
| | | I look forward to those aspects of my job that require me to use the system. | - | | | • | • | • | | |
| | | Using the system is frustrating for me. (R) | | | | • | • | • | | |
| | | Once I start working on the system. I find it hard to stop. | | | | • | • | • | | |
| | Attitude Toward Behavior | Using the system would be a good idea | ATT4 | (Peters et al., 2020); Taylor and Todd (1995) | Venkatesh et al. (2003) | • | • | • | | • |
| | | I dislike/like the idea of using the system. | - | | | • | | | | |
| | | Using the system would be wise move. | ATT5 | | | • | • | • | | • |
| | | Using the system is unpleasant/pleasant. | - | | | • | • | • | | |
| Behavioral Intention (BI) | | If this system was available to me, I would intend to use this system in the next months. | BI1 | Venkatesh et al. (2003) | - | | • | • | • | • |
| | | If this system was available to me, I predict I would use this system in the next months. | BI2 | | | | • | • | • | • |
| | | If this system was available to me, I would plan to use this system in the next months. | BI3 | | | • | • | • | • | • |
| System Transparency (ST) | | I know what will happen the next time I use the system because I understand how it behaves. | - | Madsen and Gregor (2000) | - | • | • | • | | |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | I would understand how this system will assist me with decisions I have to make. | ST1 | | | • | • | • | • | • |
| | | Although I may not know exactly how the system works. I know how to use it to make decisions about the problem. | - | | | • | • | • | | |
| | | It is easy to follow what the system does. | | | | • | • | • | | |
| | | I recognize what I should do to get the advice I need from the system the next time I use it. | | | | • | • | • | | |
| | | I would understand why this system provided the decision it did. | ST2 | Cramer et al. (2008) | | • | • | • | • | • |
| | | I would understand what this system bases its provided decision on. | ST3 | | | • | • | • | | • |
| Ability Beliefs (AB) | | This system would be competent in providing maintenance decision support. | AB1 | McKnight et al. (2002) | Cheng et al. (2008) | • | • | • | • | • |
| | | This system would perform maintenance decision support very well. | AB2 | | | • | • | • | • | • |
| | | In general, this system would be proficient providing maintenance decision support. | AB3 | | | • | • | • | • | • |
| Trust Propensity Towards AI (TP) | | It would be easy for me to trust this system. | TP1 | Lee and Turban (2001) | Cheng et al. (2008); Wang and Benbasat (2007) | • | • | • | • | • |
| | | My tendency to trust this system would be high. | TP2 | | | • | • | • | • | • |
| | | I would tend to trust this system, even though I have little or no knowledge of it. | TP3 | | | • | • | • | • | • |
| | | Trusting this system would be difficult for me. | TP4 | | Wang and Benbasat (2007) | • | • | • | • | |
| Facilitating Conditions (FC) | Perceived Behavioral Control | I have control over using the system. | - | (Taylor & Todd, 1995) | Venkatesh et al. (2003) | • | | | | |
| | | I have the resources necessary to use the system. | - | | | • | • | • | | |
| | | I have the knowledge necessary to use the system. | - | | | • | • | • | | |
| | | Given the resources, opportunities and knowledge it takes to use the system, it would be easy for me to use the system. | - | | | • | • | | | |
| | | The system is not compatible with other systems I use. | - | | | • | • | • | | |
| | Facilitating Conditions | Guidance was available to me in the selection of the system. | - | (Thompson et al., 1991) | Venkatesh et al. (2003) | • | | | | |
| | | Specialized instruction concerning the system was available to me. | - | | | • | • | • | | |
| | | A specific person (or group) is available for assistance with system difficulties. | - | | | • | • | • | | |
| | Compatibility | Using the system is compatible with all aspects of my work. | - | (Moore & Benbasat, 1991) | Venkatesh et al. (2003) | • | • | • | | |
| | | I think that using the system fits well with the way I like to work. | - | | | • | • | | | |
| | | Using the system fits into my work style. | - | | | • | | | | |
| Social Influence (SI) | Subjective Norm | People who influence my behavior think that I should use the system. | - | (Davis, 1989; Fishbein & Ajzen, 1977; Taylor & Todd, 1995) | Venkatesh et al. (2003) | • | • | | | |
| | | People who are important to me think that I should use the system. | - | | | • | • | | | |
| | Social Factors | I use the system because of the proportion of coworkers who use the system. | - | (Thompson et al., 1991) | Venkatesh et al. (2003) | • | | | | |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | The senior management of this business has been helpful in the use of the system. | - | | | • | | | | |
| | | My supervisor is very supportive of the use of the system for my job. | - | | | • | | | | |
| | | In general, the organization has supported the use of the system. | - | | | • | • | | | |
| | Image | People in my organization who use the system have more prestige than those who do not. | - | (Moore & Benbasat, 1991) | Venkatesh et al. (2003) | • | | | | |
| | | People in my organization who use the system have a high profile. | - | | | • | | | | |
| | | Having the system is a status symbol in my organization. | - | | | • | | | | |

Legend: Sequential reduction of the item collection to fit the defined use case: 1) preliminary work, 2) authors' internal discussion, 3) expert survey, 4) pre-study, 5) main study.

Table B.1 Measurement Item Collection Procedure

## Appendix B.3. Validation and Reliability Testing Results Pre-Study

*Crossloadings*

| **Factors** | ATT | BI | EE | PE | ST | AB | TP |
|---|---|---|---|---|---|---|---|
| **ATT1** | 0.819 | 0.436 | 0.521 | 0.518 | 0.180 | 0.658 | 0.580 |
| **ATT2** | 0.581 | 0.107 | 0.285 | 0.268 | 0.021 | 0.345 | 0.179 |
| **ATT3** | 0.872 | 0.269 | 0.622 | 0.482 | 0.110 | 0.565 | 0.453 |
| **BI1** | 0.498 | 0.859 | 0.500 | 0.536 | 0.325 | 0.458 | 0.589 |
| **BI2** | 0.190 | 0.911 | 0.250 | 0.321 | 0.343 | 0.264 | 0.318 |
| **BI3** | 0.229 | 0.800 | 0.276 | 0.330 | 0.361 | 0.315 | 0.275 |
| **EE2** | 0.476 | 0.208 | 0.773 | 0.313 | 0.002 | 0.480 | 0.495 |
| **EE3** | 0.366 | 0.331 | 0.758 | 0.177 | 0.214 | 0.285 | 0.198 |
| **EE4** | 0.686 | 0.437 | 0.891 | 0.452 | 0.203 | 0.520 | 0.449 |
| **PE1** | 0.327 | 0.443 | 0.207 | 0.746 | 0.396 | 0.398 | 0.371 |
| **PE2** | 0.461 | 0.497 | 0.281 | 0.846 | 0.228 | 0.432 | 0.429 |
| **PE3** | 0.509 | 0.300 | 0.342 | 0.776 | 0.129 | 0.454 | 0.347 |
| **PE4** | 0.483 | 0.219 | 0.386 | 0.739 | 0.132 | 0.342 | 0.255 |
| **PE5** | 0.530 | 0.437 | 0.376 | 0.881 | 0.305 | 0.404 | 0.348 |
| **ST1** | 0.188 | 0.225 | 0.263 | 0.211 | 0.842 | 0.269 | 0.268 |
| **ST2** | 0.091 | 0.445 | 0.235 | 0.309 | 0.890 | 0.239 | 0.105 |
| **AB1** | 0.677 | 0.421 | 0.524 | 0.522 | 0.241 | 0.918 | 0.731 |
| **AB3** | 0.521 | 0.215 | 0.328 | 0.338 | 0.222 | 0.799 | 0.558 |
| **AB3** | 0.520 | 0.399 | 0.463 | 0.356 | 0.260 | 0.707 | 0.419 |
| **TP1** | 0.637 | 0.410 | 0.503 | 0.405 | 0.191 | 0.751 | 0.929 |
| **TP2** | 0.507 | 0.436 | 0.475 | 0.440 | 0.103 | 0.632 | 0.906 |
| **TP3** | 0.213 | 0.412 | 0.335 | 0.254 | 0.267 | 0.339 | 0.584 |
| **TP4** | 0.014 | -0.180 | 0.055 | 0.083 | -0.093 | -0.107 | -0.450 |

**Fornell-Larcker Criterion**

| **Factors** | ATT | BI | EE | PE | ST | AB | TP |
|---|---|---|---|---|---|---|---|
| **ATT** | 0.768 | - | - | - | - | - | - |
| **BI** | 0.391 | 0.858 | - | - | - | - | - |
| **EE** | 0.650 | 0.427 | 0.757 | - | - | - | - |
| **PE** | 0.576 | 0.486 | 0.393 | 0.799 | - | - | - |
| **ST** | 0.156 | 0.397 | 0.285 | 0.305 | 0.866 | - | - |
| **AB** | 0.710 | 0.423 | 0.540 | 0.509 | 0.291 | 0.813 | - |
| **TP** | 0.575 | 0.495 | 0.526 | 0.443 | 0.207 | 0.720 | 0.746 |

Table B.2 Validation and Reliability Testing Results Pre-Study

## Appendix B.4. Decisions on Measurement Items after Pre-Study

We conclude that the items for *EE* are well chosen, as these have been tested and verified in many studies based on the UTAUT model. In addition, *EE* does not fail any other criteria except indicator reliability (EE1). Thus, we retain all measurement items for *EE*. We conclude that items for ST are, in principle, well chosen. However, we conclude that an additional measurement item for *ST* must be added as it lacks internal consistency (CA < 0.7). Hence, we extend our items by ST3, which is derived from Cramer et al. (2008). Item loadings for TP3 (0.58) and TP4 (-0.45) are below the threshold of 0.7. TP4 seems to be particularly problematic, as the reverse wording suggested by Cheng et al. (2008) as well as Wang and Benbasat (2007) causes convergence reliability issues. Removing TP4 leads to a higher CA (0.75), AVE (0.68), and CR (0.86). Since using reversed wording is not advised (Van Sonderen et al., 2013; Zhang et al., 2016), we decide to drop TP4. We decided to follow Lee and Turban (2001) and use the original wording in line with the other items for the main study (TP1-3). We decide to retain TP3 in the main study, as the loading is satisfactory and the construct itself is reliable if TP4 is dropped. As ATT2 performs subpar in terms of item loading, we decided to add additional items. Additionally, ATT has a low value for CA, indicating low internal consistency. Thus, we added ATT4 and ATT5 following Taylor and Todd (1995). It was recently used in a similar context by (Peters et al., 2020).

## Appendix B.5. Demographics of Pre-Study

| Characteristics | Attributes | Value | | Characteristics | Attributes | Value | |
|---|---|---|---|---|---|---|---|
| | | *Freq.* | *Percent.* | | | *Freq.* | *Percent.* |
| Gender | Male | 49 | 81.67 | Experience with intelligent systems in industrial maintenance (EXP2) | None | 7 | 11.67 |
| | Female | 9 | 15.00 | | <1 year | 22 | 36.67 |
| | Others | 2 | 3.33 | | 1-3 years | 15 | 25.00 |
| Age | <=20 | 0 | 0.00 | | 3-5 years | 16 | 26.67 |
| | 21-30 | 7 | 11.67 | | 5-10 years | 0 | 0.00 |
| | 31-40 | 11 | 18.33 | | >10 years | 0 | 0.00 |
| | 41-50 | 10 | 16.67 | Experience with AI (EXP3) | None | 0 | 0.00 |
| | 51-60 | 26 | 43.33 | | <1 year | 2 | 3.33 |
| | >61 | 6 | 0.10 | | 1-3 years | 17 | 28.33 |
| Experience in industrial maintenance (EXP1) | None | 0 | 0.00 | | 3-5 years | 22 | 36.67 |
| | <1 year | 19 | 26.67 | | 5-10 years | 15 | 25.00 |
| | 1-3 years | 9 | 15.00 | | >10 years | 4 | 6.67 |
| | 3-5 years | 8 | 13.33 | Note: Gender, Age, EXP1, EXP2, and EXP3 were used as interaction moderation. | | | |
| | 5-10 years | 4 | 6.67 | | | | |
| | >10 years | 19 | 26.67 | | | | |

Table B.3 Demographics of Pre-Study

## Appendix B.6. Validation and Reliability Testing Results for Main Study

*Fornell-Larcker Criterion Main Study*

| *Factors* | AGE | AGE-EE | AGE-PE | ATT | BI | EE | EX1-EE | EX1-PE | EXP2-EE | EXP3-PE | EXP3-EE | EXP3-PE | EPX1 | EXP2 | EXP3 | GEN | GEN-EE | GEN-PE | P*E* | ST | AB | TP |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *Age* | 1.000 | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| *AGE-EE* | 0.030 | 1.000 | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| *AGE-PE* | -0.003 | 0.515 | 1.000 | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| *ATT* | -0.026 | -0.053 | -0.008 | 0.841 | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| *BI* | 0.075 | -0.056 | -0.084 | 0.634 | 0.951 | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| *EE* | 0.024 | -0.283 | -0.071 | 0.573 | 0.434 | 0.835 | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| *EXP1-EE* | -0.019 | -0.038 | -0.022 | 0.056 | 0.086 | 0.053 | 1.000 | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| *EXP1-PE* | 0.073 | -0.026 | -0.066 | 0.076 | 0.039 | 0.019 | 0.630 | 1.000 | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| *EXP2-EE* | 0.098 | 0.043 | -0.019 | 0.127 | 0.082 | 0.036 | 0.604 | 0.387 | 1.000 | - | - | - | - | - | - | - | - | - | - | - | - | - |
| *EXP3-PE* | 0.040 | -0.056 | 0.004 | 0.110 | 0.065 | 0.037 | 0.353 | 0.561 | 0.321 | 1.000 | - | - | - | - | - | - | - | - | - | - | - | - |
| *Exp3-EE* | -0.078 | 0.039 | -0.053 | 0.027 | 0.000 | 0.005 | 0.546 | 0.347 | 0.506 | 0.520 | 1.000 | - | - | - | - | - | - | - | - | - | - | - |
| *Exp3-PE* | 0.144 | -0.024 | -0.066 | 0.039 | -0.009 | 0.007 | 0.456 | 0.713 | 0.664 | 0.528 | 0.370 | 1.000 | - | - | - | - | - | - | - | - | - | - |
| *EXP1* | 0.023 | -0.022 | 0.083 | 0.120 | 0.134 | 0.252 | 0.382 | 0.226 | 0.286 | 0.129 | 0.261 | 0.077 | 1.000 | - | - | - | - | - | - | - | - | - |
| *EXP2* | 0.079 | 0.121 | 0.149 | 0.097 | 0.059 | 0.216 | 0.303 | 0.071 | 0.235 | 0.001 | 0.168 | 0.107 | 0.571 | 1.000 | - | - | - | - | - | - | - | - |
| *EXP3* | 0.033 | -0.097 | 0.048 | 0.082 | 0.143 | 0.158 | 0.279 | 0.138 | 0.170 | 0.047 | 0.049 | 0.001 | 0.536 | 0.429 | 1.000 | - | - | - | - | - | - | - |
| *GEN* | -0.028 | -0.070 | 0.048 | 0.027 | -0.025 | 0.031 | -0.018 | 0.020 | -0.113 | 0.056 | -0.047 | -0.075 | 0.042 | 0.117 | 0.077 | 1.000 | - | - | - | - | - | - |
| *GEN-EE* | -0.052 | -0.126 | -0.163 | -0.139 | -0.146 | -0.227 | 0.066 | 0.067 | 0.163 | 0.153 | 0.244 | 0.124 | -0.016 | -0.106 | -0.044 | -0.023 | 1.000 | - | - | - | - | - |
| *GEN-PE* | 0.038 | -0.170 | -0.175 | -0.153 | -0.156 | -0.176 | 0.061 | 0.139 | 0.101 | 0.090 | 0.148 | 0.185 | 0.017 | -0.061 | 0.052 | -0.040 | 0.633 | 1.000 | - | - | - | - |
| *PE* | 0.010 | -0.073 | -0.054 | 0.720 | 0.616 | 0.596 | 0.018 | 0.004 | 0.006 | 0.038 | 0.036 | -0.127 | 0.202 | 0.139 | 0.105 | 0.053 | -0.174 | -0.191 | 0.860 | - | - | - |
| *ST* | -0.024 | -0.089 | 0.075 | 0.559 | 0.472 | 0.539 | 0.088 | 0.048 | 0.033 | 0.117 | 0.073 | -0.016 | 0.170 | 0.105 | 0.040 | -0.015 | -0.098 | -0.149 | 0.505 | 0.892 | - | - |
| *AB* | 0.025 | -0.013 | 0.118 | 0.687 | 0.552 | 0.531 | 0.000 | -0.027 | -0.018 | 0.132 | 0.026 | -0.078 | 0.115 | 0.093 | 0.105 | -0.007 | -0.173 | -0.144 | 0.594 | 0.610 | 0.905 | - |
| *TP* | -0.007 | -0.040 | -0.060 | 0.600 | 0.444 | 0.462 | 0.111 | 0.071 | 0.069 | 0.148 | 0.121 | 0.010 | 0.045 | 0.084 | 0.069 | -0.005 | -0.065 | -0.131 | 0.528 | 0.411 | 0.656 | 0.877 |

Table B.4 Fornell-Larcker Criterion Main Study

### *Crossloadings Main Study*

| *Factors* | AGE | AGE-EE | AGE-PE | ATT | BI | EE | EXP1-EE | EXP1-PE | EXP2-EE | EXP3-PE | EXP3-EE | EXP3-PE | EXP1 | EXP2 | EXP3 | GEN | GEN-EE | GEN-PE | PE | ST | AB | TP |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *AGE* | 1.000 | 0.030 | -0.003 | -0.026 | 0.075 | 0.024 | -0.019 | 0.073 | 0.098 | 0.040 | -0.078 | 0.144 | 0.023 | 0.079 | 0.033 | -0.028 | -0.052 | 0.038 | 0.010 | -0.024 | 0.025 | -0.007 |
| *ATT1* | -0.125 | -0.093 | 0.038 | 0.764 | 0.460 | 0.566 | 0.021 | -0.021 | 0.053 | 0.029 | -0.008 | 0.007 | 0.142 | 0.129 | 0.045 | -0.047 | -0.066 | -0.148 | 0.510 | 0.454 | 0.530 | 0.372 |
| *ATT3* | -0.038 | -0.107 | -0.017 | 0.856 | 0.557 | 0.490 | 0.116 | 0.137 | 0.146 | 0.166 | 0.088 | 0.080 | 0.132 | 0.166 | 0.055 | -0.010 | -0.068 | -0.110 | 0.643 | 0.521 | 0.551 | 0.495 |
| *ATT4* | 0.091 | -0.015 | -0.025 | 0.881 | 0.557 | 0.462 | 0.016 | 0.074 | 0.126 | 0.078 | 0.009 | 0.026 | 0.089 | 0.028 | 0.110 | 0.059 | -0.160 | -0.124 | 0.646 | 0.463 | 0.650 | 0.603 |
| *ATT5* | -0.036 | 0.030 | -0.015 | 0.861 | 0.554 | 0.430 | 0.032 | 0.053 | 0.095 | 0.087 | -0.004 | 0.015 | 0.047 | 0.012 | 0.063 | 0.079 | -0.166 | -0.137 | 0.613 | 0.445 | 0.576 | 0.531 |
| *BI1* | 0.026 | -0.011 | -0.085 | 0.610 | 0.953 | 0.421 | 0.028 | -0.008 | 0.059 | 0.029 | 0.009 | -0.047 | 0.132 | 0.056 | 0.120 | -0.038 | -0.141 | -0.163 | 0.591 | 0.428 | 0.506 | 0.416 |
| *BI2* | 0.155 | -0.077 | -0.063 | 0.606 | 0.931 | 0.424 | 0.121 | 0.069 | 0.091 | 0.060 | -0.030 | 0.017 | 0.141 | 0.079 | 0.142 | -0.027 | -0.130 | -0.118 | 0.595 | 0.442 | 0.540 | 0.421 |
| *BI3* | 0.032 | -0.072 | -0.092 | 0.593 | 0.968 | 0.393 | 0.094 | 0.048 | 0.084 | 0.095 | 0.022 | 0.003 | 0.108 | 0.034 | 0.146 | -0.008 | -0.147 | -0.166 | 0.572 | 0.477 | 0.527 | 0.429 |
| *EE1* | 0.017 | -0.279 | -0.053 | 0.373 | 0.273 | 0.814 | 0.060 | -0.043 | 0.019 | -0.022 | -0.047 | -0.038 | 0.212 | 0.190 | 0.182 | 0.067 | -0.180 | -0.186 | 0.423 | 0.339 | 0.291 | 0.279 |
| *EE2* | 0.001 | -0.167 | 0.006 | 0.381 | 0.232 | 0.724 | 0.099 | 0.088 | 0.080 | 0.059 | 0.035 | 0.093 | 0.236 | 0.262 | 0.153 | 0.058 | -0.135 | -0.054 | 0.346 | 0.449 | 0.450 | 0.415 |
| *EE3* | -0.042 | -0.221 | -0.109 | 0.582 | 0.466 | 0.893 | 0.026 | 0.013 | 0.012 | 0.062 | 0.050 | -0.020 | 0.186 | 0.142 | 0.084 | 0.015 | -0.271 | -0.195 | 0.619 | 0.520 | 0.529 | 0.457 |
| *EE4* | 0.107 | -0.284 | -0.061 | 0.528 | 0.422 | 0.898 | 0.013 | 0.005 | 0.020 | 0.013 | -0.035 | -0.001 | 0.224 | 0.161 | 0.140 | -0.013 | -0.154 | -0.142 | 0.546 | 0.466 | 0.465 | 0.372 |
| *PE1* | -0.034 | -0.106 | 0.002 | 0.514 | 0.447 | 0.518 | -0.021 | -0.009 | 0.008 | 0.028 | 0.075 | -0.100 | 0.224 | 0.121 | 0.072 | -0.020 | -0.097 | -0.127 | 0.838 | 0.421 | 0.449 | 0.361 |
| *PE2* | 0.012 | -0.028 | -0.048 | 0.661 | 0.505 | 0.459 | 0.075 | 0.061 | 0.068 | 0.063 | 0.056 | -0.058 | 0.162 | 0.195 | 0.131 | 0.161 | -0.160 | -0.168 | 0.855 | 0.452 | 0.536 | 0.476 |
| *PE3* | -0.020 | -0.072 | -0.048 | 0.659 | 0.545 | 0.577 | 0.009 | 0.002 | -0.034 | 0.032 | 0.059 | -0.141 | 0.181 | 0.109 | 0.082 | -0.041 | -0.164 | -0.137 | 0.895 | 0.449 | 0.544 | 0.473 |
| *PE4* | 0.058 | -0.019 | -0.071 | 0.647 | 0.608 | 0.545 | 0.034 | -0.002 | 0.025 | 0.021 | -0.025 | -0.104 | 0.144 | 0.087 | 0.060 | 0.025 | -0.131 | -0.125 | 0.853 | 0.404 | 0.509 | 0.484 |
| *PE5* | 0.018 | -0.098 | -0.060 | 0.595 | 0.529 | 0.460 | -0.026 | -0.039 | -0.043 | 0.019 | -0.001 | -0.143 | 0.168 | 0.086 | 0.104 | 0.095 | -0.188 | -0.263 | 0.858 | 0.446 | 0.508 | 0.463 |
| *ST1* | -0.032 | -0.027 | 0.045 | 0.531 | 0.481 | 0.476 | 0.034 | -0.023 | 0.008 | 0.089 | 0.062 | -0.063 | 0.096 | 0.015 | 0.018 | -0.060 | -0.132 | -0.176 | 0.469 | 0.854 | 0.600 | 0.400 |
| *ST2* | -0.023 | -0.076 | 0.111 | 0.472 | 0.401 | 0.470 | 0.040 | -0.004 | 0.013 | 0.075 | 0.025 | -0.044 | 0.159 | 0.143 | 0.045 | 0.012 | -0.105 | -0.173 | 0.430 | 0.920 | 0.524 | 0.335 |
| *ST3* | -0.008 | -0.140 | 0.046 | 0.486 | 0.373 | 0.496 | 0.166 | 0.162 | 0.070 | 0.150 | 0.109 | 0.070 | 0.206 | 0.132 | 0.046 | 0.014 | -0.020 | -0.044 | 0.448 | 0.902 | 0.501 | 0.358 |
| *AB1* | -0.024 | 0.057 | 0.154 | 0.565 | 0.463 | 0.404 | -0.029 | -0.062 | -0.076 | 0.097 | -0.013 | -0.167 | 0.104 | 0.093 | 0.096 | 0.038 | -0.153 | -0.169 | 0.507 | 0.554 | 0.899 | 0.537 |
| *AB2* | 0.055 | -0.060 | 0.052 | 0.653 | 0.514 | 0.573 | -0.007 | -0.077 | 0.019 | 0.053 | 0.008 | -0.047 | 0.124 | 0.132 | 0.111 | -0.017 | -0.157 | -0.115 | 0.563 | 0.518 | 0.888 | 0.588 |
| *AB3* | 0.035 | -0.028 | 0.115 | 0.643 | 0.519 | 0.466 | 0.031 | 0.056 | 0.004 | 0.200 | 0.070 | -0.007 | 0.086 | 0.034 | 0.079 | -0.037 | -0.159 | -0.111 | 0.543 | 0.582 | 0.926 | 0.649 |
| *TP1* | -0.051 | 0.009 | -0.031 | 0.596 | 0.449 | 0.456 | 0.097 | 0.093 | 0.083 | 0.144 | 0.139 | 0.038 | 0.057 | 0.081 | 0.067 | 0.006 | -0.074 | -0.157 | 0.519 | 0.419 | 0.628 | 0.937 |
| *TP2* | -0.006 | -0.065 | -0.049 | 0.566 | 0.402 | 0.446 | 0.113 | 0.112 | 0.110 | 0.197 | 0.147 | 0.085 | 0.039 | 0.047 | 0.050 | -0.015 | -0.065 | -0.121 | 0.497 | 0.397 | 0.606 | 0.935 |
| *TP3* | 0.060 | -0.058 | -0.091 | 0.393 | 0.300 | 0.291 | 0.080 | -0.046 | -0.036 | 0.022 | 0.006 | -0.135 | 0.017 | 0.102 | 0.066 | -0.005 | -0.024 | -0.051 | 0.355 | 0.240 | 0.477 | 0.744 |

| | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *AGE * EE* | 0.030 | 1.000 | 0.515 | -0.053 | -0.056 | -0.283 | -0.038 | -0.026 | 0.043 | -0.056 | 0.039 | -0.024 | -0.022 | 0.121 | -0.097 | -0.070 | -0.126 | -0.170 | -0.073 | -0.089 | -0.013 | -0.040 |
| *EE * EXP1* | -0.019 | -0.038 | -0.022 | 0.056 | 0.086 | 0.053 | 1.000 | 0.630 | 0.604 | 0.353 | 0.546 | 0.456 | 0.382 | 0.303 | 0.279 | -0.018 | 0.066 | 0.061 | 0.018 | 0.088 | 0.000 | 0.111 |
| *EE * EXP2* | 0.098 | 0.043 | -0.019 | 0.127 | 0.082 | 0.036 | 0.604 | 0.387 | 1.000 | 0.321 | 0.506 | 0.664 | 0.286 | 0.235 | 0.170 | -0.113 | 0.163 | 0.101 | 0.006 | 0.033 | -0.018 | 0.069 |
| *EE *EXP3* | -0.078 | 0.039 | -0.053 | 0.027 | 0.000 | 0.005 | 0.546 | 0.347 | 0.506 | 0.520 | 1.000 | 0.370 | 0.261 | 0.168 | 0.049 | -0.047 | 0.244 | 0.148 | 0.036 | 0.073 | 0.026 | 0.121 |
| *EE * GEN* | -0.052 | -0.126 | -0.163 | -0.139 | -0.146 | -0.227 | 0.066 | 0.067 | 0.163 | 0.153 | 0.244 | 0.124 | -0.016 | -0.106 | -0.044 | -0.023 | 1.000 | 0.633 | -0.174 | -0.098 | -0.173 | -0.065 |
| *PE- *AGE* | -0.003 | 0.515 | 1.000 | -0.008 | -0.084 | -0.071 | -0.022 | -0.066 | -0.019 | 0.004 | -0.053 | -0.066 | 0.083 | 0.149 | 0.048 | 0.048 | -0.163 | -0.175 | -0.054 | 0.075 | 0.118 | -0.060 |
| *PE-*EXP1* | 0.073 | -0.026 | -0.066 | 0.076 | 0.039 | 0.019 | 0.630 | 1.000 | 0.387 | 0.561 | 0.347 | 0.713 | 0.226 | 0.071 | 0.138 | 0.020 | 0.067 | 0.139 | 0.004 | 0.048 | -0.027 | 0.071 |
| *PE-*EXP2* | 0.144 | -0.024 | -0.066 | 0.039 | -0.009 | 0.007 | 0.456 | 0.713 | 0.664 | 0.528 | 0.370 | 1.000 | 0.077 | 0.107 | 0.001 | -0.075 | 0.124 | 0.185 | -0.127 | -0.016 | -0.078 | 0.010 |
| *PE- * EXP3* | 0.040 | -0.056 | 0.004 | 0.110 | 0.065 | 0.037 | 0.353 | 0.561 | 0.321 | 1.000 | 0.520 | 0.528 | 0.129 | 0.001 | 0.047 | 0.056 | 0.153 | 0.090 | 0.038 | 0.117 | 0.132 | 0.148 |
| *PE-*GEN* | 0.038 | -0.170 | -0.175 | -0.153 | -0.156 | -0.176 | 0.061 | 0.139 | 0.101 | 0.090 | 0.148 | 0.185 | 0.017 | -0.061 | 0.052 | -0.040 | 0.633 | 1.000 | -0.191 | -0.149 | -0.144 | -0.131 |

Table B.5 Crossloadings Main Study

## Appendix B.7. Variance Inflation Factor Values for Main Study

**Inner Variance Inflation Factor Values**

| | *AGE* | *AGE_EE* | *AGE_PE* | *ATT* | *BI* | *EE* | *EX1_EE* | *EX1_PE* | *EXP2_EE* | *EXP3_PE* | *EXP3_EE* | *EXP3_PE* | *EXP1* | *EXP2* | *EXP3* | *GEN* | *GEN_EE* | *GEN_PE* | *PE* | *ST* | *AB* | *TP* |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *AGE* | - | - | - | - | 1.086 | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| *AGE_EE* | - | - | - | - | 1.774 | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| *AGE_PE* | - | - | - | - | 1.529 | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| *ATT* | - | - | - | - | 2.907 | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| *BI* | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| *EE* | - | - | - | 1.625 | 2.325 | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| *EX1_EE* | - | - | - | - | 3.485 | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| *EX1_PE* | - | - | - | - | 4.606 | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| *EXP2_EE* | - | - | - | - | 3.763 | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| *EXP3_PE* | - | - | - | - | 2.117 | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| *EXP3_EE* | - | - | - | - | 2.208 | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| *EXP3_PE* | - | - | - | - | 5.245 | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| *EXP1* | - | - | - | - | 2.200 | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| *EXP2* | - | - | - | - | 1.910 | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| *EXP3* | - | - | - | - | 1.623 | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| *GEN* | - | - | - | - | 1.086 | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| *GEN_EE* | - | - | - | - | 1.924 | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| *GEN_PE* | - | - | - | - | 1.859 | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| *PE* | - | - | - | 1.774 | 2.668 | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| *ST* | - | - | - | - | 1.747 | 1.000 | - | - | - | - | - | - | - | - | - | - | - | - | 1.203 | - | 1.000 | 1.594 |
| *AB* | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | 1.594 |
| *TP* | - | - | - | 1.455 | 1.778 | - | - | - | - | - | - | - | - | - | - | - | - | - | 1.203 | - | - | - |

**Outer Variance Inflation Factor Values**

| | AGE | ATT1 | ATT3 | ATT4 | ATT5 | BI1 | BI2 | BI3 | EE1 | EE2 | EE3 | EE4 | EXP1 | EXP2 | EXP3 | GEN | PE1 | PE2 | PE3 | PE4 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *VIF* | 1.000 | 1.750 | 2.192 | 2.814 | 2.674 | 6.357 | 3.598 | 7.907 | 2.396 | 1.514 | 2.385 | 3.116 | 1.000 | 1.000 | 1.000 | 1.000 | 2.624 | 2.576 | 3.707 | 2.781 |
| | PE5 | ST1 | ST2 | ST3 | AB1 | AB2 | AB3 | TP1 | TP2 | TP3 | AGE_EE | EE_EXP1 | EE_EXP2 | EE_EXP3 | EE_GEN | PE_AGE | PE_EXP1 | PE_EXP2 | PE_EXP3 | PE_GEN |
| | 2.719 | 1.775 | 3.599 | 3.262 | 2.625 | 2.354 | 2.968 | 4.150 | 4.152 | 1.444 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |

Table B.6 Variance Inflation Factor Values for Main Study

## Appendix B.8. Distribution of Factors of Main Study

| Median and Standard Derivation of Factors | | |
|---|---|---|
| **Factor** | **Median** | **SD** |
| BI1 | 6.00 | 1.08 |
| BI2 | 6.00 | 1.03 |
| BI3 | 6.00 | 1.11 |
| ST1 | 6.00 | 0.97 |
| ST2 | 6.00 | 1.18 |
| ST3 | 6.00 | 1.12 |
| AB1 | 6.00 | 0.86 |
| AB2 | 6.00 | 0.95 |
| AB3 | 6.00 | 0.94 |
| TP1 | 5.00 | 1.09 |
| TP2 | 5.00 | 1.12 |
| TP3 | 5.00 | 1.18 |
| TP4 | 2.50 | 1.29 |
| EE1 | 6.00 | 1.08 |
| EE2 | 5.00 | 1.14 |
| EE3 | 6.00 | 0.96 |
| EE4 | 6.00 | 1.03 |
| PE1 | 6.00 | 1.25 |
| PE2 | 6.00 | 1.07 |
| PE3 | 6.00 | 1.09 |
| PE4 | 6.00 | 1.06 |
| PE5 | 6.00 | 1.05 |
| ATT1 | 5.00 | 1.06 |
| ATT2 | 5.00 | 1.18 |
| ATT3 | 6.00 | 0.99 |
| ATT4 | 6.00 | 0.82 |
| ATT5 | 6.00 | 0.86 |

Table B.7 Distribution of Factors of Main Study

# Appendix C. Stop Ordering Machine Learning Algorithms by their Explainability! A User-Centered Investigation of Performance and Explainability

## Appendix C.1. Synthesis of Common ML Algorithms Classification Schemes

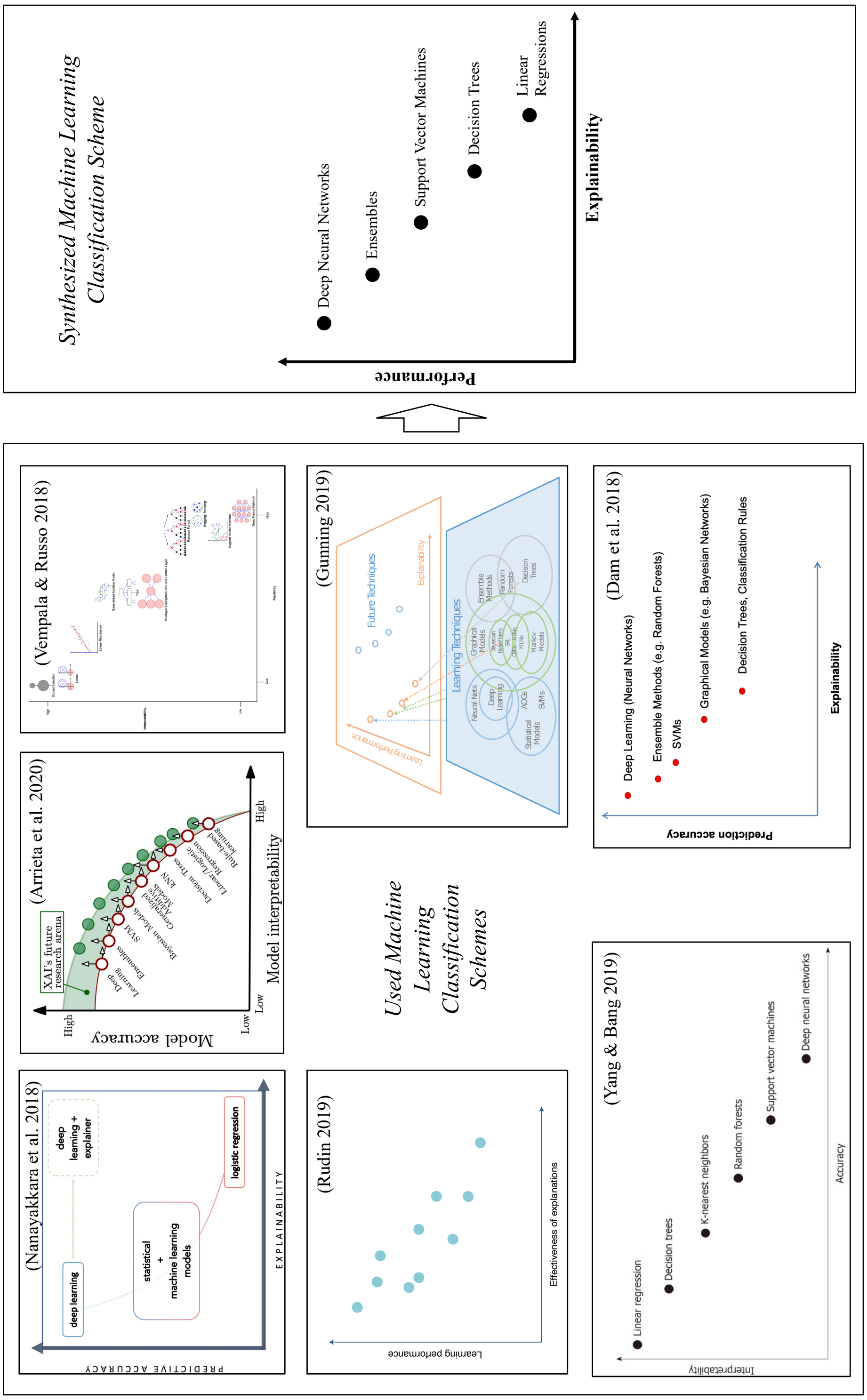


Figure C.1 Synthesis of Common ML Algorithm Classification Schemes (Arrieta et al., 2020; Dam et al., 2018; Gunning, 2019; Nanayakkara et al., 2018; Rudin, 2019; Vempala & Russo, 2018; Yang & Bang, 2019)

## Appendix C.2. Tukey’s HSD Performance Results for Experiment I

**Tukey’s HSD Results for Performance at HEART**

| *Model 1* | *Model 2* | *Difference* | *Lower* | *Upper* |
|---|---|---|---|---|
| Linear Regression | Decision Tree | -0.10423810 | -0.183724224 | -0.02475197 |
| SVM | Linear Regression | 0.18915873 | 0.109672601 | 0.26864486 |
| Random Forest | Linear Regression | 0.14090476 | 0.061418633 | 0.22039089 |
| Linear Regression | Deep Neural Network | -0.20741270 | -0.286898827 | -0.12792657 |
| SVM | Decision Tree | 0.08492063 | 0.005434506 | 0.16440676 |
| Random Forest | Decision Tree | 0.03666667 | -0.042819462 | 0.11615280 |
| Decision Tree | Deep Neural Network | -0.10317460 | -0.182660732 | -0.02368847 |
| SVM | Random Forest | 0.04825397 | -0.031232161 | 0.12774010 |
| SVM | Deep Neural Network | -0.01825397 | -0.097740097 | 0.06123216 |
| Random Forest | Deep Neural Network | -0.06650794 | -0.145994066 | 0.01297819 |

*Adjusted p-Value*

| | | | | | |
|---|---|---|---|---|---|
| Linear Regression | 1 | - | - | - | - |
| Decision Tree | 0.0041340 | 1 | - | - | - |
| SVM | 0.0000001 | 0.0303912 | 1 | - | - |
| Random Forest | 0.0000452 | 0.6971360 | 0.4406984 | 1 | - |
| Deep Neural Network | <0.0000001 | 0.0046519 | 0.9673788 | 0.1437956 | 1 |
| | Linear Regression | Decision Tree | SVM | Random Forest | Deep Neural Network |

**Tukey’s HSD Results for Performance at BRAIN**

| *Model 1* | *Model 2* | *Difference* | *Lower* | *Upper* |
|---|---|---|---|---|
| Linear Regression | Decision Tree | -0.135090446 | -0.18560311 | -0.08457779 |
| SVM | Linear Regression | 0.211848639 | 0.16133598 | 0.26236130 |
| Random Forest | Linear Regression | 0.215705797 | 0.16519314 | 0.26621846 |
| Linear Regression | Deep Neural Network | -0.449773810 | -0.50028647 | -0.39926115 |
| SVM | Decision Tree | 0.076758194 | 0.02624553 | 0.12727085 |
| Random Forest | Decision Tree | 0.080615351 | 0.03010269 | 0.13112801 |
| Decision Tree | Deep Neural Network | -0.314683364 | -0.36519602 | -0.26417070 |
| SVM | Random Forest | -0.003857158 | -0.05436982 | 0.04665550 |
| SVM | Deep Neural Network | -0.237925170 | -0.28843783 | -0.18741251 |
| Random Forest | Deep Neural Network | -0.234068012 | -0.28458067 | -0.18355535 |

*Adjusted p-Value*

| | | | | | |
|---|---|---|---|---|---|
| Linear Regression | 1 | - | - | - | - |
| Decision Tree | <0.0000001 | 1 | - | - | - |
| SVM | <0.0000001 | 0.0006182 | 1 | - | - |
| Random Forest | <0.0000001 | 0.0002936 | 0.9995189 | 1 | - |
| Deep Neural Network | <0.0000001 | <0.0000001 | <0.0000001 | <0.0000001 | 1 |
| | Linear Regression | Decision Tree | SVM | Random Forest | Deep Neural Network |

Table C.1 Tukey’s HSD Performance Results for Experiment I

## Appendix C.3. Tukey's HSD Explainability Results for Experiment I

**Tukey's HSD Results for Explainability in HEART**

| *Model 1* | *Model 2* | *Difference* | *Lower* | *Upper* |
|---|---|---|---|---|
| Linear Regression | Decision Tree | -1.38 | -1.9320503 | -0.82794965 |
| SVM | Linear Regression | -0.08 | -0.6320503 | 0.47205035 |
| Random Forest | Linear Regression | 0.87 | 0.3179497 | 1.42205035 |
| Linear Regression | Deep Neural Network | 1.51 | 0.9579497 | 2.06205035 |
| SVM | Decision Tree | -1.46 | -2.0120503 | -0.90794965 |
| Random Forest | Decision Tree | -0.51 | -1.0620503 | 0.04205035 |
| Decision Tree | Deep Neural Network | 2.89 | 2.3379497 | 3.44205035 |
| SVM | Random Forest | -0.95 | -1.5020503 | -0.39794965 |
| SVM | Deep Neural Network | 1.43 | 0.8779497 | 1.98205035 |
| Random Forest | Deep Neural Network | 2.38 | 1.8279497 | 2.93205035 |

*Adjusted p-Value*

| | | | | | |
|---|---|---|---|---|---|
| Linear Regression | 1 | - | - | - | - |
| Decision Tree | <0.0000001 | 1 | - | - | - |
| SVM | 0.9947633 | <0.0000001 | 1 | - | - |
| Random Forest | 0.0001870 | 0.0858495 | 0.0000314 | 1 | - |
| Deep Neural Network | <0.0000001 | <0.00000001 | <0.0000001 | <0.0000001 | 1 |
| | Linear Regression | Decision Tree | SVM | Random Forest | Deep Neural Network |

**Tukey's HSD Results for Explainability in BRAIN**

| *Model 1* | *Model 2* | *Difference* | *Lower* | *Upper* |
|---|---|---|---|---|
| Linear Regression | Decision Tree | -0.67326733 | -1.10541022 | -0.24112443 |
| SVM | Linear Regression | -0.05940594 | -0.49154883 | 0.3727395 |
| Random Forest | Linear Regression | 0.46534653 | 0.03320364 | 0.89748943 |
| Linear Regression | Deep Neural Network | 1.67326733 | 1.24112443 | 2.10541022 |
| SVM | Decision Tree | -0.73267327 | -1.16481616 | -0.30053037 |
| Random Forest | Decision Tree | -0.20792079 | -0.64006368 | 0.22422210 |
| Decision Tree | Deep Neural Network | 2.34653465 | 1.91439176 | 2.77867755 |
| SVM | Random Forest | -0.52475248 | -0.95689537 | -0.09260958 |
| SVM | Deep Neural Network | 1.61386139 | 1.18171849 | 2.04600428 |
| Random Forest | Deep Neural Network | 2.13861386 | 1.70647097 | 2.57075675 |

*Adjusted p-Value*

| | | | | | |
|---|---|---|---|---|---|
| Linear Regression | 1 | - | - | - | - |
| Decision Tree | 0.0002307 | 1 | - | - | - |
| SVM | 0.9957290 | 0.0000434 | 1 | - | - |
| Random Forest | 0.0275570 | 0.6806068 | 0.0084095 | 1 | - |
| Deep Neural Network | <0.0000001 | <0.0000001 | <0.0000001 | <0.0000001 | 1 |
| | Linear Regression | Decision Tree | SVM | Random Forest | Deep Neural Network |

Table C.2 Tukey's HSD Explainability Results for Experiment I

## Appendix C.4. Tukey's HSD Explainability Results for Experiment II

**Tukey’s HSD Results for Explainability in BRAIN (XAI)**

| *Model 1* | *Model 2* | *Difference* | *Lower* | *Upper* |
|---|---|---|---|---|
| How | Black-Box | 1.316327 | 0.694698 | 1.937955 |
| How-To | Black-Box | 2.428571 | 1.806943 | 3.0502 |
| What-Else | Black-Box | 3.510204 | 2.888575 | 4.131833 |
| Why | Black-Box | 3.714286 | 3.092657 | 4.335914 |
| Why-Not | Black-Box | 2.571429 | 1.9498 | 3.193057 |
| How-To | How | 1.112245 | 0.490616 | 1.733873 |
| What-Else | How | 2.193878 | 1.572249 | 2.815506 |
| Why | How | 2.397959 | 1.776331 | 3.019588 |
| Why-not | How | 1.255102 | 0.633473 | 1.876731 |
| What-Else | How-To | 1.081633 | 0.460004 | 1.703261 |
| Why | How-To | 1.285714 | 0.664086 | 1.907343 |
| Why-Not | How-To | 0.142857 | -0.47877 | 0.764486 |
| Why-Not | What-Else | 0.204082 | -0.41755 | 0.82571 |
| Why-Not | What-Else | -0.93878 | -1.5604 | -0.31715 |
| Why-Not | Why | -1.142860 | -1.76449 | -0.52123 |
| How | Black-Box | 1.316327 | 0.694698 | 1.937955 |

*Adjusted p-Value*

| | | | | | | |
|---|---|---|---|---|---|---|
| Black-Box | 1 | - | - | - | - | - |
| How | <0.0000001 | 1 | - | - | - | - |
| Why | <0.0000001 | <0.0000001 | 1 | - | - | - |
| Why-Not | <0.0000001 | 0.0000002 | 0.0000031 | 1 | - | - |
| How-To | <0.0000001 | 0.0000063 | 0.0000001 | 0.9863960 | 1 | - |
| What-Else | <0.0000001 | <0.0000001 | 0.9363242 | 0.0002666 | 0.0000127 | 1 |
| | Black-Box | How | Why | Why-Not | How-To | What-Else |

Table C.3 Tukey's HSD Explainability Results for Experiment II

# Appendix D. A Nascent Design Theory for Explainable Intelligent Systems

## Appendix D.1. Overview of Search String for Literature Review

| Database | Search string | Filter | # |
|---|---|---|---|
| Generic search string | (“Explainable AI” OR “XAI OR Machine Learning” OR “Black*” OR “Intelligent*”) AND (“User*” OR “UX” OR “UI” OR "Human*” OR “HCI” OR “Practic*” OR “Stakeholder”) AND (“Expla*” OR “Understand*” OR “Interpre*” OR “Transparency” OR “Comprehen*”) AND (“Design*” OR “Principle*” OR “Guideline” OR “Requirement”). | - | - |
| Science Direct | (“Explainable AI” OR XAI OR “Machine Learning”) AND (User) AND (Design OR Principle OR Guideline OR Requirement) | Abstract, title, author-specified keywords, research articles, review articles, practice, guidelines | 326 |
| EBSCO host | ((Explainable AI OR XAI OR Machine Learning OR Black* OR Intelligent*) AND (User* OR UX OR UI OR Human* OR HCI OR Practic* OR Stakeholder) AND (Expla* OR Understand* OR Interpre* OR Transparency OR Comprehen*) AND (Design* OR Principle* OR Guideline OR Requirement)) | No restrictions | 189 |
| IEEE Xplore | ((Explainable AI OR XAI OR Machine Learning OR Black* OR Intelligent*) AND (User* OR UX OR UI OR Human* OR HCI OR Practic OR Stakeholder) AND (Expla* OR Understand OR Interpre OR Transparency OR Comprehen*) AND (Design* OR Principle OR Guideline OR Requirement)) | Abstract, publication title, author keywords, conferences, journals, magazines | 235 |
| AISeL | abstract:( (Explainable AI OR XAI OR Machine Learning OR Black* OR Intelligent*) AND (User* OR UX OR UI OR Human* OR HCI OR Practic OR Stakeholder) AND (Expla* OR Understand OR Interpre OR Transparency OR Comprehen*) AND (Design* OR Principle OR Guideline OR Requirement)) AND title:( (Explainable AI OR XAI OR Machine Learning OR Black* OR Intelligent*) AND (User* OR UX OR UI OR Human* OR HCI OR Practic OR Stakeholder) AND (Expla* OR Understand OR Interpre OR Transparency OR Comprehen*) AND (Design* OR Principle OR Guideline OR Requirement)) | Abstract, title | 101 |
| ACM Digital Library | [[Abstract: explainable ai] OR [All: xai] OR [All: machine learning] OR [All: black*] OR [All: intelligent*]] AND [[All: user*] OR [All: ux] OR [All: ui] OR [All: human*] OR [All: hci] OR [All: practic*] OR [All: stakeholder]] AND [[All: expla*] OR [All: understand*] OR [All: interpre*] OR [All: transparency] OR [All: comprehen*]] AND [[All: design*] OR [All: principle*] OR [All: guideline] OR [All: requirement]] | Abstract,<br>journals, proceedings | 348 |

| | | | |
|---|---|---|---|
| Emerald Insight | title:”((Explainable AI OR XAI OR Machine Learning OR Black* OR Intelligent*) AND (User* OR UX OR UI OR Human* OR HCI OR Practic* OR Stakeholder) AND (Expla* OR Understand* OR Interpre* OR Transparency OR Comprehen*) AND (Design* OR Principle* OR Guideline OR Requirement))” AND (abstract: ”((Explainable AI OR XAI OR Machine Learning OR Black* OR Intelligent*) AND (User* OR UX OR UI OR Human* OR HCI OR Practic* OR Stakeholder) AND (Expla* OR Understand* OR Interpre* OR Transparency OR Comprehen*) AND (Design* OR Principle* OR Guideline OR Requirement))”) | Abstract, title, article, case study | 149 |
| Web of Science | (ALL= ((“Explainable AI” OR XAI OR “Machine Learning” OR Black OR Intelligent*) AND (User OR UX OR UI OR Human OR HCI OR Practic OR Stakeholder) AND (Expla OR Understand OR Interpre OR Transparency OR Comprehen) AND (Design OR Principle OR Guideline OR Requirement))) | Article, proceedings, paper, review | 78 |

Table D.1 Overview used Search Strings for Databases

### Appendix D.2. Iterations of the Theory-based Nascent Design Theory

**Iteration I.** To provide an overview of the topic explainable intelligent systems (EIS), we first analyzed and included an unstructured list of 20 publications. We identified these as comprehensive reappraisals of the topic area during the literature search (e.g., Amershi et al. 2019; Lim and Dey 2009; Mohseni et al. 2018). To this end, we synthesized design goals for EIS and used them as prior knowledge to derive meta design requirements (vom Brocke et al. 2020). Following the design theory development procedure of Möller et al. (2020), we then translated these meta design requirements into more output-related design requirements using the reasoning presented in the analyzed publications. Based on these design requirements, we defined an initial set of prescriptive instructions for our artifact in form of design principles and design features.

**Iteration II.** During the second iteration, we arranged and structured design (meta-) requirements, principles, and features in an explicit scheme, that is in a concept matrix according to Webster and Watson (2002). In addition, we expanded the initial literature set with 66 additional publications obtained in the literature search and analyzed the entire set of 86 publications.

**Iteration III.** Due to the large number of design elements in the third iteration cycle, we determined that specific properties were only considered relevant if they were mentioned in at least four publications. This restriction led to further condensation and summarization of individual design elements.

The following table describes an overview of the applied iterations.

| Iteration | Summary | # |
|---|---|---|
| I | Initial extraction and derivation of design (meta-) requirements and goals from the initial literature set. | MDR=3<br>DR=7<br>DP=32<br>DF=201<br>C= 20 |
| II | Analysis of the complete literature set to refine the schema. Mapping the derived design (meta-) requirements with the design principles and design features. | MDR=3<br>DR=5<br>DP=16<br>DF=62<br>C= 86 |
| III | Summary, restructuring and extension of design theory. Checking for compliance with quality criteria according to Möller et al. (2020). | MDR=3<br>DR=4<br>DP=4<br>DF=11<br>C=86 |

MDR = Meta design requirements; DR = Design requirements; DP = Design principles; DF = Design features; C = Count of contributions

Table D.2 Overview of Iterations

## Appendix D.3. Resulting Theory-based Nascent Design Theory

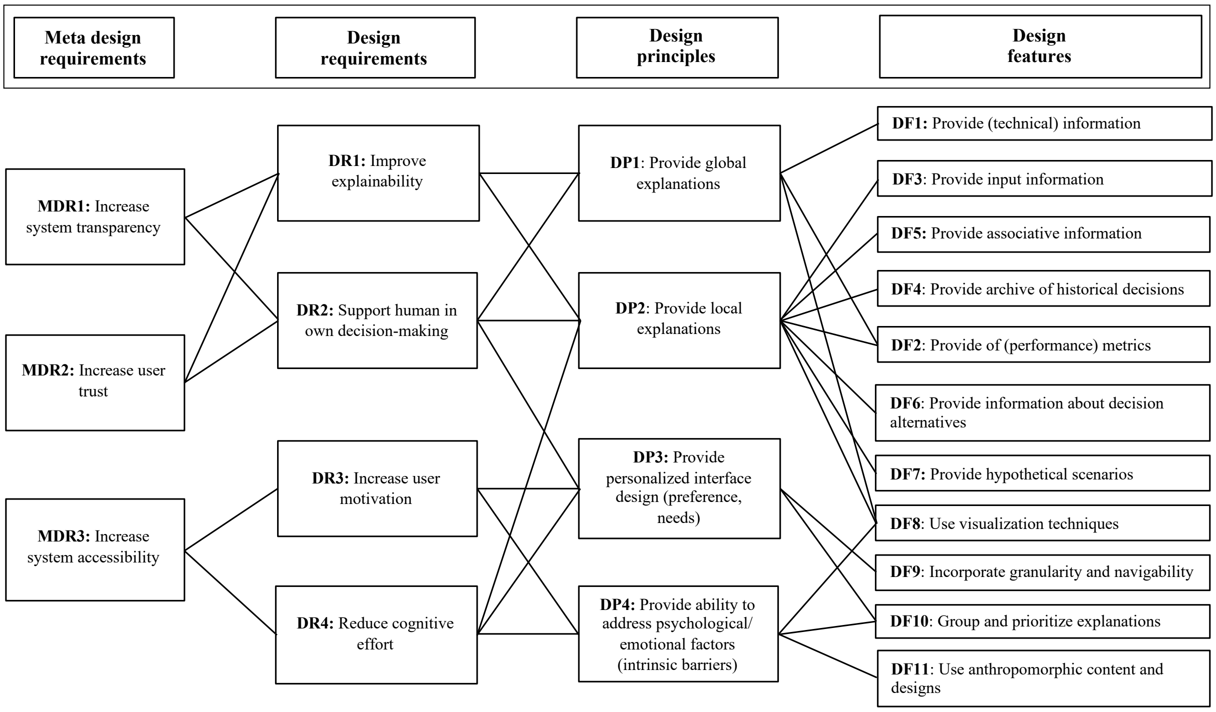

Meta design requirements
Design requirements
Design principles
Design features
MDR1: Increase system transparency
MDR2: Increase user trust
MDR3: Increase system accessibility
DR1: Improve explainability
DR2: Support human in own decision-making
DR3: Increase user motivation
DR4: Reduce cognitive effort
DP1: Provide global explanations
DP2: Provide local explanations
DP3: Provide personalized interface design (preference, needs)
DP4: Provide ability to address psychological/ emotional factors (intrinsic barriers)
DF1: Provide (technical) information
DF3: Provide input information
DF5: Provide associative information
DF4: Provide archive of historical decisions
DF2: Provide of (performance) metrics
DF6: Provide information about decision alternatives
DF7: Provide hypothetical scenarios
DF8: Use visualization techniques
DF9: Incorporate granularity and navigability
DF10: Group and prioritize explanations
DF11: Use anthropomorphic content and designs

Figure D.1 Theory-based Nascent Design Theory

## Appendix D.4. Evaluated Nascent Design Theory (First Design Cycle)

A comprehensive overview of the adjustment and evaluation of the theory-based design theory is shown in the following table. Likewise, Figure D.1 illustrates the operationalized design theory of the first design cycle. Lastly, in Table D.2 we present the concept matrix for the literature that we used in the design theory of the first design cycle.

**Iteration IV.** Within the fourth iteration, we presented our artifact within test interviews to make first adjustments to our theory-based design theory.

**Iteration V.** Within the last iteration of the first design cycle, we conducted eleven expert interviews to evaluate our design theory. First, based on their perception of potential barriers to intelligent system adoption, we carried out an initial completeness check of our meta design requirements. Second, due to the novelty of the topic, we further verified the appropriateness and completeness of our design theory through open discussions with the experts of the interview study, by asking if they would add, modify, or replace any design element.

| Iteration | Summary | # |
|---|---|---|
| IV | Restructuring theory based on two expert test interviews[1]. Revaluation of literature (cf. Table D.2). | MDR=3<br>DR=4<br>DP=4<br>DF=11 |
| V | Demonstration, enrichment, and evaluation of theory-based nascent design theory using eleven expert interviews[1]. | MDR=3<br>DR=4<br>DP=4<br>DF=11 |

MDR = Meta design requirements; DR = Design requirements; DP = Design principles; DF = Design features

[1] Please see Table 1 of the paper for more information.

Table D.3 Iteration within the Adjustment and Evaluation of the Nascent Design Theory

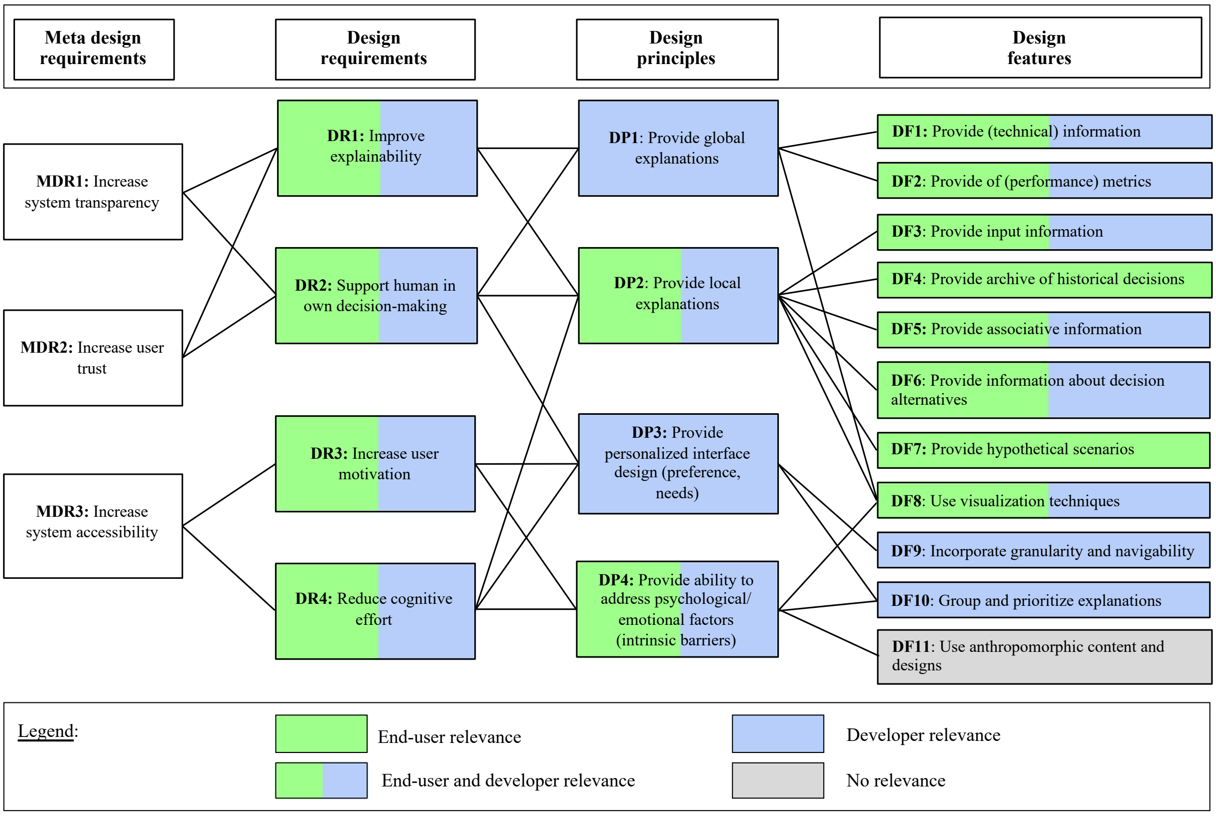

Meta design requirements
Design requirements
Design principles
Design features
MDR1: Increase system transparency
MDR2: Increase user trust
MDR3: Increase system accessibility
DR1: Improve explainability
DR2: Support human in own decision-making
DR3: Increase user motivation
DR4: Reduce cognitive effort
DP1: Provide global explanations
DP2: Provide local explanations
DP3: Provide personalized interface design (preference, needs)
DP4: Provide ability to address psychological/ emotional factors (intrinsic barriers)
DF1: Provide (technical) information
DF2: Provide of (performance) metrics
DF3: Provide input information
DF4: Provide archive of historical decisions
DF5: Provide associative information
DF6: Provide information about decision alternatives
DF7: Provide hypothetical scenarios
DF8: Use visualization techniques
DF9: Incorporate granularity and navigability
DF10: Group and prioritize explanations
DF11: Use anthropomorphic content and designs
Legend:
End-user relevance
Developer relevance
End-user and developer relevance
No relevance

Figure D.2 Evaluated Nascent Design Theory of first DSR Cycle

|  | Lim and Dey (2009) | Mohseni et al. (2021) | Amershi et al. (2019) | Meth et al. (2015) | Eurpean Commissions (2018) | Tintarev und Masthoff (2012) | van Esch (2018) | Amershi et al. (2014) | Uga (2019) | Dudley and Kristensson (2018) | Vorm et al. (2018) | Wang et al. (2019) | Bhatt et al. (2020) | Silveira et al. (2001) | Drozdal et al. (2020) | Kulesza et al. (2015) | Eiband et al. (2018) | Kulesza et al. (2013) | Pu und Chen (2007) | Chen et al. (2016) | Zhao et al. (2018) | Brennen (2020) | Kocielnik et al. (2019) | Kay et al. (2016) | Kuksenok et al. (2019) | Arrieta et al. (2020) | Lu et al. (2020) | Mueller et al. (2019) | Allahyari et al. (2011) | Arya et al. (2019) | Hind et al. (2019) | Kirsch (2017) | Kim et al. (2018) | Ras et al. (2018) | Wolf (2019, 2019a) | Glass (2008) | Hartmann (2009) | Wang et al. (2019a) | Hohman et al. (2019) | Preece et al. (2018) | Tomsett et al. (2018) | Hirsch et al. (2017) | Ferreira et al. (2020) | Nunes et al. (2012) | Eiband et al. (2018a) | Hamilton et al. (2014) | Arnold et al. (2004) | Riegelsberger und Sasse (2010) | Waytz et al. (2014) | Bohlender und Köhl (2019) | Ribeiro et al. (2016) | Rosenfeld und Richardson | Preece et al. (2018a) | Narayanan et al. (2018) | Doshi-Velez und Kim (2017 | Hall et al. (2019) | Viganò und Magazzeni (2018) | Rehse et al. (2019) | Nunes und Jannach (2017) | Brandao et al. (2019) | Nelles et al. (2016) | Cheng et al. (2019) | Shapiro (2018) | Stumpf et al. (2009) | Haynes et al. (2009) | McNee et al (2006) | Pieters et al (2011) | Oviatt et al. (2016) | Yang et al. (2018a,b) | Yang et al. (2018) | Chazette und Schneider (2020) | Bhatt et al. (2019) | Hepenstal und McNeish (2020) | Hoff und Bashir et al. (2015) | Lee und See (2004) | Wickramasinghe et al. (2020) | Ryan und Stahl (2020) | Zhou et al. (2020) | Ekman et al. (2018) | Abramoff et al. (2020) | Dove et al. (2017) |  |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Design Features** |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | ∑ |
| *Provide technical information* | x | x | x |  | x | x | x |  |  |  | x |  |  |  |  |  | x | x |  |  | x |  | x |  |  |  |  |  |  |  |  | x |  |  |  |  |  |  |  | x |  | x |  |  |  |  |  |  |  | x |  |  | x |  |  |  |  |  | x |  |  |  |  | x | x |  |  |  |  | x | x |  | x |  | x |  |  |  | x |  |  | 25 |
| *Provide input information* | x |  |  |  |  |  | x |  | x |  | x | x |  |  |  |  | x |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | x |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | x |  | x |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | x | x |  |  |  |  |  | x |  | 12 |
| *Provide (performance) metrics* | x |  | x |  |  |  | x |  | x | x |  |  |  |  |  |  |  |  |  |  |  | x |  |  |  | x |  |  |  |  |  | x |  |  |  | x |  |  |  | x |  |  |  |  |  |  |  |  |  |  |  |  |  | x |  |  |  | x | x |  |  |  |  |  | x |  |  |  |  |  |  | x |  | x | x |  |  |  | x |  |  | 20 |
| *Provide information about decision alternatives* | x |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | x |  |  |  |  |  |  |  | x |  |  |  |  |  | x | x |  |  |  |  |  |  |  | x |  |  |  |  |  |  |  |  |  |  |  |  | x |  | 7 |
| *Provide hypothetical scenarios* | x | x |  |  |  |  |  | x |  |  |  |  |  |  |  |  |  |  |  |  |  |  | x |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | x |  | x |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | x |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | 7 |
| *Provide archive of historical decisions* | x |  | x |  |  | x |  |  |  |  | x |  |  | x |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | x |  |  |  | x |  |  |  |  |  |  | 7 |
| *Provide associative information* | x |  | x |  |  |  |  |  |  |  |  |  | x |  |  | x |  |  |  |  |  | x |  |  |  |  |  | x |  |  |  | x |  |  |  |  |  |  |  |  | x |  |  |  |  |  |  |  |  |  |  |  | x |  |  |  | x |  |  |  |  |  |  |  | x |  |  |  |  |  | x |  |  |  | x |  |  |  | x |  |  | 14 |
| *Use visualization techniques* | x | x |  |  |  |  |  |  | x | x |  |  |  |  |  |  | x |  |  |  |  |  | x |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | x |  | x |  | x | x |  | x |  |  |  |  |  |  |  |  | x |  | x |  |  |  | x | x |  |  | x |  |  |  | 17 |
| *Incorporate granularity and navigability* |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | x | x | x |  | x |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | x |  |  |  |  |  |  | x |  |  |  |  |  |  |  |  |  |  |  |  |  | x |  |  | x |  |  |  |  |  |  |  |  |  |  | x |  | x |  |  |  |  |  |  | 12 |
| *Group and prioritize explanations* |  |  |  |  |  |  |  |  | x |  |  |  |  |  |  |  |  |  | x | x |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | x |  |  |  | x | x |  |  |  |  | x | x |  |  |  |  |  | x |  |  |  |  |  |  |  |  |  |  |  | x |  |  |  |  |  |  |  |  |  |  |  |  | x | 13 |
| *Use anthropomorphic content and designs* |  |  |  |  |  |  |  |  | x |  |  |  |  |  |  |  |  |  |  |  |  |  | x |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | x | x |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | x |  |  |  | x | x |  |  | x | x |  | x | 10 |
| **Design Principles** |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | ∑ |
| *Provide global explanations* | x | x | x |  | x | x | x |  | x | x |  |  | x |  |  |  | x | x |  |  | x |  | x |  |  |  | x |  |  |  |  | x |  |  |  | x |  |  |  | x | x | x | x |  |  | x | x |  |  | x |  | x | x |  | x |  |  |  | x |  |  |  |  | x | x |  |  |  | x |  | x |  | x |  |  | x | x | x | x |  |  | 40 |
| *Provide local explanations* | x | x |  |  |  | x |  |  | x |  |  | x | x |  | x | x |  |  |  |  |  | x |  |  |  | x | x | x | x | x | x | x |  | x |  | x | x |  |  |  | x |  | x |  |  | x | x | x | x | x | x | x | x |  | x | x | x | x | x | x |  |  |  |  | x |  | x |  |  |  | x |  | x |  |  | x | x | x |  |  |  | 47 |
| *Provide personalized interface design (preference, needs)* |  | x |  |  |  | x |  |  | x |  |  |  |  | x |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | x |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | x |  |  | x |  |  | x |  |  |  |  |  | x |  |  |  |  |  | x |  |  |  | x |  |  |  | 13 |
| *Provide ability to address psychological/emotional factors (intrinsic barriers)* |  | x | x |  |  |  |  |  | x | x |  |  |  |  |  |  |  | x |  |  |  |  | x |  |  |  |  |  |  |  |  |  |  |  |  | x | x |  | x |  |  | x | x | x |  | x |  |  |  | x | x |  | x |  |  | x | x |  | x |  |  | x | x |  |  | x |  | x | x |  | x | x | x | x | x | x |  | x |  |  |  | 33 |
| **Design Requirements** |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | ∑ |
| *Improve intelligibility of system's decision* | x | x | x |  | x | x | x |  |  |  | x |  |  |  |  |  | x | x |  |  | x |  | x |  |  |  |  |  |  |  |  | x |  |  |  |  |  |  |  | x |  | x |  |  |  |  |  |  |  | x |  |  | x |  |  |  |  |  | x |  |  |  |  | x | x |  |  |  |  | x | x |  | x |  | x |  |  |  | x |  |  | 25 |
| *Support human in own decision-making* |  | x |  |  |  | x |  |  | x |  |  | x |  | x |  | x |  |  |  |  | x |  | x |  |  |  |  |  |  | x |  |  |  |  |  |  |  |  |  |  |  | x |  | x |  | x |  |  |  | x | x |  | x |  |  | x | x |  | x |  |  | x |  |  |  | x |  | x |  | x | x |  | x | x |  |  |  | x | x | x | x | 31 |
| *Increase user motivation* |  | x | x |  |  |  |  |  | x | x |  |  |  |  |  | x |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | x | x |  | x |  |  | x | x |  |  |  |  |  |  |  | x |  | x |  |  | x |  |  | x |  |  | x | x |  |  |  |  | x | x |  |  | x | x |  | x | x |  | x | x |  |  | 25 |
| *Reduce cognitive effort* |  |  |  | x |  |  | x |  | x |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | x |  |  | x |  |  |  |  |  |  |  |  | x |  |  |  |  |  |  |  |  |  |  | x |  |  | 10 |
| **Meta Design Requirements** |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | ∑ |
| *Increase system transparency* | x | x | x |  | x | x |  |  | x |  | x |  |  |  |  |  | x | x |  |  | x |  | x |  |  |  |  |  |  |  |  | x |  | x |  |  |  |  |  |  |  | x |  |  |  |  |  |  |  | x |  |  | x |  |  |  |  |  |  |  |  |  |  | x |  |  |  |  |  |  | x |  | x |  | x |  |  |  |  |  |  | 22 |
| *Increase system user trust* | x |  | x |  |  |  | x |  | x | x |  |  |  |  |  |  |  |  |  |  |  | x |  |  |  | x |  |  |  |  |  | x |  |  |  | x |  |  |  | x |  |  |  |  |  |  |  |  |  |  |  |  |  | x |  |  |  | x | x |  |  |  |  |  | x |  |  |  |  |  |  | x |  | x | x |  |  |  | x |  |  | 20 |
| *Increase system accessibility* | x | x | x |  |  | x | x |  | x |  | x | x |  | x |  | x |  |  |  |  |  |  | x |  |  |  |  |  |  |  |  |  |  |  |  | x |  | x |  | x | x |  | x |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | x |  |  | x |  |  |  |  | x | x |  | x | x |  |  | x |  |  |  | 24 |

Table D.4 Concept Matrix for Nascent Design Theory

## Appendix D.5. Applied Use Case Evaluation (Second Design Cycle)

| Evaluation Criteria | Construct | Statement |
|---|---|---|
| Accessibility | AC1 | The design principles are easy for me to understand. |
| | AC2 | The design principles are easy for me to comprehend. |
| | AC3 | The design principles are intelligible to me. |
| Importance | IM1 | In my view the design principles address a real problem in my professional practice. |
| | IM2 | In my view the design principles address an - acute or foreseeable - important problem in my professional practice. |
| Novelty and insightfulness | NI1 | I find that the design principles convey new ideas to me. |
| | NI2 | I find the design principles insightful to my own practice. |
| Actability and guidance | AG1 | I think that the design principles can realistically be carried out in practice. |
| | AG2 | I think that the design principles can easily be carried out in practice. |
| | AG3 | I think the design principles provide sufficient guidance for designing an explainable intelligent system. |
| | AG4 | I think that the design principles are not restrictive when designing an explainable intelligent system. |
| Effectiveness | EF1 | I believe that the design principles can help design explainable intelligent systems. |
| | EF2 | I find the design principles useful for designing explainable intelligent systems in practice. |
| | EF3 | Compared to my current situation, I believe that the design principles would increase my productivity in using an explainable intelligent system. |
| | EF4 | Compared to my current situation, I believe that the design principles would enhance my effectiveness in using an explainable intelligent system. |

Table D.5 Adapted Questionnaire for Evaluating the Design Principles according to Iivari et al. (2021)

# References

Abdul, A., Vermeulen, J., Wang, D., Lim, B. Y., & Kankanhalli, M. (2018). *Trends and trajectories for explainable, accountable and intelligible systems: An hci research agenda*. CHI Conference on Human Factors in Computing Systems, New York City, NY, USA.

Abdul, A., Weth, C. v. d., Kankanhalli, M., & Lim, B. Y. (2020). *COGAM: Measuring and Moderating Cognitive Load in Machine Learning Model Explanations*. CHI Conference on Human Factors in Computing Systems, Honolulu, HI, USA.

Abedin, B., Meske, C., Junglas, I., Rabhi, F., & Motahari-Nezhad, H. R. (2022). Designing and Managing Human-AI Interactions. *Information Systems Frontiers*, *24*, 691-697.

Aboulian, A., Green, D. H., Switzer, J. F., Kane, T. J., Bredariol, G. V., Lindahl, P., Donnal, J. S., & Leeb, S. B. (2018). NILM dashboard: A power system monitor for electromechanical equipment diagnostics. *IEEE Transactions on Industrial Informatics*, *15*(3), 1405-1414.

Achinstein, P. (1968). *Concepts of science: A philosophical analysis*. Johns Hopkins Press.

Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on Explainable Artificial Intelligence (XAI). *IEEE Access*, *6*, 52138-52160.

Ågerfalk, P. J., Conboy, K., Crowston, K., Eriksson Lundström, J. S., Jarvenpaa, S., Ram, S., & Mikalef, P. (2022). Artificial intelligence in information systems: State of the art and research roadmap. *Communications of the association for information systems*, *50*(1), 420-438.

Alaiad, A., & Zhou, L. (2013). *Patients' behavioral intention toward using healthcare robots*. Americas Conference on Information Systems, Chicago, Illinois, USA.

Albashrawi, M., & Motiwalla, L. (2017). *When IS success model meets UTAUT in a mobile banking context: a study of subjective and objective system usage*. Swedish Artificial Intelligence Society, Karlskrona, Sweden.

Albert, W., Tullis, T., & Tedesco, D. (2009). *Beyond the usability lab: Conducting large-scale online user experience studies*. Morgan Kaufmann.

Alharbi, S. T. (2014). *Trust and acceptance of cloud computing: A revised UTAUT model*. International conference on computational science and computational intelligence, Las Vegas, NV, USA.

Alipour, K., Ray, A., Lin, X., Schulze, J. P., Yao, Y., & Burachas, G. T. (2020). *The impact of explanations on AI competency prediction in VQA*. 2020 IEEE International Conference on Humanized Computing and Communication with Artificial Intelligence (HCCAI), Irvine, CA, USA.

Allahyari, H., & Lavesson, N. (2011). *User-oriented assessment of classification model understandability*. 11th Scandinavian Conference on Artificial Intelligence, Trondheim, Norway.

Ambady, N., & Rosenthal, R. (1992). Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis. *Psychological bulletin*, *111*(2), 256-274.

Amershi, S., Cakmak, M., Knox, W. B., & Kulesza, T. (2014). Power to the people: The role of humans in interactive machine learning. *AI magazine*, *35*(4), 105-120.

Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., Suh, J., Iqbal, S., Bennett, P. N., & Inkpen, K. (2019). *Guidelines for human-AI interaction*. CHI Conference on Human Factors in Computing Systems, Glasgow, Scottland.

Anderson, A., Dodge, J., Sadarangani, A., Juozapaitis, Z., Newman, E., Irvine, J., Chattopadhyay, S., Olson, M., Fern, A., & Burnett, M. (2020). Mental models of mere mortals with explanations of reinforcement learning. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, *10*(2), 1-37.

Angelov, P., & Soares, E. (2019). Towards Explainable Deep Neural Networks (xDNN). *arXiv:1912.02523*.

Angelov, P., & Soares, E. (2020). Towards explainable deep neural networks (xDNN). *Neural networks*, *130*, 185-194.

Arnold, V., Clark, N., Collier, P. A., Leech, S. A., & Sutton, S. G. (2004). Explanation provision and use in an intelligent decision aid. *Intelligent Systems in Accounting, Finance & Management: International Journal*, *12*(1), 5-27.

Arnott, D., & Pervan, G. (2005). A critical analysis of decision support systems research. In M. Willcocks L. & Sauer L. & Lacity (Eds.), *Formulating Research Methods for Information Systems* (pp. 67-87). Palgrave Macmillan.

Arnott, D., & Pervan, G. (2008). Eight key issues for the decision support systems discipline. *Decision support systems*, *44*(3), 657-672.

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., & Benjamins, R. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information fusion*, *58*, 82-115.

Arya, V., Bellamy, R. K., Chen, P.-Y., Dhurandhar, A., Hind, M., Hoffman, S. C., Houde, S., Liao, Q. V., Luss, R., & Mojsilović, A. (2019). One explanation does not fit all: A toolkit and taxonomy of ai explainability techniques. *arXiv preprint arXiv:1909.03012*.

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.

Atkinson, K., Bench-Capon, T., & Bollegala, D. (2020). Explanation in AI and law: Past, present and future. *Artificial Intelligence*, *289*, 103387.

Axmann, B., Harmoko, H., Herm, L.-V., & Janiesch, C. (2021). *A framework of cost drivers for robotic process automation projects*. International Conference on Business Process Management: Blockchain and Robotic Process Automation Forum, Rome, Italy.

Bahari, A. (2022). Challenges and Affordances of Cognitive Load Management in Technology-Assisted Language Learning: A Systematic Review. *International Journal of Human-Computer Interaction*, 1-16.

Baird, A., & Maruping, L. M. (2021). The Next Generation Of Research On Is Use: A Theoretical Framework Of Delegation To And From Agentic Is Artifacts. *MIS Quarterly*, *45*(1b), 315-341.

Baishya, K., & Samalia, H. V. (2020). Extending unified theory of acceptance and use of technology with perceived monetary value for smartphone adoption at the bottom of the pyramid. *International Journal of Information Management*, *51*, 102036.

Bandura, A. (2001). Social cognitive theory: An agentic perspective. *Annual review of psychology*, *52*(1), 1-26.

Barda, A. J., Horvat, C. M., & Hochheiser, H. (2020). A qualitative research framework for the design of user-centered displays of explanations for machine learning model predictions in healthcare. *BMC Medical Informatics and Decision Making*, *20*(1), 1-16.

Barrouillet, P., Bernardin, S., Portrat, S., Vergauwe, E., & Camos, V. (2007). Time and cognitive load in working memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*(3), 570-585.

Baskerville, R., Baiyere, A., Gregor, S., Hevner, A., & Rossi, M. (2018). Design science research contributions: Finding a balance between artifact and theory. *Journal of the Association for Information Systems*, *19*(5), 338-376.

Baskerville, R. L., & Pries-Heje, J. (2019). Projectability in Design Science Research. *Journal of Information Technology Theory And Application*, *20*(1), 53-76.

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*, 79-82.

Bauer, K., von Zahn, M., & Hinz, O. (2023). Expl (AI) ned: The Impact of Explainable Artificial Intelligence on Users' Information Processing. *Information systems research*.

Baumgart, M., Bredebach, P., Herm, L.-V., Hock, D., Hofmann, A., Janiesch, C., Jankowski, L. O., Kampik, T., Keil, M., & Kolb, J. (2022). Plattform für das integrierte Management von Kollaborationen in Wertschöpfungsnetzwerken (PIMKoWe). In Winkelmann A. & Janiesch C. (Eds.), *Working Paper Series of the Institute of Business Management, University of Wuerzburg* (Vol. 8, pp. 1-248). OPUS Würzburg.

Benbya, H., Pachidi, S., & Jarvenpaa, S. (2021). Special Issue Editorial: Artificial Intelligence in Organizations: Implications for Information Systems Research. *Journal of the Association for Information Systems*, *22*(2), 10.

Bentele, G., & Seidenglanz, R. (2015). Vertrauen und Glaubwürdigkeit. Begriffe, Ansätze, Forschungsübersicht und praktische Relevanz. In Fröhlich R. & Szyska P. & Bentele G. (Eds.), *Handbuch der Public Relations. Wissenschaftliche Grundlagen und berufliches Handeln. Mit Lexikon* (Vol. 3, pp. 411-430). Springer.

Berente, N., Gu, B., Recker, J., & Santhanam, R. (2021). Managing artificial intelligence. *MIS Quarterly*, *45*(3), 1433-1450.

Berger, B., Adam, M., Rühr, A., & Benlian, A. (2021). Watch me improve—Algorithm aversion and demonstrating the ability to learn. *Business & Information Systems Engineering*, *63*(1), 55-68.

Bertrand, A., Belloum, R., Eagan, J. R., & Maxwell, W. (2022). *How cognitive biases affect XAI-assisted decision-making: A systematic review*. AAAI/ACM conference on AI, ethics, and society, Oxford, UK.

Bhatt, U., Andrus, M., Weller, A., & Xiang, A. (2020). Machine learning explainability for external stakeholders. *arXiv preprint arXiv:2007.05408*.

Bigley, G. A., & Pearce, J. L. (1998). Straining for shared meaning in organization science: Problems of trust and distrust. *Academy of management review*, *23*(3), 405-421.

Bilgic, M., & Mooney, R. J. (2005). *Explaining recommendations: Satisfaction vs. promotion*. International Conference on Intelligent User Interfaces: A Workshop on the Next Stage of Recommender Systems Research, San Diego, CA, USA.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.

Blanco-Justicia, A., & Domingo-Ferrer, J. (2019). *Machine learning explainability through comprehensible decision trees*. International Cross-Domain Conference for Machine Learning and Knowledge Extraction, Dublin.

Blut, M., Chong, A., Tsiga, Z., & Venkatesh, V. (2021). Meta-Analysis Of The Unified Theory Of Acceptance And Use Of Technology (UTAUT): Challenging Its Validity And Charting A Research Agenda In The Red Ocean. *Journal of the Association for Information Systems*, *23*(1), 13-95.

Bohaju, J. (2020). *Brain Tumor*. Retrieved 27.04.2022 from https://doi.org/10.34740/KAGGLE/DSV/1370629

Bohlender, D., & Köhl, M. A. (2019). Towards a characterization of explainable systems. *arXiv preprint arXiv:1902.03096*.

Boone, H. N., & Boone, D. A. (2012). Analyzing likert data. *Journal of extension*, *50*(2), 1-5.

Bove, C., Aigrain, J., Lesot, M.-J., Tijus, C., & Detyniecki, M. (2022). *Contextualization and Exploration of Local Feature Importance Explanations to Improve Understanding and Satisfaction of Non-Expert Users*. International Conference on Intelligent User Interfaces, Helsinki, Finland.

Brandão, R., Carbonera, J., de Souza, C., Ferreira, J., Gonçalves, B., & Leitão, C. (2019). Mediation Challenges and Socio-Technical Gaps for Explainable Deep Learning Applications. *arXiv preprint arXiv:1907.07178*.

Brennen, A. (2020). *What Do People Really Want When They Say They Want "Explainable AI? " We Asked 60 Stakeholders*. CHI Conference on Human Factors in Computing Systems, Virtual Conference.

Bröhl, C., Nelles, J., Brandl, C., Mertens, A., & Schlick, C. M. (2016). *TAM reloaded: a technology acceptance model for human-robot cooperation in production systems*. International conference on human-computer interaction, Toronto, Canada.

Brown, S. A., Dennis, A. R., & Venkatesh, V. (2010). Predicting collaboration technology use: Integrating technology adoption and collaboration research. *Journal of Management Information Systems*, *27*(2), 9-54.

Brunk, J., Mattern, J., & Riehle, D. M. (2019). *Effect of transparency and trust on acceptance of automatic online comment moderation systems*. Conference on Business Informatics (CBI), Moscow, Russia.

Brynjolfsson, E., & Mcafee, A. (2017). The business of artificial intelligence. *Harvard Business Review*, *7*, 3-11.

Buçinca, Z., Lin, P., Gajos, K. Z., & Glassman, E. L. (2020). *Proxy tasks and subjective measures can be misleading in evaluating explainable AI systems*. International conference on intelligent user interfaces, Cagliari, Italy.

Bunde, E. (2021). *AI-Assisted and explainable hate speech detection for social media moderators-A design science approach*. 54th Hawaii International Conference on System Sciences, Honolulu, Hawaii, USA.

Burton, J. W., Stein, M. K., & Jensen, T. B. (2020). A systematic review of algorithm aversion in augmented decision making. *Journal of Behavioral Decision Making*, *33*(2), 220-239.

Cabitza, F., Campagner, A., & Sconfienza, L. M. (2020). As if sand were stone. New concepts and metrics to probe the ground on which to build trustable AI. *BMC Medical Informatics and Decision Making*, *20*(1), 1-21.

Cai, C. J., Reif, E., Hegde, N., Hipp, J., Kim, B., Smilkov, D., Wattenberg, M., Viegas, F., Corrado, G. S., & Stumpe, M. C. (2019). *Human-centered tools for coping with imperfect algorithms during medical decision-making*. CHI Conference on Human Factors in Computing Systems, Glasgow, Scottland.

Calegari, R., Ciatto, G., & Omicini, A. (2020). On the integration of symbolic and sub-symbolic techniques for XAI: A survey. *Intelligenza Artificiale*, *14*(1), 7-32.

Carter, L., & Bélanger, F. (2005). The utilization of e-government services: citizen trust, innovation and acceptance factors. *Information systems journal*, *15*(1), 5-25.

Carvalho, T. P., Soares, F. A., Vita, R., Francisco, R. d. P., Basto, J. P., & Alcalá, S. G. (2019). A systematic literature review of machine learning methods applied to predictive maintenance. *Computers & Industrial Engineering*, *137*, 106024.

Castelo, N., Bos, M. W., & Lehmann, D. R. (2019). Task-dependent algorithm aversion. *Journal of Marketing Research*, *56*(5), 809-825.

Castiglioni, I., Rundo, L., Codari, M., Di Leo, G., Salvatore, C., Interlenghi, M., Gallivanone, F., Cozzi, A., D'Amico, N. C., & Sardanelli, F. (2021). AI applications to medical images: From machine learning to deep learning. *Physica Medica*, *83*, 9-24.

Chander, A., Srinivasan, R., Chelian, S., Wang, J., & Uchino, K. (2018). *Working with beliefs: AI transparency in the enterprise*. CEUR-WS IUI Workshops, Tokyo, Japan.

Chandra Kruse, L., Purao, S., & Seidel, S. (2022). How Designers Use Design Principles: Design Behaviors and Application Modes. *Journal of the Association for Information Systems 23*(5), 1235-1270.

Chandra, M. A., & Bedi, S. (2021). Survey on SVM and their application in image classification. *International Journal of Information Technology*, *13*(5), 1-11.

Chazette, L., & Schneider, K. (2020). Explainability as a non-functional requirement: challenges and recommendations. *Requirements Engineering*, *25*(4), 493-514.

Chen, T.-W., & Sundar, S. S. (2018). *This app would like to use your current location to better serve you: Importance of user assent and system transparency in personalized mobile services*. CHI Conference on Human Factors in Computing Systems, New York, NY, USA.

Chen, X.-B. (2013). Tablets for informal language learning: Student usage and attitudes. *Language learning & technology*, *17*(1), 20-36.

Cheng, D., Liu, G., Qian, C., & Song, Y.-F. (2008). *Customer acceptance of internet banking: integrating trust and quality with UTAUT model*. International Conference on Service Operations and Logistics, and Informatics, Beijing, China.

Cheng, H.-F., Wang, R., Zhang, Z., O'Connell, F., Gray, T., Harper, F. M., & Zhu, H. (2019). Explaining decision-making algorithms through UI: Strategies to help non-expert stakeholders. chi conference on human factors in computing systems, New York, USA.

Chin, W. W., & Newsted, P. R. (1999). Structural equation modeling analysis with small samples using partial least squares. *Statistical strategies for small sample research*, *1*(1), 307-341.

Chiu, Y.-T., Zhu, Y.-Q., & Corbett, J. (2021). In the hearts and minds of employees: A model of pre-adoptive appraisal toward artificial intelligence in organizations. *International Journal of Information Management*, *60*, 102379.

Choi, J. K., & Ji, Y. G. (2015). Investigating the importance of trust on adopting an autonomous vehicle. *International Journal of Human-Computer Interaction*, *31*(10), 692-702.

Chou, Y.-L., Moreira, C., Bruza, P., Ouyang, C., & Jorge, J. (2022). Counterfactuals and causability in explainable artificial intelligence: Theory, algorithms, and applications. *Information fusion*, *81*, 59-83.

Chromik, M., & Butz, A. (2021). *Human-xai interaction: A review and design principles for explanation user interfaces*. IFIP Conference on Human-Computer Interaction, Dublin, Ireland.

Chromik, M., & Schuessler, M. (2020). *A Taxonomy for Human Subject Evaluation of Black-Box Explanations in XAI*. Explainable Smart Systems (ExSS) & Intelligent User Interfaces for Algorithmic Transparency in Emerging Technologies (ATEC), Cagliari, Italy.

Chui, M., & Malhotra, S. (2018, 23.12.2022). *Ai adoption advances, but foundational barriers remain*. Retrieved 23.12.2022 from https://www.mckinsey.com/featured-insights/artificial-intelligence/ai-adoption-advances-but-foundational-barriers-remain

Cirqueira, D., Helfert, M., & Bezbradica, M. (2021). *Towards design principles for user-centric explainable AI in fraud detection*. International Conference on Human-Computer Interaction, Virtual Conference.

Cleven, A., Gubler, P., & Hüner, K. M. (2009). *Design alternatives for the evaluation of design science research artifacts*. Proceedings of the 4th International Conference on Design Science Research in Information Systems and Technology, Philadelphia, Pennsylvania, USA.

Cody-Allen, E., & Kishore, R. (2006). *An extension of the UTAUT model with e-quality, trust, and satisfaction constructs*. Conference on computer personnel research: Forty four years of computer personnel research: achievements, challenges & the future, New York, NY, USA.

Collins, C., Dennehy, D., Conboy, K., & Mikalef, P. (2021). Artificial intelligence in information systems research: A systematic literature review and research agenda. *International Journal of Information Management*, *60*, 102383.

Commission, E. (2019). *Ethics Guidelines for Trustworthy AI*. Retrieved 19.09.2021 from https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai

Compeau, D., Higgins, C. A., & Huff, S. (1999). Social cognitive theory and individual reactions to computing technology: A longitudinal study. *MIS Quarterly*, 145-158.

Compeau, D. R., & Higgins, C. A. (1995). Computer self-efficacy: Development of a measure and initial test. *MIS Quarterly*, 189-211.

Conati, C., Barral, O., Putnam, V., & Rieger, L. (2021). Toward personalized XAI: A case study in intelligent tutoring systems. *Artificial Intelligence*, *298*, 103503.

Cooper, H. M. (1988). Organizing knowledge syntheses. A taxonomy of literature reviews. *Knowledge in Society*, *1*(1), 104-126.

Cortez, P. (2009, 20.07.2020). *Viticulture Commission of the Vinho Verde Region*. Retrieved 20.07.2020 from archive.ics.uci.edu/ml/datasets/wine+quality

Cramer, H., Evers, V., Ramlal, S., Van Someren, M., Rutledge, L., Stash, N., Aroyo, L., & Wielinga, B. (2008). The effects of transparency on trust in and acceptance of a content-based art recommender. *User Modeling and User-adapted interaction*, *18*(5), 455.

Cui, X., Lee, J. M., & Hsieh, J. (2019). *An Integrative 3C evaluation framework for Explainable Artificial Intelligence*. American Conference on Information Systems, Cancun.

Dam, H. K., Tran, T., & Ghose, A. (2018). *Explainable software analytics*. nternational Conference on Software Engineering: New Ideas and Emerging Results, Gothenburg, Sweden.

Das, A., & Rad, P. (2020). Opportunities and challenges in explainable artificial intelligence (xai): A survey. *arXiv preprint arXiv:2006.11371*.

Davis, B., Glenski, M., Sealy, W., & Arendt, D. (2020). *Measure utility, gain trust: practical advice for XAI researchers*. Workshop on TRust and EXpertise in Visual Analytics (TREX), Salt Lake City, UT, USA.

Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly*, 319-340.

Davis, F. D., Bagozzi, R. P., & Warshaw, P. R. (1992). Extrinsic and intrinsic motivation to use computers in the workplace *Journal of applied social psychology*, *22*(14), 1111-1132.

Dellermann, D., Ebel, P., Söllner, M., & Leimeister, J. M. (2019). Hybrid intelligence. *Business & Information Systems Engineering*, *61*(5), 637-643.

Demigha, S. (2021). *Decision Support Systems (DSS) and Management Information Systems (MIS) in Today's Organizations*. European Conference on Research Methodology for Business and Management Studies, Aveiro, Portugal.

Demissie, D., Alemu, D., & Rorissa, A. (2021). An Investigation into user Adoption of Personal Safety Devices in Higher Education Using the Unified Theory of Acceptance and Use of Technology (UTAUT). *The Journal of the Southern Association for Information Systems*, *8*(1), 1-18.

DeVries, P. M., Viégas, F., Wattenberg, M., & Meade, B. J. (2018). Deep learning of aftershock patterns following large earthquakes. *Nature*, *560*(7720), 632-634.

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, *144*(1), 114-126.

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2016). Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. *Management Science*, *64*(3), 1155-1170.

Dikmen, M., & Burns, C. (2022). The effects of domain knowledge on trust in explainable AI and task performance: A case of peer-to-peer lending. *International Journal of Human-Computer Studies*, *162*, 102792.

Dominguez, V., Donoso-Guzmán, I., Messina, P., & Parra, D. (2020). Algorithmic and HCI Aspects for Explaining Recommendations of Artistic Images. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, *10*(4), 1-31.

Doran, D., Schulz, S., & Besold, T. R. (2017). What does explainable AI really mean? A new conceptualization of perspectives. *arXiv preprint arXiv:1710.00794*.

Dos Santos, D. P., Giese, D., Brodehl, S., Chon, S., Staab, W., Kleinert, R., Maintz, D., & Baeßler, B. (2019). Medical students' attitude towards artificial intelligence: a multicentre survey. *European radiology*, *29*(4), 1640-1646.

Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.

Došilović, F. K., Brčić, M., & Hlupić, N. (2018). *Explainable artificial intelligence: A survey*. 2018 41st International convention on information and communication technology, electronics and microelectronics (MIPRO), Opatija.

Dove, G., Halskov, K., Forlizzi, J., & Zimmerman, J. (2017). *UX design innovation: Challenges for working with machine learning as a design material*. CHI Conference on Human Factors in Computing Systems, Denver, Colorado, USA.

Drechsler, A., & Hevner, A. R. (2018). *Utilizing, producing, and contributing design knowledge in DSR projects*. International Conference on Design Science Research in Information Systems and Technology, Chennai, India.

Drozdal, J., Weisz, J., Wang, D., Dass, G., Yao, B., Zhao, C., Muller, M., Ju, L., & Su, H. (2020). *Trust in AutoML: exploring information needs for establishing trust in automated machine learning systems*. International Conference on Intelligent User Interfaces, New York, NY, USA.

Dudley, J. J., & Kristensson, P. O. (2018). A review of user interface design for interactive machine learning. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, *8*(2), 1-37.

Duval, A. (2019). Explainable Artificial Intelligence (XAI). *MA4K9 Scholarly Report, Mathematics Institute, The University of Warwick*.

Dwivedi, R., Dave, D., Naik, H., Singhal, S., Rana, O., Patel, P., Qian, B., Wen, Z., Shah, T., & Morgan, G. (2023). Explainable AI (XAI): core ideas, techniques and solutions. *ACM Computing Surveys*, *55*(9), 1-33.

Dwivedi, Y. K., Hughes, L., Ismagilova, E., Aarts, G., Coombs, C., Crick, T., Duan, Y., Dwivedi, R., Edwards, J., & Eirug, A. (2021). Artificial Intelligence (AI): Multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy. *International Journal of Information Management*, *57*, 101994.

Dwivedi, Y. K., Rana, N. P., Jeyaraj, A., Clement, M., & Williams, M. D. (2019). Re-examining the unified theory of acceptance and use of technology (UTAUT): Towards a revised theoretical model. *Information Systems Frontiers*, *21*(3), 719-734.

Ebers, M. (2020). *Regulating Explainable AI in the European Union. An Overview of the Current Legal Framework (s)*. Law and Informatics 2020: Law in the Era of Artificial Intelligence, Stockholm.

Ekman, F., Johansson, M., & Sochor, J. (2017). Creating appropriate trust in automated vehicle systems: A framework for HMI design. *IEEE Transactions on Human-Machine Systems*, *48*(1), 95-101.

El Bekri, N., Kling, J., & Huber, M. F. (2019). *A study on trust in black box models and post-hoc explanations*. International Workshop on Soft Computing Models in Industrial and Environmental Applications, Bilbao, Spain.

Engström, J., Markkula, G., Victor, T., & Merat, N. (2017). Effects of cognitive load on driving performance: The cognitive control hypothesis. *Human factors*, *59*(5), 734-764.

Epley, N., Waytz, A., & Cacioppo, J. T. (2007). On seeing human: a three-factor theory of anthropomorphism. *Psychological review*, *114*(4), 864.

Esfandiari, R., & Sokhanvar, F. (2016). Modified unified theory of acceptance and use of technology in investigating Iranian language learners' attitudes toward mobile assisted language learning (MALL). *Interdisciplinary Journal of Virtual Learning in Medical Sciences*, *6*(4), 93-105.

Evans, T., Retzlaff, C. O., Geißler, C., Kargl, M., Plass, M., Müller, H., Kiehl, T.-R., Zerbe, N., & Holzinger, A. (2022). The explainability paradox: Challenges for xAI in digital pathology. *Future Generation Computer Systems*.

Fahse, T. B., Blohm, I., Hruby, R., & van Giffen, B. (2022). Explanation Interfaces for Sales Forecasting. European Conference on Information Systems, Timișoara, Romania.

Fahse, T. B., Blohm, I., & van Giffen, B. (2022). *Effectiveness of Example-Based Explanations to Improve Human Decision Quality in Machine Learning Forecasting Systems*. International Conference of Information Systems, Copenhagen, Denmark.

Fan, W., Liu, J., Zhu, S., & Pardalos, P. M. (2018). Investigating the impacting factors for the healthcare professionals to adopt artificial intelligence-based medical diagnosis support system (AIMDSS). *Annals of Operations Research*, 1-26.

Ferreira, J. J., & Monteiro, M. S. (2020). What are people doing about XAI user experience? A survey on AI explainability research and practice. International Conference on Human-Computer Interaction, Copenhagen, Denmark.

Fischer, M., Heim, D., Hofmann, A., Janiesch, C., Klima, C., & Winkelmann, A. (2020). A taxonomy and archetypes of smart services for smart living. *Electronic Markets*, *30*(1), 131-149.

Fishbein, M., & Ajzen, I. (1977). Belief, attitude, intention, and behavior: An introduction to theory and research. *Philosophy and Rhetoric*, *10*(2), 178-188.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological bulletin*, *76*(5), 378.

Flick, U. (2020). Gütekriterien qualitativer Forschung. In Mey G. & Mruck K. (Eds.), *Handbuch Qualitative Forschung in der Psychologie* (Vol. 13). Springer.

Fornell, C., & Larcker, D. F. (1981). Evaluating structural equation models with unobservable variables and measurement error. *Journal of Marketing Research*, *18*(1), 39-50.

Förster, M., Klier, M., Kluge, K., & Sigler, I. (2020a). *Evaluating explainable Artifical intelligence-What users really appreciate*. European Conference on Information Systems, Virtual Conference.

Förster, M., Klier, M., Kluge, K., & Sigler, I. (2020b). *Fostering human agency: a process for the design of user-centric XAI systems*. International Conference on Information Systems, Virtual Conference.

Forsythe, L. P., Alfano, C. M., Kent, E. E., Weaver, K. E., Bellizzi, K., Arora, N., Aziz, N., Keel, G., & Rowland, J. H. (2014). Social support, self-efficacy for decision-making, and follow-up care use in long-term cancer survivors. *Psycho-oncology*, *23*(7), 788-796.

Freitas, A. A. (2014). Comprehensible classification models: a position paper. *ACM SIGKDD explorations newsletter*, *15*(1), 1-10.

Freitas, A. A., Wieser, D. C., & Apweiler, R. (2008). On the importance of comprehensible classification models for protein function prediction. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *7*(1), 172-182.

Fu, K. K., Yang, M. C., & Wood, K. L. (2015). *Design principles: The foundation of design*. International Design Engineering Technical Conferences and Computers and Information in Engineering Conference, St. Louis, MO, USA.

Fürnkranz, J., Kliegr, T., & Paulheim, H. (2020). On cognitive preferences and the plausibility of rule-based models. *Machine Learning*, *109*(4), 853-898.

Futia, G., & Vetrò, A. (2020). On the Integration of Knowledge Graphs into Deep Learning Models for a More Comprehensible AI. *Information*, *11*(2), 122-132.

Gale, W., Oakden-Rayner, L., Carneiro, G., Palmer, L. J., & Bradley, A. P. (2019). *Producing Radiologist-Quality Reports for Interpretable Deep Learning*. International Symposium on Biomedical Imaging (ISBI 2019), Venice, Italy.

García, S., Ramírez-Gallego, S., Luengo, J., Benítez, J. M., & Herrera, F. (2016). Big data preprocessing: methods and prospects. *Big Data Analytics*, *1*(9), 1-22.

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*(1), 7.

Gena, C., Brogi, R., Cena, F., & Vernero, F. (2011). *The impact of rating scales on user's rating behavior*. International Conference on User Modeling, Adaptation, and Personalization, Girona, Spain.

Gentile, D., Jamieson, G., & Donmez, B. (2021). *Evaluating human understanding in XAI systems*. CHI HCXAI Workshop, Virtual Conference.

Ghai, B., Liao, Q. V., Zhang, Y., Bellamy, R., & Mueller, K. (2021). Explainable active learning (xal) toward ai explanations as interfaces for machine teachers. *ACM on Human-Computer Interaction*, *4*(CSCW3), 1-28.

Ghanvatkar, S., & Rajan, V. (2022). *Towards a Theory-Based Evaluation of Explainable Predictions in Healthcare*. International Conference on Information Systems, Copenhagen, Denmark.

Gherheş, V. (2018). Why Are We Afraid of Artificial Intelligence (Ai)? *European Review Of Applied Sociology*, *11*(17), 6-15.

Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., & Kagal, L. (2018). *Explaining explanations: An overview of interpretability of machine learning*. International Conference on data science and advanced analytics (DSAA), Turin.

Glaser, B., & Strauss, A. (1967). Grounded theory: The discovery of grounded theory. *Sociology The Journal of the British Sociological Association*, *12*, 27-49.

Glass, A., McGuinness, D. L., & Wolverton, M. (2008). *Toward establishing trust in adaptive agents*. International conference on Intelligent user interfaces, Gran Canaria, Spain.

Glomsrud, J. A., Ødegårdstuen, A., Clair, A. L. S., & Smogeli, Ø. (2019). *Trustworthy versus explainable AI in autonomous vessels*. International Seminar on Safety and Security of Autonomous Vessels (ISSAV) and European STAMP Workshop and Conference (ESWC), Helsinki, Sweden.

Goebel, R., Chander, A., Holzinger, K., Lecue, F., Akata, Z., Stumpf, S., Kieseberg, P., & Holzinger, A. (2018). *Explainable AI: the new 42?* International Cross-Domain Conference for Machine Learning and Knowledge Extraction, Vienna, Austria.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT press.

Goodman, B., & Flaxman, S. (2017). European Union regulations on algorithmic decision-making and a "right to explanation". *AI magazine*, *38*(3), 50-57.

Górski, Ł., & Ramakrishna, S. (2021). *Explainable artificial intelligence, lawyer's perspective*. International Conference on Artificial Intelligence and Law, São Paulo, Brazil.

Gregor, S., & Benbasat, I. (1999). Explanations from intelligent systems: Theoretical foundations and implications for practice. *MIS Quarterly*, 497-530.

Gregor, S., Chandra Kruse, L., & Seidel, S. (2020). Research perspectives: the anatomy of a design principle. *Journal of the Association for Information Systems*, *21*(6).

Gregor, S., & Hevner, A. R. (2013). Positioning and presenting design science research for maximum impact. *MIS Quarterly*, *37*(2), 337-355.

Gregor, S., & Yu, X. (2002). Exploring the explanatory capabilities of intelligent system technologies. In V. Dimitrov (Eds.), *Fuzzy Logic: A Framework for the New Millennium* (pp. 288-300). Physica-Verlag.

Gretzel, U., & Fesenmaier, D. R. (2006). Persuasion in recommender systems. *International Journal of Electronic Commerce*, *11*(2), 81-100.

Grice, H. P. (1975). *Logic and conversation*. Brill.

Grice, H. P. (2019). *Logic and conversation*. Brill.

Grigorescu, S., Trasnea, B., Cocias, T., & Macesanu, G. (2020). A survey of deep learning techniques for autonomous driving. *Journal of Field Robotics*, *37*(3), 362-386.

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*, *51*(5), 1-42.

Gunning, D. (2017). Explainable artificial intelligence (xai). *DARPA/I2O*, *2*.

Gunning, D. (2019). DARPA's explainable artificial intelligence (XAI) program. *AI magazine*, *40*(2), 44-58.

Gunning, D., Stefik, M., Choi, J., Miller, T., Stumpf, S., & Yang, G.-Z. (2019). XAI—Explainable artificial intelligence. *Science Robotics*, *4*(37), 1-37.

Guo, M., Zhang, Q., Liao, X., & Chen, Y. (2019). An interpretable machine learning framework for modelling human decision behavior. *arXiv:1906.01233*.

Guo, W. (2020). Explainable artificial intelligence for 6G: Improving trust between human and machine. *IEEE Communications Magazine*, *58*(6), 39-45.

Gupta, B., Dasgupta, S., & Gupta, A. (2008). Adoption of ICT in a government organization in a developing country: An empirical study. *The Journal of Strategic Information Systems*, *17*, 140-154.

Hair, J. F., Ringle, C. M., & Sarstedt, M. (2011). PLS-SEM: Indeed a silver bullet. *Journal of Marketing theory and Practice*, *19*(2), 139-152.

Hair Jr, J. F., Hult, G. T. M., Ringle, C. M., & Sarstedt, M. (2021). *A primer on partial least squares structural equation modeling (PLS-SEM)*. Sage publications.

Hall, M., Harborne, D., Tomsett, R., Galetic, V., Quintana-Amate, S., Nottle, A., & Preece, A. (2019). *A systematic method to understand requirements for explainable AI (XAI) systems*. IJCAI Workshop on eXplainable Artificial Intelligence (XAI 2019), Macau, China.

Hamilton, K., Karahalios, K., Sandvig, C., & Eslami, M. (2014). *A path to understanding the effects of algorithm awareness*. CHI Conference on Human Factors in Computing Systems, Toronto Ontario, Canada.

Hamilton, K., Shih, S.-I., & Mohammed, S. (2016). The development and validation of the rational and intuitive decision styles scale. *Journal of personality assessment*, *98*(5), 523-535.

Hartmann, M. (2009). *Challenges in Developing User-Adaptive Intelligent User Interfaces*. Workshop on Adaptivity and User Modeling in Interactive Systems, Berlin, Germany.

Haynes, S. R., Cohen, M. A., & Ritter, F. E. (2009). Designs for explaining intelligent agents. *International Journal of Human-Computer Studies*, *67*(1), 90-110.

Hebrado, J., Lee, H. J., & Choi, J. (2011). *The Role of Transparency and Feedback on the Behavioral Intention to Reuse a Recommender System*. International Conference on Information Resources Management, Seoul, Korea.

Hebrado, J. L., Lee, H. J., & Choi, J. (2013). Influences of Transparency and Feedback on Customer Intention to Reuse Online Recommender Systems. *Journal of Society for e-Business Studies*, *18*(2).

Hein, D., Rauschnabel, P., He, J., Richter, L., & Ivens, B. (2018). *What drives the adoption of autonomous cars?* International Conference on Information Systems (ICIS), San Francisco, USA.

Heinrich, K., Graf, J., Chen, J., Laurisch, J., & Zschech, P. (2020). *Fool me Once, shame on You, Fool me Twice, shame on me: a Taxonomy of Attack and de-Fense Patterns for AI Security*. European Conference on Information Systems, Virtual Conference.

Heinrich, K., Janiesch, C., Möller, B., & Zschech, P. (2019). *Is bigger always better? Lessons learnt from the evolution of deep learning architectures for image classification*. Pre-ICIS SIGDSA Symposium, Munich, Germany.

Hemmer, P., Schemmer, M., Vössing, M., & Kühl, N. (2021). *Human-AI Complementarity in Hybrid Intelligence Systems: A Structured Literature Review*. Pacific Asia Conference on Information Systems (PACIS), Virtual Conference.

Hepenstal, S., & McNeish, D. (2020). *Explainable Artificial Intelligence: What Do You Need to Know?* International Conference on Human-Computer Interaction, Copenhagen, Denmark.

Herm, L.-V. (2023a). *Impact Of Explainable AI On Cognitive Load: Insights From An Empirical Study*. European Conference on Information Systems, Kristiansand, Norway.

Herm, L.-V. (2023b). *Supplementary Material for "Impact of Explainable AI On Cognitive Load: Insights From An Empirical Study"*. Retrieved 11.03.2023 from https://doi.org/10.23728/b2share.9814decadcfd4fb68a8963efbdc67d41

Herm, L.-V., Heinrich, K., Wanner, J., & Janiesch, C. (2023). Stop ordering machine learning algorithms by their explainability! A user-centered investigation of performance and explainability. *International Journal of Information Management*, *69*, 102538.

Herm, L.-V., & Janiesch, C. (2019). Anforderungsanalyse für eine Kollaborationsplattform in Blockchain-basierten Wertschöpfungsnetzwerken. *Working Paper Series of the Institute of Business Management, University of Würzburg*, *7*, 1-162.

Herm, L.-V., & Janiesch, C. (2021). *Towards an implementation of blockchain-based collaboration platforms in supply chain networks: A requirements analysis*. Hawaii International Conference on System Sciences (HICSS), Virtual Conference.

Herm, L.-V., Janiesch, C., & Fuchs, P. (2022). Der Einfluss von menschlichen Denkmustern auf künstliche Intelligenz-Eine strukturierte Untersuchung von kognitiven Verzerrungen. *HMD Praxis der Wirtschaftsinformatik*, *59*(2), 556-571.

Herm, L.-V., Janiesch, C., Helm, A., Imgrund, F., Fuchs, K., Hofmann, A., & Winkelmann, A. (2020). *A consolidated framework for implementing robotic process automation projects*. International Conference of Business Process Management, Virtual Conference.

Herm, L.-V., Janiesch, C., Helm, A., Imgrund, F., Hofmann, A., & Winkelmann, A. (2023). A framework for implementing robotic process automation projects. *Information Systems and E-Business Management*, *21*, 1-35.

Herm, L.-V., Janiesch, C., Reijers, H. A., & Seubert, F. (2021). *From symbolic RPA to intelligent RPA: challenges for developing and operating intelligent software robots*. International Conference on Business Process Management, Rome, Italy.

Herm, L.-V., Janiesch, C., Steinbach, T., & Wüllner, D. (2021). Managing RPA implementation projects. In Czarnecki C. & Fettke P. (Eds.), *Robotic Process Automation* (pp. 27-46). De Gruyter.

Herm, L.-V., Steinbach, T., Wanner, J., & Janiesch, C. (2022). A nascent design theory for explainable intelligent systems. *Electronic Markets*, *32*(4), 2185-2205.

Herm, L.-V., Wanner, J., & Janiesch, C. (2020). *Bridging the Architectural Gap in Smart Homes Between User Control and Digital Automation*. International Conference on Design Science Research in Information Systems and Technology, Kristiansand, Norway.

Herm, L.-V., Wanner, J., & Janiesch, C. (2022). *A Taxonomy of User-centered Explainable AI Studies*. Pacific Asia Conference on Information Systems, Virtual Conference.

Herm, L.-V., Wanner, J., Seubert, F., & Janiesch, C. (2021a). *I Don't Get It, But It Seems Valid! The Connection Between Explainability And Comprehensibility In (X)AI Research*. European Conference of Information Systems (ECIS), Virtual Conference.

Herm, L.-V., Wanner, J., Seubert, F., & Janiesch, C. (2021b). *Supplementary Material for "I Don't Get It, But It Seems Valid! The Connection Between Explainability And Comprehensibility In (X)AI Research*". Retrieved 06-04-2021 from 10.23728/b2share.b6abcb7517f64a11b16330cc683bf212

Hevner, A. R. (2021). The duality of science: Knowledge in information systems research. *Journal of information technology*, *36*(1), 72-76.

Hevner, A. R., & March, S. T. (2003). The information systems research cycle. *Computer*, *36*(11), 111-113.

Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design science in information systems research. *MIS Quarterly*, *28*(1), 75-105.

Hilton, D. (1996). Mental models and causal explanation: Judgements of probable cause and explanatory relevance. *Thinking & Reasoning*, *2*(4), 273-308.

Hind, M., Wei, D., Campbell, M., Codella, N. C., Dhurandhar, A., Mojsilović, A., Natesan Ramamurthy, K., & Varshney, K. R. (2019). *TED: Teaching AI to explain its decisions*. Conference on AI, Ethics, and Society, Honolulu, HI, USA.

Hirsch, T., Merced, K., Narayanan, S., Imel, Z. E., & Atkins, D. C. (2017). *Designing contestability: Interaction design, machine learning, and mental health*. Conference on Designing Interactive Systems, Edinburgh, UK.

Hoff, K. A., & Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust. *Human factors*, *57*(3), 407-434.

Hoffman, R. R., Johnson, M., Bradshaw, J. M., & Underbrink, A. (2013). Trust in automation. *IEEE Intelligent Systems*, *28*(1), 84-88.

Hoffman, R. R., Mueller, S. T., Klein, G., & Litman, J. (2018). Metrics for explainable AI: Challenges and prospects. *arXiv preprint arXiv:1812.04608*.

Hohman, F., Head, A., Caruana, R., DeLine, R., & Drucker, S. M. (2019). *Gamut: A design probe to understand how data scientists understand machine learning models*. CHI conference on human factors in computing systems, New York, USA.

Holmström, J., Ketokivi, M., & Hameri, A. P. (2009). Bridging practice and theory: A design science approach. *Decision sciences*, *40*(1), 65-87.

Holzinger, A., Dehmer, M., Emmert-Streib, F., Cucchiara, R., Augenstein, I., Del Ser, J., Samek, W., Jurisica, I., & Díaz-Rodríguez, N. (2022). Information fusion as an integrative cross-cutting enabler to achieve robust, explainable, and trustworthy medical artificial intelligence. *Information fusion*, *79*, 263-278.

Holzinger, A., Langs, G., Denk, H., Zatloukal, K., & Müller, H. (2019). Causability and explainability of artificial intelligence in medicine. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *9*(4), e1312.

Hosanagar, K., & Jair, V. (2018). We need transparency in algorithms, but too much can backfire. *Harvard Business Review*, *25*, 2018.

Hradecky, D., Kennell, J., Cai, W., & Davidson, R. (2022). Organizational readiness to adopt artificial intelligence in the exhibition sector in Western Europe. *International Journal of Information Management*, *65*, 102497.

Hsiao, J. H.-w., Ngai, H. H. T., Qiu, L., Yang, Y., & Cao, C. C. (2021). Roadmap of designing cognitive metrics for explainable artificial intelligence (XAI). *arXiv preprint arXiv:2108.01737*.

Hsu, C. L., Lin, J. C. C., & Chiang, H. S. (2013). The effects of blogger recommendations on customers' online shopping intentions. *Internet Research*, *23*(1), 69-88.

Hudon, A., Demazure, T., Karran, A., Léger, P.-M., & Sénécal, S. (2021). *Explainable Artificial Intelligence (XAI): How the Visualization of AI Predictions Affects User Cognitive Load and Confidence*. NeuroIS Retreat, Vienna, Austria.

Hutson, M. (2017). AI Glossary: Artificial intelligence, in so many words. *Science*, *357*(6346), 19-19.

Huysmans, J., Dejaeger, K., Mues, C., Vanthienen, J., & Baesens, B. (2011). An empirical evaluation of the comprehensibility of decision table, tree and rule based predictive models. *Decision support systems*, *51*(1), 141-154.

Hwang, W.-Y., Shih, T. K., Ma, Z.-H., Shadiev, R., & Chen, S.-Y. (2016). Evaluating listening and speaking skills in a mobile game-based learning environment with situational contexts. *Computer Assisted Language Learning*, *29*(4), 639-657.

Hyndman, R. J. (2020). A brief history of forecasting competitions. *International Journal of Forecasting*, *36*(1), 7-14.

Ibrahim, M., Louie, M., Modarres, C., & Paisley, J. (2019). *Global explanations of neural networks: Mapping the landscape of predictions*. Conference on AI, Ethics, and Society, Honolulu, Hawaii, USA.

Iivari, J., Rotvit Perlt Hansen, M., & Haj-Bolouri, A. (2021). A proposal for minimum reusability evaluation of design principles. *European Journal of Information Systems*, *30*(3), 286-303.

Imgrund, F., Herm, L.-V., Wanner, J., Janiesch, C., Fischer, M., & Winkelmann, A. (2021). *Critical Success Factors for Process Modeling Projects-Analysis of Empirical Evidence*. Pacific Asia Conference on Information Systems, Virtual Conference.

Indarsin, T., & Ali, H. (2017). Attitude toward Using m-Commerce: The Analysis of Perceived Usefulness Perceived Ease of Use, and Perceived Trust: Case Study in Ikens Wholesale Trade, Jakarta-Indonesia. *Saudi Journal of Business and Management Studies*, *2*(11), 995-1007.

Injadat, M., Moubayed, A., Nassif, A. B., & Shami, A. (2021). Machine learning towards intelligent systems: applications, challenges, and opportunities. *Artificial Intelligence Review*, *54*(5), 3299-3348.

Jackson, C. M., Chow, S., & Leitch, R. A. (1997). Toward an understanding of the behavioral intention to use an information system. *Decision sciences*, *28*(2), 357-389.

Jacovi, A., Marasović, A., Miller, T., & Goldberg, Y. (2021). *Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in AI*. FAccT: Conference on fairness, accountability, and transparency, Virtual Event, Canada.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning*. Springer.

Janiesch, C., Rosenkranz, C., & Scholten, U. (2020). An information systems design theory for service network effects. *Journal of the Association for Information Systems: forthcoming*, *21*(6), 1402-1460.

Janiesch, C., Wanner, J., & Herm, L.-V. (2021). *Design Principles for Shared Maintenance Analytics in Fleet Management*. International Conference on Design Science Research in Information Systems and Technology, Kristiansand, Norway.

Janiesch, C., Zschech, P., & Heinrich, K. (2021). Machine learning and deep learning. *Electronic Markets*, *31*, 685-695.

Janosi, A., Steinbrunn, W., Pfisterer, M., & Detrano, R. (1988). *Heart Disease Data Set* UCI Machine Learning Library. Retrieved 10.10.2021 from https://archive.ics.uci.edu/ml/datasets/Heart+Disease

Janssen, M., Hartog, M., Matheus, R., Yi Ding, A., & Kuk, G. (2022). Will algorithms blind people? The effect of explainable AI and decision-makers' experience on AI-supported decision-making in government. *Social Science Computer Review*, *40*(2), 478-493.

Jauernig, J., Uhl, M., & Walkowitz, G. (2022). People Prefer Moral Discretion to Algorithms: Algorithm Aversion Beyond Intransparency. *Philosophy & Technology*, *35*(1), 2.

Jensen, T., Albayram, Y., Khan, M. M. H., Buck, R., Coman, E., & Fahim, M. A. A. (2018). *Initial trustworthiness perceptions of a drone system based on performance and process information*. International Conference on Human-Agent Interaction, Southampton, United Kingdom.

Jesus, S., Belém, C., Balayan, V., Bento, J., Saleiro, P., Bizarro, P., & Gama, J. (2021). *How can I choose an explainer? An Application-grounded Evaluation of Post-hoc Explanations*. Conference on Fairness, Accountability, and Transparency, Virtual Conference.

Jetter, J., Eimecke, J., & Rese, A. (2018). Augmented reality tools for industrial applications: What are potential key performance indicators and who benefits? *Computers in Human Behavior*, *87*, 18-33.

Joshi, A., Kale, S., Chandel, S., & Pal, D. K. (2015). Likert scale: Explored and explained. *Current Journal of Applied Science and Technology*, *7*(4), 396-403.

Jussupow, E., Benbasat, I., & Heinzl, A. (2020). *Why are we averse towards Algorithms? A comprehensive literature Review on Algorithm aversion*. European Conference on Information Systems, Virtual Conference.

Kaiser, R. (2014). *Qualitative Experteninterviews: Konzeptionelle Grundlagen und praktische Durchführung*. Springer.

Karran, A. J., Demazure, T., Hudon, A., Senecal, S., & Léger, P.-M. (2022). Designing for Confidence: The Impact of Visualizing Artificial Intelligence Decisions. *Frontiers in Neuroscience*, *16*(883385).

Kaur, D., Uslu, S., Rittichier, K. J., & Durresi, A. (2022). Trustworthy Artificial Intelligence: A Review. *ACM computing surveys (CSUR)*, *55*(2), 1-38.

Kaur, K., & Rampersad, G. (2018). Trust in driverless cars: Investigating key factors influencing the adoption of driverless cars. *Journal of Engineering and Technology Management*, *48*, 87-96.

Kelly, C. J., Karthikesalingam, A., Suleyman, M., Corrado, G., & King, D. (2019). Key challenges for delivering clinical impact with artificial intelligence. *BMC medicine*, *17*(195).

Kenny, E. M., Ford, C., Quinn, M., & Keane, M. T. (2021). Explaining black-box classifiers using post-hoc explanations-by-example: The effect of explanations and error-rates in XAI user studies. *Artificial Intelligence*, *294*, 103459.

Khanna, R., Dodge, J., Anderson, A., Dikkala, R., Irvine, J., Shureih, Z., Lam, K.-h., Matthews, C. R., Lin, Z., & Kahng, M. (2022). Finding AI's faults with AAR/AI: An empirical study. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, *12*(1), 1-33.

Khodabandehloo, E., Riboni, D., & Alimohammadi, A. (2021). HealthXAI: Collaborative and explainable AI for supporting early diagnosis of cognitive decline. *Future Generation Computer Systems*, *116*, 168-189.

Kim, B., Khanna, R., & Koyejo, O. O. (2016). *Examples are not enough, learn to criticize! criticism for interpretability*. 30th Conference on Neural Information Processing Systems, Barcelona, Spain.

Kim, D. J. (2014). A study of the multilevel and dynamic nature of trust in e-commerce from a cross-stage perspective. *International Journal of Electronic Commerce*, *19*(1), 11-64.

Kim, J. (2019). Fear of Artificial Intelligence on People's Attitudinal & Behavioral Attributes: An Exploratory Analysis of AI Phobia. *Global Scientific Journal*, *7*(10), 9-20.

Kim, Y. J., Chun, J. U., & Song, J. (2009). Investigating the role of attitude in technology acceptance from an attitude strength perspective. *International Journal of Information Management*, *29*(1), 67-77.

Kirsch, A. (2017). *Explain to whom? Putting the user in the center of explainable AI*. International Conference of the Italian Association for Artificial Intelligence Bari, Italy.

Kizilcec, R. F. (2016). *How much information? Effects of transparency on trust in an algorithmic interface*. Conference on Human Factors in Computing Systems, New York, NY, USA.

Kloker, A., Fleiß, J., Koeth, C., Kloiber, T., Ratheiser, P., & Thalmann, S. (2022). *Caution or Trust in AI? How to design XAI in sensitive Use Cases?* Americas Conference on Information Systems, Minneapolis, USA.

Kocielnik, R., Amershi, S., & Bennett, P. N. (2019). *Will you accept an imperfect ai? exploring designs for adjusting end-user expectations of ai systems*. CHI Conference on Human Factors in Computing Systems, New York, USA.

Kock, N. (2015). Common method bias in PLS-SEM: A full collinearity assessment approach. *International Journal of e-Collaboration (ijec)*, *11*(4), 1-10.

Komiak, S. Y., & Benbasat, I. (2006). The effects of personalization and familiarity on trust and adoption of recommendation agents. *MIS Quarterly*, 941-960.

Kroll, J. A. (2018). The fallacy of inscrutability. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *376*(2133), 20180084.

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.

Kuhl, N., Lobana, J., & Meske, C. (2019). *Do you comply with AI? - Personalized explanations of learning algorithms and their impact on employees' compliance behavior*. International Conference on Information Systems (ICIS), Munich, Germany.

Kulesza, T., Stumpf, S., Burnett, M., Yang, S., Kwan, I., & Wong, W.-K. (2013). *Too much, too little, or just right? Ways explanations impact end users' mental models*. Symposium on visual languages and human centric computing, San Jose, CA, USA.

Kundisch, D., Muntermann, J., Oberländer, A. M., Rau, D., Röglinger, M., Schoormann, T., & Szopinski, D. (2022). An Update for Taxonomy Designers. *Business & Information Systems Engineering*, *64*, 421-439.

Kunreuther, H., Meyer, R., Zeckhauser, R., Slovic, P., Schwartz, B., Schade, C., Luce, M. F., Lippman, S., Krantz, D., & Kahn, B. (2002). High stakes decision making: Normative, descriptive and prescriptive considerations. *Marketing Letters*, *13*(3), 259-268.

La Cava, W., Williams, H., Fu, W., & Moore, J. H. (2019). Evaluating recommender systems for AI-driven data science. *arXiv:1905.09205*.

Laato, S., Tiainen, M., Najmul Islam, A., & Mäntymäki, M. (2022). How to explain AI systems to end users: a systematic literature review and research agenda. *Internet Research*, *32*(7), 1-31.

Lage, I., Ross, A., Gershman, S. J., Kim, B., & Doshi-Velez, F. (2018). *Human-in-the-loop interpretability prior*. Conference on Neural Information Processing Systems, Montréal.

Lakkaraju, H., & Bastani, O. (2020). "*How do I fool you?*" *Manipulating User Trust via Misleading Black Box Explanations*. Conference on AI, Ethics, and Society, New York, NY, USA.

Lakkaraju, H., Kamar, E., Caruana, R., & Leskovec, J. (2019). *Faithful and customizable explanations of black box models*. Conference on AI, Ethics, and Society, Oxford, UK.

Lambers, J., & Balzer, C. (2022). Plastics welding process data. *B2Share EUDAT - 10.23728/b2share.657bb2383ce946dcb4cab9419e1645d3*.

Lamnek, S., & Krell, C. (2010). *Qualitative Sozialforschung* (6 ed.). Beltz.

Landwehr, J. P., Kühl, N., Walk, J., & Gnädig, M. (2022). Design Knowledge for Deep-Learning-Enabled Image-Based Decision Support Systems. *Business & Information Systems Engineering*, 1-22.

Langer, M., Oster, D., Speith, T., Hermanns, H., Kästner, L., Schmidt, E., Sesing, A., & Baum, K. (2021). What do we want from Explainable Artificial Intelligence (XAI)?-A stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research. *Artificial Intelligence*, *296*, 103473.

Lankton, N. K., McKnight, D. H., & Tripp, J. (2015). Technology, humanness, and trust: Rethinking trust in technology. *Journal of the Association for Information Systems*, *16*(10), 880-918.

Lavie, N. (2010). Attention, distraction, and cognitive control under load. *Current directions in psychological science*, *19*(3), 143-148.

Le, T., Wang, S., & Lee, D. (2020). *GRACE: Generating Concise and Informative Contrastive Sample to Explain Neural Network Model's Prediction*. International Conference on Knowledge Discovery & Data Mining,

Lebovitz, S., Levina, N., & Lifshitz-Assaf, H. (2021). Is AI ground truth really "true"? The dangers of training and evaluating AI tools based on experts' know-what. *Management Information Systems Quarterly*, *45*(3b), 1501-1525.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436-444.

Lee, A. S., & Baskerville, R. L. (2003). Generalizing generalizability in information systems research. *Information systems research*, *14*(3), 221-243.

Lee, J.-H., & Song, C.-H. (2013). Effects of trust and perceived risk on user acceptance of a new technology service. *Social Behavior and Personality: an international journal*, *41*(4), 587-597.

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human factors*, *46*(1), 50-80.

Lee, M. K., & Turban, E. (2001). A trust model for consumer internet shopping. *International Journal of Electronic Commerce*, *6*(1), 75-91.

Leppink, J., Paas, F., Van Gog, T., van Der Vleuten, C. P., & Van Merrienboer, J. J. (2014). Effects of pairs of problems and examples on task performance and different types of cognitive load. *Learning and instruction*, *30*, 32-42.

Leppink, J., & Pérez-Fuster, P. (2019). Mental effort, workload, time on task, and certainty: Beyond linear models. *Educational Psychology Review*, *31*(2), 421-438.

Lewis, D. K. (1986). Causal explanation. *Philosophical Papers*, *2*, 214-240.

Liao, Q. V., Gruen, D., & Miller, S. (2020). *Questioning the AI: informing design practices for explainable AI user experiences*. CHI Conference on Human Factors in Computing Systems, New York, USA.

Liao, Q. V., & Varshney, K. R. (2022). Human-centered explainable ai (xai): From algorithms to user experiences. *arXiv preprint arXiv:2110.10790*.

Lipton, Z. C. (2018). The Mythos of Model Interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, *16*(3), 31-57.

Liu, R., Strawderman, W., & Zhang, C.-H. (2007). Complex Datasets and Inverse Problems. Tomography, Networks and Beyond. *https://arxiv.org/abs/0708.1130v1*.

Liu, S., Duffy, A., Whitfield, R., & Boyle, I. (2008). Integration of decision support systems to improve decision support performance. *Knowledge Information Systems*, *22*, 261-286.

Logg, J. M., Minson, J. A., & Moore, D. A. (2019). Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes*, *151*, 90-103.

Longo, L., Goebel, R., Lecue, F., Kieseberg, P., & Holzinger, A. (2020). *Explainable artificial intelligence: Concepts, applications, research challenges and visions*. International Cross-Domain Conference for Machine Learning and Knowledge Extraction, Dublin, Ireland.

Lotz, V., Himmel, S., & Ziefle, M. (2019). *You're my mate-acceptance factors for human-robot collaboration in industry*. International Conference on Competitive Manufacturing, Stellenbosch, South Africa.

Loyola-Gonzalez, O. (2019). Black-box vs. white-box: Understanding their advantages and weaknesses from a practical point of view. *IEEE Access*, *7*, 154096-154113.

Lu, J., Lee, D. D., Kim, T. W., & Danks, D. (2019). *Good Explanation for Algorithmic Transparency*. Conference on AI, Ethics, and Society, New York, NY, USA.

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., & Prutkin, J. M. (2020). From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*, *2*(1), 2522-5839.

Luo, Y., Tseng, H.-H., Cui, S., Wei, L., Ten Haken, R. K., & El Naqa, I. (2019). Balancing accuracy and interpretability of machine learning approaches for radiation treatment outcomes modeling. *BJR| Open*, *1*(1), 20190021.

Madsen, M., & Gregor, S. (2000). *Measuring human-computer trust*. Australasian conference on information systems, Brisbane, Australia.

Madumal, P., Miller, T., Sonenberg, L., & Vetere, F. (2019). A grounded interaction protocol for explainable artificial intelligence. *arXiv preprint arXiv:1903.02409*.

Maedche, A., Gregor, S., & Parsons, J. (2021). Mapping design contributions in information systems research: the design research activity framework. *Communications of the association for information systems*, *49*(1), 355-378.

Mahesh, B. (2020). Machine learning algorithms-a review. *International Journal of Science and Research (IJSR)*, *9*(1), 381-386.

Mahmud, H., Islam, A. N., Ahmed, S. I., & Smolander, K. (2022). What influences algorithmic decision-making? A systematic literature review on algorithm aversion. *Technological Forecasting and Social Change*, *175*, 121390.

Makridakis, S. (2017). The forthcoming Artificial Intelligence (AI) revolution: Its impact on society and firms. *Futures*, *90*, 46-60.

Malhi, A., Knapic, S., & Främling, K. (2020). Explainable agents for less bias in human-agent decision making. International Workshop on Explainable, Transparent Autonomous Agents and Multi-Agent Systems, London, UK.

Malle, B. F. (2004). *How the mind explains behavior*. MIT-Press.

Mansouri, S., Kaghazi, B., & Khormali, N. (2011). *A survey the views of the students of Gonbad Payam Noor University to mobile learning*. Conference of mobile value-added services in Iran, Iran, Tehran.

Marangunić, N., & Granić, A. (2015). Technology acceptance model: a literature review from 1986 to 2013. *Universal access in the information society*, *14*(1), 81-95.

March, S. T., & Smith, G. F. (1995). Design and natural science research on information technology. *Decision support systems*, *15*(4), 251-266.

March, S. T., & Storey, V. C. (2008). Design science in the information systems discipline: an introduction to the special issue on design science research. *MIS Quarterly*, 725-730.

Mateja, D., Bartels, E. A., Oberste, L., Herm, L.-V., & Danelski, A. (2021). *Towards a Reference Architecture for Female-Sensitive Drug Management*. Hawaii International Conference on System Sciences, Virtual Conference.

Mayer, R., & Mayer, R. E. (2005). *The Cambridge handbook of multimedia learning*. Cambridge university press.

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. *Academy of management review*, *20*(3), 709-734.

Mayr, A., Herm, L.-V., Wanner, J., & Janiesch, C. (2022). *Applications and challenges of task mining: a literature review*. European Conference on Information Systems (ECIS), Timișoara, Romania.

McKinney, S. M., Sieniek, M., Godbole, V., Godwin, J., Antropova, N., Ashrafian, 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., Ledsam, J. R., Melnick, D., Mostofi, H., Peng, L., Reicher, J. J., Romera-Paredes, B., Sidebottom, R., Suleyman, M., Tse, D., Young, K. C., De Fauw, J., &

Shetty, S. (2020). International evaluation of an AI system for breast cancer screening. *Nature*, *577*(7788), 89-94.

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 Transactions on management information systems (TMIS)*, *2*(2), 1-25.

McKnight, D. H., & Chervany, N. L. (2000). *What is trust? A conceptual analysis and an interdisciplinary model*. American Conference on Information Systems, Long Beach, California, USA.

McKnight, D. H., Choudhury, V., & Kacmar, C. (2002). Developing and validating trust measures for e-commerce: An integrative typology. *Information systems research*, *13*(3), 334-359.

McKnight, D. H., Cummings, L. L., & Chervany, N. L. (1998). Initial trust formation in new organizational relationships. *Academy of management review*, *23*(3), 473-490.

McNee, S. M., Riedl, J., & Konstan, J. A. (2006). *Being accurate is not enough: how accuracy metrics have hurt recommender systems*. CHI Conference on Human Factors In Computing Systems, Montréal Québec, Canada.

Meade, A. W., & Craig, S. B. (2012). Identifying careless responses in survey data. *Psychological methods*, *17*(3), 437-455.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)*, *54*(6), 1-35.

Meng, X.-L., Rosenthal, R., & Rubin, D. B. (1992). Comparing correlated correlation coefficients. *Psychological bulletin*, *111*(1), 172-175.

Meske, C., & Bunde, E. (2020). *Transparency and trust in human-AI-interaction: The role of model-agnostic explanations in computer vision-based decision support*. International Conference on Human-Computer Interaction, Virtual Conference.

Meske, C., & Bunde, E. (2022). Design Principles for User Interfaces in AI-Based Decision Support Systems: The Case of Explainable Hate Speech Detection. *Information Systems Frontiers*, 1-31.

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.

Meth, H., Mueller, B., & Maedche, A. (2015). Designing a requirement mining system. *Journal of the Association for Information Systems*, *16*(9), 799-837.

Mignan, A., & Broccardo, M. (2019). A deeper look into 'deep learning of aftershock patterns following large earthquakes': Illustrating first principles in neural network physical interpretability. International Work-Conference on Artificial Neural Networks, Cham.

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, *267*, 1-38.

Miller, T. (2023). Explainable AI is Dead, Long Live Explainable AI! Hypothesis-driven decision support. *arXiv preprint arXiv:2302.12389*.

Miller, T., Howe, P., & Sonenberg, L. (2017). Explainable AI: Beware of inmates running the asylum or: How I learnt to stop worrying and love the social and behavioural sciences. *arXiv preprint arXiv:1712.00547*.

Milojevic, M., & Nassah, F. (2018). *Digital Industrial Revolution with Predictive Maintenance*. Retrieved 21.12.2022 from https://www.ge.com/digital/sites/default/files/download_assets/PAC_Predictive_Maintenance_GE_Digital_Executive_Summary_2018_1.pdf

Ming, Y., Qu, H., & Bertini, E. (2018). Rulematrix: Visualizing and understanding classifiers with rules. *IEEE transactions on visualization and computer graphics*, *25*(1), 342-352.

Mohseni, S., Zarei, N., & Ragan, E. D. (2021). A multidisciplinary survey and framework for design and evaluation of explainable AI systems. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, *11*(3-4), 1-45.

Mokyr, J., Vickers, C., & Ziebarth, N. L. (2015). The history of technological anxiety and the future of economic growth: Is this time different? *Journal of economic perspectives*, *29*(3), 31-50.

Möller, F., Guggenberger, T. M., & Otto, B. (2020). *Towards a Method for Design Principle Development in Information Systems*. International Conference on Design Science Research in Information Systems and Technology, Kristiansand, Norway.

Moore, G. C., & Benbasat, I. (1991). Development of an instrument to measure the perceptions of adopting an information technology innovation. *Information systems research*, *2*(3), 192-222.

Morocho-Cayamcela, M. E., Lee, H., & Lim, W. (2019). Machine learning for 5G/B5G mobile and wireless communications: Potential, limitations, and future directions. *IEEE Access*, *7*, 137184-137206.

Motulsky, H. (2014). *Intuitive biostatistics: a nonmathematical guide to statistical thinking*. Oxford University Press.

Moyne, J., Iskandar, J., Hawkins, P., Walker, T., Furest, A., Pollard, B., & Stark, D. (2013). *Deploying an equipment health monitoring dashboard and assessing predictive maintenance*. ASMC 2013 SEMI Advanced Semiconductor Manufacturing Conference, Saratoga Springs, New York, USA.

Mualla, Y., Tchappi, I., Kampik, T., Najjar, A., Calvaresi, D., Abbas-Turki, A., Galland, S., & Nicolle, C. (2022). The quest of parsimonious XAI: A human-agent architecture for explanation formulation. *Artificial Intelligence*, *302*, 103573.

Müller, M., Ostern, N., Koljada, D., Grunert, K., Rosemann, M., & Küpper, A. (2021). Trust Mining: Analyzing Trust in Collaborative Business Processes. *IEEE Access*, *9*, 65044-65065.

Müller, O., Junglas, I., Brocke, J. v., & Debortoli, S. (2017). Utilizing big data analytics for information systems research: challenges, promises and guidelines. *European Journal of Information Systems*, *25*(4), 289-302.

Nadj, M., Knaeble, M., Li, M. X., & Maedche, A. (2020). Power to the Oracle? Design Principles for Interactive Labeling Systems in Machine Learning. *KI-Künstliche Intelligenz*, *34*, 1-12.

Nanayakkara, S., Fogarty, S., Tremeer, M., Ross, K., Richards, B., Bergmeir, C., Xu, S., Stub, D., Smith, K., & Tacey, M. (2018). Characterising risk of in-hospital mortality following cardiac arrest using machine learning: A retrospective international registry study. *PLoS medicine*, *15*(11), e1002709.

Narayanan, M., Chen, E., He, J., Kim, B., Gershman, S., & Doshi-Velez, F. (2018). How do humans understand explanations from machine learning systems? an evaluation of the human-interpretability of explanation. *arXiv preprint arXiv:1802.00682*.

Nawratil, U. (2013). *Glaubwürdigkeit in der sozialen Kommunikation*. Springer.

Nelles, J., Kuz, S., Mertens, A., & Schlick, C. M. (2016). *Human-centered design of assistance systems for production planning and control: The role of the human in Industry 4.0*. International Conference on Industrial Technology (ICIT), Taipei, Taiwan.

Neufeld, D. J., Dong, L., & Higgins, C. (2007). Charismatic leadership and user acceptance of information technology. *European Journal of Information Systems*, *16*(4), 494-510.

Nguyen, D. (2018). *Comparing automatic and human evaluation of local explanations for text classification*. 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, New Orleans, Louisiana, USA.

Nickerson, R. C., Varshney, U., & Muntermann, J. (2013). A method for taxonomy development and its application in information systems. *European Journal of Information Systems*, *22*(3), 336-359.

Nilashi, M., Jannach, D., bin Ibrahim, O., Esfahani, M. D., & Ahmadi, H. (2016). Recommendation quality, transparency, and website quality for trust-building in recommendation agents. *Electronic Commerce Research and Applications*, *19*, 70-84.

Nilsson, N. J. (2014). *Principles of artificial intelligence*. Morgan Kaufmann.

Nor, A. K. M., Pedapati, S. R., Muhammad, M., & Leiva, V. (2022). Abnormality detection and failure prediction using explainable Bayesian deep learning: Methodology and case study with industrial data. *Mathematics*, *10*(4), 554.

Nosratabadi, S., Mosavi, A., Duan, P., Ghamisi, P., Filip, F., Band, S. S., Reuter, U., Gama, J., & Gandomi, A. H. (2020). Data science in economics: comprehensive review of advanced machine learning and deep learning methods. *Mathematics*, *8*(10), 1799-1824.

Nourani, M., Roy, C., Block, J. E., Honeycutt, D. R., Rahman, T., Ragan, E. D., & Gogate, V. (2022). On the Importance of User Backgrounds and Impressions: Lessons Learned from Interactive AI Applications. *ACM Transactions on Interactive Intelligent Systems*, *12*(4), 1-29.

Nunes, I., & Jannach, D. (2017). A systematic review and taxonomy of explanations in decision support and recommender systems. *User Modeling and User-adapted interaction*, *27*(3-5), 393-444.

O'Connor, C., & Joffe, H. (2020). Intercoder reliability in qualitative research: debates and practical guidelines. *International journal of qualitative methods*, *19*, 1609406919899220.

Oh, J.-C., & Yoon, S.-J. (2014). Predicting the use of online information services based on a modified UTAUT model. *Behaviour & Information Technology*, *33*(7), 716-729.

Oliveira, T., Faria, M., Thomas, M. A., & Popovič, A. (2014). Extending the understanding of mobile banking adoption: When UTAUT meets TTF and ITM. *International Journal of Information Management*, *34*(5), 689-703.

Oppenheimer, D. M., Meyvis, T., & Davidenko, N. (2009). Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of experimental social psychology*, *45*(4), 867-872.

Österle, H., Becker, J., Frank, U., Hess, T., Karagiannis, D., Krcmar, H., Loos, P., Mertens, P., Oberweis, A., & Sinz, E. J. (2010). Memorandum zur gestaltungsorientierten Wirtschaftsinformatik. *Zeitschrift für betriebswirtschaftliche Forschung*, *6*(62), 664-672.

Otto, A. R., & Daw, N. D. (2019). The opportunity cost of time modulates cognitive effort. *Neuropsychologia*, *123*, 92-105.

Oviatt, S. (2006). *Human-centered design meets cognitive load theory: designing interfaces that help people think*. ACM International Conference on Multimedia, New York, USA.

Paas, F., Renkl, A., & Sweller, J. (2004). Cognitive load theory: Instructional implications of the interaction between information structures and cognitive architecture. *Instructional science*, *32*(1/2), 1-8.

Paas, F., Tuovinen, J. E., Tabbers, H., & Van Gerven, P. W. (2016). Cognitive load measurement as a means to advance cognitive load theory. *Educational psychologist*, *38*(1), 63-71.

Paas, F. G., & Van Merriënboer, J. J. (1993). The efficiency of instructional conditions: An approach to combine mental effort and performance measures. *Human factors*, *35*(4), 737-743.

Páez, A. (2019). The pragmatic turn in explainable artificial intelligence (XAI). *Minds and Machines*, *29*(3), 441-459.

Paravastu, N. S., & Ramanujan, S. S. (2021). Interpersonal Trust and Technology Trust in Information Systems Research: A Comprehensive Review and A Conceptual Model. *International Journal of Information Systems and Social Change*, *12*(4), 1-18.

Pawellek, G. (2016). *Integrierte Instandhaltung und Ersatzteillogistik: Vorgehensweisen, Methoden, Tools* (Vol. 2). Springer-Verlag.

Peer, E., Brandimarte, L., Samat, S., & Acquisti, A. (2017). Beyond the Turk: Alternative platforms for crowdsourcing behavioral research. *Journal of experimental social psychology*, *70*, 153-163.

Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A design science research methodology for information systems research. *Journal of Management Information Systems*, *24*(3), 45-77.

Pereira, G. T., & de Carvalho, A. C. (2019). Bringing robustness against adversarial attacks. *Nature Machine Intelligence*, *1*(11), 499-500.

Persson, A., Laaksoharju, M., & Koga, H. (2021). We Mostly Think Alike: Individual Differences in Attitude Towards AI in Sweden and Japan. *The Review of Socionetwork Strategies*, *15*(1), 123-142.

Peters, F., Pumplun, L., & Buxmann, P. (2020). *Opening the Black Box: Consumer's Willingness to Pay for Transparency of Intelligent Systems*. European Conference on Information Systems (ECIS), Virtual Conference.

Pfeuffer, N., Benlian, A., Gimpel, H., & Hinz, O. (2019). Anthropomorphic information systems. *Business & Information Systems Engineering*, *61*(4), 523-533.

Pieters, W. (2011). Explanation and trust: what to tell the user in security and AI? *Ethics and information technology*, *13*(1), 53-64.

Poole, D., Mackworth, A., & Goebel, R. (1998). *Computational Intelligence*. Oxford University Press.

Portela, F., Aguiar, J., Santos, M. F., Silva, Á., & Rua, F. (2013). *Pervasive intelligent decision support system-technology acceptance in intensive care units*. Advances in Information Systems and Technologies, Algarve, Portugal.

Poursabzi-Sangdeh, F., Goldstein, D. G., Hofman, J. M., Vaughan, J. W., & Wallach, H. (2018). Manipulating and measuring model interpretability. *arXiv preprint arXiv:1802.07810*.

Poursabzi-Sangdeh, F., Goldstein, D. G., Hofman, J. M., Wortman Vaughan, J. W., & Wallach, H. (2021). *Manipulating and measuring model interpretability*. CHI Conference on human factors in computing systems, Virtual Conference.

Power, D. J. (2008). Understanding data-driven decision support systems. *Information Systems Management*, *25*(2), 149-154.

Pratt, J. W., & Zeckhauser, R. J. (1985). Principals and agents: An overview. *Principals and agents: The structure of business*, *1*, 12-15.

Preece, A., Harborne, D., Braines, D., Tomsett, R., & Chakraborty, S. (2018). Stakeholders in explainable AI. *arXiv preprint arXiv:1810.00184*.

Pu, P., & Chen, L. (2007). Trust-inspiring explanation interfaces for recommender systems. *Knowledge-Based Systems*, *20*(6), 542-556.

Püschel, L., Röglinger, M., & Schlott, H. (2016). *What's in a smart thing? Development of a multi-layer taxonomy*. International Conference on Information Systems, Dublin, Ireland.

Putnam, V., Riegel, L., & Conati, C. (2019). Toward XAI for Intelligent Tutoring Systems: a case study. *arXiv preprint arXiv:1912.04464*.

Quijano-Sanchez, L., Sauer, C., Recio-Garcia, J. A., & Diaz-Agudo, B. (2017). Make it personal: a social explanation system applied to group recommendations. *Expert Systems with Applications*, *76*, 36-48.

Rad, M. S., Nilashi, M., & Dahlan, H. M. (2018). Information technology adoption: a review of the literature and classification. *Universal access in the information society*, *17*(2), 361-390.

Ras, G., van Gerven, M., & Haselager, P. (2018). Explanation methods in deep learning: Users, values, concerns and challenges. *arXiv:1803.07517*.

Rehse, J.-R., Mehdiyev, N., & Fettke, P. (2019). Towards explainable process predictions for industry 4.0 in the dfki-smart-lego-factory. *KI-Künstliche Intelligenz*, *33*(2), 181-187.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016a). Model-agnostic interpretability of machine learning. *arXiv preprint arXiv:1606.05386*.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016b). "*Why should i trust you?*" *Explaining the predictions of any classifier*. International conference on knowledge discovery and data mining, San Francisco, USA.

Rich, P. (1992). The organizational taxonomy: Definition and design. *Academy of management review*, *17*(4), 758-781.

Riefle, L., & Benz, C. (2021). *User-specific determinants of conversational agent usage: A review and potential for future research*. International Conference on Wirtschaftsinformatik, Duisburg, Germany.

Riegelsberger, J., & Sasse, M. A. (2010). *Ignore these at your peril: Ten principles for trust design*. International Conference on Trust and Trustworthy Computing, Berlin, Germany.

Rogers, E. M. (2010). *Diffusion of innovations*. Simon and Schuster.

Rosenfeld, A., & Richardson, A. (2019). Explainability in human-agent systems. *Autonomous Agents and Multi-Agent Systems*, *33*(6), 673-705.

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.

Rudin, C., & Radin, J. (2019). Why are we using black box models in AI when we don't need to? A lesson from an explainable AI competition. *Harvard Data Science Review*, *1*(2).

Russell, S., & Norvig, P. (2021). *Artificial intelligence: A modern approach* (4th ed.). Pearson Education, Inc.

Ryan, M., & Stahl, B. C. (2020). Artificial intelligence ethics guidelines for developers and users: clarifying their content and normative implications. *Journal of Information, Communication and Ethics in Society*, *19*(1), 61-86.

Saeed, W., & Omlin, C. (2023). Explainable ai (xai): A systematic meta-survey of current challenges and future opportunities. *Knowledge-Based Systems*, *263*, 110273.

Salleh, M., Talpur, N., & Hussain, K. (2017). *Adaptive neuro-fuzzy inference system: Overview, strengths, limitations, and solutions*. International Conference on Data Mining and Big Data, Belgrade, Serbia.

Samek, W., Wiegand, T., & Müller, K.-R. (2017). Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. *arXiv preprint arXiv:1708.08296*.

Sardianos, C., Varlamis, I., Chronis, C., Dimitrakopoulos, G., Alsalemi, A., Himeur, Y., Bensaali, F., & Amira, A. (2021). The emergence of explainability of intelligent systems: Delivering explainable and personalized recommendations for energy efficiency. *International Journal of Intelligent Systems*, *36*(2), 656-680.

Savage, N. (2022). Breaking into the black box of artificial intelligence. *Nature*.

Saxena, A., & Goebel, K. (2008). *Turbofan engine degradation simulation data set*. Retrieved 18.06.2020 from ti.arc.nasa.gov/tech/prognostic-data-repository/#turbofan

Schemmer, M., Hemmer, P., Kühl, N., & Schäfer, S. (2022). *Designing Resilient AI-based Robo-Advisors: A Prototype for Real Estate Appraisal*. Conference on Design Science Research in Information Systems and Technology, St. Petersburg, FL, USA.

Schmeck, A., Opfermann, M., Van Gog, T., Paas, F., & Leutner, D. (2015). Measuring cognitive load with subjective rating scales during problem solving: differences between immediate and delayed ratings. *Instructional science*, *43*, 93-114.

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural networks*, *61*, 85-117.

Schmidt, P., Biessmann, F., & Teubner, T. (2020). Transparency and trust in artificial intelligence systems. *Journal of Decision Systems*, *29*(4), 260-278.

Schneider, J., & Handali, J. (2019). Personalized explanation in machine learning: A conceptualization. *arXiv preprint arXiv:1901.00770*.

Schoorman, F. D., Mayer, R. C., & Davis, J. H. (2007). An integrative model of organizational trust: Past, present, and future. *Academy of management review*, *32*(2).

Schreiber, A., & Bock, M. (2019). *Visualization and exploration of deep learning networks in 3d and virtual reality*. International Conference on Human-Computer Interaction, Orlando, Florida, USA.

Seidel, S., Chandra Kruse, L., Székely, N., Gau, M., & Stieger, D. (2018). Design principles for sensemaking support systems in environmental sustainability transformations. *European Journal of Information Systems*, *27*(2), 221-247.

Shaft, T. M., & Vessey, I. (2006). The role of cognitive fit in the relationship between software comprehension and modification. *MIS Quarterly*, *30*(1), 29-55.

Shahzad, F., Xiu, G., Khan, M. A. S., & Shahbaz, M. (2020). Predicting the adoption of a mobile government security response system from the user's perspective: An application of the artificial neural network approach. *Technology in Society*, *62*, 101278.

Shapiro, V. (2018). *Explaining System Intelligence*. Retrieved 27/04/2021 from https://experience.sap.com/skillup/explaining-system-intelligence/

Sharma, R., Kumar, A., & Chuah, C. (2021). Turning the blackbox into a glassbox: An explainable machine learning approach for understanding hospitality customer. *International Journal of Information Management Data Insights*, *1*(2), 100050.

Shim, J. P., Warkentin, M., Courtney, J. F., Power, D. J., Sharda, R., & Carlsson, C. (2002). Past, present, and future of decision support technology. *Decision support systems*, *33*(2), 111-126.

Shin, D. (2020a). How do users interact with algorithm recommender systems? The interaction of users, algorithms, and performance. *Computers in Human Behavior*, *109*, 106344.

Shin, D. (2020b). User perceptions of algorithmic decisions in the personalized AI system: perceptual evaluation of fairness, accountability, transparency, and explainability. *Journal of Broadcasting & Electronic Media*, *64*(4), 541-565.

Shin, D. (2021). The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI. *International Journal of Human-Computer Studies*, *146*, 102551.

Shin, D., Zhong, B., & Biocca, F. A. (2020). Beyond user experience: What constitutes algorithmic experiences? *International Journal of Information Management*, *52*, 102061.

Shneiderman, B. (2020). Bridging the gap between ethics and practice: Guidelines for reliable, safe, and trustworthy Human-Centered AI systems. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, *10*(4), 1-31.

Shneiderman, B., & Plaisant, C. (2016). *Designing the user interface: Strategies for effective human-computer interaction* (Vol. 6). Pearson Education.

Silveira, M. S., de Souza, C. S., & Barbosa, S. D. (2001). *Semiotic engineering contributions for designing online help systems*. Annual international conference on Computer documentation, Sante Fe New Mexico USA.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., & Lanctot, M. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature*, *529*(7587), 484-489.

Silver, M. S., Markus, M. L., & Beath, C. M. (1995). The information technology interaction model: A foundation for the MBA core course. *MIS Quarterly*, *19*(3), 361-390.

Simon, H. A. (1955). A behavioral model of rational choice. *The quarterly journal of economics*, *69*(1), 99-118.

Simon, H. A. (2019). *The Sciences of the Artificial, reissue of the third edition with a new introduction by John Laird*. MIT press.

Slack, D., Hilgard, A., Lakkaraju, H., & Singh, S. (2021). *Counterfactual explanations can be manipulated*. Advances in Neural Information Processing Systems, Virtual Conference.

Slack, D., Hilgard, S., Jia, E., Singh, S., & Lakkaraju, H. (2020). *Fooling lime and shap: Adversarial attacks on post hoc explanation methods*. Conference on AI, Ethics, and Society, Honolulu, HI, USA.

Slade, E. L., Dwivedi, Y. K., Piercy, N. C., & Williams, M. D. (2015). Modeling consumers' adoption intentions of remote mobile payments in the United Kingdom: extending UTAUT with innovativeness, risk, and trust. *Psychology & Marketing*, *32*(8), 860-873.

Sokol, K., & Flach, P. (2020). *Explainability fact sheets: a framework for systematic assessment of explainable approaches*. Conference on Fairness, Accountability, and Transparency, New York, USA.

Speith, T. (2022). *A Review of Taxonomies of Explainable Artificial Intelligence (XAI) Methods*. Conference on Fairness, Accountability, and Transparency, Seoul, South Korea.

Spinner, T., Schlegel, U., Schäfer, H., & El-Assady, M. (2020). explAIner: A visual analytics framework for interactive and explainable machine learning. *IEEE transactions on visualization and computer graphics*, *26*(1), 1064-1074.

Sprague, R. H. (1980). A Framework for the Development of Decision Support Systems. *MIS Quarterly*, *4*(4), 1-26.

Storey, V. C., Lukyanenko, R., Maass, W., & Parsons, J. (2022). Explainable AI. *Communication of the ACM*, *65*(4), 27-29.

Straub, D., & Burton-Jones, A. (2007). Veni, vidi, vici: Breaking the TAM logjam. *Journal of the Association for Information Systems*, *8*(4), 223-229.

Strauss, A., & Corbin, J. (1994). Grounded theory methodology: An overview. In Denzin K. & Lincoln Y. S. (Eds.), *Handbook of qualitative research* (pp. 273-285). Sage Publications Inc.

Strohm, L., Hehakaya, C., Ranschaert, E. R., Boon, W. P., & Moors, E. H. (2020). Implementation of artificial intelligence (AI) applications in radiology: hindering and facilitating factors. *European radiology*, *30*, 5525-5532.

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.

Subramanian, G. H., Nosek, J., Raghunathan, S. P., & Kanitkar, S. S. (1992). A comparison of the decision table and tree. *Communications of the ACM*, *35*(1), 89-94.

Suleman, D., Zuniarti, I., Sabil, E. D. S., Yanti, V. A., Susilowati, I. H., Sari, I., Marwansyah, S., Hadi, S. S., & Lestiningsih, A. S. (2019). Decision Model Based on Technology Acceptance Model (Tam) for Online Shop Consumers in Indonesia. *Academy of Marketing Studies Journal*, *23*(4), 1-14.

Sultana, T., & Nemati, H. R. (2021). *Impact of Explainable AI and Task Complexity on Human-Machine Symbiosis*. American Conference on Information Systems, Virtual Conference.

Šumak, B., Polancic, G., & Hericko, M. (2010). *An empirical study of virtual learning environment adoption using UTAUT*. International conference on mobile, hybrid, and on-line learning, Washington, DC, USA.

Sun, J., Liao, Q. V., Muller, M., Agarwal, M., Houde, S., Talamadupula, K., & Weisz, J. D. (2022). *Investigating Explainability of Generative AI for Code through Scenario-based Design*. International Conference on Intelligent User Interfaces, Helsinki, Finland.

Sundar, S. S. (2020). Rise of machine agency: A framework for studying the psychology of human-AI interaction (HAII). *Journal of Computer-Mediated Communication*, *25*(1), 74-88.

Tams, S., Hill, K., de Guinea, A. O., Thatcher, J., & Grover, V. (2014). NeuroIS-alternative or complement to existing methods? Illustrating the holistic effects of neuroscience and self-reported data in the context of technostress research. *Journal of the Association for Information Systems*, *15*, 723-753.

Tausch, N., Tam, T., Hewstone, M., Kenworthy, J., & Cairns, E. (2007). Individual-level and group-level mediators of contact effects in Northern Ireland: The moderating role of social identification. *British journal of social psychology*, *46*(3), 541-556.

Tawsifur, R., Muhammad, C., & Khandakar, A. (2022). *COVID-19 Radiography Database*. Retrieved 14.09.2022 from https://www.kaggle.com/datasets/tawsifurrahman/covid19-radiography-database

Taylor, S., & Todd, P. (1995). Decomposition and crossover effects in the theory of planned behavior: A study of consumer adoption intentions. *International journal of research in marketing*, *12*(2), 137-155.

Thiebes, S., Lins, S., & Sunyaev, A. (2021). Trustworthy artificial intelligence. *Electronic Markets*, *31*(2), 447-464.

Thomas, T., Singh, L., & Gaffar, K. (2013). The utility of the UTAUT model in explaining mobile learning adoption in higher education in Guyana. *International Journal of Education and Development using ICT*, *9*(3), 71-85.

Thompson, R. L., Higgins, C. A., & Howell, J. M. (1991). Personal computing: toward a conceptual model of utilization. *MIS Quarterly*, *15*(1), 125-143.

Tintarev, N., & Masthoff, J. (2012). Evaluating the effectiveness of explanations for recommender systems. *User Modeling and User-adapted interaction*, *22*(4-5), 399-439.

Tjoa, E., & Guan, C. (2019). A survey on explainable artificial intelligence (XAI): towards medical XAI. *arXiv preprint arXiv:1907.07374*.

Tomsett, R., Braines, D., Harborne, D., Preece, A., & Chakraborty, S. (2018). Interpretable to whom? A role-based model for analyzing interpretable machine learning systems. *arXiv preprint arXiv:1806.07552*.

Turban, E., & Watkins, P. R. (1986). Integrating Expert Systems and Decision Support Systems. *MIS Quarterly*, *10*(2), 121-136.

Uga, B. (2019). *Towards Trustworthy AI: A proposed set of design guidelines for understandable, trustworthy and actionable AI.* Uppsala Universitet, Uppsala, Schweden.

Vaishnavi, V. K., & Kuechler, W. (2007). *Design science research methods and patterns: innovating information and communication technology*. Auerbach Publications.

Vale, D., El-Sharif, A., & Ali, M. (2022). Explainable artificial intelligence (XAI) post-hoc explainability methods: risks and limitations in non-discrimination law. *AI and Ethics*, 1-12.

van der Waa, J., Nieuwburg, E., Cremers, A., & Neerincx, M. (2021). Evaluating XAI: A comparison of rule-based and example-based explanations. *Artificial Intelligence*, *291*, 103404.

Van Sonderen, E., Sanderman, R., & Coyne, J. C. (2013). Ineffectiveness of reverse wording of questionnaire items: Let’s learn from cows in the rain. *PloS one*, *8*(7), e68967.

Vempala, N. N., & Russo, F. A. (2018). Modeling music emotion judgments using machine learning methods. *Frontiers in psychology*, *8*, 2239.

Venable, J., Pries-Heje, J., & Baskerville, R. (2016). FEDS: a framework for evaluation in design science research. *European Journal of Information Systems*, *25*(1), 77-89.

Venkatesh, V. (2022). Adoption and use of AI tools: a research agenda grounded in UTAUT. *Annals of Operations Research*, *308*(1), 641-652.

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.

Venkatesh, V., Thong, J. Y., & 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.

Venkatesh, V., Thong, J. Y., & 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.

Verbeke, W., Martens, D., Mues, C., & Baesens, B. (2011). Building comprehensible customer churn prediction models with advanced rule induction techniques. *Expert Systems with Applications*, *38*(3), 2354-2364.

Vessey, I. (1991). Cognitive fit: A theory-based analysis of the graphs versus tables literature. *Decision sciences*, *22*(2), 219-240.

Vidotto, G., Massidda, D., Noventa, S., & Vicentini, M. (2012). Trusting Beliefs: A Functional Measurement Study. *Psicologica: International Journal of Methodology and Experimental Psychology*, *33*(3), 575-590.

Vigano, L., & Magazzeni, D. (2020). *Explainable security*. IEEE European Symposium on Security and Privacy Workshops (EuroS&PW), Genoa, Italy.

vom Brocke, J. (2007). Design principles for reference modeling: reusing information models by means of aggregation, specialisation, instantiation, and analogy. In Fettke P. & Loos P. (Eds.), *Reference modeling for business systems analysis* (pp. 47-76). IGI Global.

vom Brocke, J., Simons, A., Niehaves, B., Riemer, K., Plattfaut, R., & Cleven, A. (2009). *Reconstructing the Giant: On the importance of rigour in documenting the literature search process*. 17th European Conference on Information Systems (ECIS), Verona, Italy.

vom Brocke, J., Simons, A., Riemer, K., Niehaves, B., Plattfaut, R., & Cleven, A. (2015). Standing on the shoulders of giants: Challenges and recommendations of literature search in information systems research. *Communications of the association for information systems*, *37*(1), 206-224.

vom Brocke, J., Winter, R., Hevner, A., & Maedche, A. (2020). Accumulation and evolution of design knowledge in design science research: a journey through time and space. *Journal of the Association for Information Systems*, *21*(3), 9.

von Esch, N. (2018). *How to Design for AI-Enabled UI*. Retrieved 24/07/2021 from https://blog.prototypr.io/how-to-design-for-ai-enabled-ui-77e144e99126

von Eschenbach, W. J. (2021). Transparency and the black box problem: Why we do not trust AI. *Philosophy & Technology*(34), 1607-1622.

Vorm, E., & Combs, D. J. (2022). Integrating Transparency, Trust, and Acceptance: The Intelligent Systems Technology Model (ISTAM). *International Journal of Human-Computer Interaction*, 1-18.

Wang, D., Yang, Q., Abdul, A., & Lim, B. Y. (2019). *Designing theory-driven user-centric explainable AI*. CHI conference on human factors in computing systems, Glasgow Scotland, UK.

Wang, J., Gou, L., Shen, H.-W., & Yang, H. (2018). Dqnviz: A visual analytics approach to understand deep q-networks. *IEEE transactions on visualization and computer graphics*, *25*(1), 288-298.

Wang, J., Ma, Y., Zhang, L., Gao, R. X., & Wu, D. (2018). Deep learning for smart manufacturing: Methods and applications. *Journal of Manufacturing Systems*, *48*, 144-156.

Wang, P., Fan, E., & Wang, P. (2021). Comparative analysis of image classification algorithms based on traditional machine learning and deep learning. *Pattern Recognition Letters*, *141*, 61-67.

Wang, W., & Benbasat, I. (2007). Recommendation agents for electronic commerce: Effects of explanation facilities on trusting beliefs. *Journal of Management Information Systems*, *23*(4), 217-246.

Wang, W., & Benbasat, I. (2016). Empirical assessment of alternative designs for enhancing different types of trusting beliefs in online recommendation agents. *Journal of Management Information Systems*, *33*(3), 744-775.

Wanner, J., Heinrich, K., Janiesch, C., & Zschech, P. (2020). *How Much AI Do You Require? Decision Factors for Adopting AI Technology*. International Conference on Information Systems (ICIS), Hyderabad, India.

Wanner, J., Herm, L.-V., Fuchs, K., Winkelmann, A., & Janiesch, C. (2021). Entscheidungsunterstützung mit KI: Eine Analyse technischer und sozialer Faktoren für die industrielle Instandhaltung in Deutschland. *INDUSTRIE 4.0 Managemen*, *2*(37), 39-43.

Wanner, J., Herm, L.-V., Hartel, D., & Janiesch, C. (2019). Verwendung binärer Datenwerte für eine KI-gestützte Instandhaltung 4.0. *HMD Praxis der Wirtschaftsinformatik*, *56*(6), 1268-1281.

Wanner, J., Herm, L.-V., Heinrich, K., & Janiesch, C. (2021). *Stop ordering machine learning algorithms by their explainability! An empirical investigation of the tradeoff between performance and explainability*. Conference on e-Business, e-Services and e-Society, Dublin, Ireland.

Wanner, J., Herm, L.-V., Heinrich, K., & Janiesch, C. (2022a). The effect of transparency and trust on intelligent system acceptance: evidence from a user-based study. *Electronic Markets*, *32*(4), 2079-2102.

Wanner, J., Herm, L.-V., Heinrich, K., & Janiesch, C. (2022b). A social evaluation of the perceived goodness of explainability in machine learning. *Journal of Business Analytics*, *5*(1), 29-50.

Wanner, J., Herm, L.-V., Heinrich, K., Janiesch, C., & Zschech, P. (2020). *White, Grey, Black: Effects of XAI Augmentation on the Confidence in AI-based Decision Support Systems*. International Conference on Information Systems, Virtual Conference.

Wanner, J., Herm, L.-V., & Janiesch, C. (2019). *Countering the fear of black-boxed AI in maintenance: towards a smart colleague*. Pre-ICIS SIGDSA Symposium, Munich, Germany.

Wanner, J., Herm, L.-V., & Janiesch, C. (2020). *How Much is the Black Box? The Value of Explainability in Machine Learning Models*. European Conference On Information Systems, Virtual Conference.

Wanner, J., Herm, L.-V., & Janiesch, C. (2021). Digitalisierungspotenziale der Instandhaltung 4.0-Von der Aufbereitung binärer Daten zum Einsatz transparenter künstlicher Intelligenz. In Meinhardt S. & Wortmann F. (Eds.), *IoT-Best Practices* (pp. 177-193). Springer.

Wanner, J., Herm, L.-V., Langer, M., Imgrund, F., & Janiesch, C. (2020). *A Moral Consensus Mechanism for Autonomous Driving: Towards a Law-compliant Basis of Logic Programming*. Wirtschaftsinformatik, Potsdam, Germany.

Wanner, J., Herm, L. V., Janiesch, C., Fuchs, K., & Winkelmann, A. (2022). *Industry 4.0 Maintenance: An Examination of the Readiness of Germany's Industrial Sector*. Conference on Business Informatics (CBI), Amsterdam, Netherlands.

Wanner, J., Popp, L., Fuchs, K., Heinrich, K., Herm, L.-V., & Janiesch, C. (2021). *Adoption barriers of AI: A context-specific acceptance model for industrial maintenance*. European Conference on Information Systems, Virtual Conference.

Ward, J. H. J. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, *58*(301), 236-244.

Waytz, A., Heafner, J., & Epley, N. (2014). The mind in the machine: Anthropomorphism increases trust in an autonomous vehicle. *Journal of experimental social psychology*, *52*, 113-117.

Webster, J., & Watson, R. T. (2002). Analyzing the past to prepare for the future: Writing a literature review. *MIS Quarterly*, xiii-xxiii.

Weerts, H. J., van Ipenburg, W., & Pechenizkiy, M. (2019). A human-grounded evaluation of shap for alert processing. *arXiv preprint arXiv:1907.03324*.

Weiber, R., & Mühlhaus, D. (2014). *Strukturgleichungsmodellierung: Eine anwendungsorientierte Einführung in die Kausalanalyse mit Hilfe von AMOS, SmartPLS und SPSS*. Springer.

Weitz, K., Schiller, D., Schlagowski, R., Huber, T., & André, E. (2019). "*Do you trust me?*" *Increasing user-trust by integrating virtual agents in explainable AI interaction design*. International Conference on Intelligent Virtual Agents, Paris, France.

Weld, D. S., & Bansal, G. (2019). The challenge of crafting intelligible intelligence. *Communications of the ACM*, *62*(6), 70-79.

Wickramasinghe, C. S., Marino, D. L., Grandio, J., & Manic, M. (2020). *Trustworthy AI development guidelines for human system interaction*. International Conference on Human System Interaction (HSI), Tokyo, Japan.

Wiegand, G., Eiband, M., Haubelt, M., & Hussmann, H. (2020). *"I'd like an Explanation for That!" Exploring Reactions to Unexpected Autonomous Driving*. 22nd International Conference on Human-Computer Interaction with Mobile Devices and Services, Oldenburg, Germany.

Wilkinson, D., Alkan, Ö., Liao, Q. V., Mattetti, M., Vejsbjerg, I., Knijnenburg, B. P., & Daly, E. (2021). Why or why not? The effect of justification styles on chatbot recommendations. *ACM Transactions on Information Systems (TOIS)*, *39*(4), 1-21.

Williams, M. D., Rana, N. P., & Dwivedi, Y. K. (2015). The unified theory of acceptance and use of technology (UTAUT): a literature review. *Journal of enterprise information management*, *28*(3), 443-488.

Winter, R. (2008). Design science research in Europe. *European Journal of Information Systems*, *17*(5), 470-475.

Wintersberger, P., Nicklas, H., Martlbauer, T., Hammer, S., & Riener, A. (2020). *Explainable automation: Personalized and adaptive uis to foster trust and understanding of driving automation systems*. Conference on Automotive User Interfaces and Interactive Vehicular Applications, Virtual Conference.

Wolf, C. T., & Ringland, K. E. (2020). Designing accessible, explainable AI (XAI) experiences. *ACM SIGACCESS Accessibility and Computing*, *6*(125), 1-1.

Xiao, B., & Benbasat, I. (2007). E-commerce product recommendation agents: use, characteristics, and impact. *MIS Quarterly*, *31*(1), 137-209.

Xie, Y., Chen, M., Kao, D., Gao, G., & Chen, X. A. (2020). *CheXplain: enabling physicians to explore and understand data-driven, AI-enabled medical imaging analysis*. CHI Conference on Human Factors in Computing Systems, New York, NY, USA.

Xinogalos, S., & Satratzemi, M. (2022). The Use of Educational Games in Programming Assignments: SQL Island as a Case Study. *Applied Sciences*, *12*(13), 6563.

Yang, G., Ye, Q., & Xia, J. (2022). Unbox the black-box for the medical explainable ai via multi-modal and multi-centre data fusion: A mini-review, two showcases and beyond. *Information fusion*, *77*, 29-52.

Yang, H.-d., & Yoo, Y. (2004). It's all about attitude: revisiting the technology acceptance model. *Decision support systems*, *38*(1), 19-31.

Yang, Y. J., & Bang, C. S. (2019). Application of artificial intelligence in gastroenterology. *World journal of gastroenterology*, *25*(14), 1666-1683.

Yao, Y., & Murphy, L. (2007). Remote electronic voting systems: an exploration of voters' perceptions and intention to use. *European Journal of Information Systems*, *16*(2), 106-120.

Zerilli, J., Bhatt, U., & Weller, A. (2022). How transparency modulates trust in artificial intelligence. *Patterns*, *4*(3), 100455.

Zhang, Q., Yang, L. T., Chen, Z., & Li, P. (2018). A survey on deep learning for big data. *Information fusion*, *42*, 146-157.

Zhang, X., Noor, R., & Savalei, V. (2016). Examining the effect of reverse worded items on the factor structure of the need for cognition scale. *PloS one*, *11*(6), e0157795.

Zhang, Y., Liao, Q. V., & Bellamy, R. K. (2020). *Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making*. Conference on Fairness, Accountability, and Transparency, Barcelona, Spain.

Zhang, Y., & Ling, C. (2018). A strategy to apply machine learning to small datasets in materials science. *npj Computational Materials*, *4*(1), 25.

Zhang, Y., Xu, F., Zou, J., Petrosian, O. L., & Krinkin, K. V. (2021). *XAI Evaluation: Evaluating Black-Box Model Explanations for Prediction*. International Conference on Neural Networks and Neurotechnologies (NeuroNT), Saint-Petersburg, Russia.

Zhao, R., Benbasat, I., & Cavusoglu, H. (2019). *Transparency in Advice-Giving Systems: A Framework and a Research Model for Transparency Provision*. IUI Workshops, Los Angeles, USA.

Zhao, X., Lynch, J. G., Jr., & Chen, Q. (2010). Reconsidering Baron and Kenny: Myths and Truths about Mediation Analysis. *Journal of Consumer Research*, *37*(2), 197-206.

Zhou, J., Gandomi, A. H., Chen, F., & Holzinger, A. (2021). Evaluating the quality of machine learning explanations: A survey on methods and metrics. *Electronics*, *10*(5), 593.

Zschech, P., Horn, R., Höschele, D., Janiesch, C., & Heinrich, K. (2020). Intelligent user assistance for automated data mining method selection. *Business & Information Systems Engineering*, *62*, 227-247.

# Erklärung

Hiermit erkläre ich gemäß § 7 Abs. 2 Punkt 2 der Promotionsordnung der wirtschaftswissenschaftlichen Fakultät der Universität Würzburg, dass ich die Dissertation mit dem Titel:

“Algorithmic Decision-Making Facilities: Perception and Design of Explainable AI-based Decision Support Systems”

eigenständig, d.h. insbesondere selbstständig und ohne Hilfe einer kommerziellen Promotionsberatung angefertigt habe. Ebenso erkläre ich gemäß § 7 Abs. 2 Punkt 8 der Promotionsordnung, dass ich diese Dissertation nicht bereits bei einem früheren Prüfungsverfahren eingereicht habe.

Würzburg, 10.05.2023

Lukas-Valentin Herm