AI explainability and governance in smart energy systems: A review

Roba Alsaigh1, Rashid Mehmood2* and Iyad Katib1

1Department of Computer Science, Faculty of Computing and Information Technology (FCIT), King Abdulaziz University, Jeddah, Saudi Arabia, 2High Performance Computing Center, King Abdulaziz University, Jeddah, Saudi Arabia

Traditional electrical power grids have long suffered from operational unreliability, instability, inflexibility, and inefficiency. Smart grids (or smart energy systems) continue to transform the energy sector with emerging technologies, renewable energy sources, and other trends. Artificial intelligence (AI) is being applied to smart energy systems to process massive and complex data in this sector and make smart and timely decisions. However, the lack of explainability and governability of AI is a major concern for stakeholders hindering a fast uptake of AI in the energy sector. This paper provides a review of AI explainability and governance in smart energy systems. We collect 3,568 relevant papers from the Scopus database, automatically discover 15 parameters or themes for AI governance in energy and elaborate the research landscape by reviewing over 150 papers and providing temporal progressions of the research. The methodology for discovering parameters or themes is based on “deep journalism,” our data-driven deep learning-based big data analytics approach to automatically discover and analyse cross-sectional multi-perspective information to enable better decision-making and develop better instruments for governance. The findings show that research on AI explainability in energy systems is segmented and narrowly focussed on a few AI traits and energy system problems. This paper deepens our knowledge of AI governance in energy and is expected to help governments, industry, academics, energy prosumers, and other stakeholders to understand the landscape of AI in the energy sector, leading to better design, operations, utilisation, and risk management of energy systems.

KEYWORDS
AI explainability, AI governance, smart energy systems, smart grid, AI trustworthiness, natural language processing (NLP), topic modelling, BERTopic

1 Introduction

Energy has fundamentally shaped the geopolitics of our world and transformed our lives in the last century (Vakulchuk et al., 2020; Blondeel et al., 2021). A look at the global past and current conflicts reveal that energy has been central to many of them involving oil, natural gas, battery minerals, among others. Energy availability enabled modern technological advancements including the ubiquity of computing and power (e.g., batteries), and transformed us into smart societies.

Traditional electrical power grids have long suffered from operational unreliability, instability, inflexibility, and inefficiency. Since power systems traditionally comprised large regional and national grids, monitoring the electrical systems of those grids and long distribution lines has been challenging causing many major electrical system failures, human lives, and hefty economic losses.

Smart grids continue to transform the energy sector with emerging technologies, renewable energy sources, decentralisation, decarbonisation, and others. We hereon will use the term “smart energy systems” for “smart grid” as a broader term that incorporates smart grids,
electrical power systems, and related business and other developments. These advancements offer many exciting opportunities such as the availability of solar, wind, hydro, and other forms of energy to organisations and homes. Development of microgrids (Hussain et al., 2019), mini-grids (Gill-Wiehl et al., 2022), community grids (Ceglia et al., 2020; Kong and Song, 2020), and supergrids (Zarauza de Rubens and Noel, 2019) have paved the way, alongside many other possibilities, for energy independence and energy trading between individuals, corporations, and nations. These smart energy systems are complex and produce massive data.

Artificial Intelligence (AI) presents an unimaginable potential for innovation, process optimisation, productivity, and other benefits in many sectors such as smart societies (Janbi et al., 2022), healthcare (Alahmari et al., 2022), education (Mehnoom et al., 2017), and transportation (Alomari et al., 2021). The energy sector is not an exception (Alkhayat et al., 2022). AI is being applied to smart energy systems to process massive and complex data in the energy sector and make smart and timely decisions. AI algorithms are black-box (Castelvecchi, 2016) needing interpretability and explainability (Doran et al., 2017; Goebel et al., 2018; Hagras, 2018; Hoffman et al., 2018; Lundberg et al., 2020) so that the decision made by AI could be explained to various stakeholders such as for regulatory and legal reasons. AI algorithms are usually imperfect or inaccurate. These AI algorithms are developed by human designers and developers trained using imperfect data and therefore they are likely to inherit bias and prejudice from them. The unregulated developments of AI have focussed on maximising efficiencies, and economic and other objectives rather than human values and priorities.

We adopt in this paper the definition of explainable AI (XAI) by NIST (National Institute of Standards and Technology) (Phillips et al., 2021) that proposes XAI systems to observe four principles, namely i) Explanation [“a system delivers or contains accompanying evidence or reason (s) for outputs and/or processes”]; ii) Meaningful [“a system provides explanations that are understandable to the intended consumer(s)”; iii) Explanation Accuracy (“an explanation correctly reflects the reason for generating the output and/or accurately reflects the system’s process”); and iv) Knowledge Limits (“a system only operates under conditions for which it was designed and when it reaches sufficient confidence in its output”). Explainability and interpretability are among many desirable characteristics to support trustworthiness in AI systems including, among others, “accuracy, privacy, reliability, robustness, safety, security (resilience), mitigation of harmful bias, transparency, fairness, and accountability” (Phillips et al., 2021). Responsibility could be another characteristic for AI trustworthiness (Yigitcanlar et al., 2021).

The lack of explainability and governability of AI had affected stakeholders’ confidence in AI systems and consequently, the uptake of AI in the energy sector has been slow. Moreover, the complexity of the design and operations space of energy systems that involves many parameters and stakeholders is on the rise and the consequent severity of risks is catastrophic due to the social, national, environmental, and geopolitical criticality of these matters.

This paper provides a review of AI explainability and governance in smart energy systems. We collected 3,568 relevant papers from the Scopus database using a specific query (see Section 2), automatically discovered 15 parameters for AI governance in smart energy systems, and group them into four macro-parameters, namely AI Behaviour and Governance, Technology, Design and Development, and Operations. We elaborate on the research landscape by reviewing over 150 papers and providing temporal progressions of the research. The methodology for discovering parameters or themes is based on “deep journalism,” our data-driven deep learning (DL)-based big data analytics approach to automatically discover and analyse cross-sectional multi-perspective information to enable better decision-making and develop better instruments for governance. We introduced the deep journalism approach (Ahmad et al., 2022) and applied it to different sectors (Alqahtani et al., 2022; Alswedani et al., 2022; Mehnoom, 2022).

The findings of this paper show that research on AI explainability in energy systems is segmented and narrowly focussed on a few AI traits (fairness, interpretability, explainability, trustworthiness) and energy system problems (stability and reliability analysis, energy forecasting, power system flexibility). The paper deepens our knowledge of AI governance in energy and is expected to help governments, industry, academics, energy prosumers, and other stakeholders to understand the landscape of AI in the energy sector, leading to better design, operations, utilisation, and risk management of energy systems.

1.1 Related works and novelty

To the best of our knowledge, this is the first comprehensive review of AI governance in the energy sector. It is a novel work due to its scope, methodology, and findings. There are several literature reviews on smart grids but they are not aimed at AI explainability or governance. We have found only two works that can be considered related to our work. A review of AI interpretability in smart grids is presented by (Xu et al., 2022) using papers collected from Google Scholar over a 5-year period. A review of AI explainability research in energy and power systems is provided by (Machlev et al., 2022) using literature from 2019 to 2022. Firstly, none of these works have used BERT (bidirectional encoder representations from transformers) to automatically discover parameters. Secondly, they do not have similar scope to ours (search query and data collection), i.e., they do not consider AI explainability and governance in its broader sense incorporating AI behaviour (governance, explainability, interpretability, responsibility, ethics, trustworthiness, and fairness) as extensively as we do (see Section 2). Thirdly, our deep journalism methodology allows us to use AI to collect a comprehensive selection of papers (a dataset) and provide a summary of a 55-year period of research on AI in energy systems.

The rest of the paper is organised as follows. Section 2 briefly describes the methodology of this work. Section 3 discusses the parameters and reviews the literature. Section 4 provides a discussion and concludes the paper.

2 Methodology and design

We briefly describe the methodology and software tool design for automatic parameter discovery here. The word limit limits us, hence the brevity, for details, see (Ahmad et al., 2022; Alqahtani et al., 2022).

We collected the data for this work from the Scopus database using the following keywords in the query: artificial intelligence, machine learning, deep learning, grid, electricity, energy, power system, governance, explainability, interpretability, responsibility, ethics, trustworthiness, and fairness. This generated 3,568 paper abstracts
published between 1967 and 2022 from various disciplines. No limits on disciplines or years were applied in collecting data. Duplicates, stop words, and irrelevant and noisy data were removed using pandas and NumPy, BERT, UMAP (uniform manifold approximation and projection), HDBSCAN (hierarchical density-based spatial clustering of applications with Noise), and class-based TF-IDF (term frequency-inverse document frequency) score were used to capture contextual relations, reduce the number of clusters, and cluster data (Grootendorst, 2021; Ahmad et al., 2022; Alqahtani et al., 2022). Finally, we used domain knowledge and a range of analysis and visualisation techniques (hierarchical clustering, topic word score, similarity matrix, term score decline) to discover parameters for AI governance in energy.

3 Parameters discovery

3.1 Overview

Table 1 lists the names of the 15 discovered parameters in Column 1 sorted by the four macro-parameters, AI Behaviour and Governance, Technology, Design and Development, and Operations. These macro-parameters will be discussed in Sections 3.2–3.5. The table provides one example research work for each parameter (Column 2) along with its research dimension (Column 3), the AI behaviour addressed by each work (Column 4) and the summary of the work (Column 5). The papers selected in the table were chosen to showcase a variety of perspectives and AI behaviours. The intention was not to rank them as the “best” works and they should not be viewed as such. Section 3.6 provides the taxonomy and temporal progression of the parameters.

3.2 Artificial intelligence behaviour and governance

This parameter is about the governance and management of AI in the energy sector by identifying the requirements to build ethical, responsible, trustworthy applications and to discuss its policies, regulations, and data privacy concerns. It captures various dimensions of AI behaviour and governance including AI responsibility and accountability in smart grids (Volkova et al., 2022), AI governance and regulations in power-related general-purpose technologies (Nitzberg and Zysman, 2022), reviewing the...
European law for the governance of AI in the electricity sector in order to allow transparent and responsible grid management (Niet et al., 2021), promoting fairness and consumer protection via the use of automated decision-making to get access to fundamental services such as electricity and telecommunications (Przhedetsky, 2021), ethics of AI and power systems.

### 3.3 Technology

#### 3.3.1 Internet of things and edge

This parameter is about the use of the Internet of Things (IoT) and edge computing for energy systems’ monitoring and efficient governance. It captures various dimensions of “IoT and Edge,” including detecting power consumption attacks for promoting vehicular edge devices reliability and AI chips’ trustworthiness (Zhu et al., 2022), enabling sustainable energy and ethical stable power applications by self-organized, learning sensor systems (Alagumalai et al., 2022), and improving the data exchange for mobile sensors to increase the energy efficiency and trustworthiness of the IoT network (Haseeb et al., 2022). Additional dimensions include efficiency in energy use via trustworthy intelligent IoT environments (Soret et al., 2022), cloud computing to monitor Wireless Sensor Network (WSNs), cloud and edge computing in energy systems applications, IoT devices in the power network, edge-cloud computing in energy monitoring, and edge computing for IoT energy systems.

#### 3.3.2 Unmanned aerial vehicles

This parameter involves the design of fair and trustworthy solutions to support the management and resource allocation in smart energy systems using Unmanned Aerial Vehicles (UAVs). It captures different dimensions of unmanned aerial vehicles, including distributed control of UAVs for enhancing the degree of coverage with limited and fair consumption of energy (Nemer et al., 2022), a fair design for multi-UAV pathway (Zhang et al., 2022), UAVs trajectory design and time allocation for fair communication in wireless NOMA-IoT networks (Zhang et al., 2022), and fair wireless communication in UAV base stations (Qin et al., 2021). Other dimensions include a fairness methodology to federated learning in vehicular edge computing (Xiao et al., 2022), resource allocation in 5G Integrated Backhaul and Access (IAB) networks to increase the trustworthiness of access links (Huang et al., 2022), and allocating resources across multiple UAVs in the IoT networks using a DL approach (Seid et al., 2021).

#### 3.3.3 Blockchain

This parameter is about improving the performance of AI applications for smart grids by integrating them with IoT and Blockchain technologies to obtain reliable and fair solutions. It captures various dimensions of “Blockchain,” including the management of electricity demand in smart grids using blockchain as a trustworthy platform (Jose et al., 2022), accountability and fairness in the energy environment using blockchain (Baashar et al., 2021), and using trustworthy blockchain-based federated learning to fight malicious devices with significant efficacy and low energy demand (Yang et al., 2022). Further dimensions include reliable, fair, and secure solutions for energy applications using blockchain (Al-Abri et al., 2022), providing data privacy and fairness via the use of blockchain and AI-powered IoT for energy management and power trading (Lin et al., 2022), a blockchain-based framework for privacy and security in energy networks, and attack detection.

### 3.4 Design and development

#### 3.4.1 Materials for energy storage and systems

This parameter involves adopting machine learning (ML) interpretable methodologies for analyzing characteristics of chemical materials, energy systems, and batteries. It captures various dimensions of “Materials for Energy Storage and Systems,” including using interpretable machine learning (IML) for estimating decomposition enthalpy that measures the durability of Chevrel phases for batteries (Singstock et al., 2021), forecasting material characteristics using IML models to provide transparency (Allen and Tkatchenko, 2022), and trustworthy approach using DL for data evaluation of battery energy storage systems (Lee et al., 2020). Additional dimensions comprise modeling and explainability of the formation energy of inorganic chemicals using DL (Huang and Ling, 2020) and developing a predictive model using IML to predict the Fermi energy level needed to build electrically conductive materials, heterostructures, and devices (Motevalli et al., 2022).

#### 3.4.2 Physics of energy systems

This parameter is about developing explainable and reliable ML models and Deep Neural Networks (DNNs) in energy systems physics. It captures various dimensions of the “Physics of Energy System,” involving developing reusable and fair intelligent systems in high-energy particle physics (Chen et al., 2021), IML methodology to improve propulsion and power systems (Longmire and Banuti, 2022), classifying and interpreting the wasted energy of high-energy jets (Du et al., 2021), ML approach in Kondo physics to optimize explainability (Miles et al., 2021), and the reliability of semiconductors (Amrouch et al., 2021).

#### 3.4.3 Sustainable energy and climate

This parameter is about investigating AI governance’s role in promoting sustainable energy and sustainable development without putting essential energy requirements at risk and considering strategies to fight climate change. It captures different dimensions of “Sustainable Energy and Climate,” including AI-powered solutions to achieve sustainable energy (Saheb et al., 2022), the aspects of water
governance in urban areas (Goulas et al., 2022), the significance of AI in promoting sustainable development in the water, food, and energy industries (D’amore et al., 2022), sustainable education and society in the energy domain (Skowronek et al., 2022), and the governance of AI to confront climate change and achieve sustainable development (Raper et al., 2022).

3.5 Operations

3.5.1 Energy markets and management

This parameter is about detecting and governing the power demand level in energy markets. It captures various dimensions of “Energy Markets and Management,” including using the IML method for the management of decentralized optimal power flow (Serna Torre and Hidalgo-Gonzalez, 2022), designing an interpretable Deep reinforcement learning (DRL) approach for transmission network expansion in wind power (Wang Y et al., 2021), and power distribution systems’ reliability, interpretability, and security (Gao and Yu, 2021). Further dimensions include optimal multi-agent energy management for interconnected energy systems in the context of a co-trading market to promote fair commerce and to maintain the privacy of entities (Sun et al., 2022), management of energy pipeline infrastructure (Belinsky and Afanasev, 2021), and applying IML and collaborative game theory for market regression analysis and its use in energy forecasting (Pinson et al., 2021).

3.5.2 Energy demand forecasting

This parameter is about data analysis to predict energy consumption and the costs for its associated services. It captures different dimensions of “Energy Demand Forecasting,” including the application of XAI in the assessment of power grid control (Kruse et al., 2022), employing Recurrent Neural Network (RNN) explainable method to predict short-term electric demands (Gürses-Tran et al., 2022), and interpretability for forecasting of probabilistic load in power network (Arora et al., 2022). Moreover, using a two-stage interpretable model for short-term energy load prediction in power management (Xie et al., 2021), improves the effectiveness of energy resource management and increases the accuracy of forecasting power consumption over the short term with IML models (Sujan Reddy et al., 2022). Additional dimensions include a multi-step interpretable probabilistic model for predicting residential energy consumption (Xu et al., 2022), short-term energy forecasting, energy consumption forecasting, load forecasting, demand forecasting, and various machine learning models for energy forecasting.

3.5.3 Solar energy systems

This parameter is about solar energy forecasting to enhance the management of power generation and propose trustworthy and explainable approaches. It captures various dimensions of “Solar Energy Systems,” including the interpretability of solar energy forecasting (Liu et al., 2022), the prediction of photovoltaic (PV) power generation that includes interpretable temporal dynamics (López Santos et al., 2022), and interpretable and trustworthy approach for global solar radiation forecasting (del Campo-Ávila et al., 2021), predicted energy output and carbon dioxide (CO₂) emissions (Bouziane and Khadir, 2022), a hybrid approach involving ML and IoT for solar radiation prediction (Ghosh et al., 2020), and irradiance in solar energy, power generation by solar energy.

3.5.4 Anomaly detection and security

This parameter is about detecting, monitoring, and classifying faults and security threats in smart energy systems using transparent and knowledge-based methods. It captures various dimensions of “Anomaly Detection and Security,” including applying XAI methods to identify conductive galloping in power grids (Sun et al., 2021), achieve transparency of fault diagnosis in electrical grids (Ardivito et al., 2022), and monitoring data poisoning attacks in smart grids using white-box and black-box analysis (Bhattacharjee et al., 2022). Other dimensions include identifying erroneous measurements in the smart grid measuring system using trustworthy data sources (Badr et al., 2022), improving distributed denial-of-service (DDoS) security of software defined networks (SDN)-based smart grids to increase security and trustworthiness (Nagaraj et al., 2021), fault detection (Landwehr et al., 2022), anomaly classification for power consumption data in smart grid (Bhattacharjee et al., 2021), anomaly attacks detection in power networks, and ML and DL models for fault detection in power systems.

3.5.5 Energy-efficient buildings

This parameter is related to adopting reliable AI models for effective management and accurate building energy consumption forecast. It captures different dimensions of “Energy-Efficient Buildings,” including the generalizability of IML for estimating building energy consumption and making buildings more energy efficient (Manren et al., 2022), classification of building energy performance certificates using XAI (Tsoka et al., 2022), and XAI approach for forecasting long-term building energy consumption (Wenning er et al., 2022). Further dimensions include providing smart recommendations based on XAI for evaluating building energy efficiency systems (Himeur et al., 2022), improving the effective use of energy by adopting an IML model to forecast room occupancy (Abdel-Razek et al., 2022), predicting energy consumption in buildings, and machine learning-based models for energy efficiency in buildings.

3.5.6 Grid reliability and stability management

This parameter is about improving energy system operations via reliable assessment models and enhancing transparency in decision-making to assure supply security and system integrity. It captures dimensions of “Grid Reliability and Stability Management,” including investigating the hazards of operating a state power grid to support responsible decision-making and management (Zhang et al., 2023) and applying ML to improve the performance of nuclear power plants. Other dimensions include discussing the ability to apply interpretable solutions (Volodin and Tolokonski, 2022) and designing a data-integration model to forecast the frequency response of a power system to help enhance the interpretability of the outcomes (Wang X et al., 2021). Moreover, reviewing the future of AI applications in the power system to support interpretability and stability (Zhao et al., 2021), improve the reliability of smart grid’s short-term voltage stability evaluation to avoid power interruptions (Luo et al., 2021), power system stability, and accuracy of ML methods for the assessment of power system, and evaluating the performance of ML methods for fault selection in power lines depending on the accuracy and explainability (Gutierrez-Rojas et al., 2022).
3.5.7 Smart city energy systems

This parameter involves using AI and IoT technologies to enhance the governance of smart cities systems and applications for energy. It captures various dimensions of “Smart City Energy Systems,” including the usage of edge AI and blockchain to manage vehicle surveillance and traffic congestion through trustworthy automobile communications and smart energy trading (Bracco et al., 2022), AI technologies in smart city governance to boost the innovative value and measurable efficiency of smart grids, electric cars, and smart buildings (Zamponi and Barbierato, 2022), and examine the most recent strategies for integrating AI and Analytics (AIA) into smart grid developments in order to enhance energy governance (Khosrojerdi et al., 2022). Additional dimensions include AI-based fairness methods for transportation localization utilizing sustainable standards (Kleisarchaki et al., 2022), smart city governance and planning using AI-based applications, such as smart transportation, smart education, and smart grid (Ashwini et al., 2022), applying DL to smart city environments and power forecasting (Naoui et al., 2021), IoT applications in smart cities such as smart transportation, smart energy, and smart governance (Ilyas, 2021), and deployment of smart energy meters for smart homes using AI-interpretable cloud analytics (Chen et al., 2019).

3.6 Taxonomy and temporal progression

Figure 1 depicts the taxonomy of AI governance of energy systems (the top) and the temporal progression of all the 15 parameters grouped into four macro-parameters. The overall intensity of research in each macro-parameter can be seen by integrating the research of its parameters.

3.7 Research in AI and energy systems (1967–2022)

Finally, we provide here an overview of research on AI and energy systems carried out between 1967 and 2022 without regard to the discovered parameters. Our methodology has afforded us a means for developing a comprehensive dataset and comprehension of research on AI in energy systems.

1967–1989: There are a total of 31 works in this period. These include AI, ethics, and human responsibility (Boden, 1987) and expert systems for reliable electric power (Siddiqi and Lubkeman, 1988). Some of the collected works are not related to energy or AI and fell under the search due to key terms such as power, responsibility, and reliability, for instance, responsible energy-rich biological compounds (Hager, 1967), ethical responsibility of automation (George, 1976), responsible decision-making in mental health organizations (Kochen, 1975), responsible AI in military applications (Beusmans and Wieckert, 1989), conserving power in robots (Selfridge and Franklin, 1990), and responsible utilization of AI (Stamper, 1988).

1990–1999: There are a total of 47 works in this period. These include AI, ethics, and human responsibility (Boden, 1987) and expert systems for reliable electric power (Siddiqi and Lubkeman, 1988). Some of the collected works are not related to energy or AI and fell under the search due to key terms such as power, responsibility, and reliability, for instance, responsible energy-rich biological compounds (Hager, 1967), ethical responsibility of automation (George, 1976), responsible decision-making in mental health organizations (Kochen, 1975), responsible AI in military applications (Beusmans and Wieckert, 1989), conserving power in robots (Selfridge and Franklin, 1990), and responsible utilization of AI (Stamper, 1988).

1990–1999: There are a total of 47 works in this period. These include the impact of AI on power plant reliability (Christie, 1990), power management and distribution for automation (Ashworth, 1990), nuclear power plant (Trovato et al., 1990), explainability of electric circuits (Kashihara et al., 1992), AI in power systems management (Janik and Gholdston, 1992; Hasan et al., 1994; Irisarri, 1996), short-term electric load forecasting (Stroemich and Thomas, 1997), reliability of nuclear weapons stockpile protection measures (Molley, 1996), AI in fault diagnosis systems for power plants (Kiupel et al., 1995), electricity transportation management (Jennings, 1995), decision tree interpretability to manage electric power utilities (Wehenkel et al., 1994), explainability for the evaluation of power system security (Boyen and Wehenkel, 1999),
and trustworthy intelligent agents to power markets (Krishna and Ramesh, 1998).

2000–2009: There are a total of 168 works in this period. These include electrical stimulation systems (Fisekovic and Popovic, 2001), ethical implications and responsibility of AI (Perri, 2001), improving quality of electricity (The engagements of the management of the French transmission system [RTE, Rseau de Transport Franais] in the matter of quality in providing electricity, no date), fair allocation of neural networks (Fidalgo et al., 2007), automatic diagnosis system for power transformers using artificial neural networks (Moreira et al., 2007), engineering ethics (Berne, 2001), and renewable energy applications (Belu, 2009).

2010–2019: There are a total of 1,142 works in this period. These include energy consumption in wireless sensor networks (Ebadi et al., 2010), reliability and efficiency of smart grids (Rocic et al., 2013), simulation of electric power markets (Vale et al., 2011), social responsibility for low energy consumption and public building energy management (Egging, 2013), smart grid IT governance (Parra et al., 2014), energy markets agent-based modelling (Lupo and Kiprakis, 2015), cybersecurity for smart grids (Yardley et al., 2015), solar power prediction (Cabrera et al., 2016), pricing systems in smart grids (Bandypadhyay et al., 2016), smart grid congestion management (MacDougall et al., 2017), electricity theft in smart grids (Yeckle and Tang, 2018), and load scheduling in smart grids (Seneviratne et al., 2019).

2020–2022: There are a total of 2,180 works during this less than a 3-year period. We can say that the last 3 years have seen a surge in AI explainability research in energy systems. The works in this period are based on the latest trends in ML and DL, explainability and interpretability, and smart approaches toward multi-source energy systems.

4 Discussion and conclusion

This paper provided a review of AI explainability and governance in smart energy systems. We automatically discovered 15 parameters and elaborated the research landscape by reviewing over 150 papers. The parameters were grouped into four macro-parameters, namely AI Behaviour and Governance, Technology, Design and Development, and Operations.

Our work supports and extends the existing literature, particularly (Machlev et al., 2022; Xu et al., 2022), that identified stability, reliability, energy forecasting, and power system flexibility as major activities in the field. This work has provided an extensive view of AI governance in energy systems and thereby has broadened and deepened the understanding of the field.

The work has identified a range of specific and broad challenges including resource allocation in wireless sensor networks with multiple UAVs (Seid et al., 2021; Zhang et al., 2022), governance of AI in power-related general-purpose technologies (Niet et al., 2021; Przedzetsky, 2021; Nitzberg and Zysman, 2022), fault detection, fault diagnosis, and anomaly detection in smart energy systems (Sun et al., 2021; Badr et al., 2022), edge computing for detecting power demand attacks (Alagumalai et al., 2022; Haseeb et al., 2022; Zhu et al., 2022), blockchain-based reliability and security (Al-Abri et al., 2022; Jose et al., 2022), governance of energy markets and energy pipeline systems (Belinsky and Afanasev, 2021; Serna Torre and Hidalgo-Gonzalez, 2022; Sun et al., 2022), forecasting short-term energy demand (Xie et al., 2021; Gunses-Tran et al., 2022), energy trading using federated learning in smart cities (Bracco et al., 2022), energy-saving edge AI applications (Khosrojerdi et al., 2022), performance optimization and stability of smart grid operations and nuclear power systems (Luo et al., 2022; Volodin and Tolokonskii, 2022), and others. All these areas are candidates for future research.

We have listed above a range of challenges related to XAI in energy systems. It is apparent that both XAI and energy systems are rich and complex fields. A proper discussion of specific limitations and challenges on the subject requires several pages. Due to the lack of space, we briefly mention below a few examples of specific challenges in XAI for energy systems. Fault detection, diagnosis, and prediction are among the most important challenges in energy systems. These are problematic due to the complexity of energy systems covering large and uninhabited geographic regions involving difficult terrains. Specific fault detection-related challenges include the management and storage issues arising due to the use of multiple data sources (solar or wind power forecasting and related faults using numerical and image data), as opposed to a single data source, for fault detection (Landwehr et al., 2022), the influence of measurement noise on fault prediction performance (Sun et al., 2021), privacy issues in fault-detection of energy systems (Gonzalez, 2022; Sun et al., 2022), security and stability-sensitive scenarios (Ardito et al., 2022), and low accuracies of AI algorithms in fault detection, diagnosis, and prediction (Wu et al., 2022). Another
increasingly important area is the security of ML and DL software (Altoub et al., 2022). The challenges in this area include, among others, data poisoning attacks and the performance of related solutions (Bhattacharjee et al., 2022), and anomaly detection methods for smart grid meter security against poisoning attacks (Bhattacharjee et al., 2021).

The parameters discovery shows that most of the research is focussed on Operations followed by research activities in Design and Technology. The least research is in AI Behaviour and Governance where much effort is needed in the future. The methods and tools to support trustworthiness (explainability and other AI traits) in AI for energy systems include, among others, visual explanation techniques using Gradient-weighted Class Activation Mapping (Grad-CAM) (Ardito et al., 2022), sequence-to-sequence RNN methods for visual explanation of short-term load forecasting (Gürses-Tran et al., 2022), the Scale-Invariant Feature Transform (SIFT) method (Singstock et al., 2021), post hoc interpretability (Allen and Tkatchenko, 2022), SHapley Additive exPlanation (SHAP) (Pinson et al., 2021; Abdel-Razek et al., 2022; Kruse et al., 2022), interpretable Tiny Neural Networks (TNN) (Longmire and Banuti, 2022), model-agnostic methods (Gürses-Tran et al., 2022), the use of Temporal Fusion Transformer (TFT) method to enhance interpretability (López Santos et al., 2022), the decision tree and Classification and Regression Tree (CART) algorithms for ML explainability (Sun et al., 2021), visual data exploration for the interpretability of fault diagnosis (Landwehr et al., 2022), a partially interpretable method using Long short-term memory (LSTM) and MLP (multilayer perceptron) for short-term load forecasting (Xie et al., 2021), and Local Interpretable Model-Agnostic Explanation (lime) (Tsoka et al., 2022). We expect that many more methods will be developed for XAI in the future.

Note that the review and analysis presented in this paper are based on the works indexed in the Scopus database. Incorporating other databases in our deep journalism tool is expected to create additional parameters and structure of research on AI in energy systems. Future work will investigate the use of our deep journalism tool with additional research databases.

**Author contributions**

RA and RM conceived, developed, analysed, and validated the study. RA developed the software. RA and RM prepared the initial draft, reviewed and edited by RM and IK. RM and IK provided
supervision, funds, resources, and contributed to the article editing.

Funding

The authors acknowledge with thanks the technical and financial support from the Deanship of Scientific Research (DSR) at the King Abdulaziz University (KAU), Jeddah, Saudi Arabia, under Grant No. RG-11-611-38.

Acknowledgments

The work carried out in this paper is supported by the HPC Center at the King Abdulaziz University.

References

Abdel-Razek, S. A., Marie, M. S., Abozinadah, A., and Elzaki, O. M. (2022). Energy efficiency through the implementation of an AI model to predict room occupancy based on thermal comfort parameters. Sustainability 14 (13), 7734. doi:10.3390/SU14137734

Ahmad, I., Alqurashi, F., Abozinadah, E., and Mehmood, R. (2022). Deep journalism and DeepJournal V1.0: A data-driven deep learning approach to discover parameters for transportation. Sustain. Switz. 14 (9), 5711. doi:10.3390/SU14095711

Al-Abri, T., Onen, A., Al-Abri, R., Hossen, A., Al-Hinai, A., Jung, J., et al. (2022). Review on energy application using blockchain Technology with an introduction in the pricing infrastructure. IEEE Access 10, 80119–80137. doi:10.1109/ACCESS.2022.3194161

Alagumalai, A., Shou, W., Mahian, O., Aghbashlo, M., Tabatabaei, M., Wongwises, S., et al. (2022). Self-powered sensing systems with learning capability. Joule 6 (7), 1475–1500. doi:10.1016/J.JOULE.2022.06.001

Alshamri, N., Alswedani, S., Alzahrani, A., Katib, I., Albeshti, A., and Mehmood, R. (2022). Musawah: A data-driven AI approach and tool to Co-create healthcare services with a case study on cancer disease in Saudi Arabia. Sustainability 14 (6), 3131. doi:10.3390/SU14063131

Alkhayat, G., Hasan, S. H., and Mehmood, R. (2022). Senergy: A novel deep learning-based auto-selective approach and tool for solar energy forecasting. Energies 15 (18), 6659. doi:10.3390/EN15186659

Allen, A. E. A., and Tkatchenko, A. (2022). Machine learning of material properties: Predictive and interpretable multilinear models. Sci. Adv. 8 (18), 7185. doi:10.1126/sciadv.abm7185

Alomari, E., Katib, I., Albeshti, A., Yigitcanlar, T., and Mehmood, R. (2021). Ikitsinf+: A big data tool with automatic labeling for road traffic social sensing and event detection using distributed machine learning. Sensors 21 (9), 2993. doi:10.3390/s21092993

Alqahtani, E., Janbi, N., Sharaf, S., and Mehmood, R. (2022). Smart homes and families event detection using distributed machine learning. Emerg. Top. Comput. Intell. 6 (3), 628–637. doi:10.1109/TECI.2021.3064028

Alosh, M., AlQurashi, F., Yigitcanlar, T., Corchado, J. M., and Mehmood, R. (2022). An ontological knowledge base of poisoning attacks on deep neural networks. Appl. Sci. 12 (21), 11053. doi:10.3390/APPL122111053

Amrouch, H., Chowdhury, A. B., Jin, W., Karri, R., Khorrami, F., Krishnamurthy, P., et al. (2021). “Special session: Machine learning for semiconductor test and reliability.” In Proceedings of the IEEE VLSI Test Symposium, San Diego, CA, USA, 25-28 April 2021. doi:10.1109/VTS50974.2021.9441052

Ardito, C., Dekdoo, Y., Noia, T. D., Sciascio, E. D., and Nazary, F. (2022). Visual inspection of fault type and zone prediction in electrical grids using interpretable spectrogram-based CNN modeling. Expert Syst. Appl. 210, 118306. doi:10.1016/J.Essa.2022.118306

Arora, P., Khosravi, A., Panigrahi, B. K., and Suganthan, P. N. (2022). Remodelling state-space prediction with deep neural networks for probabilistic load forecasting. IEEE Trans. Emerg. Top. Comput. Intell. 6 (3), 628–637. doi:10.1109/TECI.2021.3064028

Ashwini, B. P., Savithramma, R. M., and Sumathi, R. (2022). “Artificial intelligence in smart city applications: An overview,” In Proceedings - 2022 6th International Conference on Intelligent Computing and Control Systems, ICICCS, Madurai, India, 25-27 May 2022, 986–993. doi:10.1109/ICICCS53718.2022.9788152

Ashworth, B. R. (1990). Managing autonomy levels in the SSM/PMD testbed. Proc. Intersoc. Energy Convers. Eng. Conf. 1, 263–268. doi:10.1109/IECEC.1990.716891

Baishar, Y., Alkawi, G., Alkahtani, A. A., Hashim, W., Razali, R. A., and Tsong, S. K. (2021). Toward blockchain Technology in the energy environment. Sustainability 13 (16), 9008. doi:10.3390/SU13169008

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher’s note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.
