Who needs XAI in the Energy Sector? A Framework to Upgrade Black Box Explainability

Sarah Kristin Lier
Information Systems Institute, lier@iwi.uni-hannover.de

Jana Gerlach
Information Systems Institute, Leibniz Universität Hannover, gerlach@iwi.uni-hannover.de

Michael H. Breitner
Leibniz Universität Hannover, breitner@iwi.uni-hannover.de

Who needs XAI in the Energy Sector? A Framework to Upgrade Black Box Explainability

Completed Research Paper

Sarah K. Lier
Leibniz University Hannover
Königsworther Platz 1
30167 Hannover
lier@iwi.uni-hannover.de

Jana Gerlach
Leibniz University Hannover
Königsworther Platz 1
30167 Hannover
gerlach@iwi.uni-hannover.de

Michael H. Breitner
Leibniz University Hannover
Königsworther Platz 1
30167 Hannover
breitner@iwi.uni-hannover.de

Abstract

Artificial Intelligence (AI)-based methods in the energy sector challenge companies, organizations, and societies. Organizational issues include traceability, certifiability, explainability, responsibility, and efficiency. Societal challenges include ethical norms, bias, discrimination, privacy, and information security. Explainable Artificial Intelligence (XAI) can address these issues in various application areas of the energy sector, e.g., power generation forecasting, load management, and network security operations. We derive Key Topics (KTs) and Design Requirements (DRs) and develop Design Principles (DPs) for efficient XAI applications through Design Science Research (DSR). We analyze 179 scientific articles to identify our 8 KTs for XAI implementation through text mining and topic modeling. Based on the KTs, we derive 15 DRs and develop 18 DPs. After that, we discuss and evaluate our results and findings through expert surveys. We develop a Three-Forces Model as a framework for implementing efficient XAI solutions. We provide recommendations and a further research agenda.

Keywords: Explainable Artificial Intelligence, Energy Sector, Design Science Research, Topic Modeling, Design Principles, Three-Forces Model.

Introduction

AI can process data, detect patterns, and make predictions (Haag et al. 2022). Scientists, professionals, and other stakeholders utilize AI to receive support in decision-making processes (Storey et al. 2022). The implementation of AI in the energy sector is increasing, so more solutions are needed. As a part of the critical infrastructure, which includes the energy sector, the importance for the state community is high. The impairment would result in lasting supply bottlenecks, significant disruptions to public safety, or other consequences (Alova et al. 2021; Gerlach et al. 2023; Machlev et al. 2022b). The energy sector focuses on decentralization, digitalization, and decarbonization of energy systems, especially to realize and implement renewable energies. Due to the increased volume, velocity, and variety of energy-relevant data, AI can

increase the value creation of such information and mitigate the energy transition-associated complexity (Kruse et al. 2021a; Kuzlu et al. 2020; Machlev et al. 2022a). AI is used for load forecasting, power generation and management, demand side management (DSM), and electrical grid operation and control. AI has high relevance regarding decentralized smart energy solutions, e.g., smart grids (SGs) and the modernization of power grids, although AI is not always considered efficient, as traditional methods are less expensive and more sufficient (Kuzlu et al. 2020). The safe integration of renewable energies is enhanced by hardware like smart meters, energy storage devices, SGs, and software like AI and Blockchain. The modernization of power grids enables the collection of information on power quality and energy consumption data, analysis of data for such purposes as forecasting energy demand, optimizing supply and demand, reducing power generation costs and greenhouse gas emissions, improving grid stability and reliability, optimizing energy efficiency, and responding to unforeseen changes (e.g., Alova et al. 2021; Hatzakis et al. 2019; Kruse et al. 2021b; Machlev et al. 2022b). AI-based energy methods contribute to the transformed energy system by optimizing and scheduling flexibilities to increase power efficiency and consumption (Ponnusamy et al. 2021). Implementing AI solutions makes business processes in the energy sector about 28% more efficient (Benbya et al. 2020). Efficient energy distribution improves consumer welfare and energy organizations' services (Antonopoulos et al. 2020). Due to the non-transparent nature of AI methods, explainability is becoming relevant for users, developers, energy providers, decision-makers, and other stakeholders because of the limited explanations and dynamics in decision-making processes. XAI, also known as white or grey box, enables a better understanding of models and systems through interpretability and traceability (Rawal et al. 2021). Explainability is the degree to which humans understand the decisions made by AI systems. XAI in the energy sector refers to the application of AI technologies, algorithms, and models. They make accurate predictions or decisions and provide transparent and interpretable insights into how those predictions or decisions are reached. XAI aims to bridge the gap between the inherent complexity of advanced AI techniques and the need for understanding and trust in decision-making (Angelov et al. 2021; Arrieta et al. 2019; Rawal et al. 2021). XAI techniques in the energy sector enable stakeholders to comprehend and validate the rationale behind AI-driven recommendations, predictions, or optimizations related to energy production, distribution, consumption, and management. By offering comprehensible explanations, XAI enhances transparency, accountability, and stakeholder confidence in AI-powered solutions, allowing for informed and effective decision-making while minimizing the risks of biased or unexpected outcomes (Angelov et al. 2021; Arrieta et al. 2019; Machlev et al. 2022a). Current XAI literature includes programming XAI algorithms (e.g., Ren et al. 2022; Sairam et al. 2020), monoculture guidelines for XAI applications (e.g., Ignatiev 2020; Kuhn et al. 2020), or XAI challenges and requirements (e.g., Angelov et al. 2019; Rawal et al. 2021). There are general principles, guidelines, and frameworks for XAI. For example, Phillips et al. (2021) derived four XAI principles: explanation, meaning, accuracy, and knowledge limits. Some other big organizations like Google, Amazon, and Tesla also derived general principles, requirements, and guidelines for using and implementing XAI. However, these algorithms, applications, principles, requirements, and guidelines cannot be transmitted completely to using and implementing XAI in the energy sector. The energy sector represents a critical infrastructure with complex impacts. AI decisions impact the economy, environment, and society. Sustainability, environmental protection, and societal acceptance and participation are reasons for customized DPs of XAI in the energy sector (e.g., Alova et al. 2021; Kruse et al. 2021b; Kuzlu et al. 2020; Machlev et al. 2022a). To combine the fragmented scientific knowledge base, we analyze relevant literature and expert interviews to develop DPs for XAI applications in the energy sector. We address the research questions (RQs):

RQ1: *What XAI design requirements in the energy sector are known, today?*

RQ2: *Which XAI design principles in the energy sector can be deduced?*

We explain XAI methods and algorithms in the energy sector. Then, we justify and apply our research design – inspired by DSR by vom Brocke et al. (2020) and Hevner and Chatterjee (2010) – to deduce DPs. In our first research step, we address our problem formulation. In our second step, we build a knowledge base through a Mixed Methods approach and form KTs from our data set. Our knowledge base is based on a literature review with an exploratory text analysis method expanded through confirmatory expert interviews. In our third step, we derive our DRs and deduce our DPs. In our fourth step, we evaluate and adjust the DRs and DPs through additional expert interviews. After using DSR, we develop the Three-Forces Model. The Three-Forces Model is a guide for the successful implementation of AI systems. We discuss our results and findings, deduce practical and theoretical implications and recommendations and present further research directions based on our limitations.

Theoretical Background

AI includes machine learning (ML), which can learn from data without complex rules. ML is used for pattern recognition and prediction. Deep Learning (DL) is a subfield of ML characterized by accuracy and high performance. In DL, artificial neural networks are used in image or speech recognition and for predicting time series (Arrieta et al. 2019; Machlev et al. 2022a). XAI makes results from AI systems more understandable and traceable for users, developers, programmers, or other stakeholders. The definition of XAI is not standardized and generalized (Adadi and Berrada 2018). The Defense Advanced Research Projects Agency defines XAI as creating explainable models maintaining high predictive accuracy that enables people to understand, trust, and use AI effectively (Gunning and Aha 2019). Arrieta et al. (2020, p. 6) describe XAI models: "Given a certain audience, an explainable artificial intelligence produces details or reasons to make its functioning clear or easy to understand." XAI aims to explain non-transparent systems and includes more than the word "explainability" (Adadi and Berrada 2018). It unlocks an AI system's black box, providing traceable explanations. Thus, justified, ethical, and fair decisions can be made (Storey et al. 2022). The intention is to turn a non-transparent black box into a transparent white box with accuracy and explainability. The white box combines ML, human-computer interactions (HCI), and explanations from human experts in the application domain (Loyola-Gonzalez 2019). XAI is described as the interpretability of the model. Interpretability is similar to transparency, so transparency can be understood as a property of a model in which the system states are open and can be explained by how they review, analyze, or interpret (Angelov et al. 2021; Arrieta et al. 2019). Traceability describes the model property that a system is understandable to the user without explaining structures or algorithms (Rawal et al. 2021). XAI models can be subdivided into local and global explainability. Local explainability refers to individual instances of data, i.e., decisions in a model. Global explainability is based on understanding the entire model with a group of data sets as input (Angelov et al. 2021; Arrieta et al. 2019; Machlev et al. 2022a; Rawal et al. 2021). XAI models operate with algorithms and methods, e.g., local interpretable model-agnostic explanations (LIME), Shapley additive explanations (SHAP), or explain like I am 5 (ELI5), which are different in their advantages and disadvantages, e.g., accuracy. LIME approximates AI models using interpretable models. This technique is model-independent and interacts locally, facilitating understanding and approximation of global learning models (Kuzlu et al. 2020; Schönhof et al. 2021). SHAP supports the explanation of models and features to build the model. SHAP explains ML forecasts based on a feature assignment for forecasts and analyzes the distribution of feature meanings. It clarifies forecasts and global and local explanations (Górski and Ramakrishna 2021; Kuzlu et al. 2020; Schönhof et al. 2021). ELI5 assists with classification and regression models and their implementation in Python, presenting the relevance of a feature in the form of weight to explain the model. ELI5 is model-dependent and a tool in Python-based ML packages, such as extreme gradient boosting (XGBoost) (Kuzlu et al. 2020). Those models involve some explanations for the AI, but they are often lacking and non-transparent, so there is less explainability (Gerlach et al. 2023). Online appendix A aggregates various studies and works by other authors (e.g., Adadi and Berrada 2018; Arrieta et al. 2019; Machlev et al. 2022a) and demonstrates existing XAI applications in the energy sector, application areas, and contributions through XAI implementation as an overview.

Research Design and Research Methods

Design Science Research

To address our RQs, we use DSR inspired by vom Brocke et al. (2020) and Hevner and Chatterjee (2010) and follow the scheme inspired by Gregor et al. (2020). Artifacts and solutions to a problem can be created. Thus, the solution is considered the central focus. Application- and problem-oriented DSR approach can analyze constantly changing topics and consider literature and research status quo. Constantly optimizing the artifact and providing a basic understanding of the topic, the DSR approach creates an artifact to solve research problems with accompanying analysis (Hevner and Chatterjee 2010). The DSR approach consists of an iterative development through continuous adjustment and improvement of the design artifact. It is allowed to involve multiple stakeholders and to integrate theory and practice into the development and evaluation of the artifact (vom Brocke et al. 2020). To address RQ1, we use topic modeling and text mining to form clusters. Text mining and topic modeling analysis contribute to detecting patterns and improving objective research results and findings. Hard-to-recognized information is uncovered through an ML tool (Tong and Zhang 2016). Based on the clusters, we derive requirements from the associated data sets and

sort them by relevance to obtain DRs. Those DRs refer to specific criteria the designed artifact must fulfill to address the identified problem effectively. The DRs are derived from the problem context, stakeholders, and research target. They ensure the alignment of the artifact with the intended purpose (Hevner and Chatterjee 2010). For RQ2, we derive DPs from the entire literature data set and combine them with the DRs. DPs are viewed as a framework for the design and development of an artifact. They describe guidelines based on knowledge, theory, and best practices for the economy, efficiency, and usefulness of the artifact (Gregor et al. 2020). We contribute with our KTs, DRs, and DPs as a DSR artifact to level 2 (emergent design theory) according to Gregor and Hevner's (2013) DSR description. By contributing to level 2, we implied general artifacts, such as methods, models, or DPs (Gregor and Hevner 2013). Lee et al. (2015) divide artifacts into technological artifacts (e.g., software), information artifacts (e.g., messages), and social artifacts (e.g., charitable acts). Lowry et al. (2017) identified more types of IS artifacts and called for a detailed allocation. We create a context-specific artifact in the energy sector's XAI field. The DSR approach allows us to leverage DRs and DPs from the energy and other sectors to create and develop new connections. We design an artifact considering user activities. We follow the DSR scheme as Gregor et al. (2020) described to formulate DRs and DPs. We modify this scheme into 1) DP, 2) context, 3) mechanism, and 4) goal. This modification facilitates clarity, evaluation, and portability and enables clear reading of DRs and DPs (Schaffer et al. 2020). Before creating the DRs and deriving the DPs, we form KTs based on the classification of relevant literature through text mining and topic modeling. Our DSR steps are shown as numbers in Figure 1. In the first step, we define the problem of XAI in the energy sector and research needs in the relevance cycle. The relevance cycle describes the weighting of relevance from DSR, which initiated an application context. This context sets the requirements and describes acceptance criteria for evaluating the results, which must be traceable to the environment (Hevner and Chatterjee 2010). Our DSR approach is expanded by a Mixed Methods approach based on Venkatesh et al. (2013; 2016). We form our knowledge base through a literature review with exploratory text analysis in Step 2.1 and Step 2.3 and confirmatory qualitative expert interviews in Step 2.2. A Mixed Methods approach integrates quantitative and qualitative research to capture different aspects of the research subject. Complementarity is achieved quantitatively by creating patterns and trends and qualitatively through deep insights and contexts. We conduct sequential data collection. At first, we collected a literature dataset as described in Step 2.1. Then, we interviewed experts in Step 2.2. We increase the validity and reliability of our results through the resulting triangulation and create a holistic understanding of the research subject through theoretical integration (Venkatesh et al. 2013). We can develop a theory through the Mixed Methods approach, make stronger conclusions, and offer a wider selection of complementary views (Venkatesh et al. 2013; 2016). Step 2 relates to the rigor cycle, which provides the research knowledge and requires research and references. This cycle comprises the use of knowledge in terms of models and methods; thus, the applicability and generalizability of an artifact can be assessed (Hevner and Chatterjee 2010). In step 3, we create our DRs and derive our DPs in the design cycle, which describes the iterative process of developing and validating artifacts (Hevner and Chatterjee 2010). We adjust our results in the relevance and design cycle (step 4).

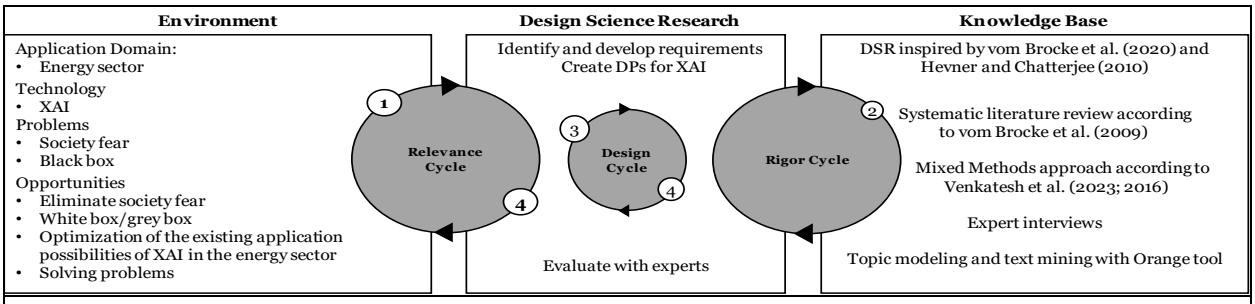

Figure 1. Research Design

Step 1: Problem Formulation

While DL models usually pursue higher accuracy with lower interpretability, XAI models aim to balance accuracy and interpretability (Adadi and Berrada 2018; Angelov et al. 2021; Arrieta et al. 2019). Systems with DL and without explainability implementation are considered black box problems. The black box problem is mainly opacity, where people neither trust nor give control to machines, and developers have fewer opportunities to intervene in a system (Zednik 2021). There are limited options to ensure data is

processed fairly and securely and whether users have rights under the general data protection regulation. This encounters dissatisfaction and fear regarding technological progress in society and inhibitions in the implementation of services due to the increased consequences (Arrieta et al. 2019; Zednik 2021). In the energy sector, accuracy in processing big data is important (Kuzlu et al. 2020). As renewable energies expand, inherent fluctuations increase, so ensuring the security of supply and grid stability are challenges (Fridgen et al. 2022). Automated control operations are required for locally expanding energy sources like photovoltaic systems. Control operations are necessary to ensure a reliable and efficient power supply. The digitization of the power supply chain is accelerating to provide control automation infrastructure. AI supports this automation to process big data heterogeneously and with high quality (Kuzlu et al. 2020; Richter et al. 2022). This leverages complex pattern recognition as a capability of AI and data analysis and interpretation (Richter et al. 2022). According to Hevner and Chatterjee (2010), DSR addresses a "wicked problem," i.e., a complex problem involving inherent flexibility, unstable requirements, and human or social capabilities. We identify XAI in the energy sector as a wicked problem due to the multidisciplinary nature of XAI, which requires an understanding of AI and energy processes, uncertain requirements, conflict of interests through different stakeholders, unpredictable impacts, and dynamic environments subjecting to constant changes by technological innovations, market fluctuations, policy decisions, or environmental changes (e.g., Arrieta et al. 2019; Kuzlu et al. 2020; Richter et al. 2022; Zednik 2021).

Step 2.1: Gathering Knowledge from Scientific Literature

A literature review is essential to understand the topic comprehensively. We follow the literature search and analysis as described by vom Brocke et al. (2009). This step addresses the quantitative research in our Mixed Methods approach, according to Venkatesh et al. (2013; 2016). We conduct a keyword-based database search in the databases SpringerLink, Web of Science, Elsevier, AIS Electronic Library, IEEE Explore, and ACM for the following search strings: "Energy" AND "Energy Market" AND "User Orient Costumer" AND ("XAI" OR "explainable AI" OR "Explainable artificial intelligence") in the period from January 2017 to May 2023. After reviewing titles and abstracts, 170 papers have been selected. We screened the full text and excluded 22 papers. In line with Watson and Webster (2020) and Webster and Watson (2002), we add 31 papers in backward, forward, author, and Google Scholar similarity searches. Finally, we included 179 papers in our final data set, which was used for our analyses with the text mining tool. Through this search, we include publications from other sectors (e.g., financial sector) and general XAI papers. This allows us to identify some DRs or DPs in our analysis, which we can adapt to the energy sector. The literature selection depended on the paper's contribution, quality of the outlet, e.g., white papers are not included, impact factors are considered, the novelty of research results, and citations, e.g., on Google Scholar.

Step 2.2: Gathering Knowledge from Experts

After the extensive literature review, we interviewed four experts, see Table 1, to enhance our knowledge. This step reflects the qualitative research in the Mixed Methods approach, according to Venkatesh et al. (2013; 2016) through the confirmatory interviews. The experts' requirements and principles were supported by scientific literature (see online appendix B) and provided insight into the practical application of XAI in the energy sector. This allowed us to take a broader view outside of theory through the scientific literature. Less argued statements from the literature could be highlighted through the expert opinions and thus included from DR or DP. The experts were interviewed online for between 30 and 60 minutes. The selection of experts allowed a wider range of perspectives and expertise on the used organizational AI solutions.

E	Expert Profile
1	Manager of the information security test center at an IT services provider.
2	Engineer for AI solutions at an organization that analyzes, forecasts, and optimizes energy data such as electricity.
3	Organization's chairman that provides XAI analytics tools to maximize the performance of photovoltaic arrays.
4	Specialist for data transmission between different energy plants.

Table 1. Part 1 of Interviewed Experts (E)

Step 2.3: Generate Key Topics and Identify Requirements

We analyze the clusters of topics by hierarchical clustering based on their keywords with the Python-based text mining tool Orange, according to Demsar et al. (2013). According to Venkatesh et al. (2013; 2016), this

is the last step of our Mixed Methods approach through the exploratory text analysis. Our selection of the top-down process allows us to identify the clusters from the data set and interpret them as KTs. KTs represent the fundamental areas of focus and exploration within a particular research context. They summarize the problem domain and efforts to create artifacts to address specific challenges (Gerlach et al. 2022b). We adapted the text mining and topic modeling approach as used by Gerlach et al. (2022b), see Figure 2. First, we cleaned the data, e.g., deleting title and abstract. In step 2, we prepared the data, e.g., deleting punctuation, and then created a keyword list with the most irrelevant words, e.g., the, therefore. Then, we created a word cloud to identify the most used words. In the fourth step, we use hierarchical clustering to transform the data set. In this process, we identified 8 clusters. The hierarchical clustering is illustrated in online appendix C. In the fourth step, we applied the topic modeling approach according to Tong and Zhang (2016) and thus formed 8 data sets, which can be identified as KTs (Gerlach et al. 2022b).

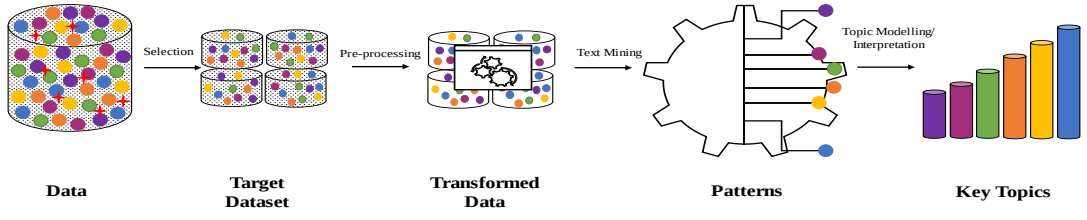

Figure 2. Text Mining and Topic Modeling adopted from Gerlach et al. (2022b)

The clusters were named on the requirements from the publications and the keywords from the text mining process. The first cluster contains all publications of the first data set. The ensuing clusters follow this structure. Figure 3 illustrates the ten most relevant words from each cluster on the vertical axis. The horizontal axis highlights the ratio of words, which varies due to the different distribution of publications.

Figure 3. Frequency of Words Charts inside the Clusters after Topic Modeling

Step 3.1: Derive Design Requirements

New data sets were created, dividing the literature into clusters. We analyzed the data sets for existing and possible requirements. These requirements were sorted based on relevance, e.g., multiple nouns, and then included as a DR. The references listed in Table 2 are examples. According to Watson and Webster (2020) and Webster and Watson (2002), the entire concept matrix is provided in the online appendix D.

DRs	KTs	Description	References	E
DR1	**KT1**	expects XAI to conform to quality standards. These quality standards must be transparent, traceable, understandable, and explainable.	Jaigirdar et al. (2020)	1-4
DR2	**KT1; KT8**	requires the integration of XAI users throughout their system. People must be able to directly influence a system's decisions for weak and strong AI.	Ehsan et al. (2022); Garcia-Magarino et al. (2019)	3
DR3	**KT1**	calls for knowledge creation, retrieval, storage, sharing, and application. The potential of AI is seen in knowledge management in predictive accuracy and calculation of sales probabilities.	Holzinger et al. (2022); Jarrahi et al. (2022)	4
DR4	**KT2; KT5; KT6**	calls to respect human intelligence and to act intelligently, e.g., through neuro-symbolic methods such as explainable neural-symbolic learning, addressing ethical, legal, and societal considerations of weak and strong (X)AI systems.	Akata et al. (2020); Holzinger et al. (2022)	
DR5	**KT1-8**	deals with explainability, transparency, interpretability, effectiveness, trustworthiness, and understanding.	Adadi and Berrada (2018); Angelov et al. (2021)	1-4
DR6	**KT2; KT7**	requires monitoring and control by XAI systems. Monitoring is significant in the energy sector in detecting and locating faults, where XAI can make energy demand/supply forecasts, identify imbalances in power grids, and detect cyber-attacks.	Machlev et al. (2022a); Richter et al. (2022); Wanner et al. (2019)	2
DR7	**KT4**	involves the incorporation of knowledge from cognitive science. It is about implementing a psychological element into XAI.	Byrne (2019); Holzinger (2018)	
DR8	**KT1; KT7**	calls for transparent risk management of XAI systems to make forecasts, safety assessments, and errors comprehensible.	Kruse et al. (2021b); Meske et al. (2022)	1, 4
DR9	**KT6; KT7; KT8**	requires flexible and reliable XAI applications. Flexibility is the ability to adapt to contexts, applications, and users. Reliability is characterized by the assessments' internal consistency and reproducibility.	Ferreyra et al. (2019); Fridgen et al. (2022)	4
DR10	**KT6; KT8**	describes the need for human-machine interfaces in XAI models to enable contextual understanding and what-if questions.	Gunning and Aha (2019); Holzinger et al. (2022)	2
DR11	**KT7; KT8**	requires XAI processing of big data. Maintaining and improving data quality is critical in avoiding introducing any bias or error.	Haag et al. (2022); Holzinger (2018)	1
DR12	**KT7; KT8**	demands resilience from XAI systems to cope with and defend against negative situations, e.g., cyber-attacks.	Richter et al. (2022)	4
DR13	**KT7; KT8**	calls to make forecast decisions through XAI models. E.g., time series data models are needed to make forecasts due to the increasing number of PV systems. Patterns and correlations can be identified from the data.	Kuzlu et al. (2020)	1, 2, 3
DR14	**KT8**	requires that XAI be contestable, i.e., humans can challenge system elements and relationships, verifiable, traceable, and predictable, i.e., to predict functional elements.	Verhagen et al. (2021)	4
DR15	**KT8**	complements **DR14** with controllability and steerability, i.e., humans must control functional system elements and steer the system's actions.	Verhagen et al. (2021)	

Table 2. Derived Design Requirements

Step 3.2: Establish Design Principles

We derive DPs based on the KTs and the DRs; see Table 3. All corresponding references for each DP are listed in the online appendix D.

DP	DRs	Description	References	E
DP1	**DR1; DR5; DR10**	considers the **6 W's** of provincial-based XAI. "Who" is involved in the system defines the responsibility for data creation, modification, and action. "What" refers to an instance to describe and create data in the AI system. "When" and	Jaigirdar et al. (2020)	

		"Where" represent activities that show the workflow in one layer and the design and implementation of changes in the other. "What" creates a system overview and facilitates risk assessment. "Why" addresses legal and political explainability and protects against black box design.		
DP2	**DR2; DR5**	deals with the realization of **user requirements** through visualization and integration. Explainable and interpretable models are created to increase user confidence in systems. The explanation depends on user requirements, needs, and characteristics identified through interviews, questionnaires, or scenarios. This can be called the desiderata of the stakeholders.	Ehsan et al. (2022); Ding (2018)	3
DP3	**DR3; DR12**	is concerned with supporting **knowledge management**, which is limited to the management of knowledge in organizations or AI as a branch of information systems with a focus on systems development. Knowledge management connects workers with the optimal number of knowledge resources at the right time to achieve the best possible outcome.	Ding (2018); Jarrahi et al. (2022)	4
DP4	**DR4; DR12**	Describes Hybrid Intelligence (**HI**) as a complement to human intellect. HI refers to the synergistic and proactive collaboration of humans and machines to achieve common goals. This combination compensates for human weaknesses and augments human capabilities so that actions can be taken and decisions made. Through interaction, AI can be used in society as it considers social, ethical, and legal issues. Most HI models are based on classification techniques that can be used for anomaly detection, energy storage system efficiency, and accurate forecast probabilities.	Akata et al. (2020); Nourani et al. (2021)	3
DP5	**DR5; DR13**	calls for **t-way combinations** to explain XAI applications' conclusions. The categories "class membership" and "non-class membership" are formed. The more feature combinations (t-values) are used, the more accurately they can be determined. This is useful in classification problems where interactions responsible for an error are sought. The t-way combinations can be viewed as a decision tree, with the tree nodes representing the t-way combinations.	Kuhn et al. (2020)	
DP6	**DR5; DR6; DR13**	requires the **redesign** or augmentation of (X)AI applications. RL allows frequency and control systems to make algorithms more understandable, find control strategies for distributed power systems, and avoid frequency and voltage instabilities. RL can be used in power planning applications, e.g., in dynamic pricing (price spikes), and to provide information about the composition of the electricity price. Smart meters can be used to estimate electricity consumption to increase the confidence of energy consumers.	Machlev et al. (2022a); Richter et al. (2022); Wanner et al. (2019)	1, 2
DP7	**DR2; DR4-7; DR10**	requires learning perceptions, skills, and **AI practices** by humans and humans by AI. Feedback interaction models allow systems to interact with humans visually, acoustically, or tactically. One example is tangible user interfaces that merge the virtual and real worlds to understand and manipulate computer systems. This allows users to participate in decision-making.	Akata et al. (2020); Ding (2018)	2-4
DP8	**DR5**	recommends the adoption of **rigorous logic-based methods**. Bias can lead to erroneous decisions. Rigorous methods can compute trustworthy explanations and validate heuristically computed explanations. Rigorous and logic-based explanations are correct for the entire feature space.	Ignatiev (2020)	
DP9	**DR5; DR7**	requires **counterfactual data and models** that provide human reasoning. The data and models are based on characteristic values and are comparative references. They provide insights into "what-if" considerations. Additively, they can go beyond the given information and offer creative solutions to problems. Creating an alternative to reality allows a better understanding of conclusions and improvements in incomplete data sets. Using counterfactual data to create a parallel space to reality, adversarial attack systems can be integrated to confuse attackers and prevent data theft.	Byrne (2019); Chen et al. (2019b)	
DP10	**DR5; DR11**	describes the method of **twinning** to achieve explainability. Comparing a non-transparent black box and a transparent white box, feature weights from the black box can be mapped into the white box, providing more transparency and explainability. It describes a hybrid system with separate modules but identical data sets and a two-part task division.	Kenny and Keane (2021)	
DP11	**DR6; DR13**	calls for **real-time decision support** of XAI systems. To achieve acceptance and human trust in AI, the system must be considered as an intelligent machine colleague that can point out automated anomalies as a comprehensible set of rules. This way, monitoring is secure, transmission is fail-safe, and reliability and quality are a given. Combined with multi-criteria	Kaya et al. (2019); Wanner et al. (2019)	1, 3, 4

		decision-making (MCDM) methods, the best possible proposals can be selected, the fuzziness in the decision-making processes can be reduced, and concrete results can be obtained. Fuzzy MCDM methods can evaluate energy sources, investments, or power plant locations in the energy sector.		
DP12	**DR5; DR6; DR14**	requires **operational data** with new levels of detail. XAI is a method that improves quantifiability, predictive accuracy, and explanation in operation and stability.	Kruse et al. (2021b)	4
DP13	**DR6; DR8; DR11**	describes the application of **context-adaptive methods** on the part of XAI. Systems can reconstruct contextual explanatory models of real-world phenomena.	Fridgen et al. (2022); Holzinger (2018)	
DP14	**DR8; DR13**	requires the **implementation** of XAI methods and algorithms. Combining generation, storage, and energy consumption, SGs improve the ability of a system's forecast decisions. The implementation increases transparency and clarity in security. The most mentioned algorithms are SHAP and LIME. To support PV forecasts, each SHAP value provides information about a feature's contribution to PV performance. The local accuracy, consistency, and results allow complete forecasts. In PV forecast systems, LIME can achieve individual forecasts of PV forecast performance with each feature.	Górski and Ramakrishna (2021); Gunning and Aha (2019)	1-4
DP15	**DR2; DR5; DR9**	calls to increase system **flexibility** through DSM to respond to inherent variability. DSM enables short-term changes in energy consumption behavior and includes all actions related to the consumer side of the energy system. DSM is one way to adjust flexibility in the energy sector and is particularly responsive to external signals, such as price signals.	Ferreyra et al. (2019); Fridgen et al. (2022)	3, 4
DP16	**DR5; DR14**	requires increasing system **explainability**, citing examples of actions based on state information and explaining decisions. System explainability must be understood to clarify revealed system elements, provide information about causality and establish relationships with other system elements.	Emmert-Streib et al. (2020); Verhagen et al. (2021)	4
DP17	**DR2; DR5; DR15**	requires system **transparency** to provide explanatory answers intended to convey knowledge of the system elements. Disclosing the external and functional system elements, the user can access, analyze, and use this information.	Adadi and Berrada (2018); Verhagen et al. (2021)	

Table 3. Deducted Design Principles

Step 4: Evaluation through Experts and Adjustment

A key DSR approach component is evaluating the understandability, transferability, and usefulness of our KTs, DRs, and DPs (Gregor et al. 2020; Hevner and Gregor 2013). We contacted three experts from organizations that market XAI services in the energy sector for evaluation. The experts were interviewed in written form. The following Table 4 summarizes key information about the surveyed experts.

E	Expert Profile
5	Vice president (data science and data engineering) for an organization that develops data science products for XAI applications in, among others, energy infrastructure.
6	Entrepreneur and AI scientist of an organization specializing in 3D image recognition and search platforms based on explainable cognitive AI, e.g., analyzing weather data and visualizing PV plant locations.
7	CEO of an organization that provides a proprietary XAI platform that enables companies to create fully XAI-compliant, accurate, understandable, and trustworthy models.

Table 4. Part 2 of Interviewed Experts

The previous Figure (online appendix B) containing the KTs, DRs, and DPs and their derivation and development procedures were sent to the experts and explained. Based on the available information, the experts were asked to assess our results and findings' understandability, transferability, and usefulness, checking for errors or incompleteness. New links through the evaluation were marked in the final Figure (online appendix E) as a blue arrow, and changed or added DRs or DPs were shown in dark gray with the signature "Evaluation." The traceability of **DP9** as the creation of counterfactual data and models can be improved and supplemented with the example of an adversarial attack system (Expert 5) and **DP11** as the implementation of real-time decision systems involving MCDM (Expert 5, 6). In addition, connections were made between **DR9** and **DP14** (Expert 7) and **DR13** and **DP11** (Expert 5). Regarding errors, it has been

noted that post-hoc methods are too unstable and lead to large changes in the explanation for small changes in continuous variables (Expert 7). Considering the model-dependent and model-independent methods, **DP18** can thus be added as a guideline to use **interpretable models** such as decision trees. **DP18** was created from the requirements of **DR9** and **DR13**. The advantage of an interpretable model consists of decisions being understandable and comprehensible for users (Expert 5). An example of interpretable and comprehensible decision trees is shown by Gerlach et al. (2022a) in their paper for efficient XAI services.

Three-Forces Model for Decision Support

We examined our DPs in more detail by categorizing the DPs' publications by topics to deduce dimensions to implement AI solutions. During the subjective clustering process, we were able to identify key interrelationships. First, the DPs deal with explainability, traceability, and certifiability, e.g., **DP10**. In connection with explainability, traceability, and certifiability, connections can also be made with ethics and privacy, e.g., **DP9** or efficiency (**DP18**). Efficiency refers not only to efficient energy but also to an efficient implementation of AI solutions. Thus, we considered valid and justifiable reasoning to use XAI applications in power systems since many XAI methods and algorithms are more complicated, costly, and provide more incomprehensible results than conventional methods. Based on our DRs and DPs in online appendix E, we can summarize the three essential dimensions to successfully implement an AI application as a Three-Forces Model. Our Three-Forces Model represents the three essential dimensions of an AI to integrate it successfully into an organization. Our model applies to other sectors and is illustrated in Figure 4 based on our DPs. The corners of the triangle illustrate the dimensions. Depending on the level at which the DPs are distributed in the model, the more influence the corresponding dimension has. **DP1** has an equal influence from all three dimensions, while **DP10** is almost exclusively associated with the dimensions of explainability, traceability, and certifiability. **DP16,** despite the results and findings from an XAI study, is least related to explainability, traceability, and certifiability. **DP9** is mainly in the dimensions of ethics and privacy as well as explainability, traceability, and certifiability. In our Three-Forces Model, there are interactions between the dimensions. The dimensions have different impacts and emphases, so parts of the DPs impact a dimension more than others DPs. In addition, the model highlighted that although a DP can be explainable, traceable, certifiable and efficient, it is not ethical and private (e.g., **DP13**). This enables an assignment of DPs and an evaluation of the DPs regarding the corresponding target.

Explainability, traceability, and certifiability: We have clarified traceability in connection with explainability; see **DR1**. Certifiability is primarily about transparency and quality assurance. Certified transparency increases trust in the systems, disclosing the data sources used and how the AI works. This also impacts the ethical and legal aspects (Byrne 2019). Quality assurance benefits from certification, ensuring requirements and reliability. In addition, certifiability ensures comparability with other AI solutions regarding the same standards and compliance with laws and regulations (Jaigirdar et al. 2020). Possibilities for certification are, e.g., the ISO, as Expert 1 explains: "Often they have certification according to ISO 27001. That is also mostly the certification of information security management, which is common. The major certification's time cycle is three years but is usually checked annually. From my point of view, this should also be the case, at least for the things that have changed. Moreover, our customers think that is good, too." Expert 3 commented on the certification according to ISO 27001: "To be certified according to international standards is the norm, but it is associated with an incredible number of requirements. You have to do a lot to get this certification. However, these companies are more competitive."

Ethics and Privacy: Society expresses concerns about the security of systems, energy stakeholders, and governments. This raises acceptance issues for the use of AI systems by society. Explainability is an essential component of ethics, while ethics is not necessarily fundamental for explainability (Arrieta et al. 2019). Significant issues rise to the forefront, particularly energy and transitional equity. The benefits of energy conservation opportunities are widely known, but collecting household privacy data by smart meters can reveal household lifestyles. Therefore, the extent to which the monitoring of electricity consumption data constitutes household surveillance needs to be clarified (Hatzakis et al. 2019). Gray (2018) stated that customer disclosure and sharing of personal data is uncomfortable for customers due to the lack of trust and transparency, inhibiting the acceptance of AI adoption in the energy sector society. In addition, there are concerns about cybercrime, as the energy sector is a part of critical infrastructure and poses a security risk due to its dependence. People's well-being is at risk in the event of a successful attack and system integrity is compromised. Thus, concerns are raised about the privacy of personal data. Data protection

concerns energy consumption data, so privacy is essential (Hatzakis et al. 2019). However, governments must set standards and guidelines so organizations can select and implement appropriate measures (Expert 1). Energy security is also considered ethically questionable, particularly threatened by interruptions in energy transmission. Concerns about the data transmission in SGs are posed concerning the algorithmic biases in power distribution, raising ethical and legal issues. Applying AI systems in the energy sector to promote climate change "may harm the environment and disadvantaged populations without sustainable energy sources and appropriate ethical oversight" (Kelly 2022, p. 9). A key factor in gaining acceptance are regulations by the government, transparent privacy policies by energy companies, certifications and education, and transparency in AI systems for energy customers (Hatzakis et al. 2019). Expert 1 explained, "Certifications are an important factor. The company must implement certain security standards, such as basic protection or ISO." Regarding data privacy, research already exists built on transparency and presented in an understandable way (e.g., Bakker et al. 2022; Gerlach et al. 2022b).

Efficiency: AI solutions can lead to an efficient SG (Antonopoulos et al. 2020) and be used in achieving such pollution control of SDGs (Richter et al. 2022). In connection with SGs, cost-effective energy systems are formed (Ponnusamy et al. 2021), which can increase the economic efficiency and efficiency of power grids, thus leading to efficient transmission and management networks, improving grid security (Ponnusamy et al. 2021). Expert 2 stated, regarding ever-newly integrated AI solutions in enterprises, "That is the AI problem – you need to know how to construct the solution to your problems." Expert 2 notes that using AI systems leads to more problems and solutions, especially since people think AI solves all problems. Integrating stakeholders with consumer needs will increase efficiency in AI systems (Ehsan et al. 2022; Garcia-Magarino et al. 2019). AI solutions in modernized energy grids need regulations to protect personal data through smart meters and user privacy (Wilcox et al. 2019; Expert 4, 6). Only smart deployment through optimal conditions of AI solutions and XAI applications can ensure efficiency that benefits a company and society (Ponnusamy et al. 2021). Expert 3 said about this, "[...] And we do use, what we call hybrid approach [...]. We combine AI but also terminist and physical models. We are mixing those two because [...] we do not need just AI models. It is not that efficient."

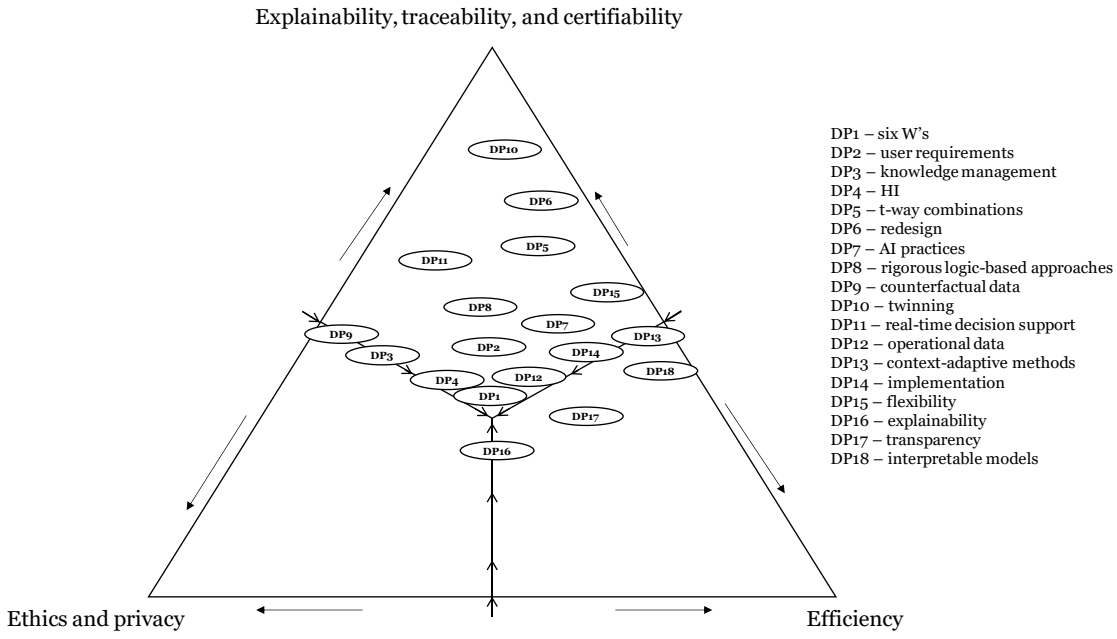

Figure 4. Three-Forces Model

Discussion, Limitations, Implications, Recommendations, and Further Research

Interpretability and explainability are two characteristics of XAI (Adadi and Berrada 2018). Interpretability is important to develop a coherent AI system that provides explainability. Thus, explainability is described

by contestability, predictability, verifiability, and traceability. Interpretability is achieved through transparency, which is provided by controllability and tractability (Verhagen et al. 2021). Expert 6 stated, "I believe that the only meaningful and reliable explanations are in inherently interpretable models, i.e., where the explanations are part of the model." A problem with XAI applications is the accuracy of systems. DL models are less integrated due to their non-transparent nature (Arrieta et al. 2019). The lack of DL models leads to lower accuracy. Accuracy is important for forecasting, decision-making, and management of high-performance systems (Arrieta et al. 2019; Rawal et al. 2021). The stakeholder cannot interpret the DL model and its decisions. Therefore, there is a trade-off between interpretability, i.e., simplicity of information about internal operations, and accuracy, i.e., a complete description (Arrieta et al. 2019). Other researchers, such as Rudin (2019), call for using interpretable models instead of explainable models to eliminate those trade-offs. Researchers have different statements and a need for more research and solutions. High performance combined with explanations for users is required to build trust in AI systems. Black box disclosure leads to debugging and traceability of the AI system and increases security vulnerabilities, allowing attackers to penetrate systems faster and more efficiently to manipulate systems, cause damage, steal data, or cause failures (Ehsan et al. 2022). **DR8** is a contradiction that is justified and necessary. A robust system is individually designed for its application domain, which is not transferable to another domain. Expert 2 highlighted, "We cannot desire that everything becomes transparent and explainable, but we as a company remain completely secretive. We must also become transparent in our security, meaning we all need standardized frameworks. […] We need standards to build the system around these standards. That is the only way we can become robust." In addition, it is important to alleviate user fear and build trust in XAI applications, as stated in **DR14**. Nourani et al. (2021) integrate an intelligent system with an exploratory task to test the impact of observations related to the system's weaknesses and strengths on the user. They verify the improvement of the explanations for the users. Weaknesses of the system in terms of low model accuracy and strengths due to high model accuracy were integrated into the experiment. As a result, the forecasts led to unintended biases in the users' mental models. Users had a better mental model of strengths due to a positive first impression of the system, although they made errors because they overestimated model accuracy and were more prone to automation errors. Users who learned about model errors early on and gained a negative impression were generally more skeptical, made fewer errors, and underestimated the model's recognition accuracy (Nourani et al. 2021; Rawal et al. 2021). User skepticism of the application and critical questioning of decisions, forecasts, or possibilities is necessary. Expert 5 noted that the name of **KT6** is confusing. The clustering results highlight the methods, models, and algorithms as DL models, which is why we include the DL model in the description of **KT6**. Expert 1 clarified the issues in compliance with legal obligations concerning AI's progress: "AI will dispatch and feed into the electricity market, etc. […] The question is how operators can then still follow their legal obligations. That will be a big challenge." In addition, Expert 4 adds that the constant connection to the Internet is to be assessed as a danger: "In the energy sector, many information systems are connected via the Internet. It is simply necessary to allow energy production plants and energy consumers to exchange data with each other. In this respect, we need special mechanisms to protect critical infrastructures connected via the Internet that cannot function without the Internet. We should take nuclear power plants as an example. They are completely disconnected from the Internet; they are more secure."

We contribute to XAI literature by providing design knowledge to upgrade explainability in the black box nature of the energy sector's AI systems. By synthesizing the fragmented XAI knowledge base, we developed a design guideline for XAI upgrading that can also be seen as a checklist in the form of the DRs. Our derived KTs highlight the hot topics in XAI research and indicate which subfields are demanded. In addition, we pave the way for further research by developing a research agenda with direct RQs. We also contribute to the DSR and data analysis methodology by applying a novel approach for deriving KTs based on a literature review proposed by Gerlach et al. (2022a). The applied text mining and topic modeling approach enables objective structuring of relevant literature through a low-code analysis tool (e.g., Orange). For practice, we support energy companies in developing, customizing, and purchasing XAI systems or upgrading the explainability of AI systems. The purchase of software can be supported by specifying requirements and principles for XAI design, while developers can target these requirements and principles. We also applied requirements and principles from other sectors to the energy sector, driving the energy transition. With our requirements, we have created added value for XAI-implementing energy companies. We discovered that 1) explainability, traceability, and certifiability, 2) ethics and privacy, and 3) efficiency are the most important dimensions for successfully implementing explainability in AI systems. Practitioners can rank their results for AI systems based on these dimensions, thus giving society more explainability. In addition,

we developed a Three-Forces Model that can innovatively simplify the development and implementation of AI in companies and drive technological advances. Researchers can evaluate and adjust their studies through the Three-Forces Model by assigning solutions to our three dimensions. Practitioners can decide "what" constitutes their AI system. The model is not only related to the energy sector. We created the model based on our DPs and looked at the interrelationships. Our Three-Forces Model can be applied to other sectors to enable the successful implementation of AI solutions. The Three-Forces Model supports assigning and categorizing DPs and DRs or general guidelines that an AI should possess. Thus, a direct and conclusive statement about the dimensions of AI is made. The three dimensions form the foundation for an AI. Principles, guidelines, and requirements for an AI can be divided into three dimensions by the Three-Forces Model, regardless of the sector or the application field of the AI. Thus, any connections and emphases between the principles, guidelines, and requirements will be visible. We have integrated our DPs into the Three-Forces Model for an application check. Our results demonstrate that our **DP1** is the most attended by the three dimensions and can be considered as the focus of all three dimensions when integrating AI in the energy sector without a specific application case. Thus, we have illustrated that the Three-Forces Model is useful and applicable.

Further research can validate our Three-Forces Model and use it for usefulness and applicability to other DPs or sectors. The energy sector-specific DPs do not allow for immediate generalization to other sectors. However, the DPs can be adapted. This leads to the further RQ: ***1)*** *"Which design principles can be transferred to other sectors and thus generalized?"* Another limitation is the small number of expert interviews for the knowledge base (Expert 1, 2, 3, 4) and the adjustment (Expert 5, 6, 7). More expert interviews can generate a broader view of the field. There are also missing interviews of other stakeholders like users, and another further RQ is: ***2)*** *"How do experts from different areas and organizations view the predominantly generic and theoretical design principles?"* It is possible to develop critical success factors and maturity models for companies from the DPs above to support companies developing and implementing XAI methods and algorithms; further RQ: ***3)*** *"What critical success factors for developing and implementing XAI methods and algorithms can be deduced?"* We noticed that DPs in the energy sector are associated with ethical and legal requirements. A maturity model can be used to measure and evaluate the ethical and legal readiness for XAI systems in the energy sector. A further RQ: ***4)*** *"How can DPs for XAI in the energy sector be consistent with ethical and legal requirements, and how can a maturity model be used to ensure that XAI systems operate in compliance?"* Our Three-Forces Model proposes a new model to support the implementation of AI and XAI solutions in the energy sector. The next step is to investigate the application of the model in practice and other sectors to prove the generalization of our Three-Forces Model.

Conclusions

To address the research needs and our RQs, we follow the DSR approach according to vom Brocke et al. (2020) and Hevner and Chatterjee (2010), including a literature review and expert surveys. To address RQ1, we analyzed 179 publications on XAI applications through text mining and topic modeling and identified 8 KTs. 15 DRs were derived based on the KTs. To address RQ2, we deduced 18 DPs. Additional expert interviews evaluated our design artifact (KTs, DRs, DPs). Based on our results and findings, we designed the Three-Forces Model to support the integration of AI Solutions, which includes 1) explainability, traceability, and certifiability, 2) ethics and policy, and 3) efficiency. We could think beyond the DSR approach and think ahead by developing the Three-Forces Model. We identify four more research gaps based on our results so other researchers are allowed to address our further RQs and develop further research so the science can address our wicked problems called XAI in the energy sector.

Acknowledgements

The research project “SiNED—Systemdienstleistungen für sichere Stromnetze in Zeiten fortschreitender Energiewende und digitaler Transformation” acknowledges the support of the Lower Saxony Ministry of Science and Culture through the 'Niedersächsisches Vorab' grant programme (grant ZN3563).

References

Adadi, A., and Berrada, M. 2018. "Peeking Inside the Black Box: A Survey on Explainable Artificial Intelligence (XAI)," *IEEE Access* (6), pp. 52138–52160.

Akata, Z., Balliet, D., Rijke, M., Dignum, F., Dignum, V., Eiben, G., Fokkens, A., et al. 2020. "A Research Agenda for Hybrid Intelligence: Augmenting Human Intellect with Collaborative, Adaptive, Responsible, and Explainable Artificial Intelligence," *Computer* (53:8), pp. 18-28.

Alova, G., Trotter, P. A., and Money, A. 2021. "A Machine-Learning Approach to Predicting Africa's Electricity Mix Based on Planned Power Plants and Their Chances of Success," *Nature Energy* (6:2), pp. 158-166.

Angelov, P. P., Soares, E. A., Jiang, R., Arnold, N. I., and Atkinson, P. M. 2021. "Explainable Artificial Intelligence: An Analytical Review," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* (11:5), # e1424.

Antonopoulos, I., Robu, V., Couraud, B., Kirli, D., Norbu, S., Kiprakis, A., Flynn, D., Elizondo-Gonzalez, S., and Wattam, S. 2020. "Artificial Intelligence and Machine Learning Approaches to Energy Demand-Side Response: A Systematic Review," *Renewable and Sustainable Energy Reviews* (130), # 109899.

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A. et al. 2020. "Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges Toward Responsible AI," *Information Fusion* (58), pp. 82-115.

Bakkar M, Bogarra S, Córcoles F, Aboelhassan a, Wang S, Iglesias J. 2022. "Artificial Intelligence-Based Protection for Smart Grids," *Energies* (15:13), # 4933.

Benbya, H., Davenport, T. H., and Pachidi, S. 2020. "Artificial Intelligence in Organizations: Current State and Future Opportunities," *MIS Quarterly Executive* (19:4).

Byrne, R. M. J. 2019. "Counterfactuals in Explainable Artificial Intelligence (XAI): Evidence from Human Reasoning," in *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, Macao, China.

Chang, X., Li, W., Ma, J., Yang, T., and Zomaya, A. Y. 2020. "Interpretable Machine Learning in Sustainable Edge Computing: A Case Study of Short-Term Photovoltaic Power Output Prediction," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, Barcelona, Spain.

Chen, M., Liu, Q., Chen, S., Liu, Y., Zhang, C. H., and Liu, R. 2019a. "XGboost-Based Algorithm Interpretation and Application On Post-Fault Transient Stability Status Prediction of Power System," *IEEE Access* (7), pp. 13149-13158.

Chen, T., Liu, J., Xiang, Y., Niu, W., Tong, E., and Han, Z. 2019b. "Adversarial Attack and Defense in Reinforcement Learning-From AI Security View," *Cybersecurity* (2:1), pp. 1-22.

Cremer, J. L., Konstantelos, I., and Strbac, G. 2019. "From Optimization-Based Machine Learning to Interpretable Security Rules for Operation," *IEEE Transactions on Power Systems* (34:5), pp. 3826-3836.

Demsar, J., Curk, T., Erjavec, A., Gorup, C., Hocevar, T., Milutinovic, M., Mozina, M., et al. 2013. "Orange: Data Mining Toolbox in Python," *Journal of Machine Learning Research* (14), pp. 2349-2353.

Ding, L. 2018. "Human Knowledge in Constructing AI Systems - Neural Logic Networks Approach Towards an Explainable AI," *Procedia Computer Science* (126), pp. 1561-1570.

Ehsan, U., Wintersberger, P., Liao, Q. V., Watkins, E. A., Manger, C., Daumé III, H., Riener, A., et al. 2022. "Human-Centered Explainable AI (HCXAI): Beyond Opening the Black-Box of AI," in *Proceedings of the Conference on Human Factors in Computing Systems*, New Orleans, USA.

Ferreyra, E., Hagras, H., Kern, M., and Owusu, G. 2019. "Depicting Decision-Making: a Type-2 Fuzzy Logic Based Explainable Artificial Intelligence System for Goal-Driven Simulation in the Workforce Allocation Domain," in *Proceedings of the 28th IEEE International Conference on Fuzzy Systems*, New Orleans, USA.

Fridgen, G., Halbrügge, S., Körner, M. F., Michaelis, A., and Weibelzahl, M. 2022. "Artificial Intelligence in Energy Demand Response: A Taxonomy of Input Data Requirements," in *Proceedings of the International Conference on Wirtschaftsinformatik*, virtual.

Garcia-Magarino, I., Muttukrishnan, R., and Lloret, J. 2019. "Human-Centric AI for Trustworthy IoT Systems with Explainable Multilayer Perceptrons," *IEEE Access* (7), pp. 125562-125574.

Garoudja, E., Chouder, A., Kara, K., and Silvestre, S. 2017. "An Enhanced Machine Learning Based Approach for Failures Detection and Diagnosis of PV Systems," *Energy Conversion and Management* (151), pp. 496-513.

Gerlach, J., Hoppe, P., Jagels, S., Licker, L., and Breitner, M. H. 2022a. "Decision Support for Efficient XAI Services - a Morphological Analysis, Business Model Archetypes, and a Decision Tree," *Electronic Markets* (32), pp. 2139–2158.

Gerlach, J., Lier, S. K., Hoppe, P., and Breitner, M. H. 2023. "Critical Success Factors for AI-driven Smart Energy Services," in *Proceedings of the 29th Americans Conference on Information Systems*, Panama City, Panama.

Gerlach, J., Scheunert, A., and Breitner, M. H. 2022b. "Personal Data Protection Rules! Guidelines for Privacy-Friendly Smart Energy Services," in *Proceedings of the 30th European Conference on Information Systems*, Timisoara, Romania.

Górski, Ł., and Ramakrishna, S. 2021. "Explainable Artificial Intelligence, Lawyer's Perspective," in *Proceedings of the 18th ACM International Conference on Artificial Intelligence and Law*, São Paulo, Brazil.

Gray, J. 2019. "Pricing and Trust: A Utilities Conundrum - Utility Week," [online] Utility Week. Available at: https://utilityweek.co.uk/pricing-trust-utilities-conundrum/ [Accessed 04 May 2023].

Gregor, S., and Hevner, A. R. 2013. "Positioning and Presenting Design Science Research for Maximum Impact," *MIS Quarterly* (37:2), pp. 337-355.

Gregor, S., Kruse, L., and Seidel, S. 2020. "Research Perspectives: The Anatomy of a Design Principle," *Journal of the Association for Information Systems* (21:6), pp. 1622–1652.

Gunning, D., and Aha, D. 2019. "DARPA's Explainable Artificial Intelligence (XAI) Program," *AI Magazine* (40:2), pp. 44-58.

Haag, F., Hopf, K., Menelau Vasconcelos, P., and Staake, T. 2022. "Augmented Cross-Selling Through Explainable AI-A Case from Energy Retailing," in *Proceedings of the 30th European Conference on Information Systems*, Timisoara, Romania.

Han, T., Chen, J., Wang, L., Cai, Y., and Wang, C. 2019. "Interpretation of Stability Assessment Machine Learning Models Based on Shapley Value," in *Proceedings of the 3rd IEEE Conference on Energy Internet and Energy System Integration*, Taiyuan, China.

Hatzakis, T., Rodrigues, R., and David, W. 2019. "Smart Grids and Ethics," *ORBIT Journal*, (22).

Hevner, A. R., and Chatterjee, S. 2010. "Design Science Research in Information Systems," in *Design Research in Information Systems*, Springer.

Holzinger, A. 2018. "From Machine Learning to Explainable AI," in *Proceedings of the 1st IEEE World Symposium on Digital Intelligence for Systems and Machines*, Kosice, Slovakia.

Holzinger, A., Saranti, A., Molnar, C., Biecek, P., and Samek, W. 2022. "Explainable AI Methods - a Brief Overview," in *Proceedings International Workshop on Extending Explainable AI Beyond Deep Models and Classifiers*.

Ignatiev, A. 2020. "Towards Trustable Explainable AI," in *Proceedings of the 29th International Joint Conference on Artificial Intelligence*, Stockholm, Sweden.

Jaigirdar, F. T., Rudolph, C., Oliver, G., Watts, D., and Bain, C. 2020. "What Information Is Required for Explainable AI?: A Provenance-Based Research Agenda and Future Challenges," in *Proceedings of the IEEE 6th International Conference on Collaboration and Internet Computing*, virtual.

Jarrahi, M. H., Askay, D., Eshraghi, A., and Smith, P. 2022. "Artificial Intelligence and Knowledge Management: A Partnership Between Human and AI," *Business Horizons* (66:1), pp. 87-99.

Kaya, İ., Çolak, M., and Terzi, F. 2019. "A Comprehensive Review of Fuzzy Multi Criteria Decision Making Methodologies for Energy Policy Making," *Energy Strategy Reviews* (24), pp. 207-228.

Kenny, E. M., and Keane, M. T. 2021. "Explaining Deep Learning Using Examples: Optimal Feature Weighting Methods for Twin Systems Using Post-Hoc, Explanation-By-Example in XAI," *Knowledge-Based Systems* (233), # 107530.

Kruse, J., Schafer, B., and Witthaut, D. 2021a. "Exploring Deterministic Frequency Deviations with Explainable AI," in *Proceedings of the IEEE International Conference on Communications, Control, and Computing Technologies for Smart Grids*, Singapore, Japan.

Kruse, J., Schäfer, B., and Witthaut, D. 2022. "Secondary Control Activation Analysed and Predicted with Explainable AI," *Electric Power Systems Research* (212), # 108489.

Kruse, J., Schäfer, B., and Witthaut, D. 2021b. "Revealing Drivers and Risks for Power Grid Frequency Stability with Explainable AI," *Patterns* (2:11), # 100365.

Kuhn, D. R., Kacker, R. N., Lei, Y., and Simos, D. E. 2020. "Combinatorial Methods for Explainable AI," in *Proceedings of the 13th IEEE International Conference on Software Testing, Verification and Validation Workshops*, Porto, Portugal.

Kuzlu, M., Cali, U., Sharma, V., and Guler, O. 2020. "Gaining Insight into Solar Photovoltaic Power Generation Forecasting Utilizing Explainable Artificial Intelligence Tools," *IEEE Access* (8), pp. 187814-187823.

Lee, A. S., Thomas, M., and Baskerville, R. L. 2015. "Going Back to Basics in Design Science: From the Information Technology Artifact to the Information Systems Artifact," *Information Systems Journal* (25), pp. 5-21.

Lee, Y. G., Oh, J. Y., and Kim, G. 2020. "Interpretation of Load Forecasting Using Explainable Artificial Intelligence Techniques," *The Transactions of the Korean Institute of Electrical Engineers* (69:3), pp. 480-485.

Lowry, P. B., Dinev, T., and Willison, R. 2017. "Why Security and Privacy Research Lies at the Centre of the Information Systems (IS) Artefact: Proposing a Bold Research Agenda," *European Journal of Information Systems* (26), pp. 546-563.

Loyola-Gonzalez, O. 2019. "Black-Box Vs. White-Box: Understanding Their Advantages and Weaknesses from a Practical Point of View," *IEEE Access* (7), pp. 154096-154113.

Machlev, R., Heistrene, L., Perl, M., Levy, K. Y., Belikov, J., Mannor, S., and Levron, Y. 2022a. "Explainable Artificial Intelligence (XAI) Techniques for Energy and Power Systems: Review, Challenges and Opportunities," *Energy and AI*, # 100169.

Machlev, R., Perl, M., Belikov, J., Levy, K. Y., and Levron, Y. 2022b. "Measuring Explainability and Trustworthiness of Power Quality Disturbances Classifiers Using XAI - Explainable Artificial Intelligence," *IEEE Transactions on Industrial Informatics* (18:8), pp. 5127-5137.

Meske, C., Bunde, E., Schneider, J., and Gersch, M. 2022. "Explainable Artificial Intelligence: Objectives, Stakeholders, and Future Research Opportunities," *Information Systems Management* (39:1), pp. 53-63.

Mitrentsis, G., and Lens, H. 2022. "An Interpretable Probabilistic Model for Short-Term Solar Power Forecasting Using Natural Gradient Boosting," *Applied Energy* (309), # 118473.

Nourani, M., Roy, C., Block, J. E., Honeycutt, D. R., Rahman, T., Ragan, E., and Gogate, V. 2021. "Anchoring Bias Affects Mental Model Formation and User Reliance in Explainable AI Systems," in *Proceedings of the 26th ACM International Conference on Intelligent User Interfaces*.

Ponnusamy, V. K., Kasinathan, P., Madurai E., R., Ramanathan, V., Anandan, R. K., Subramaniam, U., Ghosh, A., and Hossain, E. 2021. "A Comprehensive Review on Sustainable Aspects of Big Data Analytics for the Smart Grid," *Sustainability* (13:23), # 13322.

Rawal, A., McCoy, J., Rawat, D., Sadler, B., and Amant, R. 2021. "Recent Advances in Trustworthy Explainable Artificial Intelligence: Status, Challenges and Perspectives," *IEEE Transactions on Artificial Intelligence* (1:1), # 1.

Ren, C., Xu, Y., and Zhang, R. 2022. "An Interpretable Deep Learning Method for Power System Transient Stability Assessment Via Tree Regularization," *IEEE Transactions on Power Systems* (37:5), pp. 3359-3369.

Richter, L., Lehna, M., Marchand, S., Scholz, C., Dreher, A., Klaiber, S., and Lenk, S. 2022. "Artificial Intelligence for Electricity Supply Chain Automation," *Renewable and Sustainable Energy Reviews* (163), # 112459.

Ripalda, J. M., Buencuerpo, J., and García, I. 2018. "Solar Cell Designs by Maximizing Energy Production Based on Machine Learning Clustering of Spectral Variations," *Nature Communications* (9:1), # 5126.

Rudin, C. 2019. "Stop explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead," *Nature Machine Intelligence* (1), pp. 206-215.

Sahoo, S., Wang, H., and Blaabjerg, F. 2021. "On the Explainability of Black Box Data-Driven Controllers for Power Electronic Converters," in *Proceedings of the 13th IEEE Energy Conversion Congress and Exposition*.

Sairam, S., Seshadhri, S., Marafioti, G., Srinivasan, S., Mathisen, G., and Bekiroglu, K. 2022. "Edge-Based Explainable Fault Detection Systems for Photovoltaic Panels on Edge Nodes," *Renewable Energy* (185), pp. 1425-1440.

Sairam, S., Srinivasan, S., Marafioti, G., Subathra, B., Mathisen, G., Bekiroglu, K. 2020. "Explainable Incipient Fault Detection Systems for Photovoltaic Panels," *arXiv preprint arXiv:2011.09843*.

Sarp, S., Kuzlu, M., Cali, U., Elma, O., and Guler, O. 2021. "An Interpretable Solar Photovoltaic Power Generation Forecasting Approach Using an Explainable Artificial Intelligence Tool," in *Proceedings of the 12th IEEE Power and Energy Society Innovative Smart Grid Technologies Conference*, Washington, USA.

Schaffer, N., Stähler, O., and Weking, J. 2020. "Requirements and Design Principles for Business Model Tools," in *Proceedings of the 26th Americas Conference on Information Systems*. Salt Lake City, USA.

Schönhof, R., Werner, A., Elstner, J., Zopcsak, B., Awad, R., and Huber, M. 2021. "Feature Visualization Within an Automated Design Assessment Leveraging Explainable Artificial Intelligence Methods," *Procedia CIRP* (100), pp. 331-336.

Storey, V. C., Lukyanenko, R., Maass, W., and Parsons, J. 2022. "Explainable AI: Opening the Black Box or Pandora's Box?," *Communications of the ACM* (65:4), pp. 27–29.

Tong, Z., and Zhang, H. 2016. "A Text Mining Research Based on LDA Topic Modelling," in *Proceedings of the 6th International Conference on Computer Science, Engineering and Information Technology*, Warsaw, Poland.

Toubeau, J. F., Bottieau, J., Wang, Y., and Vallee, F. 2022. "Interpretable Probabilistic Forecasting of Imbalances in Renewable-Dominated Electricity Systems," *IEEE Transactions on Sustainable Energy* (13:2), pp. 1267-1277.

Venkatesh, V., Brown, S. A., Bala, H. 2013. "Bridging the Qualitative-Quantitative Divide: Guideline for Conducting Mixed Methods Research in Information Systems," *MIS Quarterly* (31:1), pp. 21-54.

Venkatesh, V., Brown, S. A., Sullivan, Y. W. 2016. "Guidelines for Conducting Mixed Methods Research: An Extension and Illustration," *Journal of the AIS* (17:7), pp. 435-495.

Verhagen, R. S., Neerincx, M. A., and Tielman, M. L. 2021. "A Two-Dimensional Explanation Framework to Classify AI As Incomprehensible, Interpretable, or Understandable," in *Proceedings of the International Workshop on Explainable, Transparent Autonomous Agents and Multi-Agent Systems*.

vom Brocke, J., Simons, A., Niehaves, B., Niehaves, B., Reimer, K., Plattfault, R., and Cleven, A. 2009. "Reconstructing the Giant: On the Importance of Rigour in Documenting the Literature Search Process," in *Proceedings of the 17th European Conference on Information Systems*, Verona, Italy.

vom Brocke, J., Winter, R., Hevner, A., and Maedche, A. 2020. "Special Issue Editorial–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), pp. 520-544.

Wang, Z., Tang, C., Sima, X., and Zhang, L. 2020. "Research On Ethical Issues of Artificial Intelligence Technology," in *Proceedings of the 2nd International Conference on Artificial Intelligence and Advanced Manufacture*, Manchester, United Kingdom.

Wanner, J., Herm, L. V., and Janiesch, C. 2019. "Countering the Fear of Black-Boxed AI in Maintenance: Towards a Smart Colleague," in *Proceedings of the Pre-ICIS SIGDSA Symposium on Inspiring Mindset for Innovation with Business Analytics and Data Science*, Munich, Germany.

Watson, R. T., and Webster, J. 2020. "Analysing the Past to Prepare for the Future: Writing a Literature Review a Roadmap for Release 2.0," *Journal of Decision Systems* (29:3), pp. 129-147.

Webster, J., and Watson, R. T. 2002. "Analyzing the Past to Prepare for the Future: Writing a Literature Review," *MIS Quarterly* (26:2), pp. xiii–xxiii.

Wilcox, T., Jin, N., Flach, P., Thumim, J. 2019. "A Big Data Platform for Smart Meter Data Analytics," *Computers in Industry* (105), pp. 250–259.

Wu, S., Le Z., Hu, W., Yu, R., and Liu, B. 2020. "Improved Deep Belief Network and Model Interpretation Method for Power System Transient Stability Assessment," *Journal of Modern Power Systems and Clean Energy* (8:1), pp. 27-37.

Zednik, C. 2021. "Solving the Black Box Problem: A Normative Framework for Explainable Artificial Intelligence," *Philosophy and Technology* (34:2), pp. 265–288.

Zhang, D., Li, C., Shahidehpour, M., Wu, Q., Zhou, B., Zhang, C., and Huang, W. 2022. "A Bi-Level Machine Learning Method for Fault Diagnosis of Oil-Immersed Transformers with Feature Explainability," in *Proceedings of the 4th International Journal of Electrical Power and Energy Systems* (134), # 107356, Hangzhou, China.

Zhang, K., Xu, P., and Zhang, J. 2020. "Explainable AI in Deep Reinforcement Learning Models: A SHAP Method Applied in Power System Emergency Control," in *Proceedings of the 4th IEEE Conference on Energy Internet and Energy System Integration*, Wuhan, China.

Zhang, K., Zhang, J., Xu, P. D., Gao, T., and Gao, D. W. 2021. "Explainable AI in Deep Reinforcement Learning Models for Power System Emergency Control," *IEEE Transactions on Computational Social Systems* (9:2), pp. 419-427.