XAI and Ethical AI in the Energy Sector
–
Guidelines to Establish Explainability and Ethics in Intelligent Black Box Systems

Masterarbeit

zur Erlangung des akademischen Grades “Master of Science (M.Sc.)“ im Masterstudiengang Wirtschaftswissenschaft der Wirtschaftswissenschaftlichen Fakultät der Leibniz Universität Hannover

vorgelegt von

Name: Lier Vorname: Sarah Kristin

[redacted] [redacted] [redacted]

Betreuerin: M. Sc. Jana Gerlach

Prüfer: Prof. Dr. Michael H. Breitner

Hannover, den 30.09.2022

Table of Contents

1. Introduction

Artificial Intelligence (AI) found its origins around 1956 through the Summer Research Project on Artificial Intelligence by computer scientist and author John McCarthy. The use of the term "Artificial Intelligence" was first applied here as a research discipline, as it was assumed that any kind of learning or other intelligent characteristics could be simulated by a machine. Over time, the targets and methods continuously changed and adapted to their environment (Dick 2019). As a key to Artificial Intelligence, Machine Learning (ML) has made great progress in recent times. ML is used in AI to process data, detect patterns in that data, and make predictions from the data (Haag et al. 2022). Processing of the enormously large and ever-growing amounts of data, accompanied by recognizing patterns and making predictions, is considered a major advance in research. Scientists, medical professionals, and others are using AI support for, among other things, automated decision making (Storey et al. 2022). Computer systems based on ML can solve an increasing number of complex AI challenges. However, solutions are proving increasingly difficult due to the predominantly non-transparent nature of these systems. Explaining these systems is becoming increasingly relevant to users, developers, and other stakeholders. Despite various research approaches over the years, there does not seem to be a solution to this yet to avoid the complexity and the non-transparency and inexplicability that accompany it (Zednik 2021).

The comprehensible reason for a decision on the part of AI is predominantly non-existent (Storey et al. 2022). Advances in ML and AI are widely applied and, in many parts, no longer require any human assistance or supervision as AI and ML have now matured through constant training to make decisions based on learned data (Rawal et al. 2021). Especially when it comes to the impact of decisions on people's lives or their environment, it is of absolute necessity to understand how and why a decision or recommendation has been made by an AI/ML.

Explainable Artificial Intelligence (XAI), also known as White Box, offers the possibility to increase the understanding of the models and systems. Within the last years, the publication of scientific papers could increase fivefold linearly, thus gaining importance of the need for its implementation (Rawal et al. 2021). The pursuit of explainability is strongly highlighted in representative surveys of the population, such as Bitkom Research (2020). Respondents were found to be critical of AI, with 85% stating that AI software should only be approved after extraordinarily thorough testing and (44%) should be banned in certain areas of application (Bitkom Research 2020).

The functioning of ML models should become comprehensible, while accuracy and high performance should at least be maintained. Non-transparent Black Box models in sensitive areas such as healthcare or other areas related to life, law, finance, or

privacy, as well as the Energy Sector, is problematic. While the interest in Deep Learning (DL) is slowly reducing due to fixing the issues at hand, for the most part, the interest in Explainable AI is increasing. XAI, in contrast to Deep Learning models, which have high accuracy but low interpretability, is expected to provide both high accuracy and high interpretability (Adadi and Berrada 2018; Angelov et al. 2021; Arrieta et al. 2020).

With the rise of the pursuit of explainability, the ethical view of AI is also becoming increasingly relevant in research, even if it is currently less mature than XAI. Ethical AI is often understood as a requirement for XAI to evolve into Responsible AI (Arrieta et al. 2020). Ethics in AI can also be categorized as Artificial General Intelligence (AGI). AGI describes the hypothetical intelligence of a program with the ability to understand intellectual tasks that can be performed by humans. It thus describes a strong or complete Artificial Intelligence. Currently, research is already being conducted on machines so that they can know how to perform actions that are far superior to human intelligence. This intelligence is referred to as Artificial Superintelligence (ASI) (Adadi and Berrada 2018). In the literature, a distinction is often made between weak and strong Artificial Intelligence. The development of weak Artificial Intelligence refers to the direct and limited accomplishment of a task, while strong Artificial Intelligence aims to mirror or surpass the intelligence of humans (Kirchschläger 2022). This thesis is dedicated to strong AI. Ethical considerations in autonomous systems are steadily increasing, but implementation and execution are not yet mature enough. Despite many kinds of research approaches and methods, few technical solutions exist for integrating ethics into AI (Yu et al. 2018). Since the research for implementing ethical values is still young, there are still very different opinions from different researchers, so no common ethics framework has been found. Many researchers describe different value criteria and try to define ethics and Ethical AI in general. However, they do not describe holistic or at least industry-specific solutions in front of different definitions (Greene et al. 2019).

Therefore, another subtarget of this thesis will be to define ethical Requirements and corresponding solution strategies for the Energy Sector. In this thesis, despite common interfaces and explanations, Ethical Artificial Intelligence and Explainable Artificial Intelligence are considered separately from each other and as two targets of science. Ethical AI is abbreviated as EAI in the figures for simplicity.

In the field of Energy Systems, the non-transparent Deep Learning has developed better algorithms for ML, so the DL methods have more advantages in the field of efficiency, noise immunity, and accuracy (Machlev et al. 2022b). In the past, energy consumption has been steadily increasing, which caused early concern in the area of energy demand, and thus more emphasis was placed on renewable energy. AI was quickly considered as a solution and integrated into systems, especially as a Smart

Grid (SG) application. Thus, AI is used for load forecasting, power generation and management, Demand Side Management (DSM), service, and also for electrical grid operation and control. AI therefore has high relevance when it comes to Smart Grids and the modernization of power grids. Energy modeling makes a crucial contribution to the creation of Smart Grids. It ensures the dynamics in Energy Systems, such as efficiency or consumption. However, the Black Box problem strongly takes hold here as well, since the way AI works is almost incomprehensible. It drastically complicates explanations and thus the accompanying dynamic decision-making processes. Especially in the Energy Sector, accuracy is of great importance, which, however, means complex systems at the same time to be able to process the increasing amounts of data (Kuzlu et al. 2020; Zhang et al. 2021b). Due to the expansion of renewable energy, the inherent fluctuations are increasing, so that ensuring security of supply as well as grid stability are becoming increasing hurdles that require solutions (Fridgen et al. 2022). Grid frequency is also increasingly becoming a major hurdle, especially for Black Box systems that are difficult to explain. Frequency deviations can lead to power generation and demand imbalances, system instabilities, or power outages (Kruse et al. 2021a). In the area of locally increasing energy sources, such as Photovoltaic (PV) systems, automated control operations are needed. This is necessary to ensure power supply in a reliable and efficient manner. To this end, the digitization of the Electricity Supply Chain (ESC) is being accelerated so that a control automation infrastructure can be ensured. In the context of the Smart Grid Architecture Model (SGAM), the ESC includes Distributed Energy Resources (DER) generation and consumption, as well as transmission and distribution. AI supports this automation to process the big data in a heterogeneous and high-quality manner. Complex pattern recognition as a capability of AI is also used in this, as well as data analysis and data interpretation. However, research is currently being conducted on how the data is processed and why exactly this result is obtained from it (Richter et al. 2022). The relevance of the thesis clearly emerges with the differentiated approach of Ethical AI and Explainable AI in their respective connections with the Energy Sector. Since XAI has already found its application in the Energy Sector, it is important to derive and optimize these application possibilities based on the requirements from the already existing application areas. For this, corresponding principles should be able to cover the requirements (Research Question (RQ)1). The use of Ethical AI in the Energy Sector is almost unexplored and hardly documented in literature. Therefore, it is necessary to derive general requirements and tailor them to the Energy Sector, if this is possible. Principles should then help to implement the ethical view of AI in the requirements (RQ2).

RQ1: What requirements for XAI already exist in the Energy Sector? What requirements for XAI can be transferred from other domains into the Energy Sector?

RQ2: How can the ethical view of AI be applied to the Energy Sector? What requirements can be derived from other sectors for the Energy Sector?

Therefore, the target of this work is to provide requirements and principles from the scientific literature for the Energy Sector for Explainable AI and Ethical AI. To achieve this target, a comprehensive literature review and analysis is first conducted according to vom Brocke et al. (2009), in which the literature analysis according to Webster and Watson (2002, 2020) is also incorporated. Here, various databases are quantitatively collected and sorted based on pre-determined keyword combinations as well as forward and backward searches. The multitude of literature will then be qualitatively assessed and incorporated for the scientific evidence in this master thesis. In the second step, the Explainable Artificial Intelligence and the Ethical Artificial Intelligence are explained. Since XAI is already applied in the Energy Sector, a corresponding Status Quo is provided here in Chapter 3.1. The third step includes the explanation of the methodology of Design Science Research (DSR) according to Hevner et al.; Hevner (2004; 2007), in order to be able to perform an analysis for the requirements and principles based on Gregor et al. (2020) and Wambsganss et al. (2020). Here, the text mining tool Orange supports filtering and sorting the multitude of scientific publications from the literature search. The derived requirements and principles are then evaluated based on companies from the database provider Crunshbase and subsequently adapted. The ensuing discussion challenges the requirements and definitions and discusses commonalities, differences, and improvements. Together with a limitation, a conclusion follows, in which case the research questions (RQ1, RQ2) can be answered successfully.

6. Conclusion

To answer the Research Questions RQ1 "What requirements for XAI already exist in the Energy Sector? Which requirements for XAI can be transferred from other sectors to the Energy Sector?" and RQ2 "How can the ethical view of AI be applied to the Energy Sector? What requirements can be derived from other sectors for the Energy Sector?" to answer, a systematic literature review was conducted at the beginning of this thesis by vom Brocke et al. (2009) and according to Webster and Watson (2002, 2020), in which 141 publications in the field of Ethical AI and 191 publications for Explainable AI have been filtered, analyzed, and synthesized. The literature review provides only a subjective view. Based on the available publications, a Theoretical Background for Ethical and Explainable AI could be created, in which the descriptions, characteristics, and, in some cases, requirements were extensively explained. In the Theoretical Background for Explainable AI, a Status Quo for the Energy Sector was given, which was not possible for Ethical AI due to a lack of publications and research contributions. After providing a comprehensive understanding of both AI systems and thereby differentiating Ethical AI and Explainable AI from each other, the heart of this thesis was described. In order to get closer to answering the existing Research Questions, the Design Science Research methodology of Hevner et al.; Hevner (2004; 2007) was explained and applied. In doing so, Guidelines for conducting the DSR were created, which cover nine points. Thereby, the Artifact represents points 2 to 5, in which Key Topics, Design Requirements and Design Principles have been developed. The Key Topics were based on the literature review conducted in Chapter 2 with 179 imported publications for XAI and 129 imported publications for Ethical AI into the text mining tool Orange. The objective view of text mining and topic modeling reduced the subjective view as the large number of publications deepened. Eight Key Topics were identified for the XAI. A total of fifteen DR and seventeen DP were then developed from the publications sorted into the KT. In the Ethical AI, four KT were formed and twelve DR and twelve DP were derived. Based on the descriptions of all KT, DR, and DP, an evaluation was then conducted by writing to companies for each of XAI and Ethical AI and asking for their assistance. For each AI system, three responses could be generated, adding one DP for XAI , one DR and three DP for Ethical AI. Similarly, changes could be made to the DR, DP, and the relationships between them. Subsequently, all results were critically reviewed and challenges or insufficient research findings were highlighted. Basically, the DR and DP could only be implemented for XAI in the Energy Sector. For implementation of DR and DP for Ethical AI in Energy Sector, the knowledge and current research contributions are not further sufficient. It was only possible to generate a connection of Ethical AI and the Energy Sector in relation to the energy supply systems in the discussion. Accordingly, there is

an urgent need for research in this area to efficiently integrate Ethical AI into the Energy Sector. Furthermore, the discussion revealed and highlighted the need for research on

- the definition of XAI and Ethical AI
- the combination of explainability and accuracy
- the optimization of transparency
- the interface of the system
- the intrinsic approaches to XAI
- the combination of White Box and Black Box models
- the context adaptive methods for XAI
- the demand side management for the Energy Sector in the context of XAI
- the RL methods of XAI
- the combination of Explainable and Ethical AI
- and the review and creation of data sets especially for Ethical AI.

For practical purposes, a comprehensive overview of requirements and principles for Explainable AI and Ethical AI has been provided in this paper. Explainable AI already offers very good approaches for applications in the Energy Sector, some of which are already mature and some of which still require optimization. Thus, it is not unlikely that XAI will find its anchor in the Energy Sector in the near future. Ethical AI, on the other hand, does not currently offer any approaches for application in the Energy Sector. However, the consideration of energy supply is a good step for an initial implementation on which to build and incorporate more ethics into AI for the Energy Sector. This work can help to do more research in the research gaps and focus, especially on maturing XAI and developing Ethical AI. By making the DR and DP general, other sectors could also benefit from the requirements and principles, while Energy Sector specialists can build on them through the descriptions of the DR and DP in XAI and use them as a research purpose. The requirements can therefore be both generalized and adapted to provide the most accurate requirements for the relevant sector of the research in question. The Principles provide a good overview of possible solution strategies and encompass new research areas through the different combinations. There are also other, in contrast Principles, which can be supplemented or added. Thus, in the end, both AI systems are not yet mature and offer efficient possibilities for application.