Electronic Markets (2022) 32:2185–2205
https://doi.org/10.1007/s12525-022-00606-3

RESEARCH PAPER

A nascent design theory for explainable intelligent systems

Lukas-Valentin Herm[1] · **Theresa Steinbach**[1] · **Jonas Wanner**[1] · **Christian Janiesch**[2]

Received: 23 May 2022 / Accepted: 14 October 2022 / Published online: 12 December 2022

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

JEL classification C6 · C8 · C9 · M15

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 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., 2022a; 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 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 et al. 2021; Sharma et al. 2021). This lack of explainability of the decision-making process leads to reduced trust

Responsible Editor: Christian Meske

✉ Lukas-Valentin Herm
lukas-valentin.herm@uni-wuerzburg.de

[1] Julius-Maximilians-Universität Würzburg, Sanderring 2, 97070 Würzburg, Germany

[2] TU Dortmund University, Dortmund, Germany

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 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 et al. 2021; 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., 2022b). 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 3 describes the used DSR methodology, including a comprehensive description of the two design cycles. Section 4 introduces the final nascent design theory and Section 5 presents an EIS real-world use case application and evaluation. We discuss the results in Section 6, before we conclude with a summary.

Research background

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 and 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 and 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 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 and Bunde 2022). For the following, we align with Herm, Heinrich, et al. (2022a) and Mohseni et al. (2021) by referring to these types of AI-based DSS or intelligent DSS as *intelligent systems*.

Artificial intelligence and intelligent systems

According to definition of Berente et al. (2021, 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., 2022a; Janiesch 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 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. 2016). In practice, this results in a complex tradeoff between the performance and the explainability of these models (Herm, Heinrich, et al., 2022a). That is, models with high predictive accuracy also tend to be more challenging to comprehend and vice versa (Herm et al. 2021). 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.

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 and 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 and 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., 2022a).

Hence, various techniques have been developed for DNNs (Adadi and Berrada 2018), showing a promising suitability for resolving the tradeoff between performance and

explainability (Arrieta et al. 2020; Herm, Heinrich, et al., 2022a). 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 et al. 2021; 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).

Related work

Apart IS-related contributions such as Förster et al. (2020) who provide a design process for user-centric XAI systems and Herm, Wanner, et al. (2022b) 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 and 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. 2019). 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., 2022a), 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.

Research methodology

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 and 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

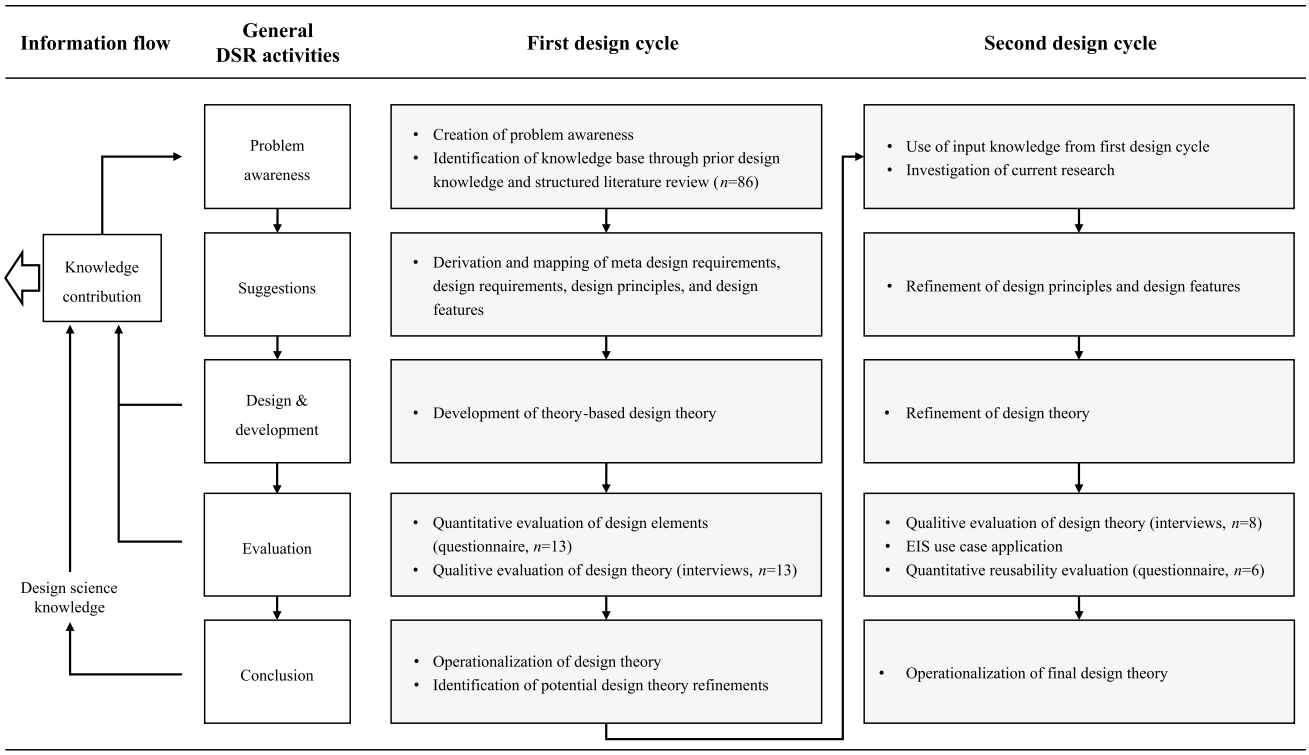

Fig. 1 Application of DSR according to Vaishnavi and Kuechler (2007)

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 and 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 Fig. 1).

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 2.3). 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

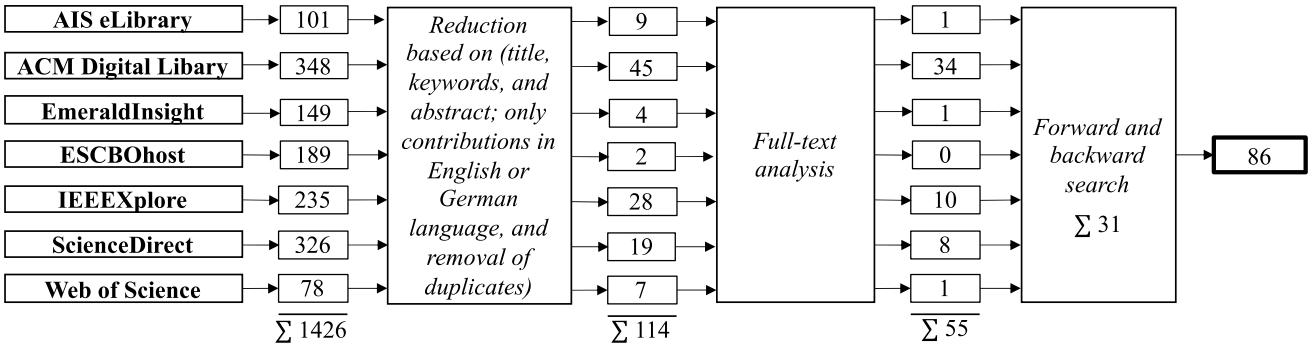

Fig. 2 Process of structured literature review according to vom Brocke et al. (2015)

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 and Kuechler 2007). Section 4 introduces and details our final nascent design theory, while Section 5 comprises the design theory instantiation and the quantitative evaluation.

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 A.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 2 summarizes the process of the literature review.

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 A 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 B). 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 1 for more information also on their demographics such as experience with AI.

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

Table 1 Overview interviewees and demographics (first design cycle)

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

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. Table 2 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 B for a graphical overview of the detailed description of the applied steps and the corresponding design theory, in Section 4 we will provide a comprehensive explanation of each element of the design theory except DF11[1].

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.

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 et al. 2021; 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

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

Table 2 Design requirements, design principles, and design features of first design cycle including their relevance

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".

(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 and 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 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.

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 5.

Table 3 Overview interviewees and demographics (second design cycle)

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

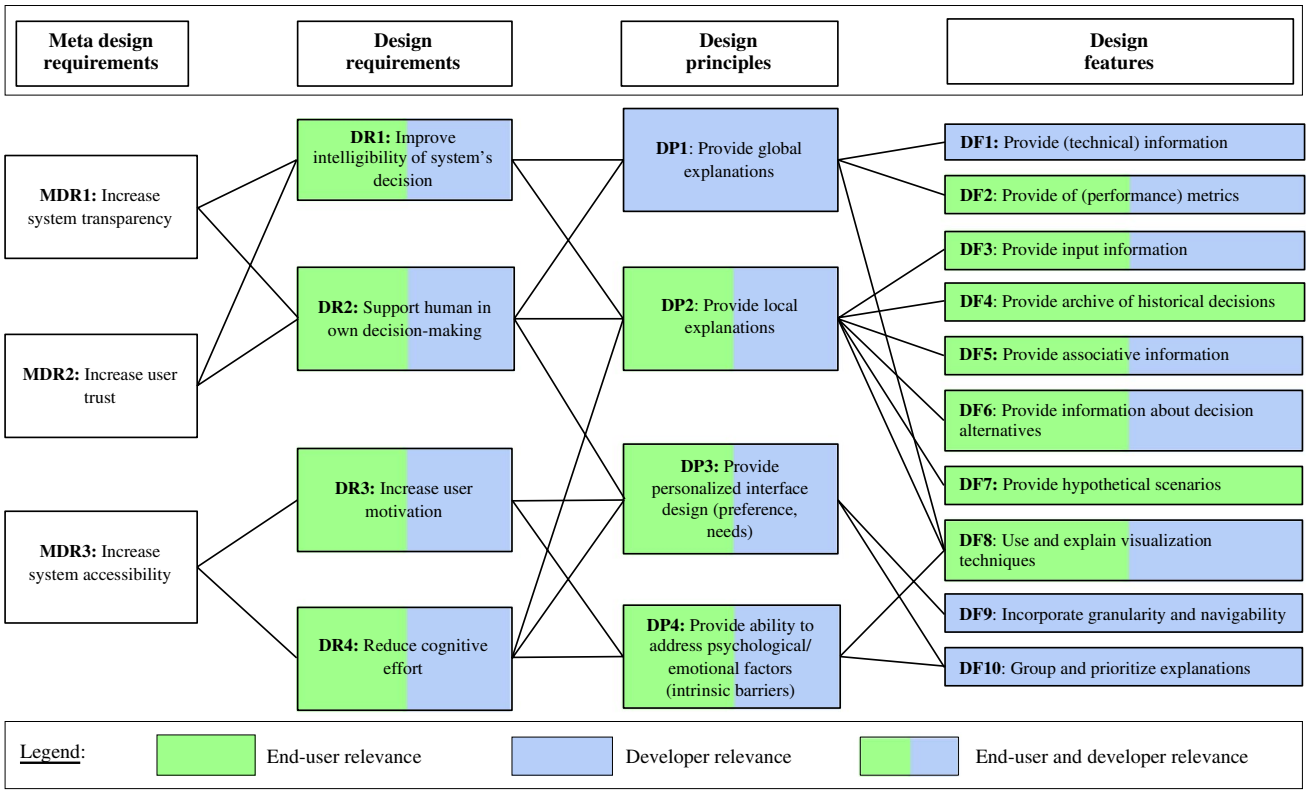

Fig. 3 Visualization of final nascent design theory

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., 2022b). 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 Fig. 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 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 and 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 et al. 2022]), 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 and 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 (Wanner et al. 2022; Carvalho et al. 2019; Thiebes et al. 2021). Especially for critical decisions, users have to rely on these results to make an informed decision (Choi and Ji 2015; Herm et al. 2021). 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 (Wanner et al. 2022; Cramer et al. 2008). 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 and Flach 2020). This also includes the adaptation of the system's recommendation (Ferreira and 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 and 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 3.2 and 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 and 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. 2019). EIS should therefore incorporate features that rise the motivation of the end-users using an EIS for decision support (Ferreira and Monteiro 2020). This could include different paradigms, as they are directly related to user expectations, leading to a well-perceived user experience (Nunes and 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).

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 and 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 Tables 4, 5, 6 and 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

Table 4 Principle of global explanations

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 5 Principle of local explanations

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 6 Principle of personalized interface design (preference, needs)

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 Principle of ability to address psychological/emotional factors (intrinsic barriers)

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

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 and 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).

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 and 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 and 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.

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 and Varshney 2021; Nunes and 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 and 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 and 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.

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 and 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 and Monteiro 2020)

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 and 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).

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 3.3 and 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 and 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 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.

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 C for the questionnaire.

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 and 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. Fig. 5), we consider our nascent design theory ready-to-use.

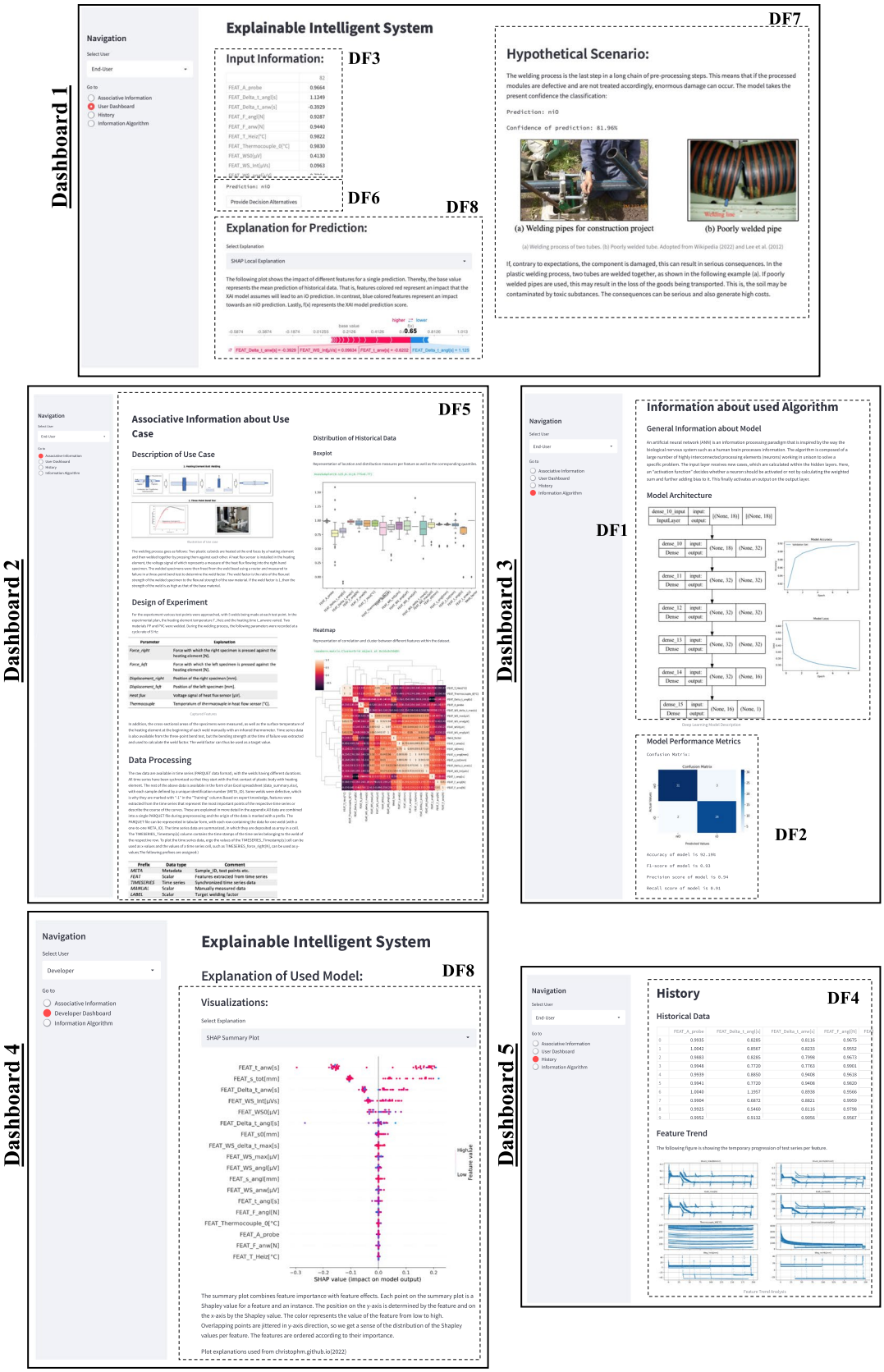

Fig. 4 Overview of the different dashboards of the EIS instantiation

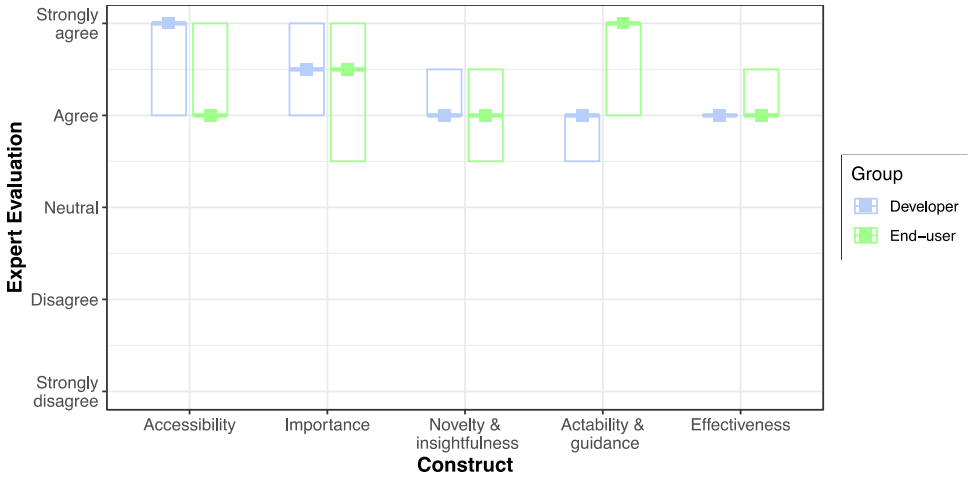

Fig. 5 Results of reusability evaluation according to Iivari et al. (2021)

Discussion of findings

Discussion and implications

Discussion. There are several contributions dealing with design approaches for EIS (e.g., Bunde 2021; Landwehr et al. 2022; Meske and 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 and 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 and Burns 2022) and demonstrate the need for an interdisciplinary design theory. The same applies for increasing the user motivation (DR3) (e.g., Ferreira and 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 and 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 and 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 and McNeish 2020), DF2 (Sun et al. 2022), DF3 (Nunes and 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 and 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 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., 2022a). 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 and 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.

Limitation and future research

Although we ensured scientific rigor by applying established DSR guidelines (Gregor and Hevner 2013; Iivari et al. 2021; Vaishnavi and 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 et al. 2021; 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 and 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 and 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.

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.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s12525-022-00606-3.

Funding Open Access funding enabled and organized by Projekt DEAL.

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, 582,* pp. 1–18. https://doi.org/10.1145/3173574.3174156

Abedin, B., Meske, C., Junglas, I., Rabhi, F., & Motahari-Nezhad, H. R. (2022). Designing and managing human-AI interactions. *Information Systems Frontiers, 1-7*. https://doi.org/10.1007/s10796-022-10313-1

Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE access, 6*, 52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052

Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., Suh, J., Iqbal, S., Bennett, P. N., Inkpen, K., Teevan, J., Kikin-Gil, R., & Horvitz, E. (2019). Guidelines for human-AI interaction. *Proceedings of the 2019 CHI conference on human factors in computing systems* (pp. 1–13). https://doi.org/10.1145/3290605.3300233

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. https://doi.org/10.1016/j.inffus.2019.12.012

Atkinson, K., Bench-Capon, T., & Bollegala, D. (2020). Explanation in AI and law: Past, present and future. *Artificial intelligence, 289*, 103387. https://doi.org/10.1016/j.artint.2020.103387

Baskerville, R. L., & Pries-Heje, J. (2019). Projectability in design science research. *Journal of Information Technology Theory And Application, 20*(1), 53–76. https://aisel.aisnet.org/jitta/vol20/iss1/3

Berente, N., Gu, B., Recker, J., & Santhanam, R. (2021). Managing artificial intelligence. *MIS Quarterly, 45*(3), 1433–1450. https://doi.org/10.25300/MISQ/2021/16274

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. https://doi.org/10.1007/s12599-020-00678-5

Bhatt, U., Andrus, M., Weller, A., & Xiang, A. (2020). Machine learning explainability for external stakeholders. *ArXiv Preprint*. https://doi.org/10.48550/arXiv.2007.05408

Boone, H. N., & Boone, D. A. (2012). Analyzing Likert data. *Journal of Extension, 50*(2), 1–5. https://tigerprints.clemson.edu/joe/vol50/iss2/48

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. *27th International Conference on Intelligent User Interfaces* (pp. 807–819). https://doi.org/10.1145/3490099.3511139

Bunde, E. (2021). AI-Assisted and explainable hate speech detection for social media moderators–A design science approach. *Proceedings of the 54th Hawaii International Conference on System Sciences* (pp. 1264–1274). http://hdl.handle.net/10125/70766

Carvalho, D. V., Pereira, E. M., & Cardoso, J. S. (2019). Machine learning interpretability: A survey on methods and metrics. *Electronics, 8*(832), 1–34. https://doi.org/10.3390/electronics8080832

Chander, A., Srinivasan, R., Chelian, S., Wang, J., & Uchino, K. (2018). *Working with beliefs: AI transparency in the enterprise*. CEUR-WS IUI Workshops. https://www.researchgate.net/publication/331970789_Working_with_Beliefs_AI_Transparency_in_the_Enterprise

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 (forthcoming).* https://doi.org/10.17705/1jais.00759

Chazette, L., & Schneider, K. (2020). Explainability as a non-functional requirement: challenges and recommendations. *Requirements Engineering, 25*(4), 493–514. https://doi.org/10.1007/s00766-020-00333-1

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*.

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. https://doi.org/10.1080/10447318.2015.1070549

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. https://doi.org/10.1016/j.inffus.2021.11.003

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.*

Cirqueira, D., Helfert, M., & Bezbradica, M. (2021). Towards design principles for user-centric explainable AI in fraud detection. *International Conference on Human-Computer Interaction.*

Conati, C., Barral, O., Putnam, V., & Rieger, L. (2021). Toward personalized XAI: A case study in intelligent tutoring systems. *Artificial intelligence, 298*, 103503. https://doi.org/10.1016/j.artint.2021.103503

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. https://doi.org/10.1007/s11257-008-9051-3

Das, A., & Rad, P. (2020). Opportunities and challenges in explainable artificial intelligence (XAI): A survey. *arXiv preprint.*https://doi.org/10.48550/arXiv.2006.11371

Dellermann, D., Ebel, P., Söllner, M., & Leimeister, J. M. (2019). Hybrid intelligence. *Business & Information Systems Engineering, 61*(5), 637–643. https://doi.org/10.1007/s12599-019-00595-2

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. https://doi.org/10.1016/j.ijhcs.2022.102792

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.

Epley, N., Waytz, A., & Cacioppo, J. T. (2007). On seeing human: a three-factor theory of anthropomorphism. *Psychological review, 114*(4), 864. https://doi.org/10.1037/0033-295X.114.4.864

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.* https://doi.org/10.1016/j.future.2022.03.009

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*. https://doi.org/10.1007/978-3-030-49760-6_4

Flick, U. (2020). Gütekriterien qualitativer Forschung. In *Handbuch qualitative Forschung in der Psychologie* (pp. 247–263). Springer. https://doi.org/10.1007/978-3-531-92052-8_28

Forster, M., Klier, M., Kluge, K., & Sigler, I. (2020). Fostering human agency: A process for the design of usercentric XAI systems. *International conference on information systems, Virtual conference proceedings* (p. 12) https://aisel.aisnet.org/icis2020/hci_artintel/hci_artintel/12

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*. https://doi.org/10.1115/DETC2015-46157

Fürnkranz, J., Kliegr, T., & Paulheim, H. (2020). On cognitive preferences and the plausibility of rule-based models. *Machine Learning, 109*(4), 853–898. https://doi.org/10.1007/s10994-019-05856-5

Glaser, B., & Strauss, A. (1967). Grounded theory: The discovery of grounded theory. *Sociology The Journal of the British Sociological Association, 12*, 27–49.

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)*. https://library.oapen.org/handle/20.500.12657/41230

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. https://doi.org/10.1609/aimag.v38i3.2741

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). https://doi.org/10.17705/1jais.00649

Gregor, S., & Hevner, A. R. (2013). Positioning and presenting design science research for maximum impact. *MIS Quarterly, 37*(2), 337–355. https://doi.org/10.25300/MISQ/2013/37.2.01

Gunning, D., Stefik, M., Choi, J., Miller, T., Stumpf, S., & Yang, G.-Z. (2019). XAI—Explainable artificial intelligence. *Science Robotics, 4*(37). https://doi.org/10.1126/scirobotics.aay7120

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. https://doi.org/10.1016/j.ijhcs.2008.09.008

Hepenstal, S., & McNeish, D. (2020). Explainable artificial intelligence: What do you need to know? In D. D. Schmorrow & C. M. Fidopiastis (Eds.), *Augmented cognition. Theoretical and technological approaches. HCII 2020. Lecture notes in computer science* (Vol. 12196). Springer. https://doi.org/10.1007/978-3-030-50353-6_20

Herm, L.-V., Heinrich, K., Wanner, J., & Janiesch, C. (2022a). Stop ordering machine learning algorithms by their explainability! A user-centered investigation of performance and explainability. *International Journal of Information Management, 102538*. https://doi.org/10.1016/j.ijinfomgt.2022.102538

Herm, L.-V., Wanner, J., & Janiesch, C. (2022b). A taxonomy of user-centered explainable AI studies (p. 9). *PACIS 2022 Proceedings*. https://aisel.aisnet.org/pacis2022/9

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 (p. 82). *ECIS 2021 Research Papers*. https://aisel.aisnet.org/ecis2021_rp/82

Hevner, A. R. (2021). The duality of science: Knowledge in information systems research. *Journal of Information Technology, 36*(1), 72–76. https://doi.org/10.1177/0268396220945714

Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design science in information systems research. *MIS Quarterly, 28*(1), 75–105. https://doi.org/10.2307/25148625

Hilton, D. J. (1996). Mental models and causal explanation: Judgements of probable cause and explanatory relevance. *Thinking & Reasoning, 2*(4), 273–308. https://doi.org/10.1080/135467896394447

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.

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. https://doi.org/10.1016/j.inffus.2021.10.007

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. https://doi.org/10.1016/j.ijinfomgt.2022.102497

Hutson, M. (2017). AI Glossary: Artificial intelligence, in so many words. *Science, 357*(6346), 19–19. https://doi.org/10.1126/science.357.6346.19

Iivari, J., Hansen, M. R. P., & Haj-Bolouri, A. (2021). A proposal for minimum reusability evaluation of design principles. *European Journal of Information Systems, 30*(3), 286–303. https://doi.org/10.1080/0960085X.2020.1793697

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. https://doi.org/10.17705/1jais.00642

Janiesch, C., Zschech, P., & Heinrich, K. (2021). Machine learning and deep learning. *Electronic Markets, 31*(3), 685–695. https://doi.org/10.1007/s12525-021-00475-2

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. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency,* New York, NY.

Kaiser, R. (2014). *Qualitative Experteninterviews: Konzeptionelle Grundlagen und praktische Durchführung*. Springer. https://doi.org/10.1007/978-3-658-02479-6

Kaur, D., Uslu, S., Rittichier, K. J., & Durresi, A. (2022). Trustworthy artificial intelligence: A review. *ACM Computing Surveys (CSUR), 55*(2), 1–38. https://doi.org/10.1145/3491209

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). https://doi.org/10.1186/s12916-019-1426-2

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. https://doi.org/10.1145/3487065

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*.

Lambers, J., & Balzer, C. (2022). Plastic welding process data. B2Share EUDAT. https://doi.org/10.23728/b2share.657bb2383ce946dcb4cab9419e1645d3

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*. https://doi.org/10.1007/s12599-022-00745-z

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. https://doi.org/10.25300/MISQ/2021/16564

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature, 521*(7553), 436–444. https://doi.org/10.1038/nature14539

Lee, A. S., & Baskerville, R. L. (2003). Generalizing generalizability in information systems research. *Information Systems Research, 14*(3), 221–243. https://doi.org/10.1287/isre.14.3.221.16560

Lewis, D. K. (1986). Causal explanation. *Philosophical Papers, 2*, 214–240.

Liao, Q. V., Gruen, D., & Miller, S. (2020). *CHI Conference on Human Factors in Computing Systems*. CHI Conference on Human Factors in Computing Systems. https://doi.org/10.1145/3313831.3376590

Liao, Q. V., & Varshney, K. R. (2021). Human-centered explainable AI (XAI): From algorithms to user experiences. *arXiv preprint*. https://doi.org/10.48550/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. https://doi.org/10.1145/3236386.3241340

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. https://doi.org/10.1007/s10115-009-0192-4

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, Cham.*

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., & Lee, S.-I. (2020). From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence, 2*(1), 56–67. https://doi.org/10.1038/s42256-019-0138-9

Mahmud, H., Islam, A. K. M. 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. https://doi.org/10.1016/j.techfore.2021.121390

Malhi, A., Knapic, S., & Främling, K. (2020). Explainable agents for less bias in human-agent decision making. In: Calvaresi, D., Najjar, A., Winikoff, M., Främling, K. (eds) *Explainable, Transparent Autonomous Agents and Multi-Agent Systems*. EXTRAAMAS 2020. Lecture Notes in Computer Science(), vol 12175. Springer, Cham. https://doi.org/10.1007/978-3-030-51924-7_8

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. https://doi.org/10.1038/s41586-019-1799-6

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. https://doi.org/10.1145/1985347.1985353

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*. https://doi.org/10.1007/s10796-021-10234-5

Meske, C., Bunde, E., Schneider, J., & Gersch, M. (2022). Explainable artificial intelligence: objectives, stakeholders, and future research opportunities. *Information Systems Management, 39*(1), 53–63. https://doi.org/10.1080/10580530.2020.1849465

Meth, H., Mueller, B., & Maedche, A. (2015). Designing a requirement mining system. *Journal of the Association for Information Systems, 16*(9), 799–837. https://doi.org/10.17705/1jais.00408

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence, 267*, 1–38. https://doi.org/10.1016/j.artint.2018.07.007

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, 11*(3–4), 1–45. https://doi.org/10.1145/3387166

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*.

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. https://doi.org/10.1016/j.artint.2021.103573

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. https://doi.org/10.3390/math10040554

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), 393–444. https://doi.org/10.1007/s11257-017-9195-0

O'Connor, C., & Joffe, H. (2020). Intercoder reliability in qualitative research: debates and practical guidelines. *International journal of qualitative methods, 19*. https://doi.org/10.1177/1609406919899220

Oviatt, S. (2006). Human-centered design meets cognitive load theory: Designing interfaces that help people think. *ACM International Conference on Multimedia*.

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. https://doi.org/10.2753/MIS0742-1222240302

Poole, D., Mackworth, A., & Goebel, R. (1998). *Computational intelligence*. Oxford University Press.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1135–1144). https://doi.org/10.1145/2939672.2939778

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence, 1*(5), 206–215. https://doi.org/10.1038/s42256-019-0048-x

Samek, W., Wiegand, T., & Müller, K.-R. (2017). Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. *arXiv preprint.*https://doi.org/10.48550/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. https://doi.org/10.1002/int.22314

Schemmer, M., Hemmer, P., Kühl, N., & Schäfer, S. (2022). Designing resilient AI-based robo-advisors: A prototype for real estate appraisal. *17th International Conference on Design Science Research in Information Systems and Technology, St. Petersburg, FL, USA*.

Schmidt, P., Biessmann, F., & Teubner, T. (2020). Transparency and trust in artificial intelligence systems. *Journal of Decision Systems, 29*(4), 260–278. https://doi.org/10.1080/12460125.2020.1819094

Schneider, J., & Handali, J. (2019). Personalized explanation in machine learning: A conceptualization. *arXiv*, 1901.00770. https://doi.org/10.48550/arXiv.1901.00770

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. https://doi.org/10.1057/s41303-017-0039-0

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. https://doi.org/10.1016/j.jjimei.2021.100050

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. https://doi.org/10.1016/j.ijhcs.2020.102551

Shin, D., Zhong, B., & Biocca, F. A. (2020). Beyond user experience: What constitutes algorithmic experiences? *International Journal of Information Management, 52*, 102061. https://doi.org/10.1016/j.ijinfomgt.2019.102061

Shneiderman, B., & Plaisant, C. (2016). *Designing the user interface: Strategies for effective human-computer interaction* (Vol. 6). Pearson Education.

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. https://doi.org/10.1002/mar.20823

Sokol, K., & Flach, P. (2020). Explainability fact sheets: A framework for systematic assessment of explainable approaches. *Conference on Fairness, Accountability, and Transparency*.

Sprague, R. H. (1980). A Framework for the development of decision support systems. *MIS Quarterly, 4*(4), 1–26. https://doi.org/10.2307/248957

Storey, V. C., Lukyanenko, R., Maass, W., & Parsons, J. (2022). Explainable AI. *Communication of the ACM, 65*(4), 27–29. https://doi.org/10.1145/3490699

Strauss, A., & Corbin, J. (1994). Grounded theory methodology: An overview. In K. Denzin & Y. S. Lincoln (Eds.), *Handbook of qualitative research* (pp. 273–285). Sage Publications Inc.

Stumpf, S., Rajaram, V., Li, L., Wong, W.-K., Burnett, M., Dietterich, T., Sullivan, E., & Herlocker, J. (2019). Interacting meaningfully with machine learning systems: Three experiments. *International Journal of Human-Computer Studies, 67*(8), 639–662. https://doi.org/10.1016/j.ijhcs.2009.03.004

Sun, J., Liao, Q. V., Muller, M., Agarwal, M., Houde, S., Talamadupula, K., & Weisz, J. D. (2022). *International Conference on Intelligent User Interfaces*. International Conference on Intelligent User Interfaces. https://doi.org/10.1145/3490099.3511119

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. https://doi.org/10.1093/jcmc/zmz026

Thiebes, S., Lins, S., & Sunyaev, A. (2021). Trustworthy artificial intelligence. *Electronic Markets, 31*(2), 447–464. https://doi.org/10.1007/s12525-020-00441-4

Turban, E., & Watkins, P. R. (1986). Integrating expert systems and decision support systems. *MIS Quarterly, 10*(2), 121–136. https://doi.org/10.2307/249031

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*. https://doi.org/10.1007/s43681-022-00142-y

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. https://doi.org/10.1016/j.artint.2020.103404

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. https://doi.org/10.1057/ejis.2014.36

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. https://doi.org/10.17705/1CAIS.03709

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. https://doi.org/10.17705/1jais.00611

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). https://doi.org/10.1007/s12525-022-00593-5

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.*

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, New York, NY.*

Xinogalos, S., & Satratzemi, M. (2022). The use of educational games in programming assignments: SQL Island as a case study. *Applied Sciences, 12*(13), 6563. https://doi.org/10.3390/app12136563

Zerilli, J., Bhatt, U., & Weller, A. (2022). How transparency modulates trust in artificial intelligence. *Patterns, 100455*. https://doi.org/10.1016/j.patter.2022.100455

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. https://doi.org/10.3390/electronics10050593

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. https://doi.org/10.1007/s12599-020-00642-3