Aspects of Explanations for Optimization-Based Energy System Models

Aspekte der Erklärungen von optimierungs-basierten Energiesystemmodellen
Zur Erlangung des akademischen Grades Doktor-Ingenieur (Dr.-Ing.)
Genehmigte Dissertation im Fachbereich Elektrotechnik und Informationstechnik von Jonas Hülsmann
Tag der Einreichung: 01. Mai 2024, Tag der Prüfung: 12. Juli 2024

1. Gutachten: Prof. Dr. Florian Steinke
2. Gutachten: Prof. Dr. Frank Jäkel
Darmstadt, Technische Universität Darmstadt

TECHNISCHE
UNIVERSITÄT
DARMSTADT

Electrical Engineering and Information Technology Department

Energy Information Networks & Systems

Aspects of Explanations for Optimization-Based Energy System Models
Aspekte der Erklärungen von optimierungs-basierten Energiesystemmodellen

Accepted doctoral thesis in the department of Electrical Engineering and Information Technology by Jonas Hülsmann

Date of submission: 01. Mai 2024
Date of thesis defense: 12. Juli 2024

Darmstadt, Technische Universität Darmstadt

Erklärungen laut Promotionsordnung

§8 Abs. 1 lit. d PromO

Ich versichere hiermit, dass von mir zu keinem vorherigen Zeitpunkt bereits ein Promotionsversuch unternommen wurde. Andernfalls versichere ich, dass der promotionsführende Fachbereich über Zeitpunkt, Hochschule, Dissertationsthema und Ergebnis dieses Versuchs informiert ist.

§9 Abs. 1 PromO

Ich versichere hiermit, dass die vorliegende Dissertation, abgesehen von den in ihr ausdrücklich genannten Hilfsmitteln, selbstständig und nur unter Verwendung der angegebenen Quellen verfasst wurde. Weiterhin versichere ich, dass die "Grundsätze zur Sicherung guter wissenschaftlicher Praxis an der Technischen Universität Darmstadtßowie die "Leitlinien zum Umgang mit digitalen Forschungsdaten an der TU Darmstadtïn den jeweils aktuellen Versionen bei der Verfassung der Dissertation beachtet wurden.

§9 Abs. 2 PromO

Ich versichere hiermit, dass die vorliegende Dissertation bisher noch nicht zu Prüfungszwecken gedient hat.

Darmstadt, 01. Mai 2024

J. Hülsmann

Kurzfassung

Die Energiewende bringt zahlreiche Herausforderungen mit sich, die eine Vielzahl von Entscheidungen erfordern. Optimierungsbasierte Energiesystemmodelle (ESM) sind nützliche Werkzeuge, um diese Entscheidungsprozesse zu erleichtern. Allerdings sind ESMs oft zu umfangreich und komplex für Entscheidungsträger wie CEOs, Politiker oder Bürger, die in der Regel nicht über Modellierungsexpertise verfügen. Erklärungen sind notwendig, um die komplexen Ergebnisse und das unintuitive Verhalten von ESM für Entscheidungsträger nutzbar zu machen und mögliche Diskrepanzen zwischen ihren Zielen und den Modellannahmen zu erkennen.

Diese Arbeit konzentriert sich auf verschiedene Erklärungsansätze für optimierungsbasierte ESM. Zunächst werden die Erklärungsansätze aus psychologischer und philosophischer Sicht betrachtet, um die grundlegenden Prozesse und Faktoren zu identifizieren, die für eine Erklärung notwendig sind und deren wahrgenommene Adäquatheit bestimmen. Basierend auf diesen grundlegenden Konzepten, bewerten wir den aktuellen Stand der Technik bei der Erklärung von ESM und identifizieren Mängel in den bestehenden Erklärungsansätzen. Wir zeigen Parallelen zwischen den Herausforderungen der Erklärbarkeit im Gebiet der ESM und dem Gebiet des maschinellen Lernens, insbesondere im Bereich der Erklärbaren Künstlichen Intelligenz (XAI), die sich auf die Entwicklung von Methoden zur Verbesserung der Erklärbarkeit von maschinellen Lernmodellen konzentriert.

Diese Arbeit beschäftigt sich mit drei verschiedenen Ansätzen aus den oben genannten Bereichen, um Erklärungen für optimierungsbasierte ESM zu verbessern. Zunächst wird ein Ansatz aus dem Bereich der XAI auf ein ESM übertragen, um zwei Defizite bestehender ESM-Erklärungen zu adressieren: den Einfluss hochdimensionaler Eingabeparameter und die Generierung von Erklärungen unterschiedlicher Komplexität für verschiedene Zielgruppen. Angesichts der Schlüsselrolle, die der Kausalität bei der Erzeugung von Erklärungen zukommt, untersuchen wir die Verwendung kausaler graphischer Modelle zur Erklärung von ESMs. Schließlich entwerfen wir einen interaktiven Ansatz, um das praktische Erlernen von ESM am Beispiel der Energiewende zu erleichtern, und evaluieren die Wirksamkeit dieses Ansatzes im Rahmen eines Universitätskurses für Masterstudierende.

Abstract

The transition towards a renewable energy system presents numerous challenges, demanding extensive decision-making processes. Optimization-based energy system models (ESMs) are valuable tools to facilitate these decisions. However, these models often prove too intricate and expansive for decision-makers, such as CEOs, politicians, or citizens, who typically lack expertise in ESMs. Consequently, there is a pressing need for explanations to bridge the gap between the complexity of ESM results and decision-makers comprehension, allowing them to discern potential disparities between their objectives and the model's assumptions.

This thesis focuses on various explanations for optimization-based ESMs. Initially, we explore explanations from psychological and philosophical perspectives to identify the fundamental concepts necessary for creating comprehensive explanations and elucidate the key factors essential for these explanations to be deemed adequate. Armed with these foundational concepts, we assess the current state-of-the-art in explaining ESMs, identifying existing shortcomings in explanation methodologies. We draw parallels between the challenges faced by the ESM domain and the field of machine learning, particularly in the domain of explainable artificial intelligence (XAI), which is dedicated to developing methods to enhance the explainability of machine learning models.

This thesis delves into three distinct approaches from the abovementioned domains to address these challenges and enhance explanations for optimization-based ESMs. Firstly, we adopt an approach from XAI to ESMs to overcome two prevalent shortcomings in ESM explanations: elucidating the impact of high-dimensional input data and tailoring explanation complexities to different target audiences. Given the key role of causality in crafting explanations, we explore the utilization of causal graphical models for explaining ESMs. Finally, we design an interactive approach to facilitate hands-on learning of ESMs, focusing on the example of energy transition, and evaluate the efficacy of this approach within the context of a graduate-level university course.

Contents

List of Abbreviations

BEV	Battery Electric Vehicle
CEO	Chief Executive Officer
CHP	Combined-heat-and-power Units
CNN	Convolutional Neural Network
CO_2	Carbon Dioxide
DAG	Directed Acyclic Graph
DM	Decision Maker
ESD	Energy System Design
ESM	Energy System Model
ETG	Energy Transition Game
IEA	International Energy Agency
LP	Linear Program
MILP	Mixed-integer Linear Program
MGA	Modelling to Generate Alternatives
POI	Point of Interest
PV	Photovoltaic
RQ	Research Question
SQL	Structured Query Language
XAI	Explainable Artificial Intelligence

1. Introduction

The global shift from traditional, non-renewable energy sources such as coal, oil, and natural gas towards renewable energy sources such as wind, solar, and hydropower is one of the important tasks of our time. The transition to renewable energy responds to the growing concern over climate change and the need to reduce greenhouse gas emissions to limit its impacts. Energy production is one of the most significant contributors to greenhouse gas emissions. Therefore a transition to renewable energy is essential for reducing our carbon footprint.

With a transition towards a renewable energy system, new challenges emerge that require a lot of decision-making. The intermittency and variability of renewable energy sources pose challenges for grid operators, who must ensure a real-time balance between energy supply and demand. A more diverse structure of energy consumers and producers, as well as the limited local potential of some renewable power plants [1], will require significant changes in storage and grid infrastructure [2].

While renewable energy is becoming increasingly cost-competitive, there are still significant upfront costs associated with building new infrastructure that have to be financed. The absence of supportive policies could slow the energy transition [3] if incentives to move away from non-renewable energy sources are lacking. The transition might be hindered even more if regulations are in place that actively limit the expansion of renewable power plants [4].

Despite the energy transition's potential environmental and economic benefits in the long run[5], it may also result in job losses in traditional energy sectors, leading to social and political backlash. Additionally, communities may resist developing new energy infrastructure, such as wind turbines and solar farms, due to concerns about their impact on local wildlife or their perceived visual or acoustic disturbance [6].

Therefore, decision-makers (DMs) who make energy-related investments face career risks, such as loss of investment or loss of reputation, if they approve energy projects that fail.

Such DMs could be CEOs who decide to invest in a new, more energy-efficient production facility, politicians who decide on new regulations, e.g. for a minimum distance between wind turbines and settlements, or citizens who invest in balcony power plants. Rejecting projects has typically fewer immediate consequences, but rejecting projects necessary to advance energy transition might be even more damaging for society then choosing unpopular projects in the long run.

The success of energy systems projects often depends on various parameters. The systems are commonly connected to other energy systems, making even small energy projects complex and hard to understand. Energy system models can serve as valuable tools to support DMs in making better-informed decisions.

Energy system models (ESMs), such as TIMES [7] or OSeMOSYS[8], provide insights into the impacts and feasibility of different energy transition pathways. These models simulate and evaluate the economic effects of various scenarios, such as different levels of renewable energy deployment or the introduction of new technologies. Moreover, ESMs can help identify the cost-optimal mix of energy sources that meet energy demand while minimizing greenhouse gas emissions. By using ESMs, regulatory interventions intended to support the energy transition, such as carbon pricing, renewable energy incentives, and energy efficiency standards, can be evaluated before implementation in the real world. These models can also offer insights into necessary energy infrastructure extensions, such as additional transmission lines or energy storage. ESMs can have different temporal and spatial scales, i.e., one ESM can help with operational planning for a single power plant, while another can help with long-term development planning for the national power grid.

However, ESMs are typically designed by domain experts for domain experts. They consist of numerous parameters, often reaching several millions, making expert knowledge essential for their construction, operation, and interpretation of results. Even for domain experts, interpreting ESM results can be challenging, as the models' behavior may be counter-intuitive under certain circumstances [9]. DMs responsible for energy infrastructure decisions are usually not domain experts but individuals with limited ESM expertise. Therefore, DMs wishing to utilize ESMs for their decision-making must either trust and rely on domain experts to interpret the results or strive to understand the models themselves [10]. In order to make informed decisions and avoid undue reliance on potentially biased domain experts, DMs should aim to understand ESMs or at least comprehend their results.

Hence, the challenge of this thesis is to develop methods that can generate explanations for ESMs that are accessible to individuals without expertise in the energy system domain, such as CEOs, politicians, or citizens. More specifically, this thesis focuses on explanation methods for optimization-based ESMs based on linear programs (LP) [7] or mixed-integer linear programs (MILP) [11]. It is essential to differentiate between optimization-based ESM and those based on statistical learning methods. Models based on statistical learning are often perceived as black boxes because their internal parameters are learned from data and are not known in advance, rendering their decision-making processes opaque. The field of explainable artificial intelligence (XAI) explores methods to clarify the decision-making of machine learning black box models, and these techniques can also be applied to learning-based ESMs enhancing their transparency [12]. In contrast, optimization-based ESMs are regarded as white-box models since domain experts select their internal parameters, and the model does not learn the parameters. However, while explaining the derivation of results for small ESM may be possible using methods such as sensitivity analysis [13], doing so becomes impractical or even infeasible within a reasonable timeframe for large models with millions of parameters and decision variables. Furthermore, the required knowledge about the mathematical principles of optimization render the understanding of such models impossible for those who lack this knowledge, despite the white-box characteristics of optimization models.

To illustrate the challenges involved in explaining optimization-based ESMs to non-experts, consider the following example.

Example 1. Consider the energy system of a single-family home equipped with a photovoltaic (PV) power plant and battery storage, as shown in Figure 1.1 (a). The system's electricity can be obtained from the grid or generated using the PV power plant to fulfill the household's electric demand. Any excess energy produced can be stored in the battery for later use or sold back to the grid. In such an energy system, questions about the capacity of the PV power plant, the battery storage capacity, or their operation plans that minimize the overall system cost are of interest. An ESM can be employed to address these questions, focusing on specific scenarios or input data. An optimization-based ESM comprises an objective function that is either minimized or maximized, along with a set of constraints that characterize component behavior and establish boundaries for their operation. In this example, an objective function might involve minimizing system costs. Constraints could involve ensuring electricity production, storage, and demand equilibrium within each time step, maintaining storage equilibrium across multiple time steps, and imposing limitations on power plant electricity production based on radiation levels in a given time series.

Figure 1.1 b) provides a schematic representation of an ESM that corresponds to the energy system of the single-family home discussed earlier. A subset of parameters for this simplified ESM is displayed on the left side. Depending on the chosen time resolution and

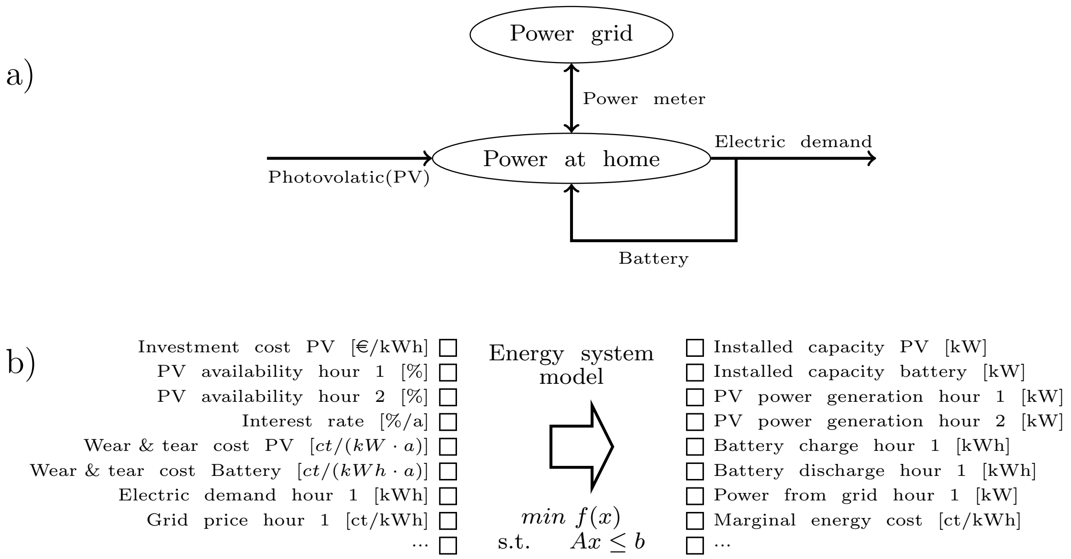
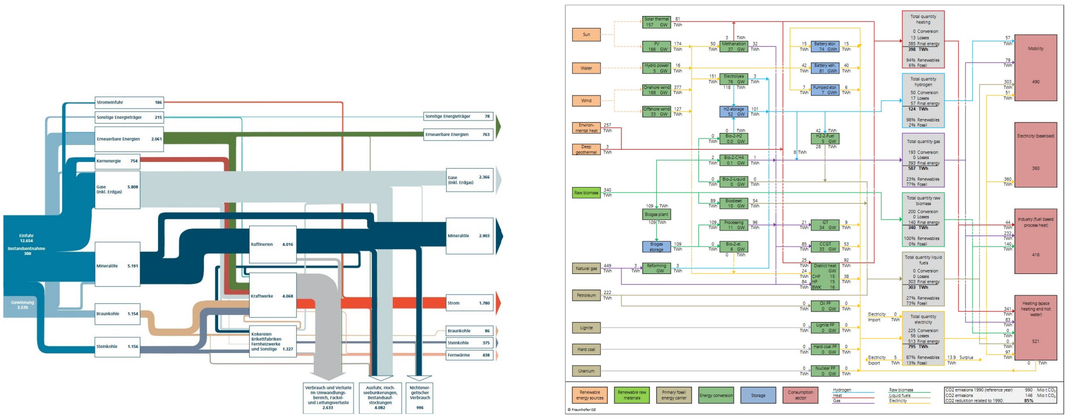

a)
Power grid
Power meter
Electric demand
Power at home
Photovolatic(PV)
Battery
b)
Investment cost PV [€/kWh] □
PV availability hour 1 [%] □
PV availability hour 2 [%] □
Interest rate [%/a] □
Wear & tear cost PV $[ct/(kW \cdot a)]$ □
Wear & tear cost Battery $[ct/(kWh \cdot a)]$ □
Electric demand hour 1 [kWh] □
Grid price hour 1 [ct/kWh] □
··· □
Energy system model
$min\ f(x)$
s.t. $Ax \leq b$
□ Installed capacity PV [kW]
□ Installed capacity battery [kW]
□ PV power generation hour 1 [kW]
□ PV power generation hour 2 [kW]
□ Battery charge hour 1 [kWh]
□ Battery discharge hour 1 [kWh]
□ Power from grid hour 1 [kW]
□ Marginal energy cost [ct/kWh]
□ ···

Figure 1.1.: a) Schematic view of an energy system for a single family home with photovoltaic electricity generation and a battery. b) Schematic view of an optimization-based energy system model for the single family home. A large number of information such as cost assumptions, efficiencies, and time series of demands, renewable availability, or model externals prices is required as input parameters for the model. An objective that is bound by a set of constraints is then optimized to create a multitude of outputs.

time horizon, the ESM for this small-scale home can comprise several thousand parameters, which rely on data related to pricing, weather patterns, demand variations, and technical specifications. The ESM is subsequently optimized by minimizing or maximizing a designated objective function, denoted as $f(x) : \mathbb{R}^{N\times 1} \rightarrow \mathbb{R}^1$. The feasible values for the decision variables $x \in \mathbb{R}^{N\times 1}$ are bounded by a set of constraints, represented by the matrix equation $Ax \leq b$, which encodes the input parameters within matrix $A \in \mathbb{R}^{M\times N}$ and vector $b \in \mathbb{R}^{M\times 1}$. In addition to determining the optimal value of the objective function, the output of the ESM includes a set of values, denoted as x^*, that correspond to the optimal solution. These values represent critical information such as cost-optimal capacities or optimal operational plans. Moreover, the marginal power costs and can also be readily derived from ESM.

Although optimization-based ESM can determine the cost-optimal values for variables such as PV or battery capacity, the solvers that obtain these solutions do not inherently



(a) Energy flow of the German energy system. Taken from [14].

(b) Schematic image of a big energy system. Taken from [15].

Figure 1.2.: Multi-modal energy systems are complex systems, that are hard to understand by non-experts. ESMs can be used for optimizing big multi-modal energy systems.

explain their decision-making process. Consequently, important questions such as "Which factors significantly impact the optimal capacity of PV power plants?" remain unanswered by the models. However, these insights are crucial for DMs to comprehend or validate a model or implement regulatory changes that affect crucial parameters.

The energy system model provided in the small single-family household example is already intricate, posing a challenge for non-experts to comprehend, despite its focus solely on electricity. In Figure 1.2a, the energy flow diagram for the German energy system is depicted, showcasing multiple modalities beyond electricity. This diagram showcases the complexity of real energy systems. A schematic representation of an ESM aiming to model a multi-modal energy system of Germany is illustrated in Figure 1.2b. Grasping the intricacies of ESMs at this level of complexity is a demanding task even for domain experts, making it virtually impossible for non-experts.

Currently, domain experts commonly employ sensitivity analysis to provide local explanations of optimization-based ESM (as well as other optimization models). Sensitivity analysis assesses the responsiveness of model outputs or system behavior to changes in input parameters. It enables the evaluation of the extent to which input variations influence specific output variables. Sensitivity analysis can also help identify a model's potential weaknesses or limitations by revealing which input parameters significantly influence the output, so domain experts can decide whether these parameters are well understood or uncertain.

However, it is essential to note that sensitivity analysis often concentrates on a restricted set of input variables, thereby excluding the consideration of all possible scenarios or combinations of input values. Consequently, the analysis may fail to capture the complete spectrum of relevant factors that can impact the output. While sensitivity analysis is useful at revealing the correlation between input variables and the output, it does not reveal the underlying causal relationships between them. Consequently, the ability to make well-informed decisions or take practical actions based solely on the analysis results may be limited.

A potential approach to increase the amount of variables that are varied without increasing the computational complexity is through marginal analysis, which involves conducting sensitivity analysis based on marginal changes. Marginal changes can be calculated as a byproduct of the LP solving process [16]. However, it is essential to note that marginals only apply for small changes, restricting the validity of explanations to local contexts. For non-experts, generating explanations based on the marginal sensitivity values of several thousand (or even several million) parameters becomes impossible without expert knowledge necessary to differentiate significant changes from irrelevant ones. Moreover, sensitivity analysis fails to capture the impact of events that simultaneously affect multiple parameters. For example, a power outage spanning several time steps can affect the model's solution that differs from the simple sum of sensitivities across those time steps. Consequently, creating explanations for optimization-based ESM using sensitivity analysis requires expert knowledge, as a comprehensive sensitivity analysis is often too complex for laypeople to handle effectively.

We observe: the domain of machine learning faces similar challenges when explaining their complex models to non-experts in their domain [17]. Recently there have been a lot of research in the field of XAI, e.g., [18]–[22]. Methods and concepts of XAI might be transferable to optimization-based ESM and thus could aid in explaining them. This transfer is motivated by the similarity between high-dimensional inputs in machine learning and ESM, where individual parameters may have minimal impact on the model's overall outcome.

In addition, some XAI explanation methods are scalable in complexity, catering to various target audiences, which is missing in the current ESM explanations. For instance, domain experts seeking to debug their models may require detailed, granular explanations to identify specific issues. On the other hand, DMs without technical expertise may desire a simplified explanation that elucidates the fundamental relationships within an ESM without requiring extensive prior knowledge. By providing explanations tailored to different audiences, we aim to enhance the accessibility and usability of ESM for a broader range of stakeholders.

We know from psychology and philosophy that causality is a crucial concept for explanation. One type of model designed to represent such causal relations are causal graphical models. Using these models as a substitute model for ESMs to represent the causal relations within the ESM could allow for easy explanations that are understandable by non-expert DMs.

A growing number of gamification can be seen in learning and educational games , called Serious Games [23]. Serious Games offer a variety of benefits in imparting knowledge compared to traditional learning methods, such as their often entertaining nature, fast feedback, competitive or cooperative elements, and the resulting engagement of learners [24]. Such a Serious Game might facilitate comprehension of concepts and challenges related to the energy transition. Enhancing players' understanding of various units of measurement of energy, demonstrating tasks and challenges within the energy and industry sectors, underscoring the significance of effective governance, and highlighting the interconnectedness of different sectors in achieving carbon reduction goals could be relevant learnings such a game could be able to convey. By better understanding magnitudes, players can answer approximate questions such as the number of CO_2 emissions produced by a coal power plant or an average household's annual energy consumption (in kilowatt-hours), encompassing thermal and electric energy requirements. Moreover, players' comprehension of CO_2 emissions, potential reduction opportunities, and associated costs in different sectors will be enhanced. Players might learn the interdependence of various stakeholders in achieving low carbon emissions through the necessity of interactions within the Serious Game without having to understand the math that models this behavior in an ESM.

1.1. Research Question

In order to create explanations for complex optimization-based ESM the following research questions (RQ) regarding explanations in general are of interest:

RQ 1. How are explanations currently created in the domain of ESM and are there major shortcomings?

RQ 2. How are explanations done in the machine learning domain and are there concepts that could improve ESM explanations?

Given there are explanation concepts in the domain of machine learning that could be beneficial for the energy system domain, the following questions emerge:

RQ 3. How can the explanatory concepts from machine learning be transferred to complex optimization-based energy system models?

RQ 4. Can the created explanations be adapted to suit different target audiences?

We will answer the following question regarding the use of directed causal graphical models to explain ESMs:

RQ 5. Can causal graphical models be used to explain ESMs?

As for using a simple, interactive energy system model as a Serious Game for explaining concepts of energy transition, the following research questions are relevant:

RQ 6. Which interactive game setup can explain the concepts of energy transition?

RQ 7. Can such a Serious Game help to achieve a set of learning goals regarding relevant concepts of power generation, important units and magnitudes, involved actors, and their conflicts of interest?

1.2. Contributions

To address the research questions above, we first review explanations and methodologies for creating them across various domains. Initially, we provide an **overview of explanations from the perspectives of psychology and philosophy**. This overview delves into the purpose of explanations, examines general definitions, explores their selection process, and identifies factors that contribute to their quality. Once we establish a foundation in the general concepts of explanations, we investigate the domain of optimization-based ESMs and conduct a **broad review** focusing on the explanations generated for these models and the corresponding methods employed. Further, we compare the explanation generation from the domain of optimization-based ESMs to explanation generation in XAI.

In response to these investigations, we then **propose a novel method for explaining ESM**, drawing inspiration from the XAI technique known as LIME (locally interpretable model-agnostic explanations) [19]. Our approach introduces LIME's concept of an interpretable feature space into the domain of ESM. We define interpretable feature spaces for two specific ESM and establish mappings to the models' inputs. By exploring variations around points of interest within the interpretable feature space and applying regression techniques to the resulting changes in the models' outputs, we generate explanations for multiple points of interest within the two ESMs. Subsequently, we compare these explanations with those produced by commonly used methods within the energy system modeling domain. Additionally, we evaluate the robustness of different mappings from the interpretable feature space to the input space for one of the ESM.

We also test an XAI approach that learns Causal Graphical Models from optimization-based ESMs. We find that there are fundamental limitations to this approach.

Regarding the research questions RQ 5 and RQ 6 of this thesis, which involves utilizing an interactive approach for explaining energy transition, we develop a **Serious Game corresponding to a country-level ESM from 2020 until 2050**. The developed Serious Game encompasses various aspects of the energy transition while constraining the decision-making scope to small decision steps. We **evaluate the Serious Game** with different groups of players and qualitatively assess their perceived learning outcomes related to different facets of the energy transition.

Parts of this thesis have been previously published by the author of this thesis:

2023
[25] Jonas Hülsmann, Julia Barbosa, and Florian Steinke. "Local Interpretable Explanations of Energy System Designs." *Energies 16.5* (2023): 2161.

2020
[10] Jonas Hülsmann, Florian Steinke, "Explaining Complex Energy Systems: A Challenge", *NeurIPS 2020 Workshop on Tackling Climate Change with Machine Learning*, 2020

The use of elements of these publications is not explicitly referred any more in the following chapters.

1.3. Structure of this Thesis

The remainder of this thesis is structured as follows. In chapter 2, we examine general concepts of explanations, as well as explanations in the domain of energy systems and potential improvements through the use of XAI. In chapter 3, we develop a method that uses XAI concepts and applies them to ESMs. For a different method of explanation, namely

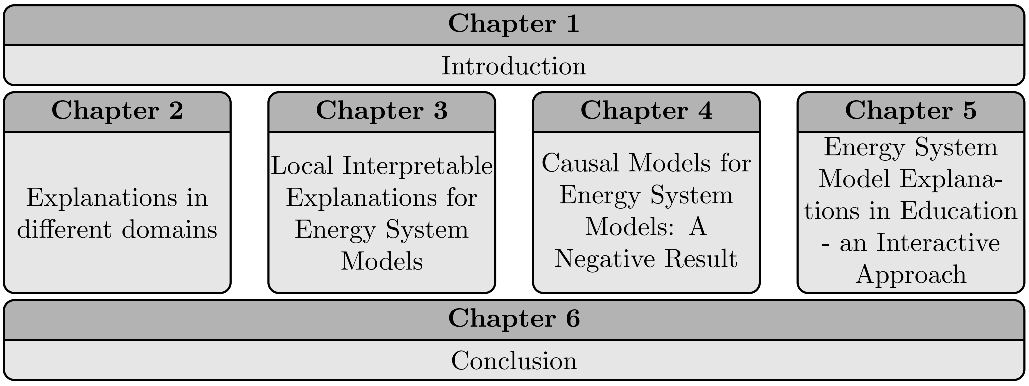

Chapter 1
Introduction
Chapter 2
Explanations in different domains
Chapter 3
Local Interpretable Explanations for Energy System Models
Chapter 4
Causal Models for Energy System Models: A Negative Result
Chapter 5
Energy System Model Explanations in Education - an Interactive Approach
Chapter 6
Conclusion

Figure 1.3.: Structure of this Thesis.

Causal Graphical Models, their benefits for explanation and their limited applicability on ESMs are discussed in chapter 4. In chapter 5, we create a Serious Game based on an ESM to explain concepts of the energy transition. A conclusion of this thesis can be found in chapter 6.

We now take a more detailed look at chapters 2 to 5.

Chapter 2 - Explanations in different domains. This chapter begins by exploring general concepts of explanations and universal principles from psychology and philosophy. It investigates the psychological aspects of explanation selection and the key factors influencing the perceived quality of explanations. This Chapter also delves into explanations in the context of optimization-based ESM and machine learning. It identifies the types of explanations commonly employed in the energy system domain, examines the methods used to facilitate their creation, and highlights potential limitations of existing approaches. Additionally, within the domain of machine learning, the Chapter explores the subfield of explainable AI, seeking concepts or methodologies that could enhance explanations within ESM.

Chapter 3 - Local Interpretable Explanations for Energy System Models. In this chapter, we introduce a method based on the concept of an interpretable feature space from XAI. This method addresses weaknesses identified in the previous chapter regarding explanation methodologies for optimization-based ESM. The chapter begins by applying the proposed method to a simple ESM, specifically a single-family house with a PV power plant, battery storage, heat pump, and heat storage. The robustness of the generated explanations is tested by evaluating different implementations of the mapping from the interpretable feature space to the input space. Subsequently, the method is applied to a more complex model of the German energy system, generating explanations for various transition pathways toward a low CO_2 emissions energy system.

Chapter 4 - Causal Models for Energy System Models: A Negative Result. This chapter discusses Structural Causal Models, and their potential benefits for creating explanations. We give proof that classical structured causal models cannot be applied to energy system model explanation. Further, we introduce an approach that allows to separate energy system models into subsystems that might be eligible for causal explanation.

Chapter 5 - Energy System Model Explanation in Education - an Interactive Approach. The chapter presents the development of an interactive simulation game designed to illustrate aspects and challenges of energy transition from the perspective of different agents. It defines the scope and limitations of the simulation game, including the various agents, their goals, and individual challenges incorporated into the game. The chapter introduces the interactive ESM, which defines the game's fundamental rules. Finally, a qualitative evaluation of the perceived learning outcomes is conducted using a test group of 25 participants.

2. Explanations in different domains

In this chapter, we delve into the topic of explanations in various contexts. We begin by exploring different concepts of explanations in psychology and philosophy in section 2.1. This section sheds light on what explanations are, how they are derived, and how their quality can be measured. Moving on, we analyze the state of the art in explaining optimization-based ESMs in section 2.2. Lastly, we examine the concept of model explanations in the domain of machine learning, also known as XAI, in section 2.3. In the discussion of this chapter, we qualitatively align the classification of XAI methods with the classification of current optimization-based ESM. While some of these XAI concepts are already being applied to data-based energy system models [12], we then demonstrate in the following chapter how the concept of XAI can also be extended to optimization-based energy system models.

2.1. Concepts of Explanation in Psychology and Philosophy

In this section, we will explore the role of explanations in psychology, including their purpose, definition, selection methods, and quality measures. The primary function of an explanation is to transfer knowledge, allowing individuals to construct mental models to understand past events, generalize properties, and make predictions about future events [26]. They serve as filters for possible causes and can change prior beliefs. Further, explanations can strengthen cognitive models and increase their stability [27]. Explanations can also facilitate the creation of shared meaning and influence an individual's emotions, thoughts, or actions [28], i.e., when explaining why somebody chooses a certain action in a situation over another action. Additionally, they can be used for a variety of purposes, including assigning blame, justifying a particular result or action, persuading someone that a given algorithm works appropriately (even if it does not), or even for mere aesthetic pleasure (e.g., explaining art) [29].

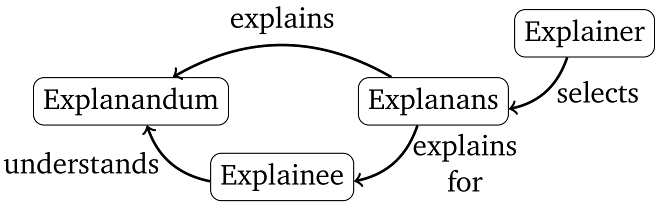


Figure 2.1.: Schematic overview of the elements involved in the cognitive process of deriving an explanation. Adapted from an image of [37].

Following Lewis's theory of causal explanation [30], the primary function of an explanation is to provide information about the causal history of an event and convey this information to someone else. While causality is a crucial element of explanation, there are also some non-causal explanations, such as answering descriptive questions like "What happened?" that ask for a sheer repetition or summarization of events without reasoning about how they come to be [31]. Causality has been a topic of interest in psychology and philosophy, with several theories proposed. One of the earliest works on causality is Hume's regularity theory of causation [32], which suggests that an event A is causally related to an event B if A is consistently followed by B. For example, the two events are causally related if turning on a switch is always followed by a light bulb starting to glow. However, Lewis [33] argues that the mere co-occurrence of events is insufficient to determine causal dependence. Instead, counterfactual reasoning is necessary to establish whether event A caused event B by considering what would happen if A had not occurred. Based on the counterfactual model, various competing definitions of causality have emerged. For instance, interventionist theories of causality state that event A is only considered a cause of event B if B can be changed by solely intervening on A [34]. Probabilistic theories suggest that A is a cause of B if bringing about A increases the probability of B occurring [35]. Other works have focused on classifying causes based on their properties, such as if they are necessary for an event to occur or if it is just sufficient for the event that one of many possible causes occur [36], or whether causes are internal or external to a system [27].

As argued by the authors of [26] and [38], explanations consist of the cognitive process of deriving the explanation, the final explanation as a product, and a social process of communicating the explanation. Figure 2.1 gives an overview of the elements involved in the cognitive process. The cognitive process involves finding possible causes for the event to be explained (explanandum) and selecting a subset from these possible causes as the explanation (explanans). Typically, an explainer creates the explanation for an explainee by selecting an appropriate explanans. Abductive inference [39], also known as "inference to the best explanation" [40], is often used to find causes. Abductive inference starts with observing the explanandum, such as a glowing light bulb. Then one or more hypotheses are made about the possible cause(s). For example, a hypothesis about a possible cause for the glowing light bulb could be a turned-on light switch, closing the electric circuit. The explainer judges the plausibility of the hypotheses and selects the "best" hypothesis as the explanation. We discuss different measures of "best" in the context of explanation selection and the various biases involved in this process in section 2.1.1.

The cognitive process of an explanation results in an explanans, i.e., the answer to a question. Questions can be classified into three categories: "What?", "How?", and "Why?" according to Pearl's *Causal Ladder* [41]. Answering "What?"-questions requires reasoning based on associations of unobserved events and observed events. For example "What event lead to the light bulb starting to glow?" can be answered by associating the observed events of a light switch being turned-on and the bulb starting to glow shortly after. "How?"-questions require interventionist reasoning to be answered, i.e., which causes would need to change to undo the event. "How could the glowing of the light bulb be prevented?" can be answered by undoing a past event, i.e., "by not turning-on the light switch." "Why?"-questions are the most challenging since they require counterfactual reasoning to consider events that could have happened. There can be different answers to the same question depending on the mode in which the question is asked. Aristotle's *Four Causes* [42] suggests that a "Why?"-question can be answered in different explanatory modes, such as material, formal, efficient, and final causes. For example, the question "Why is the light bulb glowing?" could be answered with "because it is made out of conductive metals" (material), "because the metals were turned into a thing wire that emits light when heated" (formal), "because somebody assembled the light bulb" (efficient/agent) or "because somebody needed light to read a book" (final). [43] argues that different target audiences require different explanans since they have different goals for the explanation. For instance, when creating an explanation for stakeholders of a credit scoring model, a different explanans is necessary to enable a bank employee to work with the system than for a customer to understand why a credit loan is not granted.

The process of explaining involves more than just providing an answer to a question. It also involves the social process of communicating the explanation from the explainer to the explainee, which typically occurs in the form of a conversation. Hilton's conversational model of explanation [44] emphasizes that an explanation must be relevant to the question asked and take into account the previous knowledge of the explainee. Following Grice's *maxims of conversation* [45], an explainer should only provide information that is known not to be wrong (maxim of quality) and limit the information to what is necessary for the given question (maxim of quantity). Additionally, the explainer should stick with the question asked (maxim of relevance) and communicate the answer politely and understandably (maxim of manner). Overall, a compelling explanation requires a clear and accurate answer, an understanding of the context in which the explanation is being provided, and an ability to communicate that answer effectively to the explainee. It is important to note that not all causes are relevant for explaining an event, even if a cause has a high probability of being true. For example, if asked "Why is this light bulb not working?" the cause of "it has a technical defect" has a higher probability than "its filament burnt-out" since the burnt filament is a sub-class of technical defect, but the general technical defect is probably already assumed by the explainee.

2.1.1. Explanation selection and evaluation

Both [26] and [30] highlight the importance of selecting a suitable subset of possible causes as a significant concept for an explanation. Limited human comprehension necessitates simplifying the causal structure of an explanandum. As argued by [46], an explainer can employ causal backgrounding or discounting to reduce the number of relevant causes. When discounting a cause, it is considered less relevant compared to other causes based on additional information. For instance, for the question "Why is the light bulb glowing?", both "Because she wanted to read a book." and "Because she forgot to turn it off." could be possible answers. The explainer needs to assess which answer is more plausible by using additional information, such as "She reads a lot." which makes the first answer more credible. On the other hand, a cause is backgrounded when it is considered a requirement or supporting factor for the primary cause. For instance, "Because electricity is running through it." is a part of the explanation of a glowing light bulb, but it might be backgrounded for the causal answer "Because the light switch was closed." which assumes the former.

According to [44], explanations are often framed in contrast to some other event rather than on their own. For instance, "Why is the light bulb glowing?" implies a contrasting case, such as "Why is the light bulb glowing rather than being turned off?". This influences the selection of causes since providing explanations to such contrastive questions is often easier than offering a complete causal explanation because only the causal difference between the two events needs to be explained [47]. However, it is crucial that the explainee understands the contrastive case and that the explanation only includes causes that differ between the two events [48]. For instance, the explanation "Because somebody wanted to read a book." would answer the question "Why is the light bulb glowing?" if the implied contrastive case is "..., rather than being turned off.". However, if the contrastive case is "..., rather than opening the window shutter.", the explanation is not considered good because opening the shutter and turning on a light bulb could help with reading the book.

Explainers prefer abnormal causes over normal ones as the default [49], e.g., a broken glass bulb leading to the burning out of a lightbulb's filament is preferred as an explanation over the oxygen in the atmosphere making the glass bulb necessary in the first place. Other factors that can influence the selection of causes include their necessity or sufficiency to explain an event [47], [50], perceived responsibility for the outcome [51], and whether a cause leading to an event was intentionally caused or not [52]. Additionally, explainers often choose to explain events through a conjunction of causes [53], which aligns with the *conjunction fallacy* [54] where people consider the probability of two events occurring together to be more likely than either event occurring alone if the conjunction is consistent with their prior beliefs. For example, "The light bulb is glowing because she wanted to read a book and it was dark." may be considered the "best" explanation even though it is less likely than just the darkness or the intention to read being the cause. However, when explaining failed actions, a single-cause explanation is mostly preferred [53].

According to the *theory for explanatory coherence* [55], coherence is the key criterion for an explanation to be considered "good" by the explainee. Coherence requires that an explanation is consistent with the prior beliefs of the explainee. When using multiple causes in an explanation, these causes must be coherent and related. Simpler and more general explanations, which use fewer causes and explain more events, are preferred if they are coherent [56]. While probability plays a role in selecting an explanation, the best explanation is not always the most probable cause [57]. For instance, when explaining why a light bulb is not working, saying "Because the electric circuit is open." is more probable than saying "Because the light bulb has a defect." However, the former can be assumed to be already known by the explainee so that it can be backgrounded in the explanation. People's goals can also influence how they perceive explanations, as demonstrated by [58]. Participants in their experiments rated explanations given in different explanatory modes. Those tasked with a specific goal, like classifying an organism or finding causes for certain traits preferred explanations with explanatory modes that matched their goals rather than those explanations with a divergent explanatory mode. Further, [29] argues that different domains require different types of explanation. For example, explanations in biology might differ from those in physics or sociology.

This section analyzed explanations in the context of psychology, cognitive science, and philosophy. Among these concepts, the significance of causality emerges as a pivotal insight, a theme extensively employed throughout this thesis, particularly in Chapter 4. Numerous methods and tools discussed in the subsequent sections of this chapter, specifically those addressing explanations in the domains of ESM and XAI, inherently leverage some form of abductive inference to deduce the causal factors essential for their explanations. The notion of counterfactual reasoning resurfaces in one of the methods outlined in the XAI sections of this chapter. Finally, the act of communication as a vehicle for explanation serves as the driving force behind the exploration in Chapter 5 of this thesis.

2.2. Explanations of Energy System Models - State of the Art

In this section, we present a comprehensive review of explanations within the domain of ESMs. While there are numerous reviews available on ESMs covering different aspects such as methodologies [59], available frameworks [60], correlation of input data sets [61], and the impact of climate change on energy systems, and their representation in current ESMs [62], our focus is on methods and tools domain experts use to explain the outputs of optimization-based ESMs or to enhance the models' comprehensibility. It is important to note that our review concentrates explicitly on the methods and tools employed to derive explanations rather than the explanations themselves. We exclude methodologies such as principle component analysis, correlation analysis, or aggregation approaches like those mentioned in [63], as they primarily serve to explain patterns or relationships in the input data of ESMs and simplify them rather than explaining the behavior of the ESMs. For an extensive review of time series aggregation approaches for reducing ESM complexity, we refer readers to [64]. Although specific methodologies are more common than others, we aim to cover as many different approaches for creating explanations as possible, even if they are not used very often.

Figure 2.2 shows our categorization of explanation methods used in optimization-based ESMs, referencing their uses in the energy system domain. We show a few examples of each category, but do not claim completeness of the references. We distinguish between methods that create explanations by varying the model's input parameters and those that do not involve input variations. The first class of explanatory methods, which explores the model's output sensitivity to specific parameter changes, is referred to as sensitivity analysis. Based on the type of input variation, we identify three sub-classes of sensitivity analysis:

1. Parameter Analysis: A small set of input parameters is varied extensively to observe its impact on the model's output.

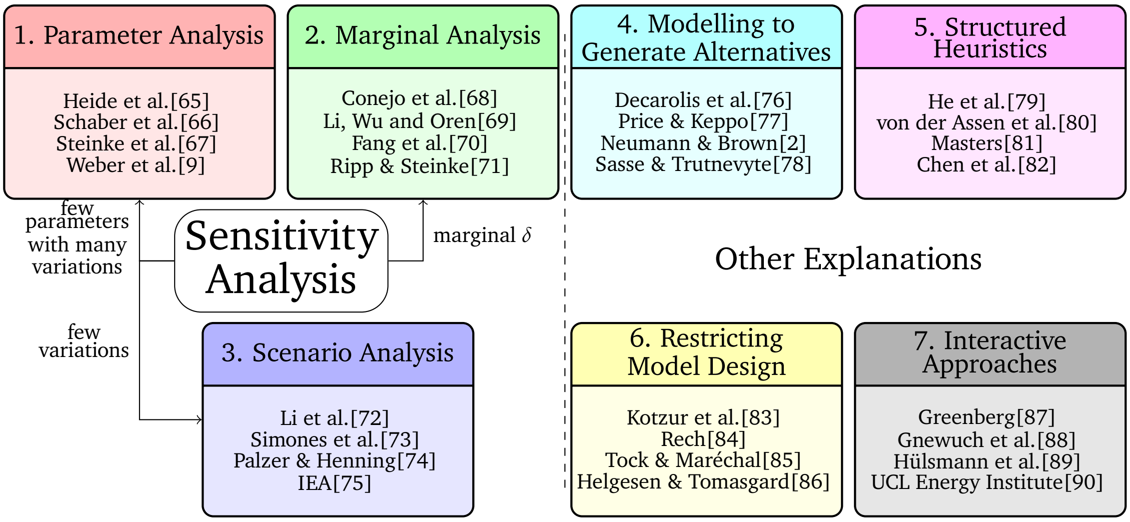


Figure 2.2.: Different classes of methodologies to create explanations for optimization-based ESMs. We differentiate between explanations created by input variations, i.e., some variation of sensitivity analysis and other explanations that do not consider input variation.

2. Marginal Analysis: The parameter sensitivity derived from the dual variables of the optimization, making the explanation only valid for marginal changes, i.e., infinitesimally small changes. However, all parameter sensitivities are available at once without additional computational effort.

3. Scenario Analysis: Multiple parameters are varied simultaneously in large steps through expert-designed scenarios to assess the model's behavior.

Regarding the explanation methods not based on input variation, we identify four classes:

4. Modelling to Generate Alternatives: This approach seeks solutions that are maximally different from an optimal solution within a tolerated slack from the target value, e.g., it minimize the use of wind turbines at a maximum 10% increase in total system cost.

5. Structured Heuristics: Certain parts of the ESM are substituted with heuristics, which can be more easily explained, although potentially sacrificing optimality in the solution.

6. Restricting Model Design: This class includes design guidelines or substitute models that aim to simplify or replace complex ESMs with models that are more understandable to DMs.
7. Interactive Approaches: Methods and tools falling under this category facilitate knowledge transfer about energy systems through interaction with the models, similar to the explanation through a conversation concept from the previous section.

2.2.1. Sensitivity Analysis

Parameter Analysis

We classify methods that assess a model's output sensitivity on a very detailed level to a limited set of parameters as parameter analysis. Unlike other sensitivity analysis methods, parameter analyses extensively inspect the effect of a model's parameters by utilizing various parameter variations or even analyzing them over a continuous interval. However, due to the computational expense of solving a model with numerous parameter variations, parameter variations are typically limited to a few parameters since those methods usually scale exponentially in the number of varied parameters. In the case of continuous analysis, simpler models that can be solved analytically for the varied parameter are often used. Varying less than three parameters allows for easy visualization the effects of parameter variations as a graph.

In [65], the possibility of a 100% renewable generation scenario in Europe is investigated. The authors compare historical normalized generation data for solar and wind power generation to the normalized power demand, aiming to find the optimal renewable generation mix. The authors employ two approaches to define the optimal mix of renewable generation. The first approach defines the optimal mix as the combination of wind and solar that minimizes the deviation between generation and demand. The second approach uses the maximum power demand and production mismatch for a given mix, corresponding to the storage capacity needed for this combination. To explain their results, the authors parameterize their simple model based on the relative share of wind generation and showcase the solution as a function of this single parameter. In a second case study, the authors investigate transition scenarios, incorporating time-independent power generation from fossil-nuclear fuels into their model. For this extended model, the authors explain the transition paths that minimize storage capacity by performing a two-parameter variation for the share of fossil-nuclear to renewable generation and wind within a renewable generation.

Similar parameter analyses are performed in [66] and [67], both utilizing the same dataset of renewable generation potential as used in [65]. Parallel to the second case study of [65], [66] analyzes the effect of the relative share of variable renewable energies (wind and solar) and the share of wind within a renewable generation. However, [66] takes this analysis further by examining the effect of these two parameters on the capacity of renewable generation, necessary backup capacities, overproduction, the sum of backup power, and several cost parameters. The authors analyze these additional model outputs for two different setups: one without a grid and one with optimal grid extensions. Furthermore, they employ a more comprehensive energy system model described in [91].

In [67], the authors introduce storage into the mix to investigate the optimal combination of storage and grid investments in achieving a fully renewable energy generation Europe while minimizing the required backup energy. The analyzed parameters are the grid size and storage capacity. The authors assume fixed renewable generation capacities of 65% of the average demand for wind and 35% for PV, as determined to be optimal in [65]. In a second scenario, the same generation share but a total of 130% of the average demand is assumed. However, the production capacities are unevenly distributed since regional potential constraints limit them. The authors use the regional distribution to model the grid size by aggregating renewable generation within a certain radius of each location. With a larger grid, a mix of wind and PV generation closer to the optimum can be accessed. In a second case study, the authors perform another parameter analysis using the same parameters but for the resulting electricity price for different types of storage.

A completely different setup is used for the parameter analysis in [9], where the authors demonstrate that ESMs can behave counter-intuitively under certain conditions. They inspect a simple ESM with two types of fuels for electricity generation: lignite and gas. Lignite has low costs but high CO_2 emissions, while gas is more expensive but emits less CO_2. The authors examine two optimization problems: The first optimization minimizes costs while meeting certain CO_2 emission restrictions. The efficiency of the gas power plant is varied as the parameter. Surprisingly, increasing the efficiency of gas power plants, thereby reducing their CO_2 output, results in less gas usage in the cost optimization. This counter-intuitive outcome occurs because the lower CO_2 emissions per energy unit of gas allow some power generation to be replaced by cheaper coal power plants, leading to additional cost savings.

In the second optimization, CO_2 emissions are minimized while adhering to certain budget limits, revealing a similar counter-intuitive effect. The authors introduce a CO_2 tax and vary its magnitude as the parameter. Raising the CO_2 tax might be expected to increase the usage of gas power generation due to its lower CO_2 emissions. However, the opposite occurs because the CO_2 tax impacts both fuels. Part of the power generation from gas must be replaced with coal power generation to satisfy the electricity demand while staying within budget limits.

Marginal Analysis

We refer to methods that create explanation based on local sensitivities derived from infinitesimal changes as marginal analyses. The examination of marginal changes in variations locally around a solution characterizes these analyses. One of the critical advantages of marginal analyses is their computation efficiency, as they often do not require solving the model multiple times, at least for common convex optimization formulations [92], but allow to extract the marginal from the solution process [68]. However, analytically calculating marginal sensitivities necessitates access to a model's constraints, which renders this approach impractical for closed-source models. Furthermore, the sensitivities extracted directly from a model's optimization process describe only the relationship of the inputs to the optimization objective. If relations to other model outputs are of interest, additional computations are required, as in [68], where the marginal input sensitivities are computed with respect to energy demand. Additionally, marginal sensitivities and any explanations derived from them apply only to infinitesimal variations around the local solution or their validity is zero.

In energy systems, marginal utilities or marginal costs, sometimes called shadow prices, play a crucial role. They are employed to determine electricity prices in specific regions and assess the electricity mix's greenhouse gas emissions over time. The method employed for determining the regional electricity price using marginal production costs is known as locational marginal pricing [93]. Locational marginal pricing is actively used by regional transmission organizations like PJM [94] and New York ISO [95]. By adopting various ESMs, locational marginal pricing can be expanded to incorporate distributed resources such as electric vehicles [69] or uncertainties in renewable energy generation [70]. Furthermore, apart from determining electricity prices in real-world energy systems, the concept of marginal pricing provides valuable insights into the behavior of the ESM utilized for these calculations. I.e., a marginal analysis effectively demonstrates each model variable's impact on the ESM's objective function.

Another application of marginal sensitivities lies in assessing greenhouse gas emissions, such as determining the emissions caused by an additional electric application. Calculating marginal emissions is often more challenging than computing local marginal prices, as emissions are typically outside the objective function of an ESM, making the direct use of the dual variables inapplicable. While [71] proposes an analytical approach for deriving marginal emissions from the ESM via the Karush-Kuhn-Tucker conditions, other publications predominantly resort to alternative methods, such as regression [96]–[98] or merit order heuristics [99].

Scenario Analysis

The class scenario analysis involves comparing model results for a finite (typically small set) of distinct input parameters and conditions, so called scenarios to provide explanations. A scenario is not restricted by the magnitude of changes like a marginal analysis, nor the number of parameters altered, as opposed to parameter analysis. Scenarios are commonly used to model uncertainties related to future events or parameters. The advantage of scenario analysis lies in its model-agnostic nature, making it applicable even to closed-source models since it does not require interaction with the model's internal functioning. Additionally, scenarios can encompass various aspects, including model inputs and different ESMs utilizing the same inputs, as demonstrated in [72].

Incorporating uncertainty into scenarios also find applications in stochastic optimization methods [100] and robust optimization [101]. Stochastic optimization seeks to find decision variable configurations that optimize expected performance based on a given distribution of scenarios. For instance, in energy systems, this could mean identifying an energy system configuration that minimizes the expected costs for a set of scenarios with a specified probability distribution. On the other hand, robust optimization aims to find a variable configuration that optimizes a target function while ensuring that a set of hard constraints remains valid for all scenarios in a given uncertainty set. For example, this could involve finding an energy system configuration that can withstand extreme scenarios without failure. Although these optimization methods share similarities with scenario analysis in using scenarios to encode uncertainty, they differ in their ultimate objectives [101]. Stochastic optimization and robust optimization focus on finding optimal solutions, whereas scenario analysis aims to explain the behavior of an ESM. Consequently, despite the everyday use of stochastic and robust optimization in the domain of energy system models, e.g., in [102] or [103], we do not consider them as methods of explanation.

Explanations derived from scenario analysis heavily depend on the scenario design, particularly the reference scenario, and necessitate an explanation of the scenarios themselves. This requirement stems from scenario analysis's capability to compare scenarios that affect entirely different sets of parameters. For example, one scenario may affect the investment costs of heat pumps, while at the same time, another may impact the efficiency of wind power plants, the investment costs for battery storage, and the operational costs for gas power plants. Comparing such scenarios can be misleading, as they involve different parameters, varied amounts of parameters, and potentially different magnitudes of effects. To improve explanations in the case of heterogeneous scenarios, one can increase the number of scenarios, introducing variations between scenarios to provide a more comprehensive picture. However, this comes with a trade-off, as the computational effort required for scenario-based methods depends on the model's runtime and the number of scenarios considered [102]. Typically, a smaller number of well-described scenarios is chosen, as demonstrated in [73], [74], and the reports by the International Energy Agency (IEA) [104], [105], [75].

For instance, the authors employ scenario analysis in [73] to explain various technology deployment paths and identify their most common technologies for an european ESM implemented in the TIMES framework [106]. They compare eight scenarios, including a business-as-usual scenario, a scenario with additional CO_2 emission restrictions, and six scenarios based on the second one, each allowing for different technology deployments or restrictions. The authors use the the scenario with the CO_2 emission restriction as their reference scenario. They show the optimal technology capacities for other scenarios relative to the reference. Moreover, they conduct a sensitivity analysis to explain the impact of different price and technical parameters for the business-as-usual scenario, effectively treating it as a scenario analysis due to the variations introduced.

Similarly, in [74], the authors investigate if a German energy system with 100% renewable energy generation can meet the energy demand. Further, the authors calculate the associated costs and determine the cost-optimal combination of technologies in their ESM [107]. They utilize three selected scenarios that build upon each other to explore and explain their model's behavior. The scenarios differ in heat demand reduction, fixed heat pump share, and limited wind-onshore and wind-offshore capacity, providing a detailed cost comparison for each technology in the scenarios.

The IEA also uses scenario analysis in their flagship reports such as "World Energy Outlook" [105], "Energy Technology Perspectives" [75], and "Net Zero by 2050" [104]. These reports include dedicated chapters explaining the scenario assumptions. The IEA employs three scenarios:

- a stated policies scenario based on existing or announced government policies
- an announced pledges scenario that includes long-term commitments made by governments worldwide
- a net-zero emissions scenario aiming for carbon neutrality by 2050

The stated policies scenario serves as a benchmark, while the announced pledges scenario highlights the "ambition gap" between commitments and the goal of limiting global warming to 1.5°C. The net-zero scenario represents a highly ambitious pathway toward carbon neutrality.

2.2.2. Other Explanation Approaches

Modelling to Generate Alternatives

A technique employed to explore and explain the behavior and flexibility of ESMs, aiming to find alternative energy system with close to optimal system costs, is known as Modeling to Generate Alternatives (MGA) [76]. In this approach, an optimal solution of an ESM serves as a starting point, and potential solutions are sought that exhibit maximum differences in specific output values, such as technology capacities. To achieve this, MGA reformulates the ESM. The original target function becomes a relaxed constraint. For instance, minimizing total system costs becomes a constraint that restricts total system costs to be not more than the optimal system cost, plus some slack. Simultaneously, MGA introduces a new optimization target, which maximizes or minimizes a specific aspect of the ESM output, such as power generation by wind power plants. All other constraints of the original ESM remain unchanged in the reformulated version. Various extremes can be found by formulating different optimization functions, allowing the flexibility of the ESM's solutions concerning relaxed optimal values to be explained. However, a drawback of this approach is the need for access to the model formulation and the computational costs, which increase depending on the number of alternatives to be generated.

In their study [76], the authors apply an iterative MGA approach to an ESM that includes both the electricity and light-duty transport sectors. The ESM is based on the Temoa open-source framework [108], using U.S. data. The iterative MGA approach employs an objective function that minimizes nonzero decision variables from the previous iteration by assigning them a higher weight and treating previous iterations' solutions as constraints. Variables that remain nonzero increase their weight in the objective function with each iteration, incentivizing their reduction in subsequent iterations. The outcome of this iterative process is a sequence of potential energy systems, with each new system being as dissimilar as possible in the decision space from the previous ones. Three scenarios, one purely cost-minimizing and two with -40% and -80% CO_2 emissions, serve as reference scenarios to compare their MGA results. The authors create variations with the iterative MGA based on the 40% emission reduction ESM, allowing for different slack levels in the objective value (1%, 2%, 5%, or 10%) and performing four iterations, generating a total of 16 MGA-generated energy systems. In a subsequent analysis, the effect of different discount rates on the decision space is explored for the 2% objective slack scenario.

A similar iterative MGA approach is employed in [77] to generate alternatives for a business-as-usual scenario and a 50% emission reduction scenario. The authors use an ESM based on the TIMES model [7], representing the global energy system across 16 aggregated regions. In contrast to [76], the authors maximize the L1 distance from the previous solution instead of increasing the weight for individual parameters. The authors apply their method for a cost slack of 1%, 5%, or 10% of their objective.

The authors [2] use a non-iterative approach to MGA to explore near-optimum energy systems aiming for an 80% emission reduction in Europe. They introduce an ESM that minimizes system costs while considering multi-period optimal power flow constraints. Their ESM is implemented in the PyPSA framework [109], focusing solely on electricity. Instead of employing an iterative search, the authors use a fixed set of 16 objectives, such as the minimization or maximization of onshore wind turbines, offshore wind turbines, both wind turbines at the same time, solar panels, gas turbines, hydrogen storage, battery storage, and transmission infrastructure. They consider different relaxations of the optimal value, ranging from 0.5% to 10%, and various greenhouse gas emission reductions of 80%, 95%, and 100%. The authors visualize feasible regions for each technology under different reduction targets and optimal value slacks. Furthermore, they illustrate their results to demonstrate no-regret grid extensions, i.e., extensions that are present in all scenarios. Lastly, the authors utilize their findings to analyze the correlation of technology capacities in the solution space of their model.

Structured Heuristics

We categorize approaches that solve ESMs not by utilizing optimization algorithms but by applying heuristics as structured heuristics. While their solutions may not always be optimal, they are often close to the optimum. The advantage of these structured heuristics lies in their ease of understanding, which can aid in explaining the results of an ESM. However, since the heuristic bypasses the optimization model, the explanation does not directly illuminate the ESM's behavior.

In the energy system domain, two commonly used structured heuristics are the merit order, sorted duration curves and, marginal abatement cost curves [80], [110], [111]. The merit order represents a list of items sorted based on a quantity measure. In energy systems, the merit order could refer to potential electric energy suppliers arranged according to their ascending marginal production costs in an electricity market. It can be visualized as a merit order curve, depicting the increase in marginal production costs over accumulated production. While this approach provides a simple model for determining electricity prices [79], it is not globally optimal as it neglects transmission costs and may require a re-dispatch to satisfy energy system constraints. A different example for using merit order curves in the energy system domain is given by [80]. The authors rank CO_2 sources by their marginal environmental impact to identify the most beneficial sources for utilization in carbon capture and utilization.

Load duration curves, also known as cumulative distribution functions of energy demands, offer another method for gaining insights about ESMs through structured data [81]. The ESM's input and output time series are sorted in descending order and represented as a graph, creating a load duration curve. This curve provides valuable information, such as peak load and capacity utilization rates, allowing quick access to critical metrics. By analyzing the shape of the load duration curve, characteristics of power plants, such as whether they are mainly used for peak load provision, can be deduced. For example, in [110], the authors use load duration curves to explain an energy system consisting of diesel generation and wind power plants.

Marginal abatement cost curves provide insights into the potential greenhouse gas emission reductions achievable through various measures, visually represented in a sorted bar plot. Each bar on the curve corresponds to a specific emission reduction measure, with the bar's width indicating the potential reduction and its height representing the cost per unit of emission reduced. These measures are arranged on the curve based on their costs per unit of emission reduction, facilitating easy identification of the costliest measures, similar to a merit order curve. However, it's important to note that while marginal abatement cost curves offer clarity on costs, they do not provide information on the implementation duration of chosen measures nor do they reveal potential synergies between measures [111].

A framework introduced in [82] combines scenario analysis with life cycle assessment to help DMs understand the environmental impact of different energy system design scenarios through a structured visualization approach. The framework consists of four sections: scenario setting, inventory analysis, visual analysis, and decision-making. In the scenario setting phase, relevant scenarios are selected, determining which technologies should be considered. The inventory analysis is then conducted to assess the potential capacity of each selected technology and its environmental impact, such as greenhouse gas emissions, through life cycle analysis. For the structured visualization the technologies are ranked from least to most environmentally harmful, and their environmental impact is visualized as a curve against a measure of interest, such as the produced energy, to show the lower bound of environmental harm. Another curve is drawn using the inverse order, i.e., the most harmful technology first, to demonstrate the upper bound of environmental harm. The area between the curves depicts all possible combinations of environmental impact with the selected set of technologies for the scenario. The distance between the curves indicates the radius of operation for a given desired output value. The framework then employs linear optimization with a two-dimensional parameterized target function to conduct a locally focused MGA at the desired output and provide additional visual insights into possible energy system designs. Using this approach, the authors demonstrate their structured visualization approach for a case study on the environmental impact of Japan's energy system.

Restricting Model Design

We classify approaches that limit the design space of the ESM, either by restricting its mathematical model or by using alternative types of models to replace ESMs as "Restricting Model Design". These approaches do not create explanations itself. However, design principles or substitute models can partially replace the need for an explanation by reducing model complexity, increasing model transparency, or moving towards a different type of model that is more familiar to a DM thus increasing their trust in it [112]. For example, while a model domain change may not make the explanation process easier, some DMs might be more familiar with economic models, potentially increasing their trust in their results.

In a study by [83], the authors provide a qualitative guideline for designing ESMs with a focus on reducing complexity. They emphasize four key design aspects to achieve this. Firstly, they suggest systematically reducing the model size by using aggregated input data and lower time resolution or simplifying the grid model, i.e., assuming a copper plate as the grid. Secondly, they recommend using convex models whenever possible to ensure a single global optimum. While linear models guarantee convexity, they may sacrifice accuracy depending on the initial formulation. The third aspect involves quantifying errors by employing relaxed problem formulations that act as bounds to the original model. Lastly, they propose decomposing an ESM into sub-problems that can be solved individually, enabling parallelization and speeding up the solution process.

For optimizing fleets of combined-heat-and-power units (CHPs), [84] presents a more specific design guideline. They advocate using a base formulation for each CHP in the fleet, focusing on power streams and energy quantities while neglecting mass flow rates, pressures, and temperatures without significant accuracy loss. In cases where non-dispatchable energy generation is part of the energy system, it can be solved independently in advance and used as an input to the ESM rather than integrated within it.

Furthermore, using well-established ESM frameworks like TIMES [7], OSeMOSYS [113], or PyPSA [8] can be considered as falling under the category of design guidelines, provided they are well-established and accepted by the DMs. Many open-source ESM frameworks offer full data transparency and cater to different use cases related to planning horizon, regional distribution, and uncertainty, making them suitable for serious applications [114]. However, setting up and configuring ESM frameworks requires domain expert knowledge. Therefore, despite the reduced need for explaining the ESM itself, explaining the configuration remains relevant, as a misconfiguration could lead to unexpected behavior.

Examples of the use of substitute models to solve problems in the energy system domain are economic models. In [85], an ESM is combined with an economic model to aid decision-making for a multi-objective energy system. More specifically, the authors use Pareto-frontiers identify the best process configurations for power plants with CO_2 capture, considering multiple competing objectives such as cost minimization, emissions reduction, and efficiency maximization. Similarly, in [115], Pareto-frontiers are utilized to visualize the cost-optimal decision pathways for reducing CO_2 emissions using various technologies. These frontiers present solutions of the ESM that offer the same value concerning a utility function determined by the DM.

In contrast, [86] combines an ESM with the Nash-Cournot competition model for power markets. The Nash-Cournot competition model seeks a Nash equilibrium, i.e., a set of decisions where no individual DM can improve their gains without changing the decisions of others. The general idea to split a problem into small agent-driven sub problems, that is found in some economic models is revisited in Chapter 4.

Interactive Approaches

The last category of explanation methods we explore is explanations by interaction. As the name suggests, interactive explanations rely on the interaction between the explainee (target audience) and the ESM itself. This approach aligns with "explanation as a form of communication," as discussed in the previous section, or more informally, "learning by doing". Interactive approaches can also enhance other explanation methods. For instance, in [87], an approach is presented to interact with sensitivity analysis and generate textual explanations based on the results. This interactive tool for sensitivity analysis is not limited to the ESM domain and can be applied to all linear models.

In the domain of ESMs, conversational agents, such as chatbots, have been utilized in [88] and in our previous work [89] to provide interactive explanations and answer user questions. In [88], the authors identify design principles for creating a chatbot focused on energy feedback. Their chatbot should be able to assist users in analyzing their energy usage and helps them identify ways to adjust their consumption behavior, e.g., to reduce peak load. The authors evaluate these design principles through expert interviews, concluding that an effective tool should provide easily accessible and understandable information through natural language, reduce the user's effort to learn sustainable energy usage, provide explanations for specific events, and be engaging for users to interact with. An implementation of a chatbot for ESM interactions can be found in [89]. This chatbot serves as a natural language interface for the ESM, enabling users to pose "what-if" questions that are transformed into input scenarios for the model. Users can interactively explore answers to their questions from the scenario results visualization.

Another form of interactive explanations comes in the form of serious games [23], which are games designed for educational purposes, training, or simulation of real-world scenarios. These games employ principles of game design, such as feedback loops, to maintain an entertaining and engaging user experience [24]. An example of a serious game in the energy domain is "Watts the deal?" [90], which focuses on peer-to-peer energy trading. Players assume the role of prosumers equipped with energy storage and rooftop solar panels. Each player must meet randomized energy demands each round, and they can generate electricity using their rooftop solar panels, purchase it from the grid, or engage in peer-to-peer trading with other players. The game offers various versions with different objectives, such as maximizing personal profit, achieving individually assigned goals, or cooperating to maximize the community's self-consumption, ensuring replayability and varied learning experiences.

2.3. Explainable AI

Much like the domain of energy system modeling, machine learning grapples with the imperative need for model explainability. While optimization-based ESMs often lack comprehensibility due to their sheer size and resultant complexity, they maintain transparency since experts meticulously select all model parameters based on data, rendering them akin to white-box models. In contrast, machine learning models use a different approach, wherein parameters are learned from data rather than being presided over by experts. Consequently, these models introduce an additional opacity layer, making them black-box models. Even domain experts often need help to discern the meaning of individual parameters within these models.

Despite their occasional super-human performance, the interpretability or explainability of machine learning models becomes an important legal factor for their employment in real-world scenarios, especially in fields such as security, health, and finance [116]. The necessity for interpretability and explainability is important to ensure that their exceptional performance does not hinge on biases within the training data. This need for understanding the models can come to the extent that in specific scenarios, models with explainable behavior may be favored over opaque models with superior performance, as discussed in [117].

Notably, despite their frequent use in the XAI community, the terms "interpretability" and "explainability" lack precise definitions, as highlighted by [117]. While some authors employ these terms interchangeably, we stick with the distinction of the therms drawn by [118]:

- **Interpretable machine learning** involves using simpler models that inherently reveal their function, such as decision trees.
- **Explainable machine learning** leverages a secondary model to help elucidate the behavior of black-box machine learning models.

[118] advocates for interpretable machine learning over post-hoc explainable machine learning for black box models. This preference arises from the inherent difficulty of ensuring fidelity in post-hoc explanations, i.e., guaranteeing that the explanation aligns faithfully with the actual behavior of the model explained.

This work aims to transfer knowledge from XAI to elucidate a class of white box models, namely optimization-based ESMs, which in principle are interpretable but often lose their interpretability due to their sheer model size and resulting complexity. Consequently, this section predominantly focuses on explanation rather than interpretability.

The remainder of this section is structured as follows:

1. **Desired properties of XAI methods and taxonomies**: We begin by discussing the desired properties of XAI methods and explore various taxonomies for classifying them.

2. **Examples of XAI methods**: Next, we look at five popular XAI methods and classify them using one of the established taxonomy approaches.

In the following section, we align existing ESM explanation approaches with the XAI taxonomy and identify open potentials for future explanation approaches.

2.3.1. Desired Properties of XAI Methods and Taxonomies

The list of desirable properties for explanations generated by machine learning methods, as outlined by [119], can be categorized into two main groups: properties related to the method itself and properties related to the quality of the explanations produced.

When it comes to the method itself, these properties encompass algorithmic complexity, the adaptability of the method to various types of models, the level of transparency (i.e., whether the method operates with model internals or treats the model as an opaque black box), and the expressive power, which refers to the type of output generated by the XAI method.

Regarding the quality of explanations, [119] identifies several desired properties. These include **stability**, which entails providing similar explanations for similar instances; **fidelity**, ensuring that explanations accurately reflect the actual behavior of the explained model; and **accuracy**, meaning that the explanations generalize to other unseen instances. While an explanation method should create stable explanations for similar instances, it also should create similar explanations for similar models (**consistency**). In the context of complex models, it may be beneficial to break down the model into distinct sub-systems and explain them individually. Thus, it is essential always to specify which part of the model is being explained (**representativeness**). Furthermore, some machine learning models can quantify their certainty about a prediction. In this regard, it is desirable for an explanation to convey whether the explained model is confident in its prediction or not (**certainty**). Similarly, when explaining instances from regions not well represented during model training (i.e., novel instances), a statement about the model's certainty in such cases becomes crucial (**novelty**). In situations where explanations involve multiple elements with varying impacts on the model's output (e.g., a set of parameters), the explanation should indicate the importance of each element (**degree of importance**). Lastly, **comprehensibility** is another crucial property. It pertains to the size and readability of the explanation, ensuring that it is digestible by the intended audience.

As the list of desired properties covers attributes applicable to all XAI methods, additional properties are required to effectively differentiate between them. In a review of different taxonomies for XAI methods provided by [120], four distinct approaches to categorizing XAI methods are identified:

1. **Function-Based Approaches**: These taxonomies categorize XAI methods based on the concepts they employ to extract information for creating explanations. Concepts include perturbations to input data, utilization of structural properties of the original machine learning model, modifications to the model's structure, use of illustrative examples, and meta-explanations that aggregate explanations from other XAI methods.
2. **Result-Based Approaches**: Result-based taxonomies differentiate XAI methods based on their yield result types. These result types may include generating surrogate models that are more comprehensible than the original machine learning model, ranking the relevance of model features, or providing input-output pairs that exemplify the model's behavior.
3. **Conceptual Approaches**: These taxonomies classify XAI methods based on dimensions such as the scope of the explanations generated, the stage of application (post-hoc or ante-hoc), the type of problem addressed by the machine learning model, and the format of the output result.
4. **Mixed Approaches**: Mixed approaches combine elements from the other three taxonomic concepts to create a comprehensive taxonomy. An example of such a mixed taxonomy is illustrated in Figure 2.3, proposed by the authors of [120]. It is grounded in conceptual approaches and incorporates result-based and function-based taxonomies as additional dimensions. Additionally, the output format employed by XAI methods serves as an extra dimension for classification.

2.3.2. Examples of explainable AI methods

While there exist some XAI approaches for unsupervised learning and reinforcement learning, such as those presented in [121] and [122], the majority of XAI methods predominantly concentrate on supervised learning, particularly in the context of classification tasks. Among the multitude of XAI techniques, we picked five notable examples that have garnered widespread recognition within the machine learning community, each amassing over 2,000 citations. We do not aim to give a complete overview of available XAI methods here, but give a quick overview of some of the most successful XAI methods. For a more comprehensive review of XAI methods, we refer the reader to [17] or [123].

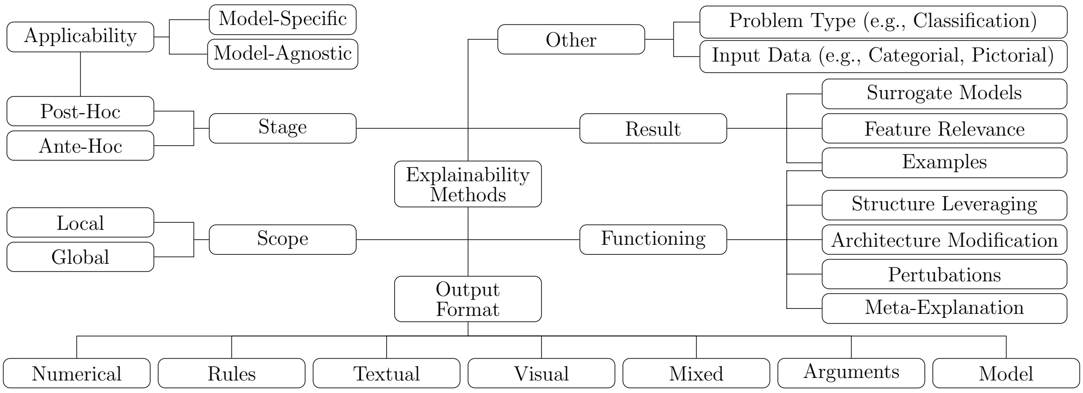


Figure 2.3.: Integral taxonomy approach from [120]

We discuss the XAI methods highlighted in the following works:

1. **Deconvolutional network** [18]: A model-specific XAI approach that utilizes the machine learning model's layered structure to highlight each layer's function.
2. **Grad-CAM** [21]: Another model-specific approach that explains which regions are relevant in image classification.
3. **LIME** [19]: The XAI approach presented in this paper is model-agnostic and can help understand any classifier by introducing an interpretable feature layer.
4. **SHAP** [20]: An approach that utilizes other approaches to create feature importance weights based on a game theory concept.
5. **Counterfactual explanations** [22]: The authors of this work introduce counterfactual explanations to the domain of machine learning in the XAI approach of this paper.

Each of the methods presented, we classify them in a figure at the start of each subsection using the taxonomy presented in Figure 2.3. Additionally, we give a quick overview of **causal models** [124]. While not exclusive to XAI, we deem causal models as important since they encode causal relations between variables of a model, thus allowing to derive explanations easily from them.

Deconvoluational network

Problem Type	Input Data	Stage	Scope	Output	Functioning	Result
Multiple	Images	Post-Hoc Model-Specific	Local	Visual	Architecture Modification	Feature Relevance

An innovative approach to visually explain the behavior of individual layers of a Convolutional Neural Network (CNN) is proposed by [18]. CNNs, a class of neural networks widely employed in tasks such as image classification, object detection, and segmentation, are comprised of a sequential arrangement of layers, each executing specific operations. The CNN layer progression unfolds as follows:

1. **Convolution**: Initially, the network conducts a convolution operation, employing filters to process either the outputs from the previous layer or, in the case of the first layer, the model's inputs.

2. **Activation**: Subsequently, the convoluted values traverse an activation function, often a rectified linear function.

3. **Pooling and Normalization**: Finally, a pooling operation, like max pooling, may be followed by optional output normalization.

The authors introduce a novel concept termed the "deconvolutional network" to map each CNN layer's activation back to an interpretable image. The approach isolates a single activation within a convolutional layer while setting all other values to zero. This activation vector serves as input for the deconvolutional network, reversing a convolutional layer's steps in a backward sequence, encompassing unpooling, rectification, and image reconstruction through transposed filters. Ultimately, a transposed version of the filters is employed to reconstruct an image that faithfully captures the features of the original input image responsible for eliciting the observed activation within the convolutional layer.

Several challenges are addressed during this process. Since pooling is a non-invertible operation, the location of relevant values must be recorded during the pooling step in the convolutional layer. The unpooling operation can effectively approximate the reversal of the pooling step, armed with this location information. Likewise, the activation function presents non-invertibility challenges, especially in the case of a rectified linear function, which discards negative values. However, due to the nature of desired image outputs, which exclusively comprise positive values representing pixel color values, an approximate inversion of the activation step becomes feasible.

Figure 2.4 showcases an explanation for an image classification by highlighting the significance of features that contribute to a specific image's classification. This explanation

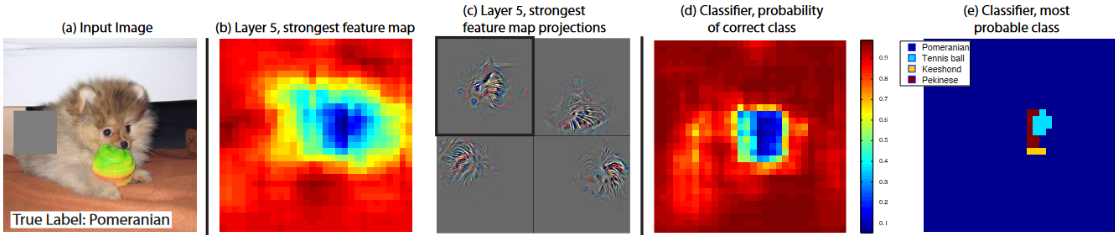

(a) Input Image
True Label: Pomeranian
(b) Layer 5, strongest feature map
(c) Layer 5, strongest feature map projections
(d) Classifier, probability of correct class
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
(e) Classifier, most probable class
Pomeranian
Tennis ball
Keeshond
Pekinese

Figure 2.4.: Explanation generated using deconversational network, taken from [18].

is generated using a deconvolutional network. The successively regions within the image under investigation are covered by a grey square (a). To unravel the contribution of these regions, the strongest response in the feature map of the final layer is recorded for various positions of the grey square (b). Subsequently, a projection of the feature map back onto the input image is generated (c). By correlating the position of the grey square, the probability of correct classification is mapped onto relevant areas of the input image (d). This process allows different areas of the image to be associated with their most probable class (e).

Grad-CAM

Problem Type	Input Data	Stage	Scope	Output	Functioning	Result
Multiple	Images	Post-Hoc Model-Specific	Local	Visual	Structure Leveraging	Feature Relevance

While methods like the previously mentioned deconvolutional network [18] offer insights into which features of an image matter to a CNN, they fall short when precisely associating these features with a particular class. In simpler terms, if, for instance, "fur" is deemed relevant for identifying a cat in an image, these methods would highlight all furry objects in the image without distinguishing them by class.

Addressing this limitation, [21] introduces the Gradient-weighted Class Activation Mapping (Grad-CAM) technique. Grad-CAM excels at pinpointing the areas within an image that hold relevance for a predicted class. Moreover, it integrates seamlessly with methods like the deconvolutional neural network, resulting in visual explanations that discriminate between classes and focus on the local features pertinent to the model's decision.

Much like the deconvolutional network approach, the Grad-CAM process begins with passing an image through the model. Next, the output of the machine learning model is

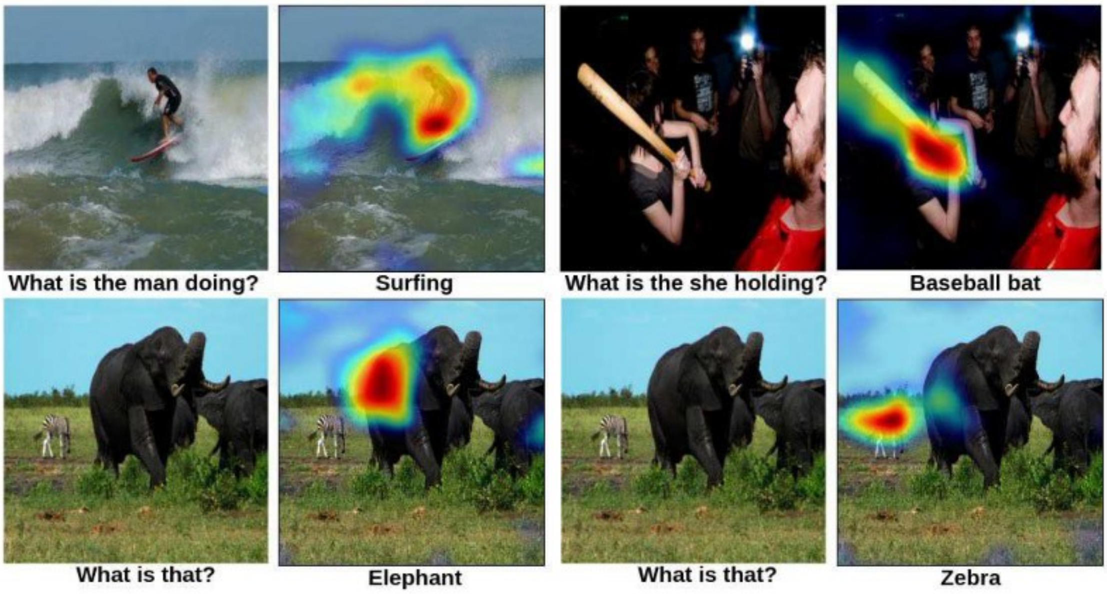


Figure 2.5.: Explanation for the result of a visual question answering task given by Grad-CAM, taken from [21].

transformed into a binary vector, with a value of one for the target class and zeros for all other classes. This binary output is then back-propagated through the model until it reaches the last convolutional layer, where the output derivative concerning this layer's features is computed. These features are subsequently weighted by their corresponding gradients, and a rectified linear activation function is employed to retain only those features that positively influence the target class.

Grad-CAM finds applicability in machine learning approaches that employ a CNN architecture for image encoding, subsequently feeding this encoded information into various architectures, such as image classification, image captioning, or visual question answering. The outcome is a coarse heat map, which can mask irrelevant regions of the input image, effectively highlighting the image elements that decisively contribute to the model's classification decision. Crucially, Grad-CAM leverages the last convolutional layer of the CNN to extract information about the image regions relevant to a chosen class. The authors deem this layer the optimal balance between sparse information representation and preserving high-level concepts that might be lost in subsequent layers.

In Figure 2.5, an exemplary explanation for visual question answering, generated by

Grad-CAM, is presented. Grad-CAM effectively highlights the specific regions of the image that correspond to the provided answer. This visual representation allows an observer to assess whether the identified important areas align with the answers given by the explained model.

LIME

Problem Type	Input Data	Stage	Scope	Output	Functioning	Result
Classification	Mixed	Post-Hoc Model-Agnostic	Local	Mixed	Petubations	Surrogate Models

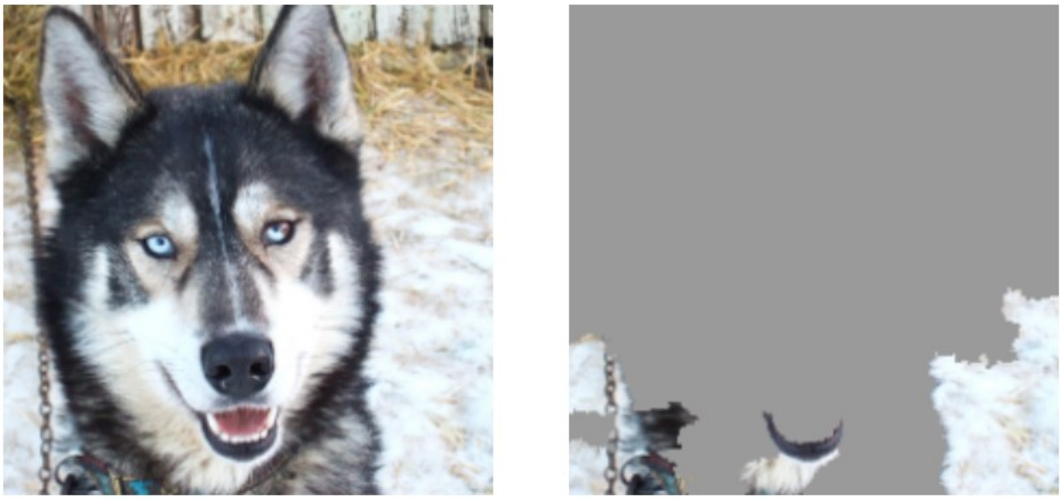
The XAI technique introduced by [19] is named "Local Interpretable Model-Agnostic Explanations," abbreviated as LIME. As the name implies, LIME is a model-agnostic approach to crafting localized explanations. The key to achieving interpretability in these explanations lies in an abstraction layer that translates model inputs into more understandable abstract features.

For instance, this abstraction layer can map individual pixels of an image into features, which can be groups of neighboring pixels that collectively convey a meaningful concept. LIME then generates explanations based on these abstract features through perturbation, such as recoloring all pixels belonging to a particular abstract feature with a single color.

LIME's workflow involves introducing input perturbations around a point of interest, such as an image to be classified, and observing the corresponding responses from the machine learning model. These perturbation-response pairs are then used to fit a linear model. To address potential variations in the characteristics of abstract features (e.g., some features may be affected by more pixels of the image than others), LIME employs a distance metric. A penalty term is also introduced to constrain the resulting linear model, promoting the retention of crucial features and simplifying the resulting explanation.

One of LIME's notable advantages is its independence from the type of classifier model or input data it is applied to, as it operates without direct interaction between the machine learning model and the XAI method. Consequently, it can be seamlessly employed with a wide range of classifier models, regardless of the input data type. Furthermore, from the derived linear model, it becomes straightforward to extract feature importance weights, shedding light on the most influential abstract features that either support or oppose the classification provided by the examined machine learning model.

In Figure 2.6, an illustrative explanation provided by LIME sheds light on the misclassification of an image. Specifically, the image of a husky is erroneously classified as a wolf


(a) Husky classified as wolf

(b) Explanation

Figure 2.6.: LIME's explanation for why a classifier wrongly classified a husky as a wolf, taken from [19].

(a). LIME serves the crucial purpose of identifying the most relevant features influencing this misclassification. As depicted in (b), the most significant feature for the model's classification of a wolf, in this particular instance, was not the showcased animal itself but rather the presence of snow in the background. If this process is repeated across multiple examples, a human observer can effectively evaluate whether the model is behaving appropriately or if it has potentially learned biases present in the training data.

SHAP

Problem Type	Input Data	Stage	Scope	Output	Functioning	Result
Classification	Mixed	Post-Hoc Model-Agnostic	Local	Numerical	Meta-Explanations	Feature Relevance

The SHAP (SHarpley Additive exPlanations) XAI framework, as introduced by [20], leverages a collection of existing XAI methods to formulate comprehensive explanations. Notably, SHAP utilizes several precursor XAI methods, including LIME [19], DeepLift [125], and layer-wise relevance propagation [126]. These methods, which serve as the

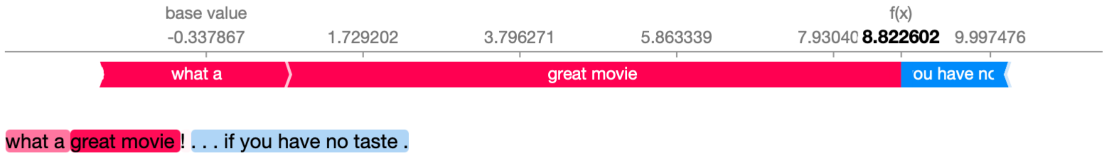


Figure 2.7.: Explanation for the result of a text classifier trained to detect the sentiment text, taken from [128].

underpinnings of SHAP, share a common approach: generating surrogate models based on simplified input features to elucidate model behavior.

The authors of SHAP establish a crucial link by demonstrating that these XAI methods exhibit inherent relationships with one another. Furthermore, they establish that there exists a theoretically unique solution that fulfills a set of essential criteria. These criteria include:

1. **Local Accuracy**: Surrogate models must precisely match the behavior of the explained machine learning model at a specified point of interest.
2. **Missingness**: Simplified features absent in the explained instance should bear no weight in the surrogate model.
3. **Consistency**: Any alterations in the importance of a feature should be accurately reflected in the weights of the surrogate model.

While the surrogate models of the earlier XAI methods typically fall short in satisfying all of these criteria, SHAP adopts a specific class of surrogate models grounded in Shapley values from game theory [127]. This choice ensures that the desired properties of local accuracy, missingness, and consistency are met consistently. The authors modify LIME and DeepLift to align with their respective functionalities and produce explanations that conform to the stipulated criteria, at the cost of some benefits of the original methods like the sparseness of the model produced by LIME.

Figure 2.7 illustrates a representative explanation provided by SHAP for a text classifier. The classifier in question is specifically designed to discern the sentiment conveyed in textual content. SHAP enables the assignment of Shapley values to each segment of the classified text, showing their contributions to the overall result. This allocation facilitates a direct assessment of the significance of individual features, akin to LIME. Nevertheless, unlike LIME, SHAP's feature importance is additive, allowing for a straightforward prediction of the impact caused by the absence of features.

Counterfactual explanations

Problem Type	Input Data	Stage	Scope	Output	Functioning	Result
Classification	Mixed	Post-Hoc Model-Agnostic	Local	Mixed	Petubations	Examples

The introduction of counterfactual explanations to the realm of machine learning is attributed to [22]. To recap counterfactual explanations briefly, they serve as a means to provide insights by presenting closely related examples that would result in a different output from a machine learning model. For instance, consider the question, "Why were the street lights turned on?" A counterfactual explanation might offer, "If there were still sunlight outside, the street lights would not have been turned on."

Counterfactual explanations are closely related to a different kind of problem in machine learning that seeks to make models more robust against adversarial perturbations of input data. While also seeking to influence model output, these perturbations strive to remain inconspicuous to human observers. Approaches to creating adversarial variations make minimal changes to multiple inputs rather than drastically altering a single input to achieve this inconspicuousness. Importantly, adversarial perturbation methods can be computed efficiently due to the differentiability of most machine learning models. This feature can also be harnessed for the efficient generation of counterfactual explanations.

The authors of this approach adopt an optimization-based strategy to compute their counterfactual explanations. They minimize the distance between a counterfactual input, denoted as x', and the original input, denoted as x, while simultaneously maximizing a penalty term that accounts for any deviation of the model's output, $f(x')$, from a desired target output, y'.

Crucially, the authors underscore the critical role of the choice of distance metric in shaping the quality of the counterfactual examples generated. Unlike the adversarial approach, they want to find variations that are interpretable, i.e., do not alter multiple inputs inconspicuously. They recommend employing an L1-norm, weighted by the inverse of the median absolute deviation in x, as the distance metric. This choice is motivated by the L1-norm's tendency to yield sparse solutions for x. At the same time, the weighting by the inverse of the median absolute deviations permits more significant deviation in x' in areas with substantial variability in x.

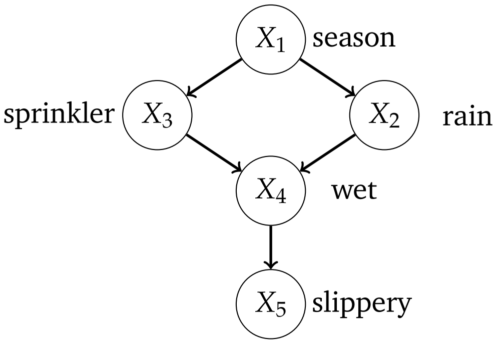


Figure 2.8.: Exemplary Causal Bayesian Network representing dependencies between five variables, from [124].

Causal Models

As highlighted in Section 2.1, causality stands out as a fundamental concept in explanations. Consequently, models that encode causal relationships between variables hold significant potential for generating explanations. These models, aptly named causal models, encompass various forms such as causal Bayesian Networks, structural causal models, and causal graphical models [124].

Figure 2.8 provides an illustrative example of a causal Bayesian Network, where each node in the network graph represents a variable, and the edges denote causal relations. In this instance, the season causally influences the probability of rain and the activation of sprinklers. Both the sprinklers and rain, in turn, impact whether the floor is wet, and a wet floor influences its slipperiness.

While conventional Bayesian Networks capture only conditional independence between variables, causal Bayesian Networks encode causality, enabling predictions about interventions. For instance, manually turning on the sprinkler renders it independent of the seasons, a distinction from selecting only samples where the sprinkler is activated in the dataset.

Deriving explanations such as "The floor is slippery because it is wet." becomes straightforward with causal models. The process of learning causal relationships through intervention is recognized as causal inference [129]. Methods focused on learning causality from data without intervening are referred to as causal discovery methods [130].

Despite the prevalence of causal language in machine learning and XAI, our knowledge suggests the absence of an XAI method utilizing causal models for explanation. However, we view causal models as a promising avenue for explanation generation and explore their application in explaining ESM in Chapter 4.

2.4. Alignment of XAI and ESM explanation methods

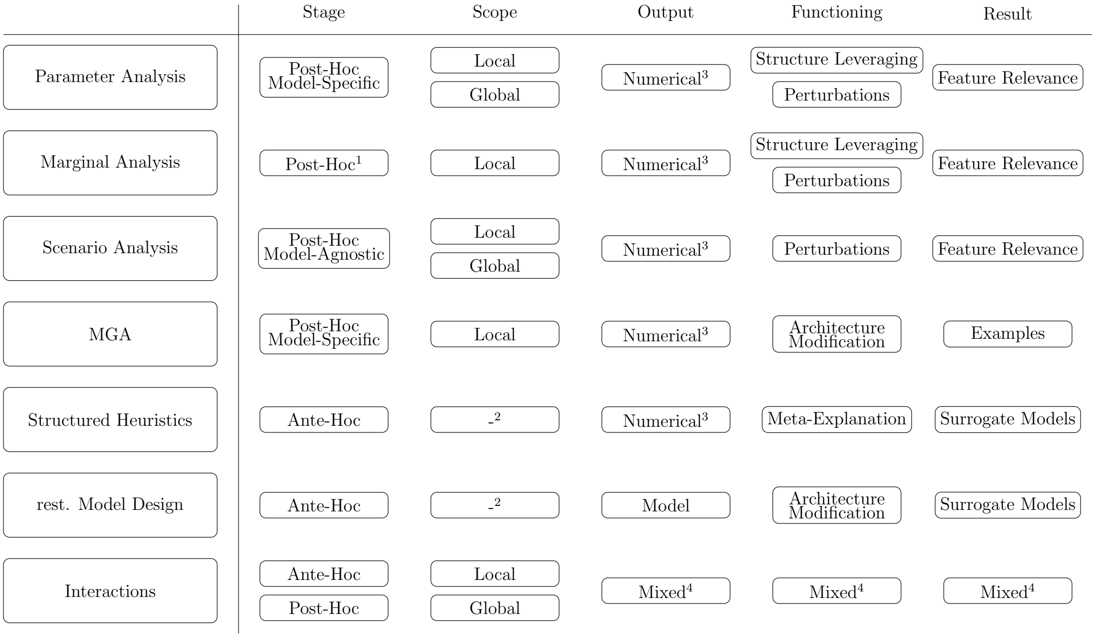
In the previous sections, we have investigated different possible XAI taxonomies. We used the integral taxonomy approach from [120] to classify the mentioned XAI examples. In this section, we use the same taxonomy to align the categories of ESM methods from Section 2.2 with the XAI domain.

Figure 2.9 shows our classification of ESM explanation categories based on XAI taxonomy criteria. It is worth noting that specific categories, such as parameter analysis, scenario analysis, and interactions, may exhibit varying scopes and functionalities depending on the specific method within each category. Further, some categories of the XAI taxonomy are skipped, namely input data and problem type. Problem type is neglected since all methods aim to explain the same problem type (i.e., optimization-based ESMs). The catagory input data depends on the concrete implementation of the ESM explanation method and thus can not be determined at the categorical level.

In the following section, we discuss the findings of this chapter and answer our first two research questions.

2.5. Discussion

In this chapter, we investigated explanations across various domains. We started with a broad perspective on explanations, as seen through the lenses of psychology and philosophy. The fundamental purpose of explanations lies in their ability to convey knowledge regarding an event, typically achieved by illustrating the event's causal history. The process of explanation can be dissected into distinct phases: first, the cognitive process of identifying potential explanations; second, the resulting explanation itself; and finally, the communication process necessary for transferring the condensed knowledge contained within the explanation.



	Stage	Scope	Output	Functioning	Result
Parameter Analysis	Post-Hoc Model-Specific	Local Global	Numerical[3]	Structure Leveraging Perturbations	Feature Relevance
Marginal Analysis	Post-Hoc[1]	Local	Numerical[3]	Structure Leveraging Perturbations	Feature Relevance
Scenario Analysis	Post-Hoc Model-Agnostic	Local Global	Numerical[3]	Perturbations	Feature Relevance
MGA	Post-Hoc Model-Specific	Local	Numerical[3]	Architecture Modification	Examples
Structured Heuristics	Ante-Hoc	-[2]	Numerical[3]	Meta-Explanation	Surrogate Models
rest. Model Design	Ante-Hoc	-[2]	Model	Architecture Modification	Surrogate Models
Interactions	Ante-Hoc Post-Hoc	Local Global	Mixed[4]	Mixed[4]	Mixed[4]

Figure 2.9.: Classification of methods for explanation in the energy system domain into the integral taxonomy approach for XAI from [120].

[1]: Marginal Analysis can be used model-specifically to access marginals from the optimization process of an optimization problem by structure leveraging, or it can be used model-agnostically by optimizing the model multiple times with small perturbations.

[2]: These approaches change/ limit the model itself and therefore scope does not apply for them.

[3]: Even so the output is purely numerical it is commonly post-processed and a visual or textual explanation is provided instead, i.e., line charts, time series plots or colored maps.

[4]: Depending on the design of the interactive approach, multiple outputs, functions, and results are possible. For example, a language agent could perturb the input, modify its architecture or extract information from its structure depending on the user's request.

During the cognitive phase of uncovering possible causes, multiple explanations often surface. Explainers employ various factors, such as backgrounding and discounting, to sift through these potential causes and determine which ones are crucial for the event. It is noteworthy that, in the process of selecting explanations, implicit default scenarios are frequently embedded in the question, necessitating consideration.

In the second section, we looked into the current state-of-the-art regarding explanations for optimization-based ESMs. We examined the advantages and disadvantages of existing methods in this domain. Our analysis led us to categorize these methods into seven distinct categories, three of which are a form of sensitivity analysis. These subcategories, namely parameter, marginal, and scenario analysis, evaluate the ESM's responsiveness to input changes. Parameter analysis, although valuable for comprehending the impact of individual parameters on the model in detail, has limitations, including high computational costs or applicability primarily to scenarios involving a limited number of parameters and simpler ESMs. Marginal analysis circumvents the parameter limitation by extracting marginal sensitivities from the ESM's solution process. However, it requires access to the model or its solution process, which may only sometimes be available, and its explanations have limited local validity. Sensitivity analysis, in contrast, does not necessitate access to the ESM and is unrestricted in terms of the number of parameters or the magnitude of variation. Nonetheless, this freedom comes at the expense of explanations being highly contingent on the design of the scenarios, necessitating the need to justify and clarify the scenario choices made.

Three ESM explanation methods that are not sensitivity-based are MGA, structured heuristics, and model design restrictions. MGA is instrumental in exploring alternative solutions near the optimal solution of an ESM, thus shedding light on the flexibility of the solution space. However, it mandates comprehensive access to the model's structure and strong domain knowledge for effective implementation. Techniques such as structured heuristics and model design restrictions enhance ESM understandability. They achieve this by replacing the complex optimization process to find optimal solutions with more accessible heuristics or constraining ESM complexity during the design phase.

Lastly, interactions constitute another category of ESM explanation methods, focusing on facilitating communication processes that allow the target audience to engage with the model. This engagement can be achieved by a conversational agent (e.g., a chatbot) or even a serious game. While interaction-based approaches hold great promise, as they can combine and leverage various methods based on user requests, they are also the most challenging to implement. They require domain expertise, a deep understanding of the target audience, access to the model, and proficiency in other domains, such as large language models or game design.

The second section answers the first part of our first research question RQ 1. For the second part of the research question, we identified the following shortcomings in the current ESM explanation methods. A significant limitation in all ESM explanation methods categories is their handling of high-dimensional input data, mainly structured data like time series. Time series data holds substantial importance within ESMs, where it finds application in various contexts, such as solar radiation profiles, electricity demand patterns, or wind speed fluctuations. The challenge arises when dealing with alterations to individual time steps within these time series. Often, such changes exert minimal influence on the optimal solution, whereas modifications to broader patterns can bear significant consequences. None of the abovementioned approaches effectively address these pattern-level variations, except for scenario analysis if the pattern change is explicitly formulated as a scenario. Nevertheless, scenario analysis encounters issues regarding the comparability of scenarios, given that they may exhibit considerable dissimilarities in nature and characteristics.

Our investigation into XAI methods from the domain of machine learning is motivated by the shared challenges in achieving explainability faced by both machine learning models and large optimization-based ESMs and in line with the first half of our second research question RQ 2. Notably, machine learning models frequently deal with the interpretability of their outputs, especially when handling high-dimensional and structured input data – an issue that mirrors the shortcomings identified in ESM explanation methods.

Given the vast landscape of XAI, we initiated our exploration by establishing fundamental criteria for adequate explanations, along with an overview of different taxonomies employed to classify XAI approaches. Subsequently, we examine five prominent XAI methods, outline their functioning, and classify them according to one of the discussed XAI taxonomies. Specifically, we inspected deconvolutional networks and Grad-CAM, two techniques geared towards explaining Convolutional Neural Network (CNN)-based machine learning models frequently employed for image-related tasks such as image classification and virtual question answering. Additionally, we explored three model-agnostic XAI methods, namely LIME, SHAP, and "counterfactual explanations." These methods offer versatile approaches to explain general classification tasks, each distinguished by its unique methodology and yielding slightly distinct resulting explanations.

Now that we have gained insights into several XAI methods, it is imperative to address an outstanding second part of our research question RQ 2: can these concepts be applied to enhance explanations for ESMs? In pursuit of this objective, we can draw valuable lessons from the XAI taxonomy and the specific XAI methods we explored.

To begin, we can leverage the XAI taxonomy to generalize and align them with the categories we have established for ESM explanations as shown in the previous section. A notable observation from the alignment is that the outputs of most ESM explanation methods tend to be numerical, often necessitating manual post-processing to enhance their interpretability. A potential avenue for future research lies thus in automating or standardizing these post-processing procedures.

Additionally, we can gain insights from the XAI methods themselves for a potential transfer to the ESM explanation setting. While Deconvolutional Networks and Grad-Cam are tailored to the nuances of CNNs and their specific structures, rendering them less applicable to ESM explanations, the model-agnostic nature of LIME, SHAP, and "counterfactual explanations" offers promise for potential adaptation. For instance, LIME employs the concept of an abstraction layer to render high-dimensional inputs interpretable for human observers. LIME is adaptable to different types of input data in the context of different classification models, i.e., a different abstraction layer is used to make text classification interpretable then for image classification but all inputs with the model are the same (letters or pixels). On the other hand, ESMs encompass diverse input data types within a single model, e.g., financial data, weather time series, or technical parameters of power plants. An abstraction layer tailored to ESMs would need to be more complex to handle the combination of various input data types effectively. However, an abstraction layer could offer a means to explain complex inputs like time series data. Furthermore, LIME employs a regression with a penalty term favoring sparse explanations. Incorporating a similar auto-selective regression approach in ESM explanations could reduce complexity and customize explanations for different target audiences. SHAP, an extension of LIME, operates on a comparable concept. Although it sacrifices the auto-selective penalty term in LIME to achieve specific properties, its underlying principles remain similar. Hence, we are going to explore the use of LIME for ESM explanation in Chapter 3.

Lastly, "counterfactual explanations" aim to generate examples that emphasize why a model classifies an input as it does rather than as a different class. These examples closely resemble the original input but lead to a different classification. Exploring a similar approach in the context of ESMs could be interesting. However, it is essential to note that "counterfactual explanations" are rooted in adversarial attacks within the machine learning domain. On the other hand, adversarial attacks on ESMs have not been extensively explored. Nevertheless, a preliminary exploration of this research area has been conducted in a student's work [131].

In the subsequent sections of this work, we narrow our focus to a singular concept from each domain discussed in this chapter. In the forthcoming chapter, we undertake the task of adapting the principles of LIME for application in the context of ESMs. Additionally, we explore the integration of the concept of causality, a pivotal element in the realm of explanations, with ESMs in Chapter 4. Lastly, in Chapter 5, we create an interactive approach rooted in ESMs aimed at elucidating pertinent concepts related to energy transition for an audience of students.

3. Local Interpretable Explanations for Energy System Models

One application of ESMs is extension planning [132]. ESMs allow domain experts to optimize the energy system design (ESD) with respect to system costs or CO_2 emissions. All such models are based on a multitude of parameter data. Required data include time series, technical and economic parameters, and legal and physical limitations. For example, for modeling wind power, historic weather time series for the wind speeds, the currently installed capacity of wind power plants per model region, the specific costs for new installations, and the maximum possible capacity allowed by regulation are required. When no data are available, assumptions have to be made. The computed ESDs are highly dependent on these data and assumptions, which is problematic since the DMs may face personal consequences if the ESD they commit to does not perform well. Hence, explanations are required to help DMs understand the ESDs proposed from ESMs and decide on the ones best suited for their needs.

As shown in the previous chapter, a tool often used by domain experts to explain ESMs' results is sensitivity analysis, which creates explanations based on how a system reacts to individual parameter changes. The term sensitivity analysis is used for methods that vary the input of a model and observe the change in outputs. We categorise sensitivity analysis approaches into parameter analysis, marginal analysis and scenario analysis.

A major downside of sensitivity analysis is their inability to give adequate explanations on high-dimensional input data such as time series, where small changes to individual parameters have often only minor impact on the ESD. Changes for multiple time steps, however, can have a significant impact. For example, suppose the renewable energy potential is increased for a single time step. In that case, it is unlikely that the cost-optimal storage capacity changes since additional storage capacities are expensive when used only once. If the renewable increase covers several well-distributed time steps, storage capacity can be reused and becomes more attractive.

A similar problem occurs in the domain of machine learning, e.g., in image classification where changes to individual pixels of the image should have no impact on the classification result, but changes to a set of pixels may change the result. A popular explainable AI method to overcome these limitations is "locally interpretable model-agnostic explanations" (LIME) [19]. LIME, designed for application to classification problems, addresses the shortcoming of traditional sensitivity analysis regarding high input dimensions by introducing an interpretable abstraction layer for the input features. It then aims to find the interpretable features whose variations are most relevant for changing the predicted class label. This feature is then considered most relevant for a given class decision.

In this chapter, we propose to transfer the LIME idea from the machine learning domain to the ESM community given the parallels between the two described problem settings to answer our research question RQ 3. While explainable AI methods have been used in energy and power systems before, explaining machine learning-based models employed in power grids, the energy sector, or energy management in buildings, see e.g., [133]–[135] and the review [12], we use explainable AI methods here for optimization-based ESMs, not for statistical learning-based methods.

We apply the transferred method to two ESMs with different complexities: one for a building ESD and one for a nationwide ESD model in order to tackle RQ 4. By making ESMs better understandable to non-expert decision-makers, we hope that the work supports informed decision-making in the transition towards low-carbon energy systems.

The remainder of the chapter is structured as follows: An exemplary ESM for finding the optimal ESD of a building with a renewable energy supply is presented in Section 3.1. It serves as a running example throughout this work. We then explain the concept of LIME and show how the methodology can be transferred from the machine learning domain to ESMs in Section 3.2. Explanations for the optimal design of the exemplary energy system are derived and discussed in Section 3.3. Explanations of a more complex model, i.e., a model of the German energy system, are presented in Section 3.4. A discussion is given in Section 3.5.

3.1. An Exemplary Building Energy System

This chapter uses a building energy system as a running example. Its structure is shown in Figure 3.1. The building is characterized by an electric and a thermal demand. Electric energy is locally generated by a photovoltaic (PV) plant or can be bought from the external electricity grid. Electricity can be stored in a battery for later use and converted into heat using a heat pump. Heat can be stored in heat storage, such as a hot water tank.

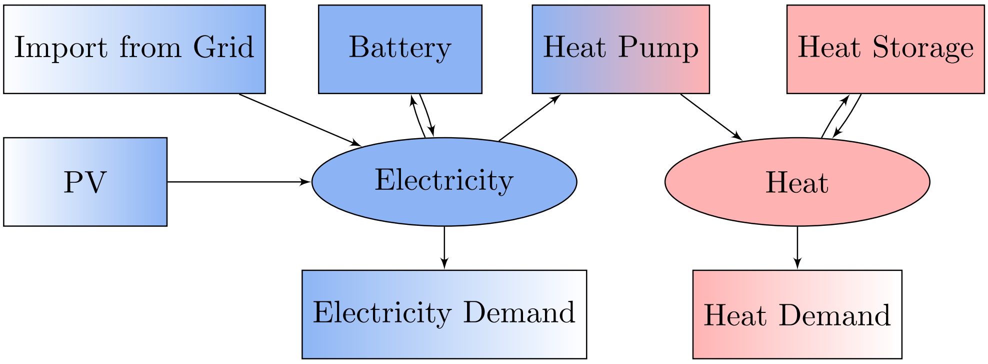


Figure 3.1.: Schematic overview of exemplary energy system model of a building with photovoltaic (PV) power plant. Energy commodities are displayed as ellipses and energy conversion processes as rectangles. Arrows of a conversion process denote which energy commodities can be transformed into each other, while colors encode different types of energy. Conversion processes without input or output represent energy demands or energy imports into the model domain.

The energy system is modeled using a bottom-up approach based on linear programming [136], extending the model of [10] by a heat sector. The model's objective is to minimize the total system costs, including investment and operational contributions. Optimization variables are the scheduling decisions of all components, e.g., when to charge or discharge the battery. Furthermore, the system's design is optimized by model-endogenously choosing cost-optimal capacities for the battery and heat storage. In contrast to [10], the PV power plant capacity is assumed to be known. The model adheres to a set of constraints. These include the power balances for electricity and heat at every time step. Other conditions describe the storage level's dependency on charging and discharging decisions. Three sets of inequalities limit the available PV energy per time step to an externally given, weather-dependent time series, ensure feasible storage levels, and constrain the heat pump's power output by capacity. The detailed mathematical equations are given in Appendix A.

3.2. Proposed Explanation Methodology

In this section, we present our methodology to derive explanations for ESMs that go beyond the scope of traditional sensitivity analyses. As our approach is based on the LIME [19] method from the explainable machine learning domain, we first describe LIME in its original context and then introduce our proposed methodology in parallel.

3.2.1. LIME for Machine Learning

LIME introduces an interpretable abstraction layer for the input dimensions of the machine learning task and combines it with a modified sensitivity analysis. As the name "local interpretable model-agnostic explanation" suggests, LIME creates explanations for non-experts for any classifier $f : \mathcal{X} \to [0,1]^L$ given a specific point of interest (POI) $\bar{\mathbf{x}} \in \mathcal{X}$, where $\mathcal{X} \subseteq \mathbb{R}^d$ is the original input space of the classification problem and L is the number of class labels f can predict. The created explanations are local, i.e., only valid for variations close to the POI. To this end, an interpretable feature space $\mathcal{Z} \subseteq \{0,1\}^{d'}$ is created for the original, high-dimensional input space of the classification problem $\mathcal{X} \subseteq \mathbb{R}^d$ with $d' \ll d$. The vector $\bar{\mathbf{z}} \subseteq \mathcal{Z}$ is a binary vector encoding the presence or absence of interpretable features related to the POI $\bar{\mathbf{x}}$. First, I variations $\mathbf{z_i}$ in the interpretable feature set are made around $\bar{\mathbf{z}}$. Second, a function $h^{\bar{\mathbf{x}}} : \mathcal{Z} \to \mathcal{X}$ maps the interpretable variations $\mathbf{z_i}$ back to the input space, i.e., $\mathbf{x_i} = h^{\bar{\mathbf{x}}}(\mathbf{z_i})$. Each variation $\mathbf{x_i}$ is then weighted by $\pi_{\bar{\mathbf{x}}}(\mathbf{x_i})$ based on its similarity to $\bar{\mathbf{x}}$. Finally, LIME determines the most important interpretable features, i.e., the explanation, for a given class label by solving the regularized least-squared regression

$$\arg\min_{g \in \mathbb{G}} \sum_{i=1}^{I} \pi_{\bar{\mathbf{x}}}(\mathbf{x_i})(f(\mathbf{x_i}) - g(\mathbf{z_i}))^2 + \Omega(g), \tag{3.1}$$

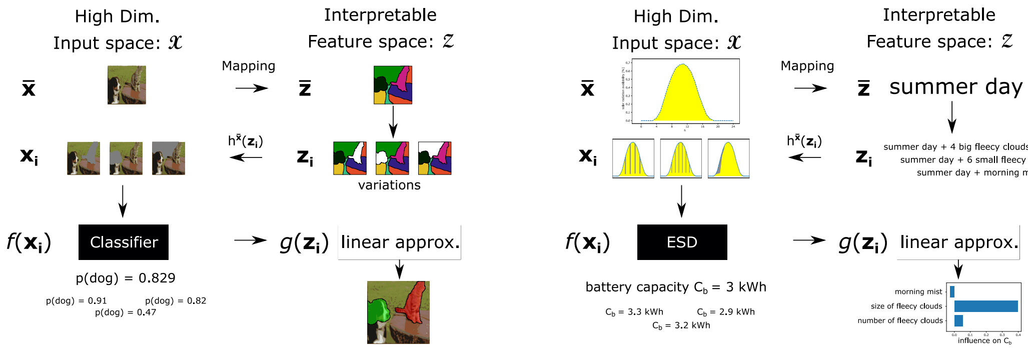
where the model $g : \mathcal{Z} \to \mathbb{R}$ is a model in the class of all linear models $\mathbb{G} \subseteq \mathbb{R}^{\mathcal{Z}}$ and $\Omega(g)$ is a complexity measurement of $g \in \mathbb{G}$, for example, the number of non-zero weights of g. Interpretable features with non-zero weighting are then deemed explanations for the classifier's local behavior around the POI $\bar{\mathbf{x}}$. By choosing $\Omega(g)$ appropriately, the complexity of the explanation can be limited as desired.

A visualization of the concept of LIME applied to image classification is found in Figure 3.2a, where an interpretable explanation for the classifier f's prediction of the "dog" class is created by LIME. The original input space $\mathcal{X}$ contains vectors with the color values of the individual pixels of an image. Clusters of pixels, the so-called super-pixels, are used to define the interpretable feature space $\mathbb{Z}$. Each interpretable feature encodes the presence or absence of a super-pixel. Variations $\mathbf{z_i}$ around the interpretable representation $\bar{\mathbf{z}} = (1, \ldots, 1)$ of the image are made. Those variations are mapped back to the input space $\mathcal{X}$ by replacing all pixels belonging to a non-selected super-pixel with a neutral color, i.e., a 50% grey value. The resulting inputs $\mathbf{x_i}$ are classified by f, and a linear model g is fitted to the resulting "dog" class probabilities by solving the problem defined in Equation (3.1). The magnitudes of the weights in g represent the impacts of the super-pixels on f and can thus be used to explain f locally. The complete explanation is a list of super-pixels in decreasing order based on their absolute impact on the class label. Shown in green is the super-pixel with the highest weight for predicting the class "dog"; shown in red is the super-pixel with the highest weight for not predicting the class "dog".

3.2.2. Proposed Method for Explaining ESM

We now want to explain to non-experts an output value from the cost-optimal solution of an ESM. To this end, we propose a method based on the LIME concept, where instead of a class probability given by the classifier, we explain an output value of a cost-optimizing ESD. The POI is a vector $\bar{\mathbf{x}}$ in the ESM's input space $\mathcal{X}$. We then identify interpretable features $\mathcal{Z}$ that may or may not be part of the original model parameters but are assumed to be potentially relevant for an explanation. Since we are interested in quantitative explanations and not only qualitative ones, instead of the binary encoding used by LIME for the interpretable features space, we define $\mathcal{Z} \subseteq \mathbb{R}^{d'}$. Similar to LIME, variations in the interpretable features are mapped to the model's inputs via a mapping $h^{\bar{\mathbf{x}}} : \mathcal{Z} \rightarrow \mathcal{X}$ between the interpretable features space $\mathcal{Z}$ and the input space $\mathcal{X}$. The ESM is optimized for each variation. A linear model $g : \mathcal{Z} \rightarrow \mathbb{R}$ is fitted to the results of the ESM's runs by minimizing the objective in Equation (3.1). As the complexity measure $\Omega(g)$, we chose the number of non-zero weights of g. The linear model g is finally used as an interpretable explanation for the ESM: a weight in g corresponding to an interpretable feature represents that feature's influence on the optimal ESD.

A visual representation of the methodology proposed for the explanation of an ESM is shown in Figure 3.2b. The illustration is based on our exemplary building energy system from Section 3.1. An interpretable explanation for the cost-optimal battery capacity designed by the ESM is created for a given set of input parameters, the POI $\bar{\mathbf{x}}$. In the presented case, the POI is a summer day's solar radiation time series. An interpretable feature space $\mathcal{Z}$ is then created. Interpretable features considered potentially relevant for



(a) LIME's explanation methodology for image classifiers.

(b) Proposed explanation methodology for an ESM for optimal ESD.

Figure 3.2.: (**a**): The LIME method: For an image classifier where individual pixels of an input image $\bar{\mathbf{x}}$ have no explanatory power, an abstraction layer with interpretable features $\mathcal{Z}$ (here super-pixels) is created. Variations $\mathbf{z_i}$ around the interpretable representation of the input image $\bar{\mathbf{z}}$ are made and mapped back to the input space $\mathcal{X}$ via $h^{\bar{\mathbf{x}}}$, i.e., by replacing each pixel belonging to a super-pixel with a single neutral color. The classifier f is applied to the modified inputs. LIME fits a linear model $g(\mathbf{z})$ to the classifier's outputs and uses g to explain the behavior of g locally around $\bar{\mathbf{x}}$. (**b**): Proposed Method: For a high-dimensional input $\bar{\mathbf{x}}$, e.g., a radiation time series, interpretable features $\bar{\mathbf{z}}_i$ that could be relevant for explanation are identified. Variations $\mathbf{z_i}$ around $\bar{\mathbf{z}}$ are made and mapped back to the input space $\mathcal{X}$ by $h^{\bar{\mathbf{x}}}$. The ESM f is optimized for each input variation $\mathbf{x_i}$. A linear model $g(\mathbf{z_i})$ is fitted to the output of the ESM optimization, e.g., the cost-optimal battery capacity C_b. The weights of the linear model g represent the relevance of the interpretable features for the optimal ESD derived from the ESM f locally around $\bar{\mathbf{x}}$.

an explanation are the specific battery investment cost, the number and shape of "fleecy clouds" occurring during the day, and the existence of "morning mist". Interpretable variations $\mathbf{z_i}$ around the POI are made, and the mapping $h^{\bar{\mathbf{x}}} : \mathcal{Z} \to \mathcal{X}$ maps the interpretable variations to the models' input space $\mathcal{X}$, resulting, e.g., in variations $\mathbf{x_i}$ in the solar radiation time series. The ESM is then optimized for each variation. A linear model g is fitted to approximate the cost-optimal battery capacity C_b of the ESD model's results. The weights of the resulting linear model g are taken as an interpretable explanation for the ESD model. A weight in g corresponding to an interpretable feature represents that feature's influence on the ESD model's output. In the given example, the most relevant interpretable feature for the ESM's cost-optimal battery capacity is the size/duration of the "fleecy clouds".

We show the application of this method to an exemplary building energy system in Section 3.3 and a model of the German energy system in Section 3.4. These examples include the explicit definition of the interpretable feature space $\mathcal{Z}$ and the mapping $h^{\bar{x}}$.

3.3. Experimental Demonstration: The Exemplary Building Energy System

In this section, we apply our proposed method to create an explanation of the cost-optimal battery capacity for our exemplary building energy system introduced in Section 3.1. To this end, we first define a set of interpretable features that can impact the cost-optimal battery capacity. They consist of cloud characteristics, the PV surplus with respect to the load, and the specific battery investment cost. These features are described in Section 3.3.1. Section 3.3.2 describes details of the first implementation of our methodology to the exemplary building energy system. Section 3.3.3 shows results for our proposed method on a simplified version of the same exemplary energy system model, which includes only the electricity sector. We create explanations around two points of interest with different battery costs and test their robustness towards different feature mappings $h^{\bar{x}}$. Explanations for the complete version of the exemplary building model, i.e., including the heat sector, are then given in Section 3.3.4.

3.3.1. Interpretable Features

The defined interpretable features $\mathcal{Z} \subseteq \mathbb{R}^{d'}$ for this demonstration are shown in Table 3.1b. They consist of the number of clouds, the size of clouds, the existence of morning mist, the storable PV surplus, specific battery investment costs, and the specific heat storage investment costs. The number of clouds n_c refers to the number of individual clouds in a simulation period, i.e., one day here. The cloud size s_c describes the duration of an individual cloud, and it is measured in terms of lost energy, i.e., the cloud clips the PV power plant's output to zero until the energy amount s_c is lost. The interpretable feature morning mist m_m describes a reduced PV power production in the early hours of the simulated day, and we implement it by reducing the solar radiation to zero during the m_m

time steps following the first time $av_{PV}(t) > 0$. The storable PV surplus s_{PV} measures the energy available for storage. It is defined as the sum over all modeled time steps $t \in T$ of the positive difference between the available PV supply $av_{PV}(t)$ and electricity demand $D_e(t)$, i.e.,

$$s_{PV} = \sum_{t=0}^{T} \max(0, av_{PV}(t) - D_e(t)) + n_c s_c, \tag{3.2}$$

$n_c s_c$ accounts for the lost energy from clouds, and its addition keeps s_{PV} independent from the interpretable cloud features.

Table 3.1a shows the input space $\mathcal{X} \subseteq \mathbb{R}^d$ of the exemplary building energy system model. Note that the specific battery investment costs and specific heat storage investment costs, which are part of $\mathcal{Z}$, also belong to the input space $\mathcal{X}$ and thus do not need to be mapped. A mapping $h^{\bar{\mathrm{x}}} : \mathcal{Z} \rightarrow \mathcal{X}$ is required for the four remaining features, i.e., number of clouds, size of clouds, morning mist, and storable PV surplus, targeting the solar radiation availability time series $av_{PV} \in \mathcal{X}$. First, $h^{\bar{\mathrm{x}}}$ takes the solar radiation time series and finds a multiplicator for which $\sum_{t=0}^{T} \max(0, av_{PV}(t) - D_e(t))$ is equal to the desired storable PV surplus. Then, the starting time step for each one of the n_c clouds is calculated. Starting points can be calculated by assuming clouds to be distributed at equal distances from one another or uniformly at randomly selected hours where $av_{PV}(t) > 0$. In our experiments, clouds either have a fixed sizeor each cloud has its size determined based on a Gaussian distribution with its mean equal to the interpretable feature and a fixed variance of 0.1 kWh. Mapping the clouds to the solar radiation availability av_{PV} affects the storable PV surplus s_{PV}. Hence, we calculate the difference of the s_{PV} after the mapping to the desired s_{PV} and distribute the difference equally to all time steps of av_{PV} where $av_{PV}(t) > 0$.

3.3.2. ESM Implementation

We solve the exemplary building energy system model for a single day with a time-step duration of 10 min, resulting in 144 time steps. To prevent unnatural storage depletion at the end of the optimization horizon, we define the storage levels of $t = 0$ and $t = 144$ to be equal. We assume a lifetime of 10 years for all technical components such as the battery and distribute their investment costs evenly over their lifetime. A constant grid electricity price p_e of 0.25 €/kWh is assumed.

An example of a solar radiation availability time series created by our mapping and the resulting cost-optimal battery scheduling is shown in Figure 3.3. Three clouds of randomized size and distribution are added to the original solar radiation availability

Table 3.1.: (**a**): Inputs of the exemplary building energy system. Time-dependent parameters are vectors with entries for every modeled time step. This model needs a total of $2T + 4$ values as its input, with T as the number of time steps. (**b**): The six interpretable features for explaining the cost-optimal battery capacity of the exemplary building energy system.

(**a**)

Input Parameters	**Symbol**	**Unit**	**Number of Values**
Specific battery investment costs	p_b	€/kWh	1
Solar radiation availability	$av_{PV}(t)$	kWh	T
Demand time series	$D_e(t)$	kWh	T
Grid electricity price	p_e	€/kWh	1
Specific heat storage investment costs	p_{HS}	€/kWh	1
Heat pump, coefficient of performance	COP	-	1

(**b**)

Interpretable Feature	**Symbol**	**Unit**	**Number of Values**
specific battery investment costs	p_b	€/kWh	1
cloud size	s_c	kWh	1
number of fleecy clouds	n_c	-	1
morning mist	m_m	-	1
storable PV surplus	s_{PV}	kWh	1
Specific heat storage investment costs	p_{HS}	€/kWh	1

time series. The solar radiation availability data correspond to the historical data of Darmstadt, Germany, obtained at [137] for 2019. Random days within May, June, and July are selected for our mapping. We assume a constant electricity demand of 1 kW for the electricity-only model. For model runs with randomized cloud distribution and size, we solve the ESM fifteen times and take the average of the results.

Standardized load profiles for German households [138], [139] are used for the heat and electricity demands for the full model in Section 3.3.4. The time series based on those profiles are scaled to have a total electricity demand equal to the electricity-only model (24 kWh per day). The total heat demand is twice as large as the electricity demand (48 kWh per day). We assume the heat pump to have a coefficient of performance of 3. The heat pump's power rating is large enough to cover the heat demand in every time step, making the heat storage optional from a pure heat balancing perspective. We use the mapping $h^{\bar{\mathbf{x}}}$ as described above with randomly distributed clouds and random cloud size.

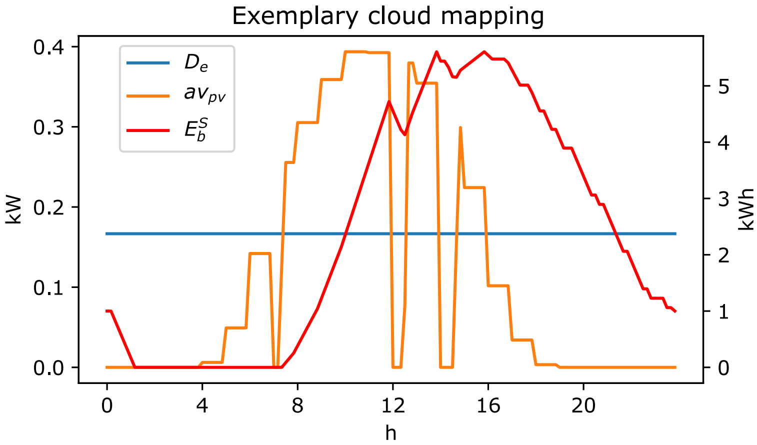


Figure 3.3.: Exemplary solar radiation time series (yellow line) created by the mapping for randomly sized and randomly distributed clouds. The resulting cost-optimal battery storage levels (red line), which provide for the constant electricity demand (blue line), are shown.

To deal with the scale heterogeneity of our interpretable features, we normalize them before applying the regression in Equation (3.1). To normalize the input and interpretable feature vectors, we group their entries. One group contains, for example, all entries in $\mathbf{x_i}$ that describe the solar radiation availability time series. We subtract the smallest value from each group member and divide by the largest difference within the group. We define $\tilde{\mathbf{x}}_\mathbf{i}, \tilde{\mathbf{z}}_\mathbf{i}$ as the normalized versions of $\mathbf{x_i}$ and $\mathbf{z_i}$ respectively, and $\tilde{\mathbf{y}}_\mathbf{i}$ as the normalized output values to be explained from the ESM $f(\mathbf{x_i})$, in this case, the cost-optimal battery capacities. We then rewrite Equation (3.1) to create the explanation for the building energy system as

$$\arg\min_{g \in G} \sum_{i=1}^{I} \pi_{\tilde{\mathbf{x}}}(\tilde{\mathbf{x}}_\mathbf{i})(\tilde{\mathbf{y}}_\mathbf{i} - g(\tilde{\mathbf{z}}_\mathbf{i}))^2 + \Omega(g). \tag{3.3}$$

For the distance metric $\pi_{\tilde{\mathbf{x}}}(\mathbf{x_i})$, we use an exponential kernel with a radial basis function on the normalized inputs, i.e.,

$$\pi_{\tilde{\mathbf{x}}}(\tilde{\mathbf{x}}_\mathbf{i}) = exp\left(\frac{\|\tilde{\mathbf{x}}_\mathbf{i} - \tilde{x}\|^2}{2\sigma^2}\right), \tag{3.4}$$

with $\tilde{x}$ as the normalized input vector at the POI. The vector σ is the standard deviation of all normalized input variations in the experiment.

For the complexity measurement $\Omega(g)$, we chose the number of non-zero weights of g, allowing only one weight to be non-zero. For the implementation, we choose $\Omega(g) = \alpha \|\mathbf{w_z}\|_1$, with $\mathbf{w_z}$ as the weights vector of the linear function g, i.e., $g(\tilde{\mathbf{z}}_\mathbf{i}) = \mathbf{w_z}\tilde{\mathbf{z}}_\mathbf{i}$. The parameter α is increased until $\mathbf{w_z}$ has only one non-zero entry. This LASSO path procedure is described in [140]. We will refer to the feature corresponding to this non-zero entry as the most relevant interpretable feature.

The python implementation of our experiments can be found on `https://github.com/pe0nd/LIME_for_ESD` (accessed on 20 January 2023).

3.3.3. Explanation Results: Electricity Only

We first omit the heat sector of the exemplary building energy system to facilitate the manual validation of the results. For our exemplary building energy system of a building with a PV power plant, we create an explanation for the cost-optimal battery capacity for two points of interest by comparing the most relevant interpretable features determined by our approach.

The first POI $\bar{\mathbf{x}}$ has low specific battery investment costs p_b of 600 €/kWh, a storable PV surplus of 5 kWh, 5 clouds, and a cloud size of 0.5 kWh. We refer to this POI as *cheap battery*. The input vector for the second POI is identical except for a higher specific battery investment cost of 1200 €/kWh. We refer to this second POI as *expensive battery*.

Table 3.2 shows that for *cheap battery*, the last feature that remains non-zero is the cloud size s_c; i.e., s_c is the most relevant interpretable feature for explaining the cost-optimal battery capacity at this POI. Applying our approach to the *expensive battery* results in the PV surplus s_{PV} being the most relevant interpretable feature affecting the cost-optimal battery capacity.

Verifying why the different most relevant interpretable features at each POI is a good explanation considering the different uses for battery storage. First, suppose that the specific costs of battery capacity are high. In this case, building a small battery is cost-optimal to mitigate the fluctuations in electricity production caused by the clouds during the day. The optimal battery capacity for this purpose corresponds to the energy lost by a single cloud. In contrast, if the specific battery investment costs are low compared to the electricity prices, it is cost-optimal to build a large battery. This battery stores

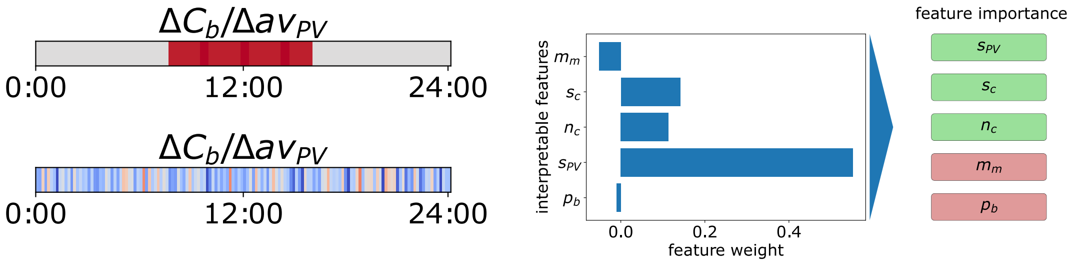
Table 3.2.: Output of the proposed methodology for the electricity-only ESM. The most relevant interpretable features for the cost-optimal battery capacity using different cloud mappings $h^{\bar{\mathbf{x}}}$ are shown. Interpretable features are the cloud size s_c, the number of clouds n_c, the storable PV surplus s_{PV}, specific battery investment costs p_b, and morning mist m_m.

Point of Interest $\bar{\mathbf{x}}$	**Equal Cloud Dist.**		**Random Cloud Dist.**	
	Equal s_c	**Random s_c**	**Equal s_c**	**Random s_c**
cheap battery	s_{PV}	s_{PV}	s_{PV}	s_{PV}
expensive battery	s_c	s_c	s_c	s_c

electric energy for nighttime, which otherwise would be curtailed from the PV power plant production during the daytime. In this case, the storable PV surplus is the dominant feature for determining the battery capacity.

Table 3.2 also shows the most relevant interpretable features at the points of interest for different implementations of the mapping function $h^{\bar{\mathbf{x}}}$, i.e., fixed or random cloud size and equally or randomly distributed clouds. The most relevant interpretable features are not affected by the different mapping functions. Hence, the explanations for these points of interest are robust against different implementations of the mapping function.

Figure 3.4 compares our proposed methodology and traditional sensitivity analysis. To this end, we provide in Figure 3.4a the effect of changes in the solar radiation availability time series av_{PV} on the cost-optimal battery capacity C_b for the POI *cheap battery*. The top plot shows the sensitivity analysis results for a fixed demand and a deterministic feature mapping with evenly distributed clouds of fixed size. The bottom plot displays the sensitivity results for a standardized load curve, namely demand and randomized cloud placement and size. For the deterministic scenario, the period between 7:40 and 16:00 turns out to be crucial for determining the optimal battery capacity. This is plausible since it is when the PV production exceeds the demand, resulting in surplus PV energy. However, this explanation is not easily deduced from the sensitivity results in the randomized setting. On the other hand, the proposed methodology provides clear and interpretable results as shown in Figure 3.4b. The fitted linear model provides weights for the interpretable features that are better understandable for both experts and non-experts.



(a) Sensitivities of C_b towards av_{PV}

(b) Proposed Approach

Figure 3.4.: Comparison of sensitivity analysis and the proposed methodology for explaining the cost-optimal battery capacity C_b at the POI *cheap battery*. (**a**): Sensitivity of C_b towards changes in solar radiation availability av_{PV}. Red values indicate a positive sensitivity, and blue values indicate a negative sensitivity. Top: deterministic cloud mapping and constant demand. Bottom: cloud mapping with random cloud sizes s_c, random cloud distribution, and a standardized load profile. (**b**): Weights of the fitted linear model. Green implies positive influence, and red implies negative influence. The features are ordered by absolute weight magnitude on the right.

3.3.4. Explanation Results: Including the Heat Sector

We now examine the exemplary building energy system, including the heat sector, i.e., with a heat demand, a heat storage, and a heat pump, and explain the cost-optimal battery capacity C_b and the cost-optimal heat storage capacity C_{HS}. The heat storage can offer temporal flexibility to the system if the heat pump converts excess electricity from the PV power plant into heat. This additional flexibility might allow part of the battery storage to be replaced by heat storage.

We investigate four different points of interest $\bar{\mathbf{x}}$ to check if the additional flexibility provided by the heat storage changes the explanation for the cost-optimal battery capacity. All points of interest map to interpretable features with a storable PV surplus of 9 kWh, 5 clouds, and a mean cloud size of 0.5 kWh. We consider a larger storable PV surplus compared to Section 3.3.3 since the additional heat demand has to be provided by the heat pump, which increases the electricity demand. The examined points of interest differ in their specific battery investment costs p_b and specific heat storage investment costs p_{HS}. We refer to the points of interest by their specific investment costs for heat storage and the battery. The specific heat storage investment costs are either *cheap* ($p_{HS} = 50$ €/kWh) or *expensive* ($p_{HS} = 200$ €/kWh). Specific battery investment costs are *cheap* ($p_b = 600$ €/kWh) or *expensive* ($p_b = 1200$ €/kWh).

The most relevant interpretable features determined by our methodology for all POIs are shown in Table 3.3. The storable PV surplus is the most relevant feature for explaining the cost-optimal heat storage capacity, no matter the specific battery investment cost if heat storage is *cheap*. The cost-optimal C_b for the points of interest with *cheap heat storage* is mostly explained by the storable PV surplus s_{PV} or the cloud size s_c, but not the specific heat-storage investment costs p_{HS}. One may anticipate that the specific heat storage investment costs would be the most relevant interpretable feature, as incorporating heat storage with heat pumps presents a more cost-effective way of utilizing electric energy generated by PV production, which could potentially replace battery storage in the exemplary building energy system. However, it is important to note that heat storage is not a complete substitute for battery storage in the exemplary building energy system, as heat cannot be converted back into electricity. As a result, battery storage cannot be fully replaced by heat storage.

Table 3.3.: The most relevant interpretable features for cost-optimal battery capacity C_b and heat storage capacity C_{HS}. Interpretable features are the cloud size s_c, the number of clouds n_c, the storable PV surplus s_{PV}, specific battery investment costs p_b, morning mist m_m, and specific heat storage investment costs p_{HS}.

Point of Interest $\bar{x}$		C_b	C_{HS}
cheap heat storage,	*cheap battery*	s_{PV}	s_{PV}
cheap heat storage,	*expensive battery*	s_c	s_{PV}
expensive heat storage,	*cheap battery*	p_{HS}	p_{HS}
expensive heat storage,	*expensive battery*	s_c	s_{PV}

For POI *expensive battery* and *expensive heat storage*, Table 3.3 shows that the cloud size s_c is the most relevant interpretable feature for C_b. The storable PV surplus S_{PV} is the most relevant interpretable feature for C_{HS}. At this POI, expensive heat storage is still relatively cheap compared to expensive battery storage. Hence it will be used to store most of the storable PV surplus. Keeping a small battery capacity is cost-optimal for storing electricity fluctuations caused by clouds. The POI *expensive heat storage, cheap battery* has the specific heat storage investment costs p_{HS} as the most relevant interpretable feature for explaining

C_b and C_{HS}. At this POI, the specific investment costs of heat storage capacity and battery capacity are close to balance; i.e., a change in relative investment costs shifts between battery and heat storage.

These examples show that the behavior of even this simple energy system model is not always intuitive. However, our method is able to create explanations in the form of the most relevant interpretable features for C_b and C_{HS}.

3.4. Experimental Validation: Country-Wide Model

In this section, we employ our approach to create an explanation for different German energy system transition paths towards low-carbon-emitting technologies, e.g., heat pumps and battery electric vehicles (BEVs). Section 3.4.1 introduces the ESM used. Next, we define interpretable features in Section 3.4.2. Finally, we show the explanation created in Section 3.4.3.

3.4.1. ESM Implementation

We use the German energy system model presented in [141]. The model is based on Germany's 2016 production capacities and energy demands as an initial condition and takes the heat, electricity, and transport sector into account. The objective function is cost minimization in a time horizon until 2050. We only simulate even years using a sparse time step selection of 8 weeks per simulated year to reduce computation time. Due to linearly decreasing CO_2 limits in all feasible solutions, the initial energy system has to change.

The ESM has the cost-optimal operation and extension plan as its output. We refer to this cost-optimal extension plan of a technology as its transition path. For comparing different transition paths by a single value, we explain the aggregated use of a technology cp by its energetic use as modeled years $s^{cp} = \sum_y \sum_t E^{cp}(t,y)$, with $E^{cp}(t,y)$ as the energy output and t as time and y as years. Note that a technology that is deployed earlier will typically provide more total energy than one that is deployed later, but this could be offset if the later-deployed technology is adopted at a faster rate.

3.4.2. Interpretable Features

We use three interpretable features: the fossil fuel price, the correlation of PV availability with heat demand, and the correlation of wind availability with heat demand.

The German ESM uses five fuels: coal, gas, oil, lignite, and biomass. We define a change in fossil fuel price to be the change in prices of all fuels, except biomass; i.e., for a 10% increase in fossil fuel price, the costs of coal, gas, oil, and lignite increase by 10%. Hence, we define the mapping $h^{\bar{\mathbf{x}}}_{fuel} : \mathbb{R}^1 \to \mathbb{R}^{4Y}$ with Y as the number of years of the optimization horizon since prices are fixed within a year in this model.

We define $h^{\bar{\mathbf{x}}}_{wind}$ and $h^{\bar{\mathbf{x}}}_{PV}$ to map the correlation to the input availability time series of wind and PV, i.e., $\mathbb{R}^1 \to \mathbb{R}^T$ with T being the set of time steps within a modeled year. The mapping takes the availability time series of wind (or PV) at the POI $\bar{\mathbf{x}}$ and alters them to increase or decrease their correlation towards the heat demand without changing their full load hours. First, we determine the correlation of the wind (or PV) availability time series to the heat demand. If the correlation is below the desired level, the time step with the highest heat demand is determined, as well as the time steps with the highest wind (or PV) availability. For the availability time series, the values of those two time steps are switched. Since the highest wind (or PV) availability is now in the same time step as the highest heat demand, the correlation of the two time series increases slightly. We continue this sorting process with the next highest values until the desired correlation, and thus the simultaneity of demand and production availability is reached. If the correlation of an availability time series and the heat demand is above the desired correlation, the lowest heat demand time series is used for the value swapping; i.e., the highest wind (or PV) availability will appear when the heat demand is at its minimum.

The correlation of the wind's onshore and offshore availability time series with the heat demand at $\bar{\mathbf{x}}$ is about 0.2 for both of them. For PV availability at $\bar{\mathbf{x}}$, the correlation with heat demand is about -0.33. We create two variations of the time series each for wind onshore, wind offshore, and PV production availability. The first set of time series created has its correlation increased by 0.2, and the second set of time series has its correlation decreased by 0.2. For the distance metric, we use an exponential kernel as in Equation (3.4) on the availability time series and the price vector.

3.4.3. Explanation Results: Energy Transition Paths

Figure 3.5 shows the cost-optimal heat provision by heat pumps, the cost-optimal transport provided by BEVs, and the cost-optimal electricity production from wind power for different input variations.

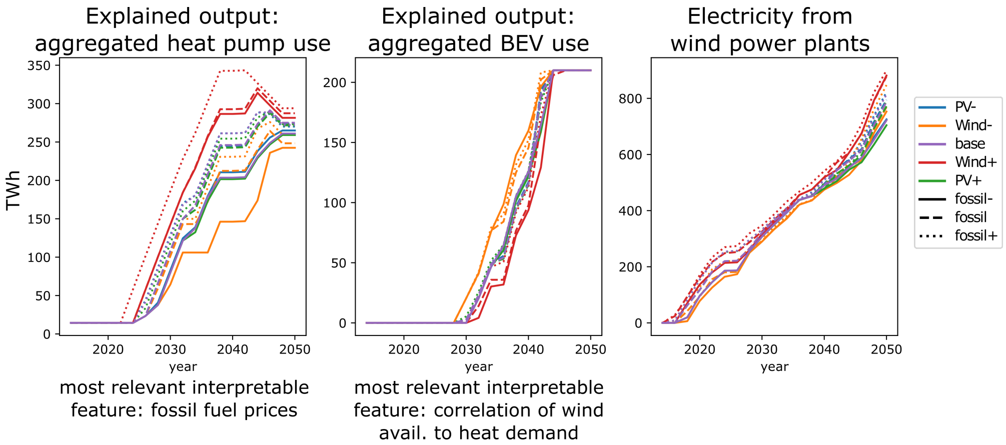


Figure 3.5.: Yearly energy output of different technologies for different interpretable input variations in the German energy system model. Each line is a different variation with the color encoding the correlation of the PV or wind availability time series to the heat demand and the line style encoding the fossil fuel price. On the left is heat output by heat pumps. In the middle is propulsion output from battery electric vehicles. On the right is electricity output of wind onshore and offshore power plants.

The most relevant interpretable feature for the transition speed towards heat pumps is fossil fuel prices. The importance of fossil fuel prices can be seen in the left graph of Figure 3.5. For variations with expensive fossil fuel prices, heat pumps are used earlier and to a greater extent than for cheap fossil fuel prices.

This explanation makes sense when considering the available technologies for providing heat within the model. Heat is produced by burning biomass, gas, or oil, or by using electricity to power resistive heaters or heat pumps. Resistive heaters have lower investment costs but are more expensive in the long term compared to heat pumps because of their lower coefficient of performance. Therefore, when fossil fuel prices rise, the cost-optimal solution to meet the heat demand is heat pumps. Additionally, electricity produced by wind power plants can be used more efficiently by heat pumps when wind power production is better aligned with heat demand; however, this effect is weaker than the fossil fuel price change.

The transportation energy provided by BEV for different input variations is shown in the center graph of Figure 3.5. The most relevant interpretable feature for explaining the transition towards BEVs is the correlation between wind power availability and heat demand. If the wind power availability is less correlated with the heat demand, a transition towards BEV happens earlier. If the wind power availability is correlated more with the heat demand, the transition happens in later years.

It may seem counterintuitive that less simultaneity between wind power production and heat demand would impact the transport sector, but it makes sense. When wind power, the cheapest source of renewable energy in this model, is less able to provide the required heat, more heat is generated by burning fossil fuels. This leads to an increase in CO_2 emissions. The additional CO_2 emissions needed for heating require savings in other sectors, as emissions must be kept below their limit. This explains an earlier transition to battery-electric vehicles (BEVs) away from combustion vehicles, as the transport sector is the cheapest option for reducing emissions in this setup. Furthermore, the right graph of Figure 3.5 shows that large capacities of wind power plants are built in later years of the model, even when the simultaneity of wind availability with heat demand is low. This indicates the need for CO_2 reduction.

3.5. Discussion

In this chapter we have shown how to transfer the concept of LIME to different ESMs for optimal ESD (RQ 3). When transferring the concept of LIME to energy systems, two challenges arise that have to be considered by domain experts: defining interpretable features and choosing the proper distance metric between different input variations. The definition of interpretable features is challenging since they have to be independent of one another. Dependent interpretable features will also be related in an explanation. For example, consider the exemplary building energy system from Section 3.3.3 and the POI with low battery investment costs. Instead of the storable PV surplus, we use the total energy availability $\sum_{t=0}^{T} av_{PV}(t)$ of the PV power plant as an interpretable feature. The explanation for the cost-optimal battery capacity will equally depend on cloud size, the number of clouds, and the total energy output availability since the energy output is affected by the number of clouds times the cloud size.

The distance metric weights the changes in the interpretable features based on the changes in the actual model inputs. For machine learning classifiers, model inputs are homogeneous; e.g., all inputs of an image classifier are pixels. The distance between interpretable variations can be determined by summing up the distance between individual input changes. However, finding an appropriate distance metric remains a challenge, as noted in [142]. For energy system models, the inputs are heterogeneous, e.g., cost parameters that are part of the model's objective function or availability time series in the model's constraints. Interpretable features that affect multiple constraints, such as by altering an availability time series, are considered more distant than those that affect only a single parameter of the objective because their input distances are simply summed. However, it should be noted that the objective often has a greater impact on the model's outcome than changes in the constraints, which is not considered by the distance metric. Investigating the effect of different distance metrics on the stability of the explanation could be an interesting area for future research.

When compared to sensitivity analysis the proposed LIME-based methodology offers several benefits regarding RQ 4. First, the number of explaining factors is significantly reduced. This is beneficial for discussions with experts and non-experts. The number of parameters that will be part of the explanation is based on a hyper-parameter of the method. Hence, even with a high-dimensional input space, the complexity of the explanation can be adjusted to the demands of the target audience by changing the hyper parameter. Second, the sensitivity results for each individual input dimension make it non-trivial to extract the underlying determining reasons. For example, considering the setup in Figure 3.4, the color map derived from sensitivity analysis could hardly encode factors such as cloud size in an obvious fashion.

Another benefit, when comparing the LIME-based approach to sensitivity analysis concerns computation times and ease of implementation. While sensitivity analysis for linear programs can be made efficient by exploiting the KKT optimality conditions [143], this is not implemented or easily accessible in many existing ESM frameworks. If sensitivities for high-dimensional inputs then have to be computed externally via numeric differentiation, the effort quickly becomes infeasible. In contrast, for the proposed methodology, the effort can be adapted by changing the number of selected interpretable features and the number of parameter variations used.

4. Causal Graphical Models for Energy System Models: A Negative Result

In Chapter 2, we presented causality as a pivotal concept in explanations within the realms of psychology and philosophy. This chapter tries to apply the causality concept to optimization-based ESMs (RQ 5), specifically to linear programs – a mathematical optimization model underpinning numerous ESMs. Intuitively it would make sense to causally argue step-by-step with ESMs. To illustrate, consider a scenario where a homeowner must determine optimal capacities for a rooftop PV power plant and for a battery storage unit devoid of optimization tools. In this context, the progression would probably involve deciding the PV capacity first. External factors such as roof size and average electricity demand influence the capacity of the rooftop PV power plant. Subsequently, one would maybe determine the optimal battery capacity based on the derived PV capacity (or the energy it generates). Conversely, deciding on the battery capacity before determining the PV power plant's capacity appears counterintuitive to the DMs since the optimal PV power plant fitting the derived battery might collide with external factors such as the roof size. Understanding these causal relationships is crucial for explaining behavior when model modifications occur. For instance, if constraints on PV capacity arise due to new building regulations, a causal model would readily reveal that this change necessitates a corresponding adjustment of the optimal battery storage capacity.

Causality and causal exploration is a central research topic in statistics for over three decades [130]. However, deducing causal relations from linear programs poses a novel challenge. Figure 4.1 illustrates a possible conceptual framework for deriving causal explanations from linear programs, wherein a causal graphical model is derived from an ESM or any linear program. This model would then serve as the foundation for extracting causal information, enabling the creation of explanations that elucidate the behavior of the ESM.

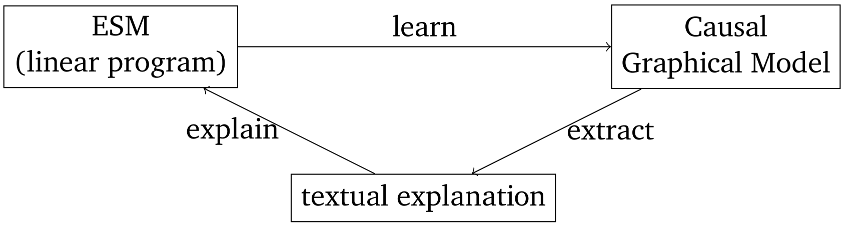


Figure 4.1.: Idea: Finding a causal graphical model for an ESM could help explaining its behaviour since textual explanations can be easily extracted from causal graphical model.

The two main contributions of this chapter are:

- We demonstrate that learning a directed causal graphical model from a linear program is not impossible in many important cases.
- We present an approach based on undirected graphical models that allows to split linear programs into causally independent subparts.

The subsequent sections of this chapter are structured as follows: First, we introduce directed causal graphical models and the concept of interventions. Section 4.2 explores the adaptation of the intervention concept to linear programs. In section 4.3, we unveil a fundamental issue with causality in linear programs – rendering it impossible to represent them as a directed causal graphical model in important cases. Section 4.4 provides an initial approach to salvaging the concept of causality for explaining linear programs through undirected graphical models. Finally, we discuss implications of this chapter and open challenges in section 4.5.

4.1. Causal Graphical Models & Interventions

A graphical model for a probability distribution can be written as a graph $G = (\boldsymbol{V}, \boldsymbol{E})$ with edges $\boldsymbol{E}$ and vertices $\boldsymbol{V}$ [129]. Each vertex $i \in \boldsymbol{V}$ holds a random variable X_i. We denote the vector of all random variables X_i as $\boldsymbol{X}$. Throughout this chapter we assume, that probability densities for all involved distributions of $\boldsymbol{X}$. An edge $\boldsymbol{E}$ connects a pair of vertices, e.g., vertex i and j which can be written as (i, j). If two vertices are connected by an edge they are called adjacent. A path in G connects two (not necessarily adjacent) vertices and can be written as a sequence of neighbouring vertices. For example, given a path from vertex i to vertex j passing vertex k this is $p = (V_i, \ldots, V_k, \ldots, V_j)$.

Causality is a directed relationship between two instances, i.e., flipping a switch causes a light bulb to glow but the glow of the light bulb does not cause the switch to flip. Therefore, a directed graph is suited to model such dependencies. A graph is directed if all edges in a graph are directed. An edge (i, j), is directed if there is no edge (j, i), which we do note as $i \rightarrow j$. If both edges (i, j) and (j, i) exist the edge is called undirected; if neither exists, the vertices i and j are referred to as non-adjacent. For a directed edge (i, j), we call vertex i the parent of j (j the child of i). Additionally, if there is a directed path from i to j, i is deemed an ancestor of j and j is a descendant of i. A directed graph that lacks cycles – meaning that for all vertex pairs (k, l), there is no directed path from k to l if there exists a directed path from l to k – is specifically known as a directed acyclic graph (DAG).

A DAG can serve as a representation for the joint probability distribution of all $\boldsymbol{X}$ as expressed by the factorization of its probability density function:

$$p(\boldsymbol{X}) = \prod_{i \in \boldsymbol{V}} p(X_i | X_{\text{parents}(i)}). \tag{4.1}$$

When a variable X_i lacks parents, it is termed *exogenous* and can be condensed to $p(X_i)$ utilizing $p(\cdot)$ as an arbitrary probability density function. Conversely, X_i is considered *endogenous* when it possesses parents. For a joint distribution represented by a DAG, the conditional independence of X_j from X_i, given a set of vertices S, can be assessed when X_j is d-separated [124] from X_i by S. This property, known as the *Markov property*, is articulated as:

$$X_i \perp\!\!\!\perp_G X_j | S \implies X_i \perp\!\!\!\perp X_j | S. \tag{4.2}$$

Here, d-separation [124] is denoted as $\perp\!\!\!\perp_G$. Variables X_i and X_j within a DAG are deemed d-separated if every path between them is blocked by a set S (which excludes X_i and X_j) or any subset of S. In order for a path to be blocked by S, S has to contain a node X_k, that fulfills one of the following properties[129]:

1. $X_k \in S$ for paths with the following structures:
   - $X_{k-1} \rightarrow X_k \rightarrow X_{k+1}$ or
   - $X_{k-1} \leftarrow X_k \leftarrow X_{k+1}$ or
   - $X_{k-1} \leftarrow X_k \rightarrow X_{k+1}$
2. neither X_k, nor $X_{\text{descendants}(k)}$ is in S for paths with the structure:

- $X_{k-1} \rightarrow X_k \leftarrow X_{k+1}$

If all causal independencies that can be derived from $p(\boldsymbol{X})$, can also be derived from G, then G is considered *faithful*, expressed as:

$$X_i \perp\!\!\!\perp X_j | S \implies X_i \perp\!\!\!\perp_G X_j | S. \tag{4.3}$$

It is worth noting that a graphical model may exhibit the Markov property while lacking faithfulness. This mismatch between graphical model and probability density occurs especially when two effects cancel each other out. To an observer this cancellation implies independence between two variables where there is none [129].

A Markovian and faithful DAG corresponding to the joint distribution $p(\boldsymbol{X})$ is a visual representation of the distributional interdependence among variables in $\boldsymbol{X}$. However, this graphical model of the joint distribution is not causal yet. To make the graphical models capable of supporting causal reasoning, it must be able to predict the effects of interventions [129].

An *intervention* is carried out on variable X_i when its value is set externally, independent of the values of its ancestors. For instance, intervening on X_i and setting it to a value x_i allows the measurement of the effect of this intervention on a second variable, say X_j. The intervention removes all links to X_i in the graphical model, rendering it an exogenous variable. The intervention results in a new DAG that is a subgraph of the original model. Following the definitions of [124], this intervention is denoted as $p(X_j|\text{do}(X_i = x_i))$. Unlike conditioning on X_i (i.e., $P(X_j|X_i = x_i)$), in an intervention, the distribution of X_i is replaced with a new distribution. The difference in the effect on X_j due to interventions with different values $x_i' \neq x_i''$, i.e., $p(X_j|\text{do}(X_i = x_i')) - p(X_j|\text{do}(X_i = x_i''))$, is termed the *causal effect* of X_i on X_j [144].

A question arises: how can a graphical causal model be inferred if the mutual dependencies of the random variables of $\boldsymbol{X}$ are unknown? A field of research addressing this challenge is known as *causal discovery learning*. Various approaches to identify causal relations from data exist [130]. A practical and straightforward method involves experiments that intervene on a variable and observe the consequent effects on another [145]. Real-world experiments often have difficulties in precisely targeting the desired value without inadvertently affecting or conditioning others [124]. However, in our theoretical setup of ESMs, all model parameters are accessible. In the next section, we introduce interventions on linear programs, a prerequisite for conducting experiments in causal discovery.

4.2. Interventions in Linear Programs

Our objective is to translate the behavior seen in a causal graphical model to linear programs. In a graphical model, an intervention eliminates the dependence of a variable on other factors, effectively severing all edges pointing to the node representing that variable. In a causal graphical model, a distribution like $p(X_i|X_{\text{parents}(i)})$ is deleted and substituted with a distribution assigning probability 1 to a specific value, i.e., $p(X_i \neq x'_i) = 0$ with x'_i as the target value with probability 1. In contrast, within a linear program, the values of X_i are not probabilistically determined but rather by a feasible set and an objective function. Unlike a probabilistic distribution, there is generally no single equation in the optimization model that can be eliminated to render X_i independent of other variables. Nevertheless, a mechanism exists to impose a specific value on a variable in a linear program, i.e., by adding a new constraint to enforce the desired value.

Definition 4.2.1 (Interventions on variables of linear programs)**.** *Consider the cost vector $\boldsymbol{c} \in \mathbb{R}^{n\times 1}$ and a variable vector $\boldsymbol{x} \in \mathbb{R}^{n\times 1}$ that describe a linear program of the form:*

$$\min_{\boldsymbol{x}} \boldsymbol{c}^T \boldsymbol{x}. \tag{4.4}$$

$$s.t. \quad \boldsymbol{A}\boldsymbol{x} \leq \boldsymbol{b}. \tag{4.5}$$

Constraints are described by a matrix $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ and a vector $\boldsymbol{b} \in \mathbb{R}^{m\times 1}$. An intervention $do(X_i = x_i^{'})$ on $X_i \in \boldsymbol{x}$ in the context of the described linear program is performed by adding a new constraint to the linear program that enforces X_i to be $x_i^{'}$, i.e.,

$$X_i = x_i^{'}. \tag{4.6}$$

Note that interventions must match the constraints, i.e., they should not render the linear program infeasible. For instance, if the linear program includes a non-negativity constraint, such as $X_i \geq 0$, an intervention setting $X_i = -1$ would result in infeasibility.

Moreover, the vectors $\boldsymbol{b}$ and $\boldsymbol{c}$, along with the matrix $\boldsymbol{A}$, are considered exogenous variables. Since these exogenous variables are not part of the optimization process, they cannot have parent nodes within the definition of the graphical model for the linear program (assuming the absence of hidden cofounders in real-world data within the scope of this work). When intervening on exogenous variables, the process involves simply changing their values. However, alterations to $\boldsymbol{b}$ and $\boldsymbol{A}$ potentially introduce infeasibilities, much like interventions on $\boldsymbol{x}$.

4.3. Linear Programs as Directed Causal Graphical Models

Now that we can intervene in the linear program, we can assess variable independence and gain insights into their causal relationships. However, the following theorem demonstrates that linear programs do not consistently exhibit causal behavior.

Theorem 4.3.1. *The dependencies of linear programs with interventions cannot, in general, be modeled by a causal graphical model.*

Proof via counter-example.

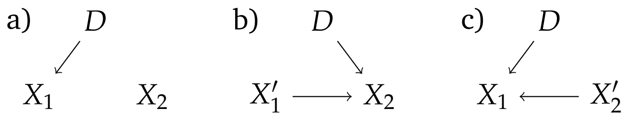


Figure 4.2.: Optimization based ESMs (i.e., LPs) cannot be explained causally since interventions change their behaviour.

Consider the following minimization target with optimization variables X_1 and X_2:

$$\min_{X_1, X_2}(X_1 + 2X_2) \tag{4.7}$$

subject to the constraints:

$$X_1 + X_2 = D \tag{4.8}$$

$$X_1, X_2 \geq 0, \tag{4.9}$$

where $D \geq 0$. The optimal solution to this optimization problem is $\boldsymbol{X}^* = (D, 0)$. X_1 directly depends on D in the optimal solution. Figure 4.2 a illustrates a causal graphical model, derived from observations of X_1 and X_2 as D is changed.

If a) would be the accurate causal graphical model reflecting the actual causal relations of the linear program, then all variables must be independent once we intervene on X_1. Consequently, we introduce an intervention to X_1 according to Definition 4.2.1 (e.g., $do(X_1 = a)$, with $0 < a < D$). This intervention changes the optimal solution to $\boldsymbol{X}^* = (a, D - a)$. From this optimal solution, we observe that X_1 no longer depends on its parent D, as intended by the intervention. However, X_2 now depends on D and the intervention value a. We derive the causal graphical model shown in Figure 4.2 b), where $'$ denotes the intervention on a variable.

One could argue that the causal relation from $D \to X_2$ was present from the beginning but was negated by the effect of $X_1 \to X_2$. To test this hypothesis, we intervene on X_2, which should not affect the causal relation in the model since it is a node with no children. By removing the intervention on X_1 and instead intervening on X_2 (e.g., by adding $X_2 = b$ as a constraint and setting $0 < b < D$), we observe the optimal solution to be $\boldsymbol{X}^* = (D - b, b)$. Figure 4.2 c) shows the resulting causal graphical model.

This last intervention reveals $X_2 \to X_1$, which contradicts the initial assumption of $X_1 \to X_2$. In total, a directed causal graphical model cannot explain the circular dependence inherent in this linear program, as any graph able to represent the pairwise dependencies would violate the non-circularity condition of the DAG. □

The example above illustrates a standard structure found in energy system models. The generic variables X_1 and X_2 could represent the electricity generated by two power plants. Equation 4.8 might describe an electric power balance equation, ensuring that the electric demand D is met by the combined generation of the power plants. This example highlights that, even if one were to argue that a change in electricity output from the first power plant (e.g., due to a defect) would cause the second power plant to generate electricity as a consequential effect, this linear argumentation does not hold at the level of the actual optimization. The power plants are optimized simultaneously, and their outputs are interdependent, making directed linear causation not a valid model.

It is essential to clarify that we do not assert that causality in linear programs is impossible in all cases. However, there is no (or at least no non-circular) causation within the optimization variables. In practice, even though it may not be intuitive for a human to choose a battery capacity before determining a PV capacity, the solver used to address the linear program does not rely on this intuition. Instead, the values are derived simultaneously (the solver internally follow some iterative order).

4.4. Linear Programs as Cyclic Graphical Models

Although we have demonstrated that linear programs cannot always be transformed into directed causal graphical models due to their cyclic dependencies, we can structure harness the derived insights about LPs by using cyclic graphical models of probability distribution. Cyclic graphical models can be beneficial in two ways:

1. **Create cyclic explanations:** Instead of the causal explanations we initially sought, we can develop explanations incorporating cyclic relationships. While these explanations may not be as easy to understand as strictly causal ones since no non-iterative construction from first principles is possible, they can still be used to create partial explanations.
2. **Utilize the Markov property to segment the graphical model:** By leveraging a concept called *Markov blanket* [146], we can partition the graphical model into sections that have minimal dependency on other sections. Segmentation allows for a more focused analysis of specific aspects of the linear program.

For the first use of circular graphical models, i.e., a direct interpretation, we need a way to interpret cyclic causality. As emphasized by [147], cyclic causality can be conceptualized as an equilibrium process or deconstructed into a series of interactions dependent on the actions of previous steps. An illustrative example of a model featuring circular dependence is an economic supply and demand model, often approached and comprehended as an equilibrium process. An explanation within this context might unfold as follows. If the price of a good is too high, the supply of that good would surpass its demand. This surplus in supply prompts prices to decrease, subsequently boosting demand. The heightened demand, in turn, elevates prices again (although to a lesser extent than the prior decrease), and this cyclical process persists until supply and demand converge at the optimal price. Similarly, an example of a circular model that can be explained well if transformed into a time series is the model of a thermostat. In this model, the heat generated by a heater influences the room temperature, which impacts the thermostat that regulates the heat generated. When viewed in discrete time steps, the current heat produced depends on the past room temperature, while the current room temperature depends on both the past room temperature and the past heat produced by the heater. This transformation renders the model non-cyclic. Similar chains of reasoning can be applied to explain ESMs. For instance, a change in the power output of a gas power plant in a preceding time step could increase the power output of a coal power plant in the current time step to meet the required power demand.

The second approach to construct explanations using cyclic graphical models involves the utilization of Markov blankets, as defined by [146]. In a graphical model, a Markov blanket for a set of nodes X comprises a minimal set of nodes Z necessary to render X independent of all other variables in the model, corresponding to Equation 4.2.

We introduce a concept parallel to a Markov blanket for a linear program. Consider a linear program defined by equations 4.4 and 4.5. We can modify this linear program

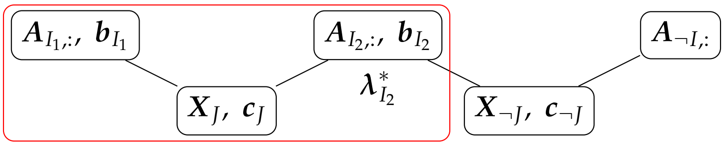


Figure 4.3.: Structure of the constraint of an optimization problem. The explanation of X_J only requires looking at constraints $A_{I_1,:}$ and $A_{I_2,:}$ that contain X_J. For $A_{I_2,:}$ that also contains variables that are not in J, the dual variables can be used to locally explain X_J without the need of $X_{\neg J}$.

slightly, transforming it into a quadratic program by introducing a quadratic term, as shown in the following equation:

$$\min_{\boldsymbol{x}} \boldsymbol{c}^T \boldsymbol{x} + \rho \|\boldsymbol{x}\|^2 \text{ s.t. } \boldsymbol{A}\boldsymbol{x} \leq \boldsymbol{b}. \tag{4.10}$$

By weighting the quadratic term $\|\boldsymbol{x}\|^2$ with a small $\rho > 0$, the solution to this problem remains similar to the original LP solution (if it is unique). However, the quadratic term ensures strict convexity, guaranteeing a unique solution and continuity of the solution with respect to changes in $\boldsymbol{A}$, $\boldsymbol{b}$, and $\boldsymbol{c}$. The optimal solution $\boldsymbol{x}^*$ can be characterized as the unique solution of [92]:

$$\boldsymbol{x}^* = \arg\min_{\boldsymbol{x}} \boldsymbol{c}^T \boldsymbol{x} + \rho \|\boldsymbol{x}\|^2 + \boldsymbol{\lambda}^* (\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}), \tag{4.11}$$

where $\boldsymbol{\lambda}^* \in \mathrm{R}^m$ represents the vector of optimal dual variables. For a fixed $\boldsymbol{\lambda}^*$ we can compute the optimal values for (groups of) primal variables with an index set J independently of the remaining variables. This is

$$\boldsymbol{x}_J^* = \arg\min_{\boldsymbol{x}_J} \boldsymbol{c}_J^T \boldsymbol{x}_J + \rho \|\boldsymbol{x}_J\|^2 + \boldsymbol{\lambda}^{*,T} \boldsymbol{A}_{:,J} \boldsymbol{x}_J. \tag{4.12}$$

Here, we only require the optimal dual variables connected to J through $\boldsymbol{A}_{:,J}$. Let I_1 encompass all rows of $\boldsymbol{A}$ referring exclusively to variables in J, and let I_2 include rows referencing variables in both J and other variables, as illustrated in Figure 4.3. Consequently, we can alternatively compute:

$$\boldsymbol{x}_J^* = \arg\min_{\boldsymbol{x}_J} \boldsymbol{c}_J^T \boldsymbol{x}_J + \rho \|\boldsymbol{x}_J\|^2 + \boldsymbol{\lambda}_{I_2}^{*,T} \boldsymbol{A}_{I_2,J} \boldsymbol{x}_J \text{ s.t. } \boldsymbol{A}_{I_1,J} \boldsymbol{x}_J \leq \boldsymbol{b}_{I_1}. \tag{4.13}$$

This approach allows large linear programs like ESMs to be divided into smaller segments for explanation. If $\boldsymbol{\lambda}_{I_2}^{*,T}$ is known, the set J can be isolated from the rest of the problem, i.e., I_2 acts like a Markov blanket. Such segmentation *enables goal-oriented explanations* similar to causal explanations; for instance, if the aim is to explain specific components of the ESM, they can be separated from the rest of the model and mainly explained independently. Note, the explanations are only locally valid since they are dependent on $\boldsymbol{\lambda}_{I_2}^{*,T}$, i.e., the explanation is only valid given that the coupling variables are known. Note, Instead of the optimal dual variables $\boldsymbol{\lambda}_{I_2}^{*,T}$ the optimal solution $\boldsymbol{x}^*_{\neg J}$ can be used to segment the LP, by using the duality of the primal and dual problems [92].

Consider a market-based ESM modeling an energy market with participants selling and buying electricity. Given the coupling variable, such as the price for electricity, each participant can be treated as an independent optimization problem and explained autonomously by other participants. Since the optimization problem for a single market participant is likely much smaller than the entire problem, established methods like sensitivity analysis can be employed to provide explanations efficiently.

One drawback of the segmentation approach of optimization problems for explanation is the potential loss of coupling effects. If market participants, for instance, can influence the electricity price through their behavior, their response to an explanation created through this method might differ.

Another potential drawback lies in the conversion from a linear to a quadratic program. However, despite this conversion to a quadratic program, the solution time should be mostly unaffected, especially when solvers employ interior point methods, which many solvers do.

A question up for future research is how the segmentation process can be improved to get smaller explanation-focused segments. Currently, we assume the choice of J to be predefined. However, selecting J is challenging since adding a variable x_{J_2} connected to J via $A_{I_2,:}$ may enhance the explanation but also make it more complex.

Consider a scenario where explanations are sought for all variables $\boldsymbol{x}$ of an optimization problem individually, i.e., $J_i := \{X_i\}$ for all $i \in \boldsymbol{x}$. In this extreme case, we would end up with a bipartite graph defined by $\boldsymbol{A}$, with rows and columns as two distinct sets of nodes and the values of $\boldsymbol{A}$ representing the edges between nodes. The explanation for each variable would be trivial since it could be determined by $\boldsymbol{\lambda}^*$ of all neighbors of X_i. Conversely, where all variables are in J, i.e., $J := \{\boldsymbol{x}\}$, everything would be included in the explanation. Striking the right balance regarding the number of clusters and determining which variables belong to each cluster is a challenge for future work.

4.5. Discussion

In this Chapter, we discussed causality, a pivotal concept in explanations, and investigated the application of causal discovery to linear programs. We demonstrated the transferability of interventions, a vital tool in causality, to linear programs. Nevertheless, linear programs inherently lack a clear causal direction between optimization variables A and B, as the optimization aims to discover the optimal mix of A and B. Consequently, a change in A induces a change in B, and vice versa, which answers our fifth RQ: causal graphical models cannot be used to explain ESMs because the relations in an ESM violate the acyclicity required by directed causal graphical models.

Hence, we looked into graphical models with cycles and how they could be beneficial for explaining ESMs. Grahical models with cyclical causation can sometimes be transformed into directed graphical causal models. This transformation involves unraveling circular dependencies by introducing a series of actions or a time dimension. For instance, A_t depends on B_{t-1} and B_t depends on A_{t-1}. Another approach interprets circularity as an outcome of equilibrium or optimization processes, yielding explanations that, while less easy to understand than non-cyclic causality, still provide insights.

To this end, we can leverage a concept from directed graphical models – Markov blankets. These blankets render groups of variables in a directed graphical model independent from others, given the set of their parents, children, and the parents of their children. Utilizing the optimal solution of a linear program, we can partition the program into distinct clusters for explanation. Given the optimal dual variables these clusters are mutually dependent. This approach proves beneficial in disregarding unimportant variables during explanations breaking down complex models into smaller, more manageable components. However, determining the optimal clusters remains an open topic for future research.

5. Energy System Model Explanation in Education - an Interactive Approach

In Chapter 2, we presented interactive approaches as one of the methodical categories for explaining ESMs. These approaches include tools such as chatbots or Serious Games offering a "hands-on" means to comprehend energy systems. Current approaches within Serious Games for ESM are rare and the few that exist, focus on singular aspects of ESMs, such as local energy trading [90] or power grid extensions [148]. In this chapter, we introduce an interactive approach to a top-down ESM to address RQ 6. Unlike existing interactive ESM explanations, our approach diverges from specific sectoral or problem-focused settings within an energy system, providing instead a more comprehensive view on the energy transition.

Referred to as the Energy Transition Game (ETG), our proposed approach targets college-level education, serving to educate students about the challenges associated with energy transition. The primary design objective of the ETG is not to intricately simulate an energy system but rather to reduce the complex topic of energy transition into fundamental concepts graspable within a few hours.

The ETG forms part of a three-part interdisciplinary university course for graduate students, comprising a lecture, seminar, and the ETG. To assess the effectiveness of the ETG in enhancing students' understanding of energy transition (RQ 7), we conducted a survey designed to isolate the learning outcomes attributable to the game from those derived from the lecture and seminar components of the course.

The ETG is the work of multiple authors with the following contributions: Concept & Game Design: Jonas Hülsmann, Prof. Dr. Stefan Niessen; Implementation: Jonas Hülsmann, Alexander Wagner; Supervision: Jonas Hülsmann, Prof. Dr. Stefan Niessen; Evaluation: Jonas Hülsmann.

The remainder of this chapter is structured as follows: In Section 5.1, we introduce the learning objectives of the course that the ETG aims to enhance. Section 5.2 outlines the game's rules, an illustrative round, and deliberate constraints imposed to mitigate its complexity. The implementation of the ETG is expounded upon in Section 5.3. Following this, Section 5.4 outlines the methodology employed to evaluate the ETG and presents key findings. Subsequently, in Section 5.5, we discuss the results of this chapter.

5.1. Learning goals

We designed the ETG as an integral component of the university course "Energiewende gestalten" (designing energy transition), catering to graduate students from the electrical engineering and political science departments. The course endeavors to achieve four primary learning objectives:

1. **Units and Magnitudes:** Students are expected to comprehend the varying orders of magnitude relevant to energy matters. Additionally, they should distinguish between energy and power, grasp the scale of CO_2 emissions, and develop a rough understanding of the costs associated with the energy transition.

2. **Processes of Power Generation:** Students should understand the processes involved in power generation's operational and expansion planning. Students should learn about upcoming challenges, such as the uncertainties inherent in renewable generation, and learn rudimentary strategies for cost-effective power generation planning.

3. **Agents, Institutions, and Potentials:** The ETG is supposed to familiarize participants with the diverse stakeholders shaping the energy transition landscape. We emphasize understanding the roles of institutions and other actors impacting the transition and identifying emission reduction potentials within crucial groups involved.

4. **Conflicts:** Recognizing various stakeholders' divergent interests and interdependence is important for understanding their behaviour regarding the energy transition. Students should be equipped to identify and articulate conflicts of interest among significant sectors involved in the energy transition.

To realize these learning objectives, we structured the course into three distinct parts: a lecture series, seminars, and the ETG. The lecture segment comprises six interdisciplinary sessions covering the energy transition's technical, economic, and political dimensions. Seminars expose students to recent research studies, where they collaborate in small groups to dissect, critically evaluate, and present key findings.

The choice of a Serious Game as the final segment stems from its ability to provide an immersive learning experience, as noted in the literature [149]. Such games entertain and engage participants [24], ultimately enhancing learning outcomes compared to traditional methods [150].

In particular, the ETG is designed to encourage active application and revision of knowledge across diverse facets of the energy transition. By assuming the roles of crucial agents, students navigate cooperative or competitive dynamics aiming at realizing a net-zero energy system by 2050. Each role has distinct goals, and highlights a different energy transition dimensions and allowing for player immersion.

5.2. An interactive energy transition game

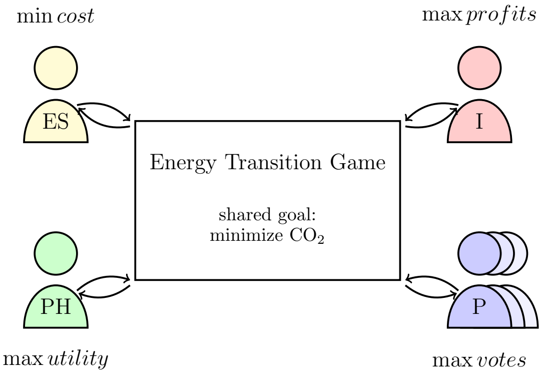


Figure 5.1.: Explaining the energy transition with a Serious Game. Different players are assigned a role with individual goals and interact with the other player through the game, which presents them with a common goal: to achieve a net-zero energy system.

Figure 5.1 provides an overview of the ETG concept. Players assume one of four roles: energy sector (ES), industry (I), private household (PH), or politics (P), each with distinct objectives. Ideally played by four to six participants, each role is assigned once in a four-player game, with politics potentially assigned multiple times in larger games.

The primary aim of the ETG is to achieve a net-zero energy system by 2050, spanning six rounds from 2020 to 2050, i.e., each round symbolizes five years. In addition to this main goal, players can influence specific parameters within the game, enabling them to pursue their objectives.This interactive setup aims to give participants a holistic understanding of energy system dynamics, emphasizing interactions and interdependencies.

Consequently, the game's design necessitates collaboration and interaction among players to fulfill their respective goals. The decisions of one player depend on the actions of others. For instance, the energy sector's provision of electricity and heat impacts the industry's manufacturing choices and profitability.

To mitigate the complexity of the energy transition, we streamlined decision-making by reducing it into discrete actions, sequenced one at a time. Visual aids and action previews assist decision-making, while roles are scaled to depict typical units and magnitudes relevant to each, aligning with course learning goals. For example, the yearly electricity generation, the emitted CO_2 as well as the variables costs of a coal power plant are within the right magnitude compared to a real world scenario. On the other hand, the sum of generated energy mirrors the German energy system's dimensions, where not a single agent but multiple entities are involved in power generation. To account for individual roles and the system's aggregate, actions taken by a player are amplified by a leveraging factor. For instance, there are 100 million private households in the ETG, players get to play one private household, to learn important magnitudes of the role. To depict the influence on the whole ETG, the actions taken are multiplied by 100 million.

A trading system enables resource exchange, allowing participants to focus on their objectives while engaging in trade and facilitating player interaction. A competitive election element simulates political dynamics, with elected parties influencing the game through regulatory adjustments like taxes, subsidies, or constraints.

Subsequent sections detail each role's objectives, tasks, and CO_2 emission management abilities, followed by an outline of a typical game round. Lastly, we outline limitations of the ETG design, that are taken to simplify the complex, overwhelming reality to a level suitable for the audience of graduate student from electrical engineering and political science.

5.2.1. Roles

Energy Sector

The primary objective of the energy sector is to supply electricity and heat at minimal cost. Achieving this goal demands prudent planning, balancing each power plant's fixed annual expenses and variable operational costs. Notably, energy cannot be stored between rounds, necessitating careful resource management to ensure continuous supply with no or minimal overproduction.

Various energy sources are at the disposal of the energy sector, ranging from fossil fuels like coal, gas, and uranium to renewable sources such as wind and solar radiation. While fossil fuels incur distinct costs associated with their respective power plants, renewable sources are treated as resources in the game, subject to varying availability each round. Within the game, wind energy is further categorized into on-shore and off-shore, with both types having a correlated availability. The power plants are designed to resemble the cost of their real world counterpart and are normalized by their energy output, i.e., have all the same output but varying inputs. Renewable energy power plants operate autonomously, generating electricity without player intervention, whereas conventional power plants can be controlled by adjusting their used output capacity in increments of 0.5 GW.

At the start of the game, the energy sector inherits a power plant setup resembling Germany's 2020 infrastructure. Over time, plants age and retire, necessitating the construction of new facilities to maintain capacity. Players can introduce new power plants into the system without upfront costs, although each new installation incurs a four-round annuity payment reflective of its type.

Figure 5.2 depicts the interface of the energy sector role within the ETG. At the top are the global variables shared by all players. The current availability of renewable energy sources is displayed on the top left. The top center showcases the CO_2 emissions with a suggested limit for the round, diminishing linearly to zero by 2050 to track emission reduction progress. The top right corner indicates the current game year. Adjacent to the year display, an undo button allows players to revert the current round, subject to unanimous agreement. Below the CO_2 display is the role name, "Energy Sector," and a leveraging factor denoted by the number ten.

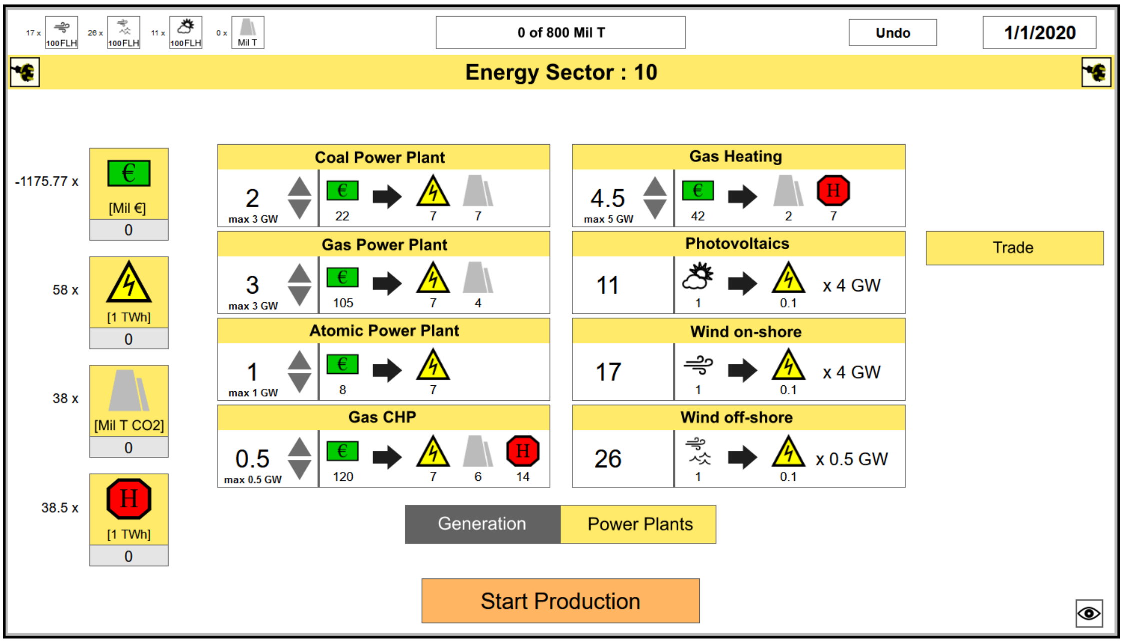


Figure 5.2.: Overview of the game interface for the energy sector. The top row of items displays global parameters, such as the availability of renewable energy sources, CO_2 emissions, and the current year of the game. Below the global parameters is the name of the role and its leverage factor. Below the role name is the actual game board of the role. For the energy sector, the board shows the stockpiles of different resources on the left, the current power plants in the middle, a second tab to switch to the building section below the power plants, a button to start power generation, and a button to access the trade menu on the right.

The main playing field is divided into the resource stockpile on the left, the power plant section in the center, and a trade window button on the right. The resource stockpiles include money, electric energy, CO_2 emissions, and heat energy, each represented by a corresponding symbol and unit of measurement (e.g., [Mil €] for money). A preview to the left of each stockpile illustrates the potential change resulting from power generation. The grey section of each stockpile shows any units in store that are finalized, e.g., after the power generation.

Eight rectangles in the power plant section represent different power plant types, each labeled at the top. Below the label, the resource transformation rate indicates how resources can be converted into energy or emissions. Players can adjust the capacity used for generation using the grey arrows up to the maximum capacity built for the specific type of power plant. Renewable power plants operate at maximum capacity based on available resources. Switching to the "Power Plants" tab reveals the remaining power plant lifetimes and allows the construction of new facilities. The "Start Production" button activates the chosen operation plan, locking it in once initiated.

On the right, the trade button provides access to the trade menu, enabling the energy sector to offer electricity to private households and the industry and heat energy to the industry.

Industry

The industry's primary goal is to maximize profitability by supplying goods to private households. Due to the different product makespans, strategic planning becomes is necessary. Coordination with private households to forecast demand and collaboration with the energy sector to secure necessary resources are essential to industry strategy. Upgrading production facilities not only reduces CO_2 emissions but also lowers production costs, aligning with the industry's profit-maximization objective.

In the ETG, the industry plays a crucial role in producing three types of goods essential for private households: consumption goods, investment goods, and transportation. The industry relies on acquiring electric and heat energy from the energy sector to produce the required goods. Unlike the energy sector's power plants, industry production facilities do not have a fixed lifespan or limitations. However, there is a delay in the production of consumption and investment goods, occurring one and two rounds after initiation. Transportation, like electricity and heat energy, is produced immediately and cannot be stored, but consumption and investment goods can be stored for future use.

The industry can enhance its production capabilities through investment in research, altering resource intake, and minimizing emission output. Research is facilitated by a technology tree offering various upgrades, some of which may require prior updates for implementation. Upgrades applied in the current round affect production facilities from the subsequent round onward.

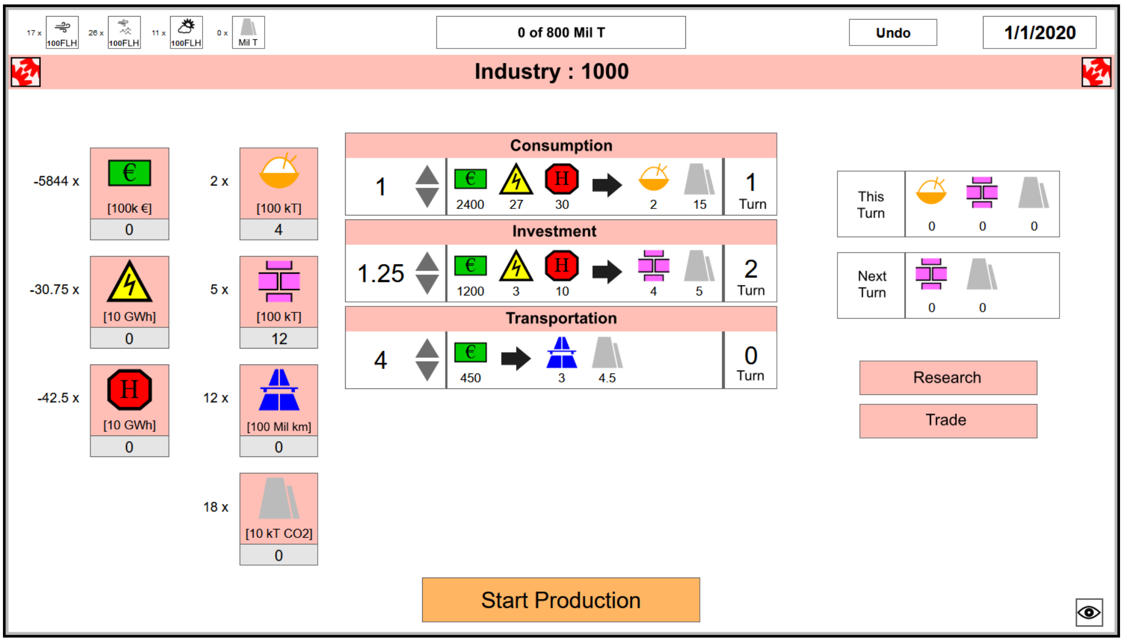


Figure 5.3.: Overview of the Industry interface. At the top of the interface are the global parameters, the role name, and the leverage factor. To the left of the interface are your resource supplies. In the middle are the production facilities for the industry's three goods, with a button to start production below them. The right side of the interface shows the goods that will be available in future rounds. Below the future goods are two buttons. One button takes you to the research tree, and the other opens the trade window.

Figure 5.3 illustrates the interface of the ETG for the industry role. At the top, global game parameters are displayed, including currently available renewable energy sources, the CO_2 emission counter, the undo button, and the current game year. Below these parameters, the role name "Industry" and its leveraging factor (set at 1000) are indicated.

On the left side, the industry's resource stockpiles are depicted, featuring symbols representing various resources, units, current stock levels, and previews of anticipated changes resulting from current production settings. Featured resources are from top to bottom and left to right: money, electric energy, heat energy, consumption goods, investment goods, transportation, and CO_2 emissions.

The center section showcases production facilities, with each rectangle representing a facility producing a different type of good. A red banner atop each facility denotes the produced good. Below, resource intake and outcomes are detailed, alongside production duration. Players adjust production levels using the gray arrows, with consumption and transportation adjustable in steps of one and investment goods in steps of 0.25.

The right side displays future goods expectations from previous round production, alongside buttons to access the research tree and trade window. The industry can purchase electric and heat energy from the energy sector and sell consumption goods, investment goods, and transportation to private households.

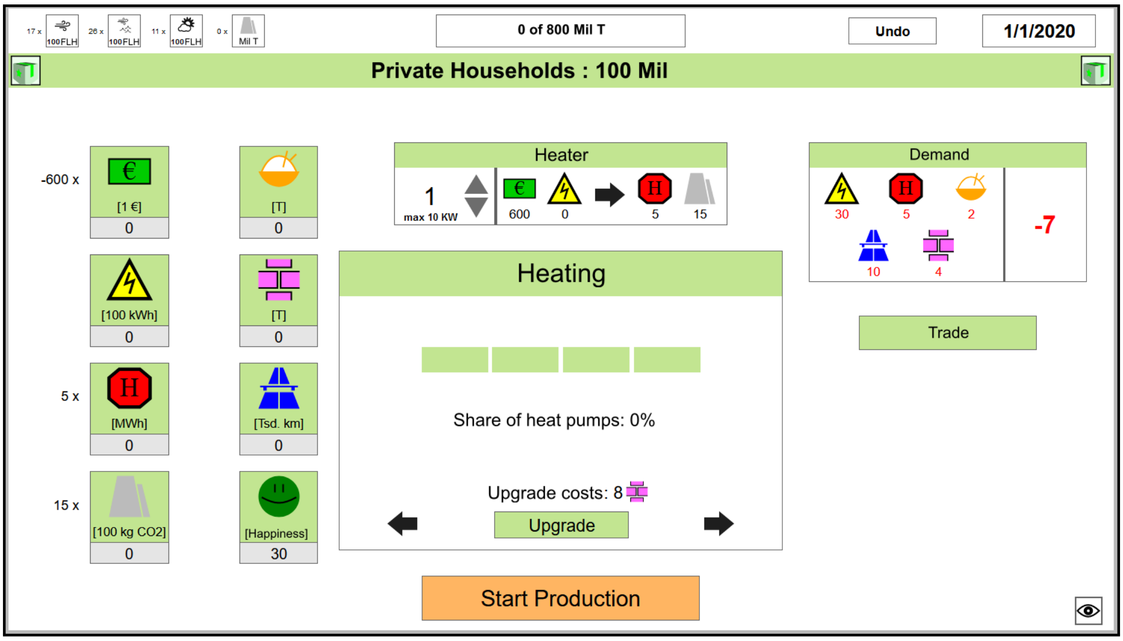
Private Households

The primary objective for private households is to maximize welfare in the form of happiness. Happiness is achieved when a set demand is met each round, which is necessary to provide for their basic needs. This demand encompasses electric energy purchasable from the energy sector, consumption goods, investment goods, transportation available from the industry, and heat energy produced by the private households themselves. Private households receive a fixed income each round to procure these goods from other roles.

Players controlling private households have three options to influence their demand and mitigate CO_2 emissions: implementing building insulation, upgrading heaters to heat pumps, and altering lifestyle choices. Both building insulation and heater upgrades require investment goods procured from the industry and feature five and four upgrade steps, respectively. Insulation steps reduce heat demand incrementally, while heater upgrades reduce heating costs and CO2 emissions while necessitating additional electric energy for operation. Lifestyle adjustments, available each round, can slightly decrease demand for investment goods and transportation at the expense of happiness or increase happiness at the cost of increased demand for these goods.

Figure 5.4 depicts the interface of the ETG for private households. Like other roles, the top section displays global parameters such as renewable energy sources, CO_2 emissions, the undo button, and the current game year. Below these parameters, the role name is shown alongside the leveraging factor, set at 100 million for private households.

The private household stockpiles on the left side of the interface include money, electric energy, heat energy, CO_2 emissions, consumption goods, investment goods, transportation,



Share of heat pumps: 0%

Upgrade costs: 8

Figure 5.4.: Overview of the private households game interface. At the top of the interface are the global parameters, the role name, and the leverage factor. To the left are the Households' resource supplies. The middle area shows the Household's Heater at the top and a selection of different Household options to either upgrade their Heater, invest in insulation, or adjust their lifestyle. The right side of the interface shows the demand of the households and a button to open the trade window.

and happiness. Each stockpile features an icon, unit measurement, current level, and a preview of the effect of current production settings.

In the central section, the interface features the heater, akin to a power plant, with up and down arrows allowing selection of production quantity. Unlike the power plants, the heater does not age but can be upgraded, with options toggling between upgrading the heater, building insulation, and lifestyle changes.

The right side displays the demand: 30 units of electric energy (3 MWh), five units of heat energy (5 MWh), two units of consumption goods (2 tons), ten units of transportation (10,000 kilometers), and four units of investment goods (4 tons). A display indicates the current change in happiness, with categories turning green when demand is met, and the expected decrease in happiness is lowered. Once all demands are fulfilled, the happiness increases instead of being decreased.

Below the demand display, a button opens the trade window, enabling households to purchase required goods from the industry and energy sector.

Politics

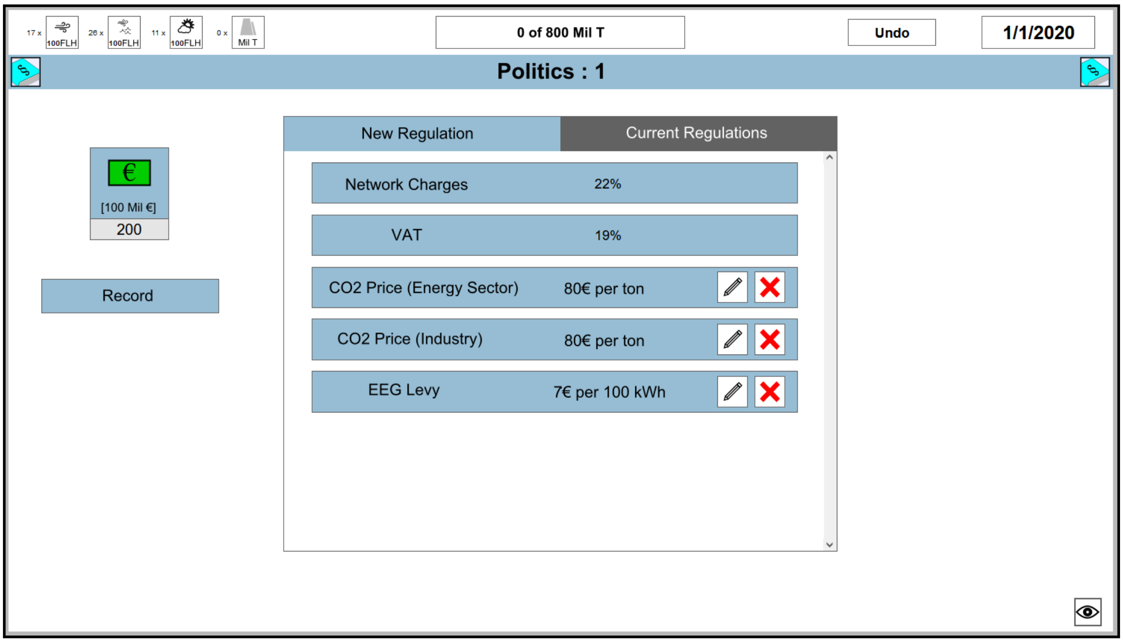
The role of politics in the ETG is unique in that it can be undertaken by multiple players, setting it apart from other roles. The goals of politics vary slightly depending on the number of players involved.

In the scenario where multiple players assume the role, they engage in electoral competition to secure a position in the government, where they wield the authority to introduce new regulations in the ETG, such as taxes, subsidies, and limits. Each player's objective is to attain governmental participation for as long as possible by maximizing votes, thereby convincing most other players to support them. The voting process occurs offline, i.e., the game interface does not feature the voting process. Competing politicians present election manifestos outlining the benefits their governance would bring to individual roles. Following manifesto presentations, other roles cast their votes through preference voting, with each player having three votes to distribute among the politicians. Coalition building may be necessary if no politician secures an absolute majority. The government, once formed, can enact new regulations or modify existing ones to please their voters.

Alternatively, if only one player takes on the role of politics, there is no electoral competition. Consequently, their individual goal shifts to balancing wealth, represented by resources like money, among the various roles within the state as effectively as possible. Politics can achieve this by implementing new game regulations through subsidies and taxes.

Taxes and subsidies impact the state's financial status, with the government balancing income and expenses to avoid accruing debt.

Figure 5.5 showcases the interface of the ETG for the role of politics. Like all roles, at the top of the interface are the global parameters displaying the availability of renewable energy sources, the CO_2 display, the undo button, and the current game year. The role name and leveraging factor are also indicated.



Network Charges	22%
VAT	19%
CO2 Price (Energy Sector)	80€ per ton
CO2 Price (Industry)	80€ per ton
EEG Levy	7€ per 100 kWh

Figure 5.5.: Overview of the Politics interface. The top of the interface shows the global game parameters, the role name, and the leverage factor. The left side of the interface shows the stockpile with the state's budget. Below the state's budget, a button allows you to open a record of the state's income and expenditures for the previous round. The center of the interface displays the current rules and allows you to edit, delete, or add rules.

Politics in the ETG solely manages money as a resource; hence, only one stockpile is shown on the left side of the interface. Below the money stockpile, a button opens a window displaying the role's income and expenditures from the previous round.

In the center of the interface, current regulations are displayed. This interface offers a selection of named regulations, including CO_2 taxes, EEG levies, production-related taxes, coal limits, subsidies for renewable power plants, and nuclear power plant capacity. Additionally, the government can provide general subsidies to specific roles or issue offline regulations, such as power plant type restrictions. However, specific regulations, like network charges and value-added tax, are fixed and cannot be altered. The game begins with the two fixed regulations and three variable rules. Variable rules can be edited using the pen icon or deleted with the red "X". New regulations can be added via the "New

Regulation" tab, where they are categorized into subsidies and taxes. The coal limit regulation is included in the tax section for interface simplicity. When selecting a new regulation, a brief dialog describes the regulation and prompts for a value input, such as the level of the CO_2 emission tax.

5.2.2. Exemplary round

The following eight steps outline a typical round of the ETG. While this structure is recommended for early rounds, for later rounds, the energy sector and the industry can perform steps 4 and 6 concurrently with other players to speed up gameplay.

1. The households determine their required goods and energy for the round. The demand may change based on lifestyle choices or upgrades to heaters or insulation from previous rounds. Once determined, households request the required goods from the industry and energy sector. All politicians begin to draw up their electoral manifesto which has to be finished before step eight. The current government also decides what new taxes, subsidies, or laws will come into play in the next round.
2. With knowledge of the private households' requests, the industry plans its production, considering that some goods have delayed availability. The industry then requests the required energy for current production from the energy sector.
3. The energy sector, now aware of the complete demand for electricity and heat, creates an operational plan and schedules power plants accordingly. Once energy is available, the energy sector negotiates with households and the industry on the price of electricity and heat.
4. The energy sector can invest in new power plants or replace old ones that have reached the end of their lifespan. New power plants become available in the next round.
5. After receiving the requested electricity and heat, the industry begins production. The industry adjusts its production plan accordingly if the energy sector can only provide partial energy. Once goods are produced, the industry offers them to households at a negotiated price.
6. The industry invests in research, affecting required energy, costs per produced good and emitted CO_2. These upgrades become available in the next round.

7. With goods from the industry and energy sector, households fulfill their basic requirements, except for heat. The private households can produce the required heat using their heater. Surplus investment goods may be used to upgrade insulation or heaters for future rounds. The private households gain or lose happiness depending on lifestyle and requirement fulfillment.

8. After the energy sector, industry, and households finish their rounds, the election phase begins if more than one politician is in the game. Each politician presents their electoral manifesto and may make political pledges. Other roles then vote for each politician. Politicians may engage in political bargaining to form a new government if no outright majority is achieved. If only one politician is present, this phase is skipped, and that politician remains in power.

5.2.3. Limitations of the ETG Concept

Several design limitations were implemented to make the ETG suitable for educating students about the highly complex topic of energy transition, but without overwhelming them. Here are some major limitations and the reasons behind them:

- **Missing Technologies:** Many technologies in the energy sector, such as energy storage and carbon capture and storage, were omitted to reduce complexity.
- **Transmission Grids and Balancing Power:** Despite their importance, transmission grids and balancing power were excluded to prevent further complexity, as the energy sector is already one of the more complex roles.
- **Renewable Energy Source Uncertainties:** Uncertainties in renewable power generation, like complete unavailability during short time frames (dark doldrums), were not depicted due to mismatching with the game's yearly time resolution, although they are potential real-world issues.
- **Simplified extension planning:** Various simplifications were made, such as assuming all conventional power plants have the same availability of 7000 hours a year, simplifying building times of one round and simplified lifetimes of eight rounds, and omitting building limitations for power plants. Further, players can build power plants by capacity in 0.5 GW steps instead of building individual power plants. The cost of building new power plants was adjusted to match the normalized lifetime, i.e., the cost of power plants with shorter lifetimes in a real world scenario were increased to meet the game lifetime.

- **Transport Sector:** The transport sector was integrated into the industry to reduce complexity.
- **Production Facilities and Times:** Production ramp-up times were neglected, allowing high production peaks in one round without negative effects, contrary to real-world scenarios.
- **Technological Advancements:** The success of research leading to technological advancements in the industry sector is not guaranteed, unlike in the game, where costs, effects, and the order of technologies are predefined.
- **Fixed Income of Private Households:** Private households receive a fixed income each round, simplifying planning but neglecting real-world effects like inflation and unemployment.
- **Heating:** Heating methods were simplified, ignoring technologies like district heating and solar thermal energy to reduce complexity. Further, the heater has to be partially upgraded in 25% steps to better reflect the slow transition to different heating systems in the real world scenario.
- **Incomplete Markets:** The energy sector and industry act as monopolists in the game, with no competition or mechanism to prevent abusive pricing. Players are advised not to abuse this monopoly status and choose fair prices, but no mechanism inside the game would punish malicious behavior.
- **Imports and Exports:** Players cannot import or export goods, simplifying the market dynamics.
- **Debts and Interest Rates:** There are no limits on debt and interest rates within the game, simplifying financial calculations.
- **Voting System:** The voting system in the game does not accurately represent reality, as it gives equal voting power to all sectors despite differences in size and influence.
- **Available Regulations:** The game offers only a small set of predefined regulations, limiting specific regulatory actions. However, players are allowed to agree on offline regulations.

Additionally, leverage factors were not chosen to represent the actual number of agents of each sector in Germany but rather to maintain realistic orders of magnitude for the energy and financial units, thereby facilitate one-to-one unit trading between roles.

5.3. Implementation

The development of the ETG began in 2019 intending to create a physical board game. However, the outbreak of the Corona pandemic halted progress on the board game, leading to a shift towards developing the game as an online game instead. Initially, the ETG was crafted as a single-player prototype using JavaScript. As the project evolved, this prototype was transformed into a multiplayer experience employing a client-server architecture to facilitate data coordination among multiple players. A student assistant implemented the client-server structure. Opting for a web-based approach was a strategic decision to ensure student accessibility. This approach accommodates participants with varying hardware configurations, computing power, and operating systems, enabling a broader audience to engage with the game seamlessly.

Figure 5.6 illustrates the client-server structure of the ETG implementation. Players utilize a web browser as their client to connect to a website hosted on the game server. The server furnishes the client with the game's frontend, which was developed using JavaScript and HTML.

A lobby system facilitates player interaction within the frontend, allowing them to assemble with other participants and commence a game. The progress of each game is saved after every round and stored within an SQL database. Should players wish to pause an ongoing game, the lobby system can resume unfinished sessions later.

Moreover, the game data preserved within the database can be visualized and utilized for in-depth analysis, offering insights into the reasons behind a group's success or failure in navigating the energy transition within the game.

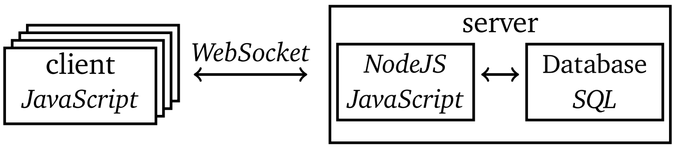


Figure 5.6.: ETG's client-server structure. Players access the game through a web browser. The game logic is implemented in JavaScript in the frontend and backend. Game data between different players of a game is transferred using the WebSocket protocol. A SQL database is used to store game data, either to continue paused games or to perform data analysis on finished games.

Communication between the frontend and backend occurs via the WebSocket protocol. The backend, which operates on a Node.js server [151], is coded in JavaScript, completing the client-server architecture of the ETG implementation.

5.4. Evaluation: Did the ETG improve the understanding of energy transition?

To assess the effectiveness of the ETG in meeting the learning objectives (RQ 7), we conducted a comprehensive survey consisting of three parts. In this section, we outline the methodology employed in the survey, which entailed an online questionnaire, and present the key findings. For the English translation of the questionnaire and a detailed list of all results, please refer to Appendices B and C, respectively.

5.4.1. Methodology

A three-part survey was chosen to evaluate the effectiveness of the ETG, aligning with the structure of the course "Energiewende gestalten" (designing energy transition), which comprises three components: a lecture series, a seminar, and the ETG itself. The first survey was conducted prior to the start of the lecture series to establish a baseline of participants' knowledge. Subsequently, the second survey took place after the completion of the lecture and seminar but before the ETG, aiming to gauge the knowledge acquired from these instructional components. Finally, the third survey was administered after the ETG to assess if the learning goals of the ETG (Chapter 5.1) are met.

Each survey entailed an online questionnaire comprising 35 to 39 questions. Four additional questions were included after the first survey, for which data is only available for the second and third surveys. These questions were organized into seven thematic blocks. Five of the blocks correspond to the learning goals (with agents and institutions being a different block than their emission reduction potentials). One block asks participants about their opinion on energy transitions and another asks them to answer more complex questions about specific challenges in energy transition.

Participants were recruited from students enrolled in the "Energiewende gestalten" course, which targets graduate students from electrical engineering and political sciences departments. The course and the questionnaire were conducted in German. Participation was voluntary and uncompensated, with participants fully informed about their data protection rights and the survey's purpose. All data collected were anonymized, with no personal information obtained regarding participants' field of study or semester. The ethics review committee of the Technical University of Darmstadt approved the conduct of the surveys.

Sixty-one valid questionnaires were collected, with validity determined by participants' completion of the questionnaire and agreement to the data protection and survey information. Of these valid questionnaires, 25 were completed in the first, 22 in the second, and 14 in the final survey. Note, that there is probably some kind of sampling bias since only students interested in the topics of the course tend to stay until the end of the course.

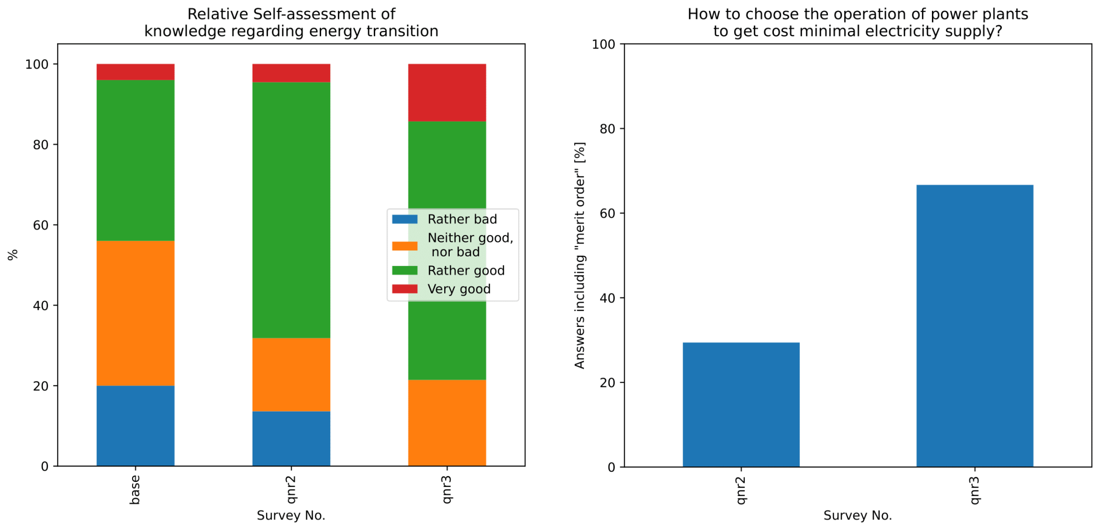
5.4.2. Results

For the presentation of the results, the survey numbers presented in each figure refer to the surveys as follows:

- Survey No. base = first survey (before start of the course)
- Survey No. qnr2 = second survey (after lecture and seminar, before ETG)
- Survey No. qnr3 = third survey (after ETG)

As part of the general questions in the questionnaire, participants were asked to rate their own knowledge of the energy transition. Figure 5.7a illustrates the distribution of self-assessed knowledge across the different surveys. Notably, in the initial survey, no participant rated their knowledge as "very bad," yet a majority (56%) assessed their knowledge as "rather bad" or "neither good nor bad." However, this percentage decreased to 32% in the second survey, indicating a 24% increase in participants rating their knowledge as "rather good" or better. In the third survey, the proportion of participants rating their knowledge as "rather good" remained consistent at 64% compared to the second survey. Interestingly, no participants rated their knowledge as "rather bad" in the third survey, and there was an increase in the share of participants rating their knowledge as "very good." It is worth noting that, due to the diminishing number of participants in each subsequent survey, the notable increase from 4% to 14% in "very good" ratings should be interpreted with caution.

One of the questions asked participants to explain how minimal cost for electricity supply can be achieved. During the lecture, the concept of merit order was introduced as a heuristic approach to achieving minimal cost for electricity generation. Figure 5.7b depicts the percentage of responses containing explicit mentions of the merit order or describing



(a) Self assessment of participants regarding their knowledge about energy transition.

(b) Merit order mentioned as a way to heuristically minimize the electricity price.

Figure 5.7.: Survey results to measure the impact of the ETG on learning objectives. (a) Relative change in participants' self-assessment of their knowledge of the energy transition over the course of the three surveys. The proportion of participants who rated their own knowledge as "rather good" or better increased after the lecture and seminar. After the ETG, the proportion of "very good" participants increased, and no participants rated their knowledge as "rather bad".
(b) Merit order or a description of it as responses to minimize electricity costs. While 29% of the participants in the second survey were able to name merit order as a way to minimize the cost of generating electricity, this percentage increased to 64% after the ETG.

its principles as a strategy for minimizing electricity generation costs. It is important to note that this question was introduced after the first survey; hence, no data is available for the base survey. Despite the merit order concept being explicitly covered in the lecture, only 29% of participants could name or describe it as a method for cost minimization. However, after participating in the ETG, 64% of participants referenced the merit order in their responses. Notably, no other approaches that would result in minimal cost were mentioned in the responses.

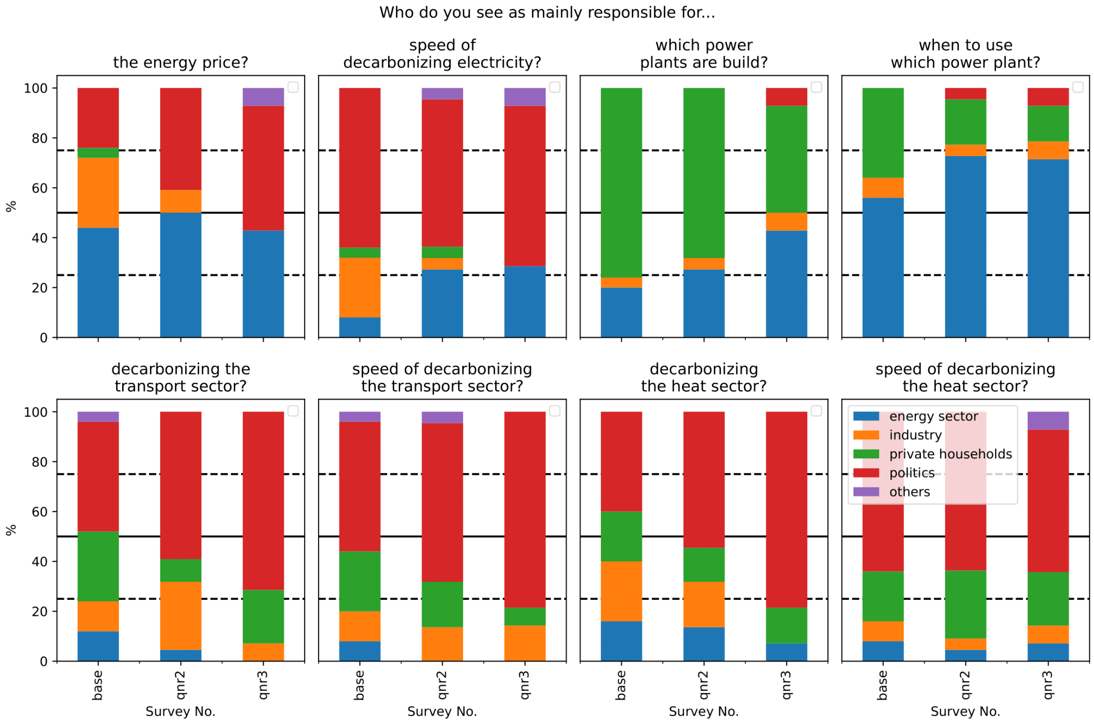


Figure 5.8.: Change in perceived responsibility. The participants see politics as more responsible in most fields after they attended the lecture and the seminar. This trend is continued after playing the ETG.

Figure 5.8 presents the results of a series of questions that asked participants to identify the sector primarily responsible for various aspects of the energy transition. Beginning with the plot in the top row on the far left, participants were asked to identify the primary entity responsible for determining electricity prices. In the initial survey, 44% of participants attributed the responsibility to the energy sector, while only 24% held politics accountable. Following the lectures and seminar in the second survey, the percentage attributing responsibility to politics increased to 40%, with the energy sector's share rising to 50%. After participating in the ETG in the final survey, 50% of participants viewed politics as primarily responsible for electricity prices, while the energy sector's share slightly decreased to 43%.
Moving on with the plot to the right, participants were asked about their responsibility for the speed of decarbonizing electricity generation. The majority identified politics as

responsible, and this opinion remained consistent across all surveys.
The third plot in the top row illustrates participants' views on who should be responsible for deciding which types of power plants to build. Initially, 76% of participants attributed this responsibility to private households. However, this share decreased to 68% in the second survey and further to 43% in the final survey, aligning with the energy sector's share.
Regarding the responsibility for choosing which power plants to use, approximately 58% of participants initially assigned this responsibility to the energy sector. This percentage increased to around 70% in both the second and third surveys.
Questions about the responsibility for decarbonizing the transport and heat sectors were posed in the bottom row. In both cases, the majority of participants viewed politics as primarily responsible. This opinion increased from 44% to 71% for the transport sector and from 40% to 79% for the heat sector across the surveys.

Overall, the trends observed in the third survey align with those seen when comparing the second survey (after the lectures and seminar) to the first survey (at the beginning of the course). Therefore, it can be inferred that the ETG amplified ideas taught in the lecture, leading to changes in the opinions of the majority. However, the responsibilities for the speed of decarbonizing electricity generation and the heat sector remained the same. Despite this, politics was consistently viewed as mainly responsible for various aspects of the energy transition, with the proportion of participants holding this view increasing throughout the surveys.

Please note that there is no definitive right or wrong answer regarding the primary responsibilities in the energy transition, as it necessitates collaboration among various stakeholders. For instance, the decarbonization of the heat sector primarily falls within the purview of private households, as they predominantly rely on decentralized fossil heating systems and must transition to alternative options such as heat pumps. However, for private households to adopt these new heating systems, they must first be made available by the industry. The industry is responsible for manufacturing and supplying these heating systems to private households. Moreover, given the high cost of acquiring new heating systems, private households may only be willing to invest if incentivized or mandated by governmental policies. Furthermore, municipal heat planning is the responsibility of local energy suppliers in Germany, as they are tasked with coordinating and implementing strategies for heat supply within their respective jurisdictions.

Participants were asked to identify three conflicts of interest between different actors involved in the energy transition. Figure 5.9 displays the conflicts of interest between the energy sector and politics as word clouds for each survey. A word cloud visually represents

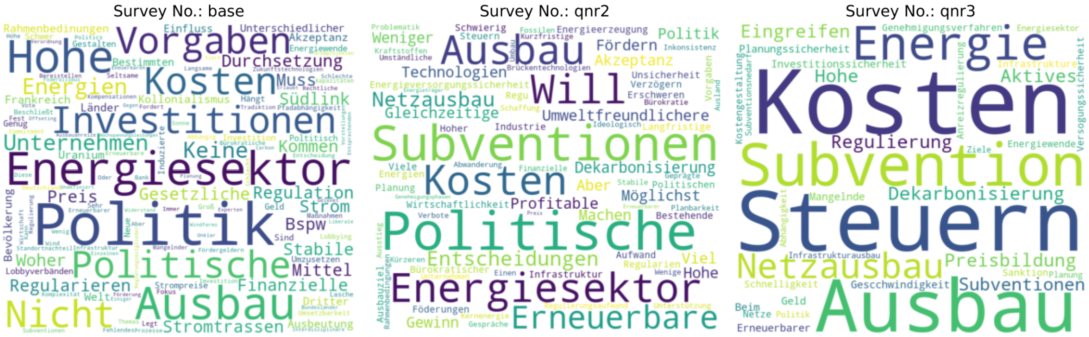


Figure 5.9.: Major conflicts between policy and the energy sector, according to survey respondents. Word size is related to the number of mentions. Participants were limited to 3 responses. The diversity of answers decreased. The answers were more focused on the main conflicts.

words with sizes corresponding to their frequency in the responses. Words with three letters or fewer were omitted. Since the questionnaire was in German, the word clouds are also in German.

Observing the word clouds, it is noticeable that the number of words decreases with each succeeding survey. While this could be attributed to the reduced number of participants, the content and prominence of more common words also change. In the first survey, the most common words were "Politik" (politics) and "Energiesektor" (energy sector). However, in the final survey, prominent words include "Kosten" (cost), "Steuern" (taxes), "Ausbau" (expansion), and "Subvention" (subsidy). This shift in common words towards slightly more specific terms suggests a change in participants' perceptions and priorities regarding conflicts of interest between the energy sector and politics throughout the surveys.

For measuring the improvement in the participants ability to deal with units and orders of magnitudes in the context of energy, they were asked to fill in a fill-in-the-blank text. While participants initially averaged two correct responses out of 8 fill-in-the-blank prompts, the lecture increased this average to 3.3 correct answers, with a slight further increase to 3.5 after gameplay.

5.5. Discussion

This chapter introduced our comprehensive approach to explaining energy systems within the context of energy transition through interactive means. We began by delineating specific learning goals to enhance players' understanding of crucial short and long-term processes within energy systems while also imparting knowledge about various stakeholders, institutions, and their interconnectedness, including conflicts of interest inherent in energy transition dynamics. With these objectives in mind, we addressed RQ 6 by devising a Serious Game tailored for four to six players, each assuming one of four distinct roles with individual objectives and responsibilities pertinent to energy transition. Notably, the energy sector's role was crafted to illuminate operational and expansion planning facets within ESMs. The game was implemented as a web-based application to ensure accessibility across various devices.

To assess the efficiency of our chosen design in achieving the learning goals (RQ 7), we conducted a three-part survey involving 25 students who participated in the game as part of a graduate-level university course. The survey, comprising an online questionnaire, was administered at different stages of the course. Results indicated a significant enhancement in participants' self-assessed knowledge throughout the course. Initially, over 50% of participants rated their knowledge of energy transition as "neither good, nor bad" or "rather bad," decreasing to below 30% after the lecture and seminar, and further to 21% after gameplay, with none rating their knowledge as "rather bad" anymore.

Regarding understanding energy generation processes, the proportion of participants identifying the merit order as a heuristic for minimizing electricity generation costs surged from 30% to 70%. Participants also provided insights into the attribution of responsibility to various actors for specific energy transition aspects and tasks. While perceptions shifted after the lecture and seminar for most queried fields and tasks, the game served to amplify these trends in all but one instance. In the sole exception, where participants were asked about responsibility for operational planning ("...when to use which power plant?"), the game did not substantially impact participant opinions.

When questioned about the primary conflicts between different stakeholders, the variety of responses diminished, yet the provided answers became more focused on crucial conflicts, with commonly cited keywords surfacing more frequently.

However, specific learning goals showed less significant improvement through gameplay. For instance, the objective of imparting crucial orders of magnitudes and relevant units within the energy sector saw only marginal advancement.

Overall, the game garnered positive participant feedback, which was evident in their enhanced knowledge self-assessment. However, given the relatively small sample size and the dropout rate among participants, which resulted in a decline in questionnaire responses in later phases of the survey, the results deserve further evaluation in future research efforts.

6. Conclusion & Outlook

The transition toward a renewable energy system presents challenges, necessitating careful decision-making. Optimization-based ESMs serve as valuable tools to support DMs in making informed decisions. However, DMs often lack expertise in ESMs and face significant personal consequences if they make incorrect decisions, which could prevent them from making risky but necessary decisions. Given the inherent complexity of energy systems and their corresponding ESMs, explanations are necessary to help DMs comprehend ESM results and facilitate well-informed decision-making.

To comprehend the creation of explanations, we initially examined the philosophical and psychological perspectives and recognized causality as a vital concept. For our first RQ, we identified seven categories of state-of-the-art ESM explanations, three of which involve some form of sensitivity analysis. One major limitation of all the methods found is their inability to provide sufficient explanations for non-expert DMs regarding the relationship between ESM results and large, related input parameter sets, such as time series. To address this issue, we looked to the field of machine learning, which faces similar challenges in explaining model behaviors based on high-dimensional input parameter sets. We examined popular XAI techniques that enhance the interpretability of machine learning model outcomes and identified concepts that could be beneficial in explaining ESMs (see RQ 2).

As demonstrated by our analysis of explanations from various perspectives, explainability is a broad research field with multiple aspects. This thesis focuses on three of these aspects, which we address through the development of three approaches: a causal approach, an interactive approach, and an approach based on the XAI domain.

Starting from the XAI approach, we have transferred the concept of an interpretable abstraction layer to improve the ESM explanation concerning high-dimensional input data (RQ 3). The developed approach is based on the XAI method LIME [19] and allows to generate explanations of different complexity that can serve different target audiences (RQ 4).

We attempted to learn directed causal graphical models from an ESM for the causal approach. Having a causal model of the ESM would enable us to create straightforward causal explanations for the ESM. However, we discovered that it is not always possible to use a directed causal graphical model to depict ESM behavior for some crucial cases due to cyclical dependencies (RQ 5). However, we have demonstrated how the Markov blanket can divide ESMs into smaller sub-problems that can be explained independently of others, in a circular fashion.

For the interactive approach, we presented the setup of a serious game that we designed to explain different concepts relevant to the energy transition (RQ 6). We surveyed to evaluate the game's ability to improve several learning objectives when used in the context of a university course for graduate students (RQ 7). We found that after playing the game, participants felt more confident about their knowledge of the energy transition and had stronger opinions about the responsibilities for various tasks in the energy transition. Participants also improved their ability to identify critical conflicts between actors relevant to the energy transition and to identify merit order as a heuristic for minimizing electricity generation costs.

While this thesis sheds light on three aspects to improve explanations of optimization-based ESMs, there are many other aspects and tasks that require further research. For the LIME-based approach, a challenge that remains open for future research is the optimal choice of a distance metric to weight input variations. Although directed, causal explanations are unsuitable for ESMs, breaking down complex ESMs into smaller, less complex parts that are easier to explain is promising. However, further investigation is necessary. Rather than creating a Serious Game, exploring the possibility of combining a large language model with ESM to develop an interactive chatbot may be worthwhile. This chatbot would enable DMs to interact with ESMs and their explanations, thereby challenging their understanding or exploring alternative scenarios without relying on domain experts. Additionally, adversarial attacks, a known issue in the machine learning domain, may become relevant for ESM in the near future. Adversarial attacks, i.e., changes to the model's input space that are not detectable by the human eye, can significantly affect the model's outcome. These changes may be hidden within time series data and can convince decision-makers to make choices that benefit the attacker rather than the decision-maker. The possibility of adversarial attacks on optimization-based ESMs highlights once again the importance of explanations in the ESM domain.

Curriculum Vitae

04/19 - 04/24	Ph.D. Candidate Energy Information Networks & Systems Lab, Technical University of Darmstadt, Germany
10/17 - 04/19	M.Sc. Business Administration/Industrial Engineering – specializing in Electrical Engineering Technical University of Darmstadt, Germany Master Thesis: Techno-economic feasibility study for the provision of balancing power via pooling of small scale units
10/12 - 10/17	B.Sc. Business Administration/Industrial Engineering – specializing in Electrical Engineering Technical University of Darmstadt, Germany

Publications

2023

J. Hülsmann, J. Barbosa, and F. Steinke, "Local Interpretable Explanations of Energy System Designs." *Energies 16.5* (2023): 2161.

2021

J. Hülsmann, L. J. Sieben, M. Mesgar, and F. Steinke, "A Natural Language Interface for an Energy System Model," *2021 IEEE PES Innovative Smart Grid Technologies Europe (ISGT Europe)*, Espoo, Finland, 2021, pp. 1-5

2020

J. Hülsmann, F. Steinke, "Explaining Complex Energy Systems: A Challenge", *NeurIPS 2020 Workshop on Tackling Climate Change with Machine Learning*, 2020

2019

J. Hülsmann, C. Ripp, and F. Steinke, "Balancing power potential of pools of small-scale units", *2019 16th International Conference on the European Energy Market (EEM)*, Ljubljana, Slovenia, 2019, pp. 1-6

Supervised Theses

Master's Theses

- Sebastian Deppisch, *Calculation of PV profiles based on smartphone image data*, 2024
- Pascal Uetz, *Investigation of the Vulnerability of Energy System Models to Adversarial Attacks*, 2022
- Jonas Seng, *Discovery of Causal Graphs in Energy System Models*, 2021
- Lennart Sieben, *Natural Language Interface for an Energy System Design Tool*, 2020
- Julian Reckter, *Dynamic Pricing for Electric Vehicle Charging*, 2020

Project Seminars

- Mohamed Ghanmi, *Mechanisms of Survivable Virtual Network Embedding*, 2019

Bibliography

[1] W. Krewitt and J. Nitsch, "The Potential for Electricity Generation from On-Shore Wind Energy under the Constraints of Nature Conservation: A Case Study for Two Regions in Germany," *Renewable energy*, vol. 28, no. 10, pp. 1645–1655, 2003.

[2] F. Neumann and T. Brown, "The near-optimal feasible space of a renewable power system model," *Electric Power Systems Research*, vol. 190, p. 106 690, 2021.

[3] S. D'Alessandro, T. Luzzati, and M. Morroni, "Energy Transition Towards Economic and Environmental Sustainability: Feasible Paths and Policy Implications," *Journal of cleaner production*, vol. 18, no. 4, pp. 291–298, 2010.

[4] T. Tröndle, D. Süsser, and J. Lilliestam, "Ohne Windenergie keine Energiewende. Die 1000 Meter-Abstandsregelung macht Windenergieausbau unmöglich und stellt damit den Kohleausstieg in Deutschland in Frage," 2019.

[5] N. H. Stern, *The Economics of Climate Change: the Stern Review*. cambridge University press, 2007.

[6] A. Rahman, O. Farrok, and M. M. Haque, "Environmental Impact of Renewable Energy Source Based Electrical Power Plants: Solar, Wind, Hydroelectric, Biomass, Geothermal, Tidal, Ocean, and Osmotic," *Renewable and Sustainable Energy Reviews*, vol. 161, p. 112 279, 2022.

[7] R. Loulou, U. Remme, A. Kanudia, A. Lehtila, and G. Goldstein, "Documentation for the times model part ii," *Energy Technology Systems Analysis Programme*, 2005.

[8] M. Howells, H. Rogner, N. Strachan, *et al.*, "Osemosys: The open source energy modeling system: An introduction to its ethos, structure and development," *Energy Policy*, vol. 39, no. 10, pp. 5850–5870, 2011.

[9] J. Weber, H. U. Heinrichs, B. Gillessen, *et al.*, "Counter-intuitive behaviour of energy system models under co2 caps and prices," *Energy*, vol. 170, pp. 22–30, 2019.

[10] J. Hülsmann and F. Steinke, “Explaining complex energy systems: A challenge,” en, Workshop: Tackling Climate Change with ML, 34th Conference on Neural Information Processing Systems (NeurIPS 2020), virtual Conference, 2020. [Online]. Available: `http://tubiblio.ulb.tu-darmstadt.de/124133/`.

[11] H. Ren and W. Gao, “A MILP Model for Integrated Plan and Evaluation of Distributed Energy Systems,” *Applied energy*, vol. 87, no. 3, pp. 1001–1014, 2010.

[12] R. Machlev, L. Heistrene, M. Perl, *et al.*, “Explainable Artificial Intelligence (XAI) techniques for energy and power systems: Review, challenges and opportunities,” *Energy and AI*, p. 100 169, 2022.

[13] W. Tian, “A review of sensitivity analysis methods in building energy analysis,” *Renewable and sustainable energy reviews*, vol. 20, pp. 411–419, 2013.

[14] A. E. e.V., *Energieflussbild der bundesrepublik deutschland 2021 energieeinheit petajoule (pj)*, `https://ag-energiebilanzen.de/daten-und-fakten/energieflussbilder/`, [Online; accessed January-2024], 2023.

[15] H.-M. Henning, “Pathways for Transforming the German Energy System by 2050-Methodology and Results of a Comprehensive System Simulation and Optimization,” 2017.

[16] C. Ripp and F. Steinke, “Sensitivity Analysis of Linear Programming Economic Dispatch Models,” in *International ETG-Congress 2019; ETG Symposium*, VDE, 2019, pp. 1–6.

[17] R. Dwivedi, D. Dave, H. Naik, *et al.*, “Explainable ai (xai): Core ideas, techniques, and solutions,” *ACM Computing Surveys*, vol. 55, no. 9, pp. 1–33, 2023.

[18] M. D. Zeiler and R. Fergus, “Visualizing and Understanding Convolutional Networks,” in *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13*, Springer, 2014, pp. 818–833.

[19] M. T. Ribeiro, S. Singh, and C. Guestrin, “"Why should I trust you?" Explaining the Predictions of Any Classifier,” in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 2016, pp. 1135–1144.

[20] S. M. Lundberg and S.-I. Lee, “A Unified Approach to Interpreting Model Predictions,” *Advances in neural information processing systems*, vol. 30, 2017.

[21] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, “Grad-CAM: Visual Explanations From Deep Networks via Gradient-Based Localization,” in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618–626.

[22] S. Wachter, B. Mittelstadt, and C. Russell, "Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR," *Harv. JL & Tech.*, vol. 31, p. 841, 2017.

[23] Y. Zhonggen, "A Meta-Analysis of Use of Serious Games in Education over a Decade," *International Journal of Computer Games Technology*, vol. 2019, 2019.

[24] T. Anastasiadis, G. Lampropoulos, and K. Siakas, "Digital Game-Based Learning and Serious Games in Education," *International Journal of Advances in Scientific Research and Engineering*, vol. 4, no. 12, pp. 139–144, 2018.

[25] J. Hülsmann, J. Barbosa, and F. Steinke, "Local Interpretable Explanations of Energy System Designs," *Energies*, vol. 16, no. 5, p. 2161, 2023.

[26] T. Lombrozo, "The Structure and Function of Explanations," *Trends in cognitive sciences*, vol. 10, no. 10, pp. 464–470, 2006.

[27] F. Heider, *The Psychology of Interpersonal Relations*. Psychology Press, 1982.

[28] B. F. Malle, *How the Mind Explains Behavior: Folk Explanations, Meaning, and Social Interaction*. MIT press, 2006.

[29] F. C. Keil, "Explanation and understanding," *Annual review of psychology*, vol. 57, p. 227, 2006.

[30] D. Lewis, "Causal explanation," 1986.

[31] C. Ginet, "In defense of a non-causal account of reasons explanations," *The Journal of Ethics*, vol. 12, pp. 229–237, 2008.

[32] D. Hume, *An enquiry concerning human understanding: A critical edition*. Oxford University Press on Demand, 2000, vol. 3.

[33] D. Lewis, "Causation," *The journal of philosophy*, vol. 70, no. 17, pp. 556–567, 1974.

[34] J. Woodward, *Making Things Happen: A Theory of Causal Explanation*. Oxford university press, 2005.

[35] P. Menzies and H. Price, "Causation as a Secondary Quality," *The British Journal for the Philosophy of Science*, vol. 44, no. 2, pp. 187–203, 1993.

[36] H. H. Kelley, "Causal Schemata and the Attribution Process.," in *Preparation of this paper grew out of a workshop on attribution theory held at University of California, Los Angeles, Aug 1969.*, Lawrence Erlbaum Associates, Inc, 1987.

[37] Carlos Zednik, *Disentangling xai concepts: Explanation, interpretation, and justification*, Presentation given at the "Whitebox Symposium on Explainability", 2023.

[38] T. Miller, “Explanation in artificial intelligence: Insights from the social sciences,” *Artificial intelligence*, vol. 267, pp. 1–38, 2019.

[39] C. S. Peirce, *Collected Papers of Charles Sanders Peirce*. Harvard University Press, 1974, vol. 5.

[40] G. H. Harman, “The Inference to the Best Explanation,” *The philosophical review*, vol. 74, no. 1, pp. 88–95, 1965.

[41] J. Pearl and D. Mackenzie, *The Book of Why: the New Science of Cause and Effect*. Basic books, 2018.

[42] R. J. Hankinson, *Cause and Explanation in Ancient Greek Thought*. Clarendon Press, 1998.

[43] C. Zednik, “Solving the Black Box Problem: A Normative Framework for Explainable Artificial Intelligence,” *Philosophy & technology*, vol. 34, no. 2, pp. 265–288, 2021.

[44] D. J. Hilton, “Conversational Processes and Causal Explanation.,” *Psychological Bulletin*, vol. 107, no. 1, p. 65, 1990.

[45] H. P. Grice, “Logic and Conversation,” in *Speech acts*, Brill, 1975, pp. 41–58.

[46] D. Hilton, “Social Attribution and Explanation.,” 2017.

[47] P. Lipton, “Contrastive Explanation,” *Royal Institute of Philosophy Supplements*, vol. 27, pp. 247–266, 1990.

[48] G. Hesslow, “The Problem of Causal Selection,” *Contemporary science and natural explanation: Commonsense conceptions of causality*, pp. 11–32, 1988.

[49] D. J. Hilton and B. R. Slugoski, “Knowledge-Based Causal Attribution: The Abnormal Conditions Focus Model.,” *Psychological review*, vol. 93, no. 1, p. 75, 1986.

[50] J. Woodward, “Sensitive and Insensitive Causation,” *The Philosophical Review*, vol. 115, no. 1, pp. 1–50, 2006.

[51] H. Chockler and J. Y. Halpern, “Responsibility and Blame: A Structural-Model Approach,” *Journal of Artificial Intelligence Research*, vol. 22, pp. 93–115, 2004.

[52] D. J. Hilton and L. M. John, “The Course of Events: Counterfactuals, Causal Sequences, and Explanation,” in *The psychology of counterfactual thinking*, Routledge, 2007, pp. 56–72.

[53] J. Leddo, R. P. Abelson, and P. H. Gross, “Conjunctive Explanations: When Two Reasons are Better than One.,” *Journal of Personality and Social Psychology*, vol. 47, no. 5, p. 933, 1984.

[54] A. Tversky and D. Kahneman, “Extensional Versus Intuitive Reasoning: The Conjunction Fallacy in Probability Judgment.,” *Psychological review*, vol. 90, no. 4, p. 293, 1983.

[55] P. Thagard, “Extending Explanatory Coherence,” *Behavioral and brain sciences*, vol. 12, no. 3, pp. 490–502, 1989.

[56] S. J. Read and A. Marcus-Newhall, “Explanatory Coherence in Social Explanations: A Parallel Distributed Processing Account.,” *Journal of Personality and Social Psychology*, vol. 65, no. 3, p. 429, 1993.

[57] D. J. Hilton, “Mental Models and Causal Explanation: Judgements of Probable Cause and Explanatory Relevance,” *Thinking & Reasoning*, vol. 2, no. 4, pp. 273–308, 1996.

[58] N. Vasilyeva, D. A. Wilkenfeld, and T. Lombrozo, “Goals Affect the Perceived Quality of Explanations.,” in *CogSci*, 2015.

[59] S. Collins, J. P. Deane, K. Poncelet, *et al.*, “Integrating Short Term Variations of the Power System into Integrated Energy System Models: A Methodological Review,” *Renewable and Sustainable Energy Reviews*, vol. 76, pp. 839–856, 2017.

[60] P. Lopion, P. Markewitz, M. Robinius, and D. Stolten, “A Review of Current Challenges and Trends in Energy Systems Modeling,” *Renewable and sustainable energy reviews*, vol. 96, pp. 156–166, 2018.

[61] J. Jurasz, F. Canales, A. Kies, M. Guezgouz, and A. Beluco, “A Review on the Complementarity of Renewable Energy Sources: Concept, Metrics, Application and Future Research Directions,” *Solar Energy*, vol. 195, pp. 703–724, 2020.

[62] J. Cronin, G. Anandarajah, and O. Dessens, “Climate Change Impacts on the Energy System: A Review of Trends and Gaps,” *Climatic change*, vol. 151, pp. 79–93, 2018.

[63] L. Hoettecke, S. Thiem, and S. Niessen, “Enhanced Time Series Aggregation for Long-Term Investment Planning Models of Energy Supply Infrastructure in Production Plants,” in *2021 International Conference on Smart Energy Systems and Technologies (SEST)*, IEEE, 2021, pp. 1–6.

[64] M. Hoffmann, L. Kotzur, D. Stolten, and M. Robinius, “A Review on Time Series Aggregation Methods for Energy System Models,” *Energies*, vol. 13, no. 3, p. 641, 2020.

[65] D. Heide, L. Von Bremen, M. Greiner, C. Hoffmann, M. Speckmann, and S. Bofinger, “Seasonal optimal mix of wind and solar power in a future, highly renewable europe,” *Renewable Energy*, vol. 35, no. 11, pp. 2483–2489, 2010.

[66] K. Schaber, F. Steinke, P. Mühlich, and T. Hamacher, "Parametric study of variable renewable energy integration in europe: Advantages and costs of transmission grid extensions," *Energy Policy*, vol. 42, pp. 498–508, 2012.

[67] F. Steinke, P. Wolfrum, and C. Hoffmann, "Grid vs. storage in a 100% renewable europe," *Renewable Energy*, vol. 50, pp. 826–832, 2013.

[68] A. J. Conejo, E. Castillo, R. Minguez, and F. Milano, "Locational Marginal Price Sensitivities," *IEEE Transactions on Power Systems*, vol. 20, no. 4, pp. 2026–2033, 2005.

[69] R. Li, Q. Wu, and S. S. Oren, "Distribution Locational Marginal Pricing for Optimal Electric Vehicle Charging Management," *IEEE Transactions on Power Systems*, vol. 29, no. 1, pp. 203–211, 2013.

[70] X. Fang, B.-M. Hodge, E. Du, C. Kang, and F. Li, "Introducing Uncertainty Components in Locational Marginal Prices for Pricing Wind Power and Load Uncertainties," *IEEE Transactions on Power Systems*, vol. 34, no. 3, pp. 2013–2024, 2019.

[71] C. Ripp and F. Steinke, "Modeling Time-dependent CO_2 Intensities in Multi-modal Energy Systems with Storage," *arXiv preprint arXiv:1806.04003*, 2018.

[72] C. Z. Li, Y. M. Shi, and X. H. Huang, "Sensitivity analysis of energy demands on performance of cchp system," *Energy Conversion and Management*, vol. 49, no. 12, pp. 3491–3497, 2008.

[73] S. Simoes, W. Nijs, P. Ruiz, A. Sgobbi, and C. Thiel, "Comparing policy routes for low-carbon power technology deployment in eu–an energy system analysis," *Energy Policy*, vol. 101, pp. 353–365, 2017.

[74] A. Palzer and H.-M. Henning, "A comprehensive model for the german electricity and heat sector in a future energy system with a dominant contribution from renewable energy technologies–part ii: Results," *Renewable and Sustainable Energy Reviews*, vol. 30, pp. 1019–1034, 2014.

[75] IEA, *Energy Technology Perspectives 2023*, 2023.

[76] J. F. DeCarolis, S. Babaee, B. Li, and S. Kanungo, "Modelling to generate alternatives with an energy system optimization model," *Environmental Modelling & Software*, vol. 79, pp. 300–310, 2016.

[77] J. Price and I. Keppo, "Modelling to generate alternatives: A technique to explore uncertainty in energy-environment-economy models," *Applied energy*, vol. 195, pp. 356–369, 2017.

[78] J.-P. Sasse and E. Trutnevyte, “Distributional trade-offs between regionally equitable and cost-efficient allocation of renewable electricity generation,” *Applied Energy*, vol. 254, p. 113 724, 2019.

[79] Y. He, M. Hildmann, F. Herzog, and G. Andersson, “Modeling the Merit Order Curve of the European Energy Exchange Power Market in Germany,” *IEEE Transactions on Power Systems*, vol. 28, no. 3, pp. 3155–3164, 2013.

[80] N. von der Assen, L. J. Müller, A. Steingrube, P. Voll, and A. Bardow, “Selecting CO2 Sources for CO2 Utilization by Environmental-Merit-Order Curves,” *Environmental science & technology*, vol. 50, no. 3, pp. 1093–1101, 2016.

[81] G. M. Masters, *Renewable and Efficient Electric Power Systems*. John Wiley & Sons, 2013.

[82] I.-C. Chen, Y. Kikuchi, Y. Fukushima, H. Sugiyama, and M. Hirao, “Developing Technology Introduction Strategies Based on Visualized Scenario Analysis: Application in Energy Systems Design,” *Environmental Progress & Sustainable Energy*, vol. 34, no. 3, pp. 832–840, 2015.

[83] L. Kotzur, L. Nolting, M. Hoffmann, *et al.*, “A Modeler’s Guide to Handle Complexity in Energy Systems Optimization,” *Advances in Applied Energy*, vol. 4, p. 100 063, 2021.

[84] S. Rech, “Smart Energy Systems: Guidelines for Modelling and Optimizing a Fleet of Units of Different Configurations,” *Energies*, vol. 12, no. 7, p. 1320, 2019.

[85] L. Tock and F. Maréchal, “Decision Support for Ranking Pareto Optimal Process Designs under Uncertain Market Conditions,” *Computers & Chemical Engineering*, vol. 83, pp. 165–175, 2015.

[86] P. I. Helgesen and A. Tomasgard, “An Equilibrium Market Power Model for Power Markets and Tradable Green Certificates, including Kirchhoff’s Laws and Nash-Cournot Competition,” *Energy Economics*, vol. 70, pp. 270–288, 2018.

[87] H. J. Greenberg, “The ANALYZE rulebase for supporting LP analysis,” *Annals of Operations Research*, vol. 65, no. 1, pp. 91–126, 1996.

[88] U. Gnewuch, S. Morana, C. Heckmann, and A. Maedche, “Designing Conversational Agents for Energy Feedback,” in *Designing for a Digital and Globalized World: 13th International Conference, DESRIST 2018, Chennai, India, June 3–6, 2018, Proceedings 13*, Springer, 2018, pp. 18–33.

[89] J. Hulsmann, L. J. Sieben, M. Mcsgar, and F. Steinke, “A Natural Language Interface for an Energy System Model,” in *2021 IEEE PES Innovative Smart Grid Technologies Europe (ISGT Europe)*, IEEE, 2021, pp. 1–5.

[90] U. E. Institute, *Watts the Deal?* https://wattsthedeal.org, [Online; accessed June-2023], 2018.

[91] K. Schaber, F. Steinke, and T. Hamacher, “Transmission Grid Extensions for the Integration of Variable Renewable Energies in Europe: Who benefits where?” *Energy Policy*, vol. 43, pp. 123–135, 2012.

[92] S. P. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge university press, 2004.

[93] F. Schweppe, R. Tabors, M. Caraminis, and R. Bohn, “Spot Pricing of Electricity,” 1988.

[94] PJM, *Regional Transmission Organization*, Website: https://www.pjm.com/ (accessed June 2023).

[95] New York ISO, *Regional Transmission Organization*, Website: https://www.nyiso.com/ (accessed June 2023).

[96] S. P. Holland, M. J. Kotchen, E. T. Mansur, and A. J. Yates, “Why Marginal CO2 Emissions are not Decreasing for US Electricity: Estimates and Implications for Climate Policy,” *Proceedings of the National Academy of Sciences*, vol. 119, no. 8, e2116632119, 2022.

[97] A. Hawkes, “Estimating marginal co2 emissions rates for national electricity systems,” *Energy Policy*, vol. 38, no. 10, pp. 5977–5987, 2010, ISSN: 0301-4215. DOI: https://doi.org/10.1016/j.enpol.2010.05.053.

[98] A. Hawkes, “Long-run marginal co2 emissions factors in national electricity systems,” *Applied Energy*, vol. 125, pp. 197–205, 2014, ISSN: 0306-2619. DOI: https://doi.org/10.1016/j.apenergy.2014.03.060.

[99] K. R. Voorspools and W. D. D’haeseleer, “An evaluation method for calculating the emission responsibility of specific electric applications,” *Energy Policy*, vol. 28, no. 13, pp. 967–980, 2000, ISSN: 0301-4215. DOI: https://doi.org/10.1016/S0301-4215(00)00080-X.

[100] J. C. Spall, “Stochastic Optimization,” *Handbook of computational statistics: Concepts and methods*, pp. 173–201, 2012.

[101] A. Ben-Tal, L. El Ghaoui, and A. Nemirovski, *Robust Optimization*. Princeton university press, 2009, vol. 28.

[102] A. Perera, V. M. Nik, D. Chen, J.-L. Scartezzini, and T. Hong, “Quantifying the Impacts of Climate Change and Extreme Climate Events on Energy Systems,” *Nature Energy*, vol. 5, no. 2, pp. 150–159, 2020.

[103] B. Zeng and L. Zhao, “Solving Two-Stage Robust Optimization Problems Using a Column-and-Constraint Generation Method,” *Operations Research Letters*, vol. 41, no. 5, pp. 457–461, 2013.

[104] IEA, *Net Zero by 2050 - A Roadmap for the Global Energy Sector*, 2021.

[105] IEA, *World Energy Outlook 2022*, 2022.

[106] S. Simoes, W. Nijs, P. Ruiz, *et al.*, “The JRC-EU-TIMES Model,” *Assessing the long-term role of the SET Plan Energy technologies*, 2013.

[107] H.-M. Henning and A. Palzer, “A comprehensive model for the german electricity and heat sector in a future energy system with a dominant contribution from renewable energy technologies—part i: Methodology,” *Renewable and Sustainable Energy Reviews*, vol. 30, pp. 1003–1018, 2014.

[108] K. Hunter, S. Sreepathi, and J. F. DeCarolis, “Modeling for Insight Using Tools for Energy Model Optimization and Analysis (Temoa),” *Energy Economics*, vol. 40, pp. 339–349, 2013.

[109] T. Brown, J. Hörsch, and D. Schlachtberger, “PyPSA: Python for Power System Analysis,” *arXiv preprint arXiv:1707.09913*, 2017.

[110] E. Gavanidou, A. Bakirtzis, and P. Dokopoulos, “A Probabilistic Method for the Evaluation of the Performance and the Reliability of Wind-Diesel Energy Systems,” *IEEE transactions on energy conversion*, vol. 8, no. 2, pp. 197–206, 1993.

[111] A. Vogt-Schilb and S. Hallegatte, “Marginal Abatement Cost Curves and the Optimal Timing of Mitigation Measures,” *Energy Policy*, vol. 66, pp. 645–653, 2014.

[112] K. Siau and W. Wang, “Building trust in artificial intelligence, machine learning, and robotics,” *Cutter Business Technology Journal*, vol. 31, no. 2, pp. 47–53, 2018.

[113] R. Loulou and M. Labriet, “Etsap-tiam: The times integrated assessment model part i: Model structure,” *Computational Management Science*, vol. 5, no. 1-2, pp. 7–40, 2008.

[114] M. Groissböck, “Are Open Source Energy System Optimization Tools Mature Enough for Serious Use?” *Renewable and Sustainable Energy Reviews*, vol. 102, pp. 234–248, 2019.

[115] O. D. Doleski, T. Kaiser, M. Metzger, S. Niessen, and S. Thiem, “Digitale Dekarbonisierung,” 2021.

[116] R. Marcinkevičs and J. E. Vogt, “Interpretability and Explainability: A Machine Learning Zoo Mini-Tour,” *arXiv preprint arXiv:2012.01805*, 2020.

[117] Z. C. Lipton, "The mythos of model interpretability," *Queue*, vol. 16, no. 3, pp. 31–57, 2018.

[118] C. Rudin, "Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead," *Nature machine intelligence*, vol. 1, no. 5, pp. 206–215, 2019.

[119] M. Robnik-Šikonja and M. Bohanec, "Perturbation-based explanations of prediction models," *Human and Machine Learning: Visible, Explainable, Trustworthy and Transparent*, pp. 159–175, 2018.

[120] T. Speith, "A Review of Taxonomies of Explainable Artificial Intelligence (XAI) Methods," in *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 2022, pp. 2239–2250.

[121] G. Montavon, J. Kauffmann, W. Samek, and K.-R. Müller, "Explaining the Predictions of Unsupervised Learning Models," in *International Workshop on Extending Explainable AI Beyond Deep Models and Classifiers*, Springer, 2020, pp. 117–138.

[122] I. Lage, D. Lifschitz, F. Doshi-Velez, and O. Amir, "Exploring Computational User Models for Agent Policy Summarization," in *IJCAI: proceedings of the conference*, NIH Public Access, vol. 28, 2019, p. 1401.

[123] A. Das and P. Rad, "Opportunities and Challenges in Explainable Artificial Intelligence (XAI): A Survey," *arXiv preprint arXiv:2006.11371*, 2020.

[124] J. Pearl, *Causality*. Cambridge university press, 2009.

[125] A. Shrikumar, P. Greenside, and A. Kundaje, "Learning Important Features Through Propagating Activation Differences," in *International conference on machine learning*, PMLR, 2017, pp. 3145–3153.

[126] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek, "On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation," *PloS one*, vol. 10, no. 7, e0130140, 2015.

[127] L. S. Shapley *et al.*, "A Value for n-person Games," 1953.

[128] Scott M. Lundberg et. al, *SHAP github ReadMe*, `https://github.com/shap/shap/blob/master/README.md`, [Online; accessed January-2024], 2019.

[129] J. Peters, D. Janzing, and B. Schölkopf, *Elements of Causal Inference: Foundations and Learning Algorithms*. The MIT Press, 2017.

[130] C. Glymour, K. Zhang, and P. Spirtes, "Review of Causal Discovery Methods based on Graphical Models," *Frontiers in genetics*, vol. 10, p. 524, 2019.

[131] P. Uetz, “Investigation of the Vulnerability of Energy System Models to Adversarial Attacks,” Master’s thesis, Technical University of Darmstadt, 2022.

[132] L. Gacitua, P. Gallegos, R. Henriquez-Auba, *et al.*, “A Comprehensive Review on Expansion Planning: Models and Tools for Energy Policy Analysis,” *Renewable and Sustainable Energy Reviews*, vol. 98, pp. 346–360, 2018.

[133] C. Fan, F. Xiao, C. Yan, C. Liu, Z. Li, and J. Wang, “A novel methodology to explain and evaluate data-driven building energy performance models based on interpretable machine learning,” *Applied Energy*, vol. 235, pp. 1551–1560, 2019.

[134] M. Chen, Q. Liu, S. Chen, Y. Liu, C.-H. Zhang, and R. Liu, “Xgboost-based algorithm interpretation and application on post-fault transient stability status prediction of power system,” *IEEE Access*, vol. 7, pp. 13 149–13 158, 2019.

[135] M. Chaibi, E. Benghoulam, L. Tarik, M. Berrada, and A. E. Hmaidi, “An Interpretable Machine Learning Model for Daily Global Solar Radiation Prediction,” *Energies*, vol. 14, no. 21, p. 7367, 2021.

[136] A. Herbst, F. Toro, F. Reitze, and E. Jochem, “Introduction to energy systems modelling,” *Swiss journal of economics and statistics*, vol. 148, no. 2, pp. 111–135, 2012.

[137] S. Pfenninger and I. Staffell, *Renewables.ninja*, Available online: `https://www.renewables.ninja/` (accessed on 04 March 2022).

[138] BDEW, Bundesverband der Energie- und Wasserwirtschaft, *Standartlastprofile strom*, Available online: `https://www.bdew.de/energie/standardlastprofile-strom/`(only in German language) (accessed on 15 March 2022).

[139] BDEW, Bundesverband der Energie- und Wasserwirtschaft, *Standartlastprofile gas*, Available online: `https://www.bdew.de/energie/standardlastprofile-gas/`(only in German language) (accessed on 15 March 2022).

[140] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, “Least angle regression,” *The Annals of statistics*, vol. 32, no. 2, pp. 407–499, 2004.

[141] J. Barbosa, C. Ripp, and F. Steinke, “Accessible modeling of the german energy transition: An open, compact, and validated model,” *Energies*, vol. 14, no. 23, 2021, ISSN: 1996-1073. [Online]. Available: `https://www.mdpi.com/1996-1073/14/23/8084`.

[142] C. Molnar, *Interpretable machine learning - A Guide for Making Black Box Models Explainable*, Second. Independently published (28. Februar 2022), 2022.

[143] S.-C. Fang and S. Puthenpura, *Linear optimization and extensions: theory and algorithms*. Prentice-Hall, Inc., 1993.

[144] P. R. Rosenbaum and D. B. Rubin, “The central role of the propensity score in observational studies for causal effects,” *Biometrika*, vol. 70, no. 1, pp. 41–55, 1983.

[145] J. Pearl, “The Art and Science of Cause and Effect,” in *Shaping Entrepreneurship Research*, Routledge, 2020, pp. 446–474.

[146] J. Pearl, *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan kaufmann, 1988.

[147] A. Hyttinen, F. Eberhardt, and P. O. Hoyer, “Learning linear cyclic causal models with latent variables,” *The Journal of Machine Learning Research*, vol. 13, no. 1, pp. 3387–3439, 2012.

[148] *What’s Netz?* `https://www.eon.com/de/c/whatsnetz.html`, [Online; accessed January-2024].

[149] M.-T. Cheng, Y.-W. Lin, H.-C. She, and P.-C. Kuo, “Is immersion of any value? whether, and to what extent, game immersion experience during serious gaming affects science learning,” *British Journal of Educational Technology*, vol. 48, no. 2, pp. 246–263, 2017.

[150] M. Bakhuys Roozeboom, G. Visschedijk, and E. Oprins, “The effectiveness of three serious games measuring generic learning features,” *British journal of educational technology*, vol. 48, no. 1, pp. 83–100, 2017.

[151] O. J. Foundation, *Node.js*, Website: `https://nodejs.org` (accessed February 2024).

ChatGPT, Grammarly, and DeepL Write were employed throughout this thesis to enhance writing quality, ensuring grammar, spelling, and style accuracy.

A. An exemplary Energy System Design Model of a Building

In this supplementary section, we present the technical details of the building energy system model used as a running example in Chapter 3. Below are the equations for the simple PV house energy system. The equations are a modified version of the equations from [10]. Blue variables or equations are only present if the model considers the heat sector.

The objective of the energy system is to minimize system costs. The system costs consists of operational costs for buying energy E_e and investment costs for the battery capacity C_b and heat storage capacity C_{HS}, respectively, at price π_e, π_b, and π_{HS}.

$$\min_{E,Cap}(\sum_{t=0}^{T}\pi_e \times E_e(t)) + \pi_b \times C_b + \pi_{HS} \times C_{HS}. \tag{A.1}$$

Note that the PV and heat pump capacity is assumed to be fixed and thus is not part of the objective function.

A set of constraints restricts the objective. The first constraint is the electric power balance of the system,

$$E_{PV}(t) + E_b^{out}(t) + E_e(t) = D_e(t) + E_b^{in}(t) + E_{HP}^{in}(t), \forall t. \tag{A.2}$$

On the left side of the equation is the energy provided by the PV power plant $E_{PV}(t)$, the energy taken from the battery storage $E_b^{out}(t)$, and the energy bought from the grid $E_e(t)$. The on the right side is the energy demand $D_e(t)$, the energy put into the battery $E_b^{in}(t)$ and the electricity consumption of the heat pump $E_{HP}^{in}(t)$. Both sides of the equation must be balanced for every time step t to ensure that the electricity supply equals the demand. Like the electric power balance, a heat power balance constraint is defined for every time step t if the energy system considers the heat sector.

$$E_{HP}^{out}(t) + E_{HS}^{out}(t) = D_h(t) + E_{HS}^{in}(t), \forall t \tag{A.3}$$

with $E^{out}_{HP}(t)$ as the heat output of the heat pump, $E^{out}_{HS}(t)$ as the heat taken from the heat storage, $D_h(t)$ as the heat demand and $E^{in}_{HS}(t)$ as the heat put into the heat storage.

The battery storage level is defined by

$$E^S_b(t) = E^S_b(t-1) + loss_b \times E^{in}_b(t) - E^{out}_b(t), \forall t \tag{A.4}$$

where the current energy stored in the battery $E^S_b(t)$ is given by the energy in the previous time step $E^S_b(t-1)$ reduced by the energy taken from the battery $E^{out}_b(t)$ and increased by the energy added in the current time step $E^{in}_b(t)$. A power inverter loss is considered for the battery storage by applying a loss $loss_b$ to the energy stored in the battery. We chose $loss_b$ to be 0.95 to represent an inverter loss of 5%. The upper and lower limits of the battery level are defined as

$$0 \leq E^S_b(t) \leq C^S_b, \forall t. \tag{A.5}$$

For heat storage, the same logic is applied to the battery.

$$E^S_{HS}(t) = loss_{HS} \times E^S_{HS}(t-1) + E^{in}_{HS}(t) - E^{out}_{HS}(t), \forall t \tag{A.6}$$

The current Energy stored in the heat storage $E^S_{HS}(t)$ is given by the energy stored at the previous time step $E^S_{HS}(t-1)$ decreased by the heat taken from the storage $E^{out}_{HS}(t)$ and increased by the energy added to the storage $E^{in}_{HS}(t)$. It is assumed that part of the heat stored is lost every time step, which is described by $loss_{HS}$. We chose $loss_{HS}$ to represent a heat loss of 1% per hour of the energy stored. The heat storage is restricted by its capacity C^S_{HS} and an additional bound UB_{HS} on the maximum storage size that can be installed.

$$0 \leq E^S_{HS}(t) \leq C^S_{HS} \leq UB_{HS}, \forall t \tag{A.7}$$

We chose $UB_{heat\_storage}$ to be 46.6 kWh or $1m^3$ of water with a maximum heat difference of 40°K. In order to prevent storage depletion of the battery and the heat storage at the end of the optimized time interval, we define $t = T$ as the time step prior to $t = 0$ in Equations A.4 and A.6.

In this energy system, the PV capacity is an input parameter, not part of the optimization. Hence, the PV energy production $E_{PV}(t)$ can be described with the time series $av_{PV}(t)$. The time series $av_{PV}(t)$ limits the possible energy output at each time step, i.e., a percentage of output at each time step times the PV capacity.

$$0 \leq E_{PV}(t) \leq av_{PV}(t), \forall t \tag{A.8}$$

The limits of the heat pump are defined as:

$$0 \leq E_{HP}^{out}(t) \leq COP \times E_{HP}^{in}(t) \leq 2 \times D_h^{max}, \forall t \tag{A.9}$$

where, $E_{HP}^{out}(t)$ is the heat output and $E_{HP}^{in}(t)$ the electricity consumption of the heat pump. The heat output of the heat pump is bound by the electricity consumption times the coefficient of performance COP of the pump. We chose a COP of 3 in our experiments. Furthermore, we assume that the heat pump's power is limited by twice the highest heat demand value D_h^{max}. This limitation was made to prevent production spikes since there are no costs for the heat pump's capacity.

Finally, selling energy to the grid is not considered in this model. Hence, the energy bought is limited to be positive by the following equation.

$$0 \leq E_{buy}(t), \forall t \tag{A.10}$$

B. Questionnaire to Evaluate the Energy Transition Game

In this appendix the questionnaire which was posed to the participants of the course "Energiewende gestalten" is described. The original questionnaire is in the German language since the course is also offered in the German language. We provide the English translation of the questions here to be inline with the language of this thesis.

The questionnaire starts with some general clarifications about the survey itself and a declaration of data security. The general clarification contains the type of study at hand (online questionnaire), the procedure (3 separate questionnaire that should be answered at three different points in time), the duration (about 20 minutes per questionnaire), the potential use (measuring of the learning success aiming to aide in improving the quality of the course) and the potential risks for the participants of the study (non above the risks of daily life). The declaration of data security contains the necessary formulations to be in line with the European General Data Protection Regulations (GDPR). The necessary formulations contain what type of personal data is collected (field of study and semester), how and where the data is stored (anonymised and not longer then 5 years on university servers), as well as a clarification of the rights each participant has according to the GDPR. Further, the participants are sophisticated that their participation is voluntary and that they can abort the survey at any time without any consequences. Participants have to declare that the read the general information and the data security declaration before they can proceed with the study. Finally, participants could opt-in into allowing their free text answers to be directly cited if they did not contain any sensitive data.

In the next part of the questionnaire, participants were asked to fill in some meta data, i.e., their field of study and their current semester of studying. Further, they were asked to generate a manual hash code that can be used to measure individual improvements by identifying the later questionnaires without breaking their anonymity. This generation of a manual hash was optional.

The next part of the questionnaire starts with some general questions about the participants opinion towards the energy transition:

$A01$: **Is the energy transition reasonable?**

□ Yes □ No

$A02$: **Is the energy transition technical feasible?**

□ Yes □ No

$A03$: **Do you think that the energy transition will be succeed?**

□ Yes □ No

$A04$: **Where do you think is the biggest challenge for energy transition?**

□ technical aspects □ economical aspects □ political aspects

$A05$: **How would you rate your own understanding of the energy transition?**

□ very bad □ rather bad □ neither bad, nor good □ rather good □ very good

The next block of questions contained some free text questions about different topics related to energy transition. This block oaf questions were added after the first online interview, hence, no data is available for the first interview.

$B01$: **Why is the energy transition in the heat sector relevant?**

$B02$: **Which measures can help to achieve energy transition in the heat sector?**

$B03$: **Which problems can occur if the decarbinization of the industry is done to hasty?**

$B04$: **How should the operation of power plants be chosen to achieve cost optimal power production?**

The free text questions about different parts of energy transition are followed-up by some fill-in-the-blank-text, a question where items should be put in an increasing order and some more complicated questions about measures of energy transition that require calculation or reasioning to be solved. All tasks/questions of this question block are dedicated to physical units and magnitudes. Participants were asked to fill in physical units into the following fill-in-the-blank-text:

*C*01: An average photovoltaic system for a single-family home has a peak output of about 5 to 10 ______.

*C*02: A direct current high-voltage line has a transmission capacity of up to 5 ______.

*C*03: The average annual energy consumption of a German household (2 people) is about 3000 ______.

*C*04: A AAA battery stores approximately one ______ of energy.

*C*05: As this power is not consumed throughout the entire duration of the ironing process, an average energy requirement of about 1.5 ______ is obtained.

*C*06: An iron has a power rating of 2 to 3 ______.

*C*07/*C*08: Coal-fired power plants typically have a capacity of 100 ______ up to one ______.

In *C*09 participants should rank different options in ascending order.

*C*09: **Sort the following means of electric power generation by their CO_2 intensity from lowest intensity to highest:**

- German average (as of 2022)
- Power from coal
- Power from lignite
- Power from Gas (60% efficency)
- Power from waste combustion

Questions *C*10 – *C*12 each consist of two parts. First the question and a field where the requested result should be written down and a free text field where participants are asked to argue or show some calculation how they came to the result they gave. For clarity reasons, we omit the second part of each question here.

*C*10: **How many households can an average wind power plant supply with electricity?**

*C*11: **What price do households have to pay for electricity so a wind power plant can be economic after 10 years?**

*C*12: **If every person in Germany would require an additional kWh of energy each year (distributed evenly throughout the year), how much additional power plant capacity is needed to provide this energy?**

A small block of regarding the merit order as one of the concepts taught in the course contains the following two questions.

*D*01: **If possible, please explain the merit order.**

*D*02: **What are arguments why the merit order is reasonable?**

The participants are asked about all agents that they know that are related to energy transition.

*E*01: **Name in bullet points all agents/institutions that you know of, that are somehow involved in the energy transition, respectively can influence it.**

Next, participants should name their opinion which sector they deem mainly responsible for different aspects of the energy transition. Possible answers form which they can choose are: energy sector, industry, private households, politics, or others.

Who do you think is mainly responsible for...

*F*01: **the electricity price?**

*F*02: **the speed of the decarbonization of the electricity generation?**

*F*03: **the decision which power plants will be build?(expansion planing)**

*F*04: **the decision which power plants are actually used?(operational planning)**

*F*05: **the implementation of the energy transition in the transport sector?**

*F*06: **the speed of the decarbonization of the transport sector?**

*F*07: **the implementation of the energy transition in the heat sector?**

*F*08: **the speed of the decarbonization of the heat sector?**

Regrading the four explicitly named sectors the questionnaire further asks how CO_2 emissions can be reduced in each sector

*F*09: **Name in bullet points all measures/potentials that you can think of, how the energy sector can reduce its CO_2 emissions.**

*F*10: **Name in bullet points all measures/potentials that you can think of, how the industry can reduce their CO_2 emissions.**

*F*11: **Name in bullet points all measures/potentials that you can think of, how private households can reduce their CO_2 emissions.**

*F*12: **Name in bullet points all measures/potentials that you can think of, how the politic can directly or indirectly influence CO_2 emissions.**

The final block of questions of the survey is dedicated to conflict of interests regrading energy transition. The survey gives the following conflict of interest between private households and politics, potentially caused by the energy transition, as an example: *Private households want to pay low taxes so they have more money they can spend on other goods. On the other hand, the government wants to have high income trough taxes to finance political measures.* Participants are asked to name up to three conflicts of interest between the following parties.

*G*01: **Name the in your opinion most important conflicts of interest regarding the energy transition between the energy sector and the industry.**

*G*02: **Name the in your opinion most important conflicts of interest regarding the energy transition between the energy sector and politics.**

*G*03: **Name the in your opinion most important conflicts of interest regarding the energy transition between private households and the industry.**

C. Full Questionnaire Results

In this appendix, we show the full results of the questionnaire posed to the participants of the course "Energiewende gestalten". The course consists of a lecture, and a seminar as well as the simulation game presented in this thesis. A total of 85 questionnaires were started by the participants, distributed over three different surveys at three different points in time. Of the 85 started questionnaires 61 have been completed and are valid, i.e., participants have read and agreed on the general clarifications and the declaration of data security. The first survey has 25 valid questionnaires and was conduced before the start of the first lecture. A total of 22 valid questionnaires were collected for the second survey after the lecture and the seminar, but before the first play of the simulation game. The final survey was conduced after the participants had played the simulation game and contained 14 valid questionnaires. We refer to the three surveys by in the remained of this appendix and in all figures by their survey number: base (first survey), qnr2 (second survey), and qnr3 (final survey). We refer the reader to the question numbers from Appendix B and omit repeating the questions to simplify the text.

All participants have been recruited from the students of the course "Energiewende gestalten". Their participation was voluntarily and was not rewarded by any means. Of the valid questionnaires, 34 were answered by participants that studies something in the field of electrical engineering. The most common field of study from that group was "Energy Science and Engineering" with 25 questionnaires. A total of 27 valid questionnaires were finished by participants studying in the field of political sciences. The most common field of study from the group of politic science students was "Governance and Public Policy" (17 questionnaires). The average semester of the participants was 2.67. No other personal information, e.g., age or sex, have been collected.

The top part of Figure C.1 shows the answers to questions $A01$ – $A03$ where the participants were asked if they think the energy transition is reasonable, technical feasible and if it will become reality. The results are split by the three surveys conducted. The bottom part of

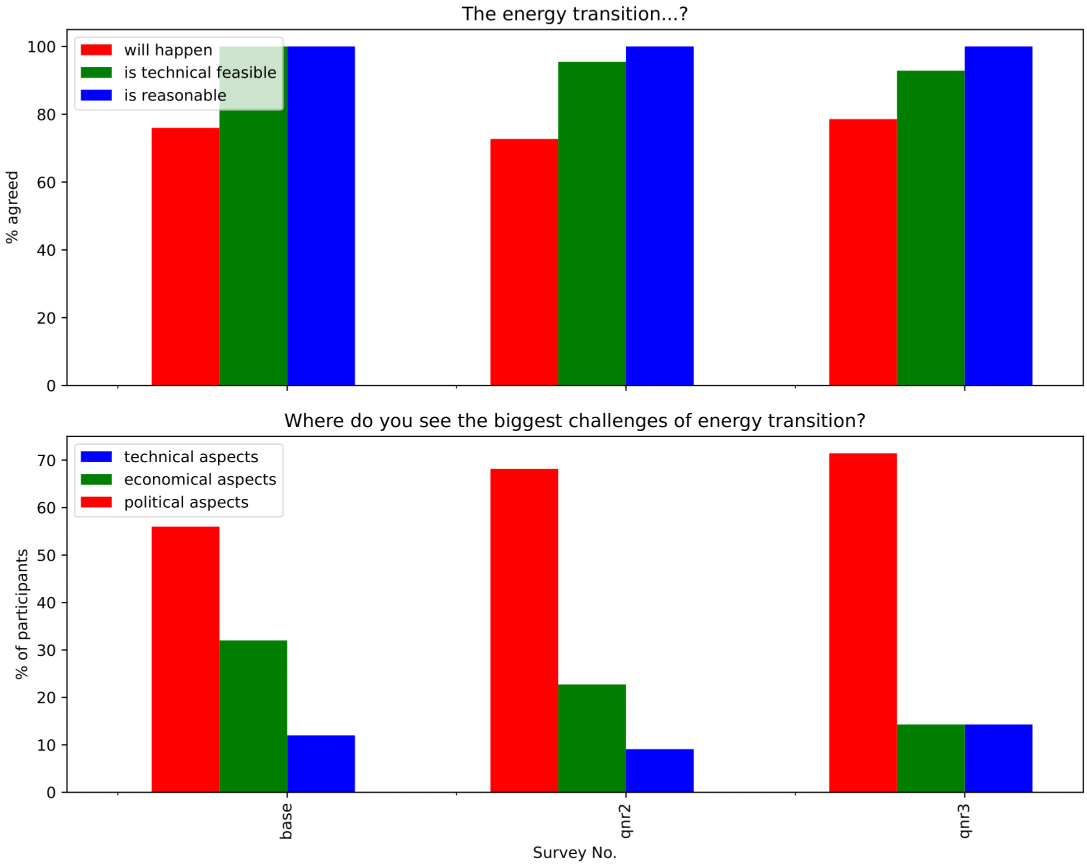


Figure C.1.: Answers to questions $A01$ – $A03$ (top) and $A04$ (bottom). Personal opinion on energy transition and biggest challenge.

Figure C.1 shows the results of question $A04$. Participants had to rate, where they think the biggest challenge of energy transition is.

Figure C.2 shows how the participants assessed their own knowledge regarding the energy transition. While in the first two surveys, some participants assessed their knowledge of energy transition as "rather bad", none of the participants in the last survey did. Further, the share of participants who think their knowledge of energy transition is "very good" increased.

Figure C.3 shows the categorized answers of the open text questions regarding energy

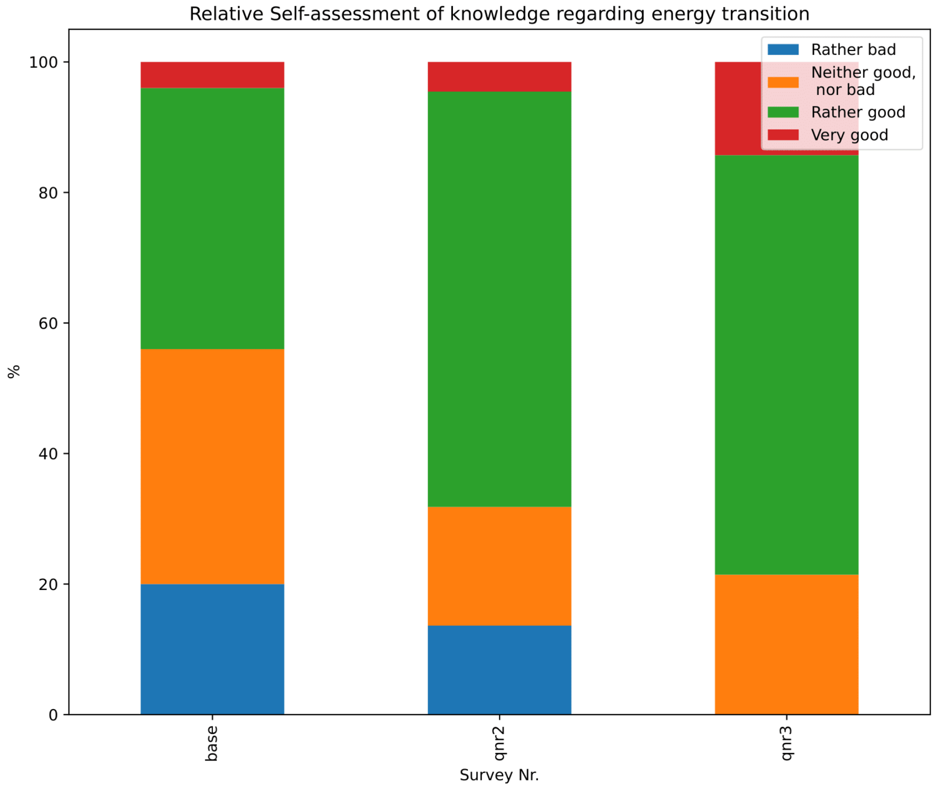
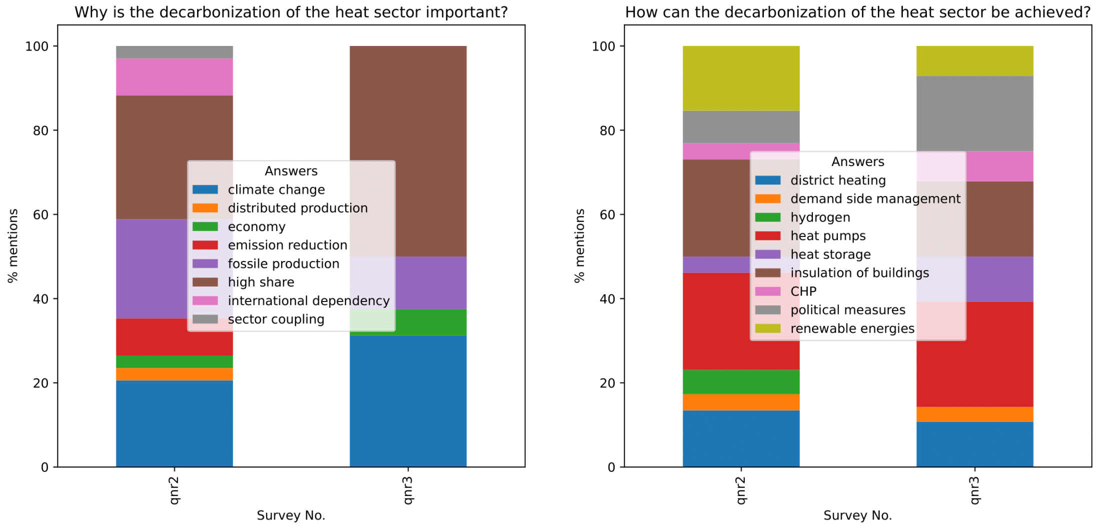


Figure C.2.: Answers to questions $A05$. Self Assessment of knowledge about energy transition.

transition in the heat sector. The answers were manually categorized. Naming multiple categories was possible. In Figure C.3a, the given answers to question $B01$ are shown. The following categories were identified:

The energy transition of the heat sector...

- is important to achieve **climate change** goals.
- allows for **distributed production**.
- is important for a stable and prospering **economy**.



(a) *B*01: Relevance of the heat sector.

(b) *B*02: Measures for decarbonizing heat.

Figure C.3.: Results of the open questions regarding the energy transition of the heat sector.

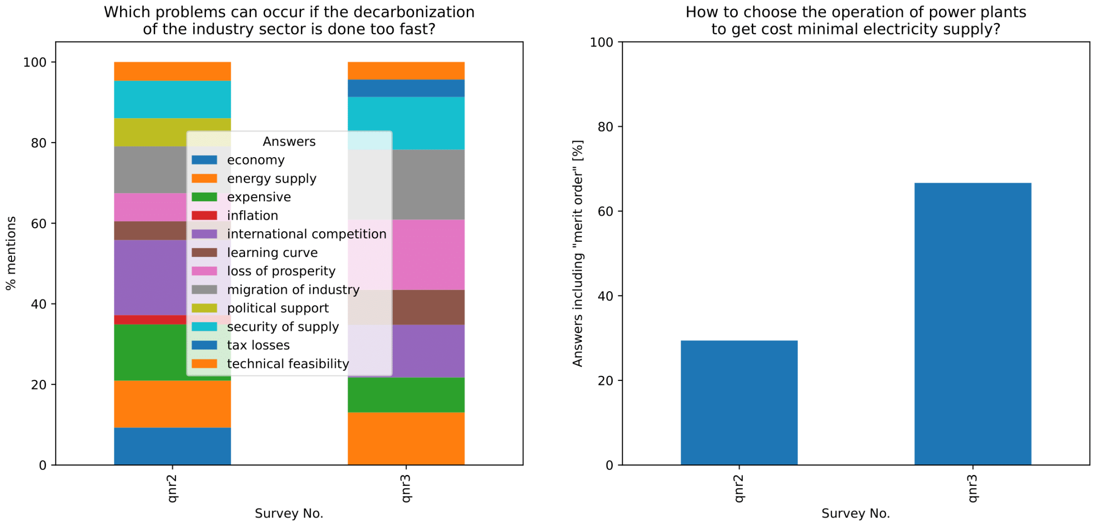
- helps with the **emissions reduction**.
- is necessary since the heat sector still has a high share of **fossil production**.
- is important since the heat sector itself has a **high share** of the final energy demand
- is important to reduce **international dependency**, i.e., on gas import.
- is important since it increases **sector coupling**.

Figure C.3b shows the categorized answers to question *B*02. The following categories were identified:

The decarbonization of the heat sector can be achieved by...

- using more **district heating**.
- applying **demand side management**, i.e., shifting or reducing heat demand.
- using more **hydrogen**.
- electrification of the heat production, i.e., using **heat pumps**.

- build/using more **heat storage**.
- better **insulation of buildings** to reduce the heat demand.
- using more combined heat and power (**CHP**).
- **political measures**, i.e., ban new oil heaters or subsidize heat pumps.
- using more **renewable energies**, like solarthermics.



(a) *B*03: Hasty industry decarbonization. (b) *B*04: Cost optimal operation of power plants.

Figure C.4.: Results of the open questions regarding industry decarbonization and cost optimal operation of power plants.

Figure C.4 shows the answer to to the remaining two open text questions from block B, i.e., *B*03 and *B*04. In Figure C.4a, the categorized answers of question *B*03 are shown. Multiple categories per answer were possible and the following categories were identified:

A hasty decarbonization of the industry sector can lead to...

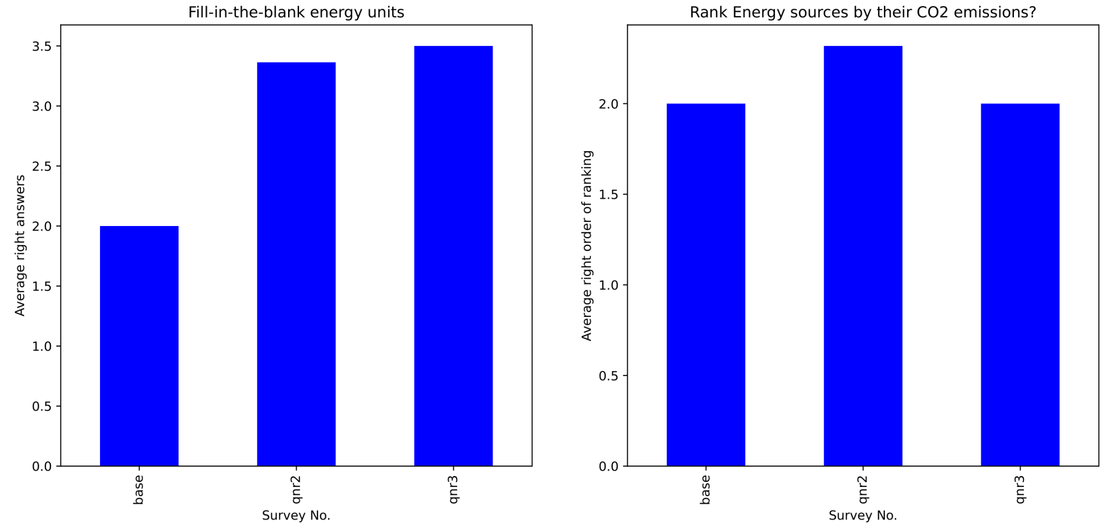
- an **economy** breakdown.
- a shortage of **energy supply**.
- high costs for the industry, i.e., is **expensive**.

- a high **inflation**.
- the local industry having a disadvantage in the **international competition**, e.g., due to high energy prices.
- lost opportunities due to better technologies in later stages in development (**learning curve**).
- a **loss of prosperity**, as a consequence of job losses and harm to the economy.
- a **migration of industry** to other states with less restrictions.
- a loss of **political support**, e.g., public protests.
- loss of **security of supply** (of produced goods) and in the process to higher prices.
- to **tax losses** since the income generated by the CO_2 taxes is lower due to lower emissions.
- no problems since **technical feasibility** limits the speed of the decarbonization.

Figure C.4b shows the percentage of answers to question $B04$ that either include the phrase "merit order" or a description of its functioning. The scheme of categorizing all answers was omitted for $B04$ since the question aimed at validating if the participants learned about the merit order as one of the vital concepts of the lecture.

Figure C.5 shows the results of the fill-in-the-blank-text questions ($C01$ – $C08$) and the task where participants were asked to order different means of electricity generation by their CO_2 intensities. For the fill-in-the-blank-text the average right answers are shown in Figure C.5a. In the first questionnaire participants were to fill in 2 out of 8 blanks correctly. The average right answers increase for the later surveys to 3.3 and 3.5 correct answers. Figure C.5b shows the average of a score for the correct ranking of the CO_2 intensities. The score has a base value of zero and was increased by one if the intensity of the named method of electricity generation is higher then the intensity of the one named before. I.e., a score of two means that twice the successor had a higher intensity then its predecessor. The best possible score is four, if all fife means of electricity generation are in the right order.

Figure C.6 shows the results of questions $C10$ – $C12$, where participants were asked to give a numeric answer first and then argue how they had come to the result. The asked questions did not give any details so it was up to the participants to take some assumptions. The top part of the figure shows the percentage of correct answers. An answer was deemed as correct, if it was in the same order of magnitude then the expected answer based on



(a) $C01 - C08$. Fill-in-the-blank energy units.

(b) $C09$. Electricity generation by CO_2 intensity.

Figure C.5.: Results of questions $C01 - C09$. Energy units and CO_2 intensities.

the assumptions taught in the lecture. The lower part of the figure shows the percentage of calculations used as arguing for the correctness of the own answer. Note, that only the presence of a calculation as a justification was tracked, not if the calculation was actually correct.

Figure C.7 shows a bar plot of the results for question $D01$ and $D02$ on the merit order. The left part of the figure shows the percentage of right answers given to $D01$. The correctness was checked manually. An answer was deemed correct if it explained that the generation is sorted in ascending order by its marginal cost of production. The left part of the figure shows the percentage of reasonings with valid arguments for the use of the merit order. Reasons deemed valid are a minimal energy price, the incentives for building power plants with low variable costs (e.g. renewable) to get accepted, and the highest marginal cost as income for everybody as a means to finance the investment.

The bar plot in Figure C.8 shows the average number agents and institutions that are related to the energy transition named by participants in the three different surveys.

Figure C.9 shows several bar plots that show the sector which was perceived as mainly responsible for the statements $F01 - F08$. The figure starts with $F01$ in the top left plot

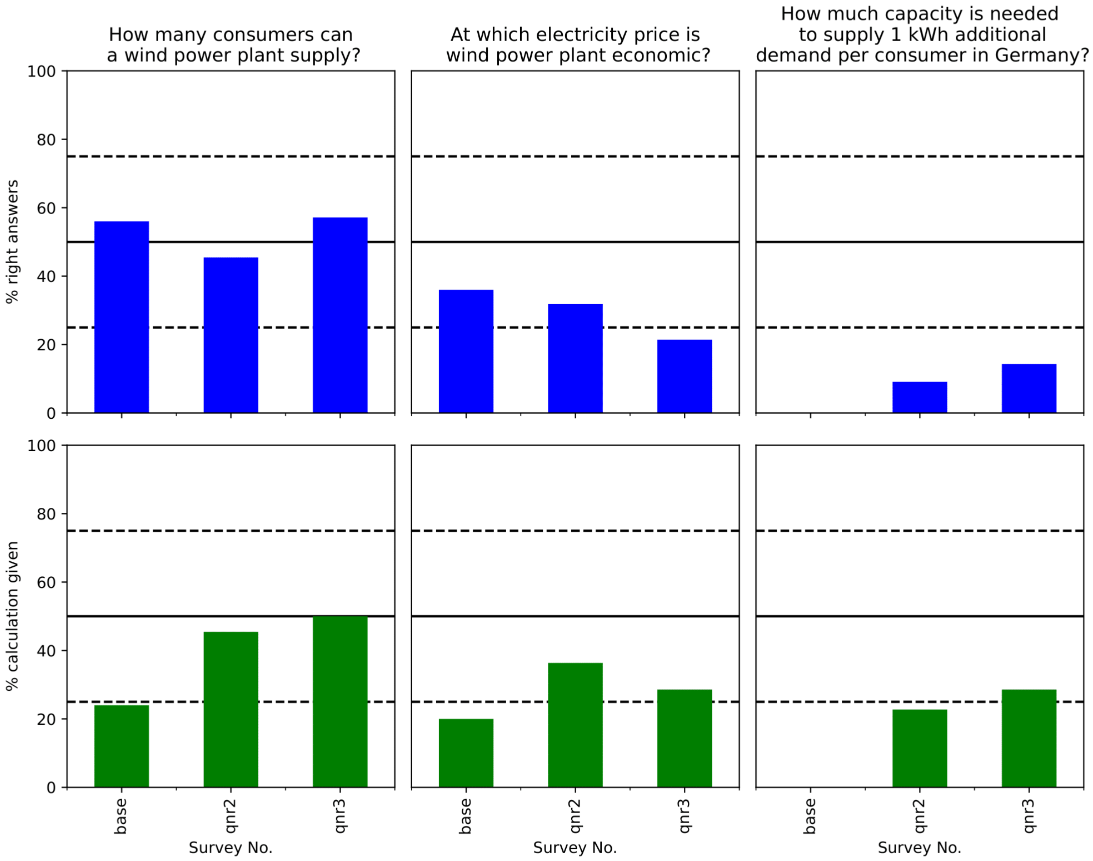


Figure C.6.: Results of questions $C10$ – $C12$: working with energy units. Answers is within right order of magnitude (top) and reasoning is based on calculations (bottom).

and continues with a plot for every question until $F08$ in the bottom right. Each plot shows the relative share of the group seen as mainly responsible for the given part of the energy transition in each question grouped by the different surveys. For $F01$, energy sector and politics are perceived as most responsible. The share of politics cannibalizes a small share of industry and the energy sector and politics end up with roughly the same share in the last survey. According to the participants, politics is mainly responsible for $F02$ in all surveys. A minority of opinions switch from industry as the main contributor to the energy sector over the course of the studies. Roughly 80% see the private households responsible

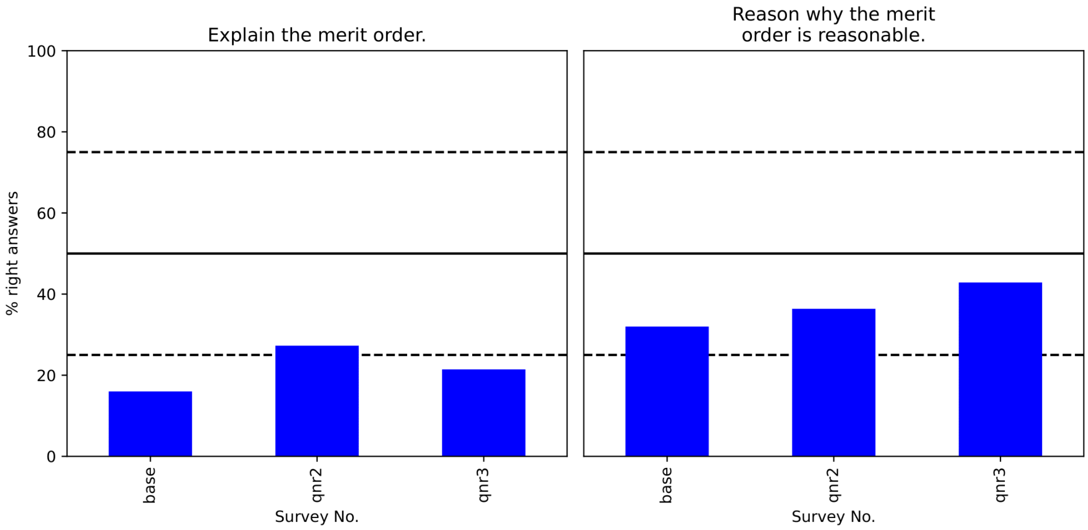


Figure C.7.: Results of questions $D01$ and $D02$. Percentage of right explanations of the merit order (left) and good arguments for its reasonableness (right).

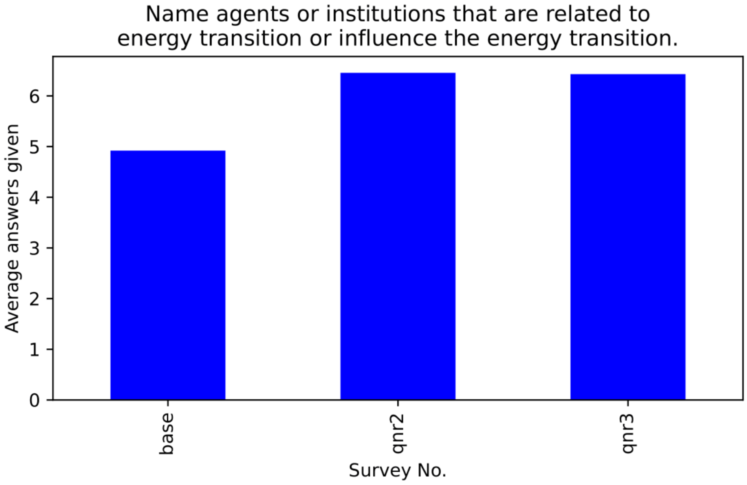


Figure C.8.: Results of questions $E01$. Agents and institutions of the energy transition.

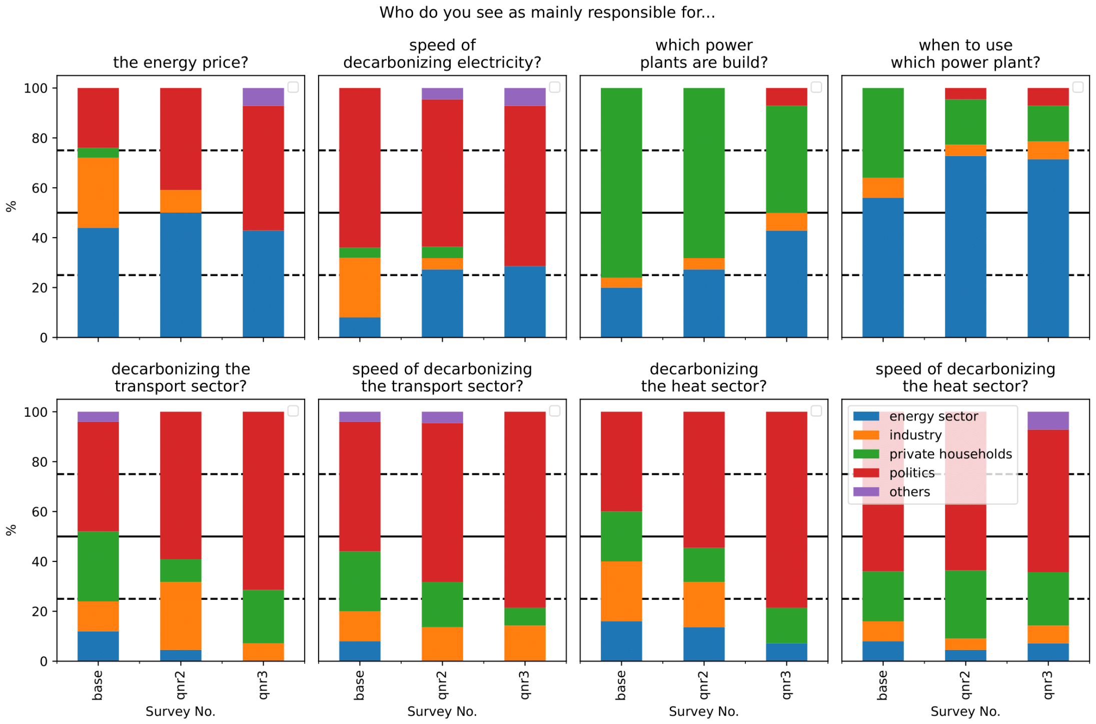


Figure C.9.: Results to questions $F01$ – $F08$. Perceived responsibility in the energy transition grouped by the different surveys.

for $F03$ in the first survey. The opinion on the role of the private households decreases to be only about 40% in the last survey with increasing responsibilities seen with the energy sector. Further, the energy sector is perceived mainly responsible for $F04$ with shares of 55% – 70%.

In the bottom half of Figure C.9, i.e., questions $F05$ – $F08$, politics is seen as mainly responsible. While the share of participants that see politics as responsible increases over the course of the different surveys for $F05$, $F06$, and $F07$, it remains about the same for $F08$.

Figure C.10 lists the average emission reduction potentials named by the participants for the different sectors, i.e., questions $F09$ – $F12$.

The Figures C.11, C.12, and C.13 show each three word clouds, one for each survey. In

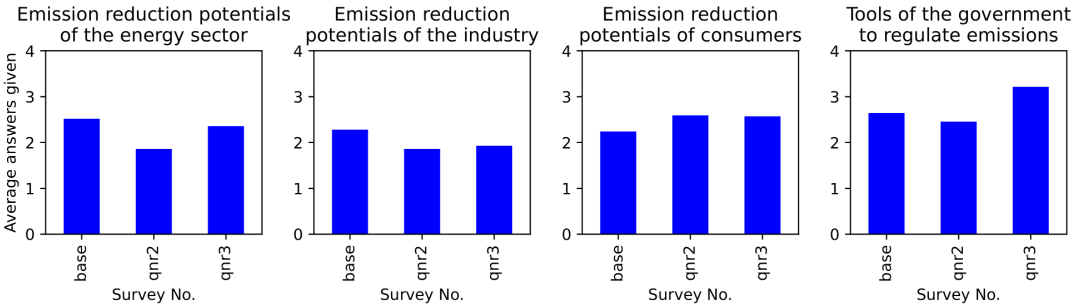


Figure C.10.: Results for questions $F09$ – $F12$. Average emission reduction potentials named for different sectors.

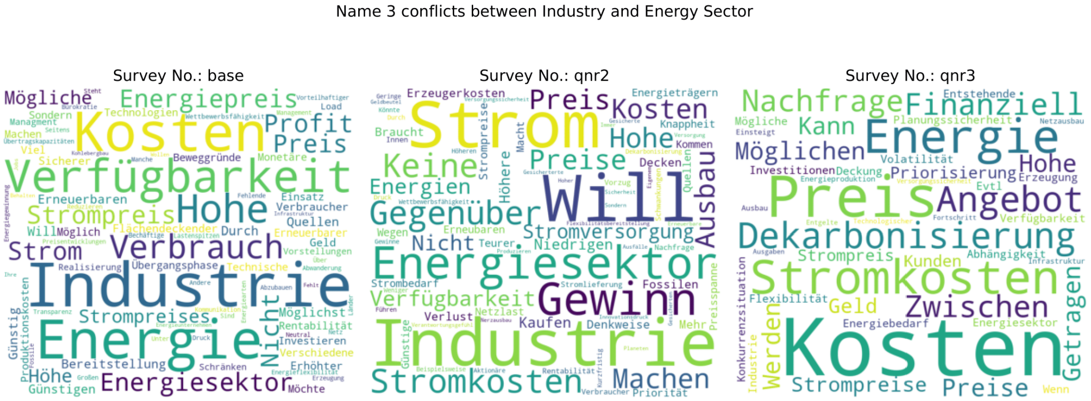


Figure C.11.: Word cloud of named conflicts of interests between the energy sector and industry ($G01$).

the word cloud the size of each word is proportional to the amount it was mentioned in the named conflicts of interest of the respective question. Words with 3 letters or less were excluded. The original questionnaires were asked in the German language, hence, the words in the word clouds are also German. Figure C.11 shows the word clouds of the conflicts of interest between the energy sector and the industry ($G01$). Figure C.12 depicts the word clouds of the conflicts of interest between the energy sector and politics. The word clouds of the conflicts of interest between private households and industry are shown in Figure C.13.

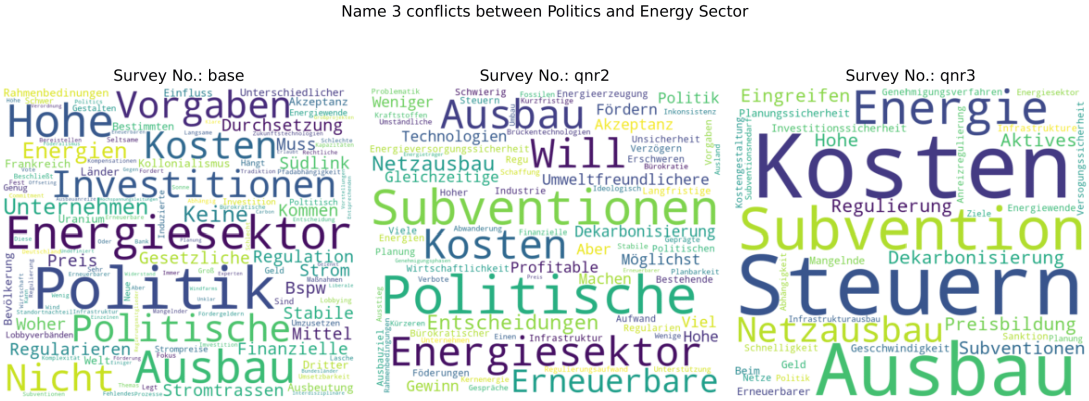


Figure C.12.: Word cloud of named conflicts of interests between the energy sector and politics ($G02$).

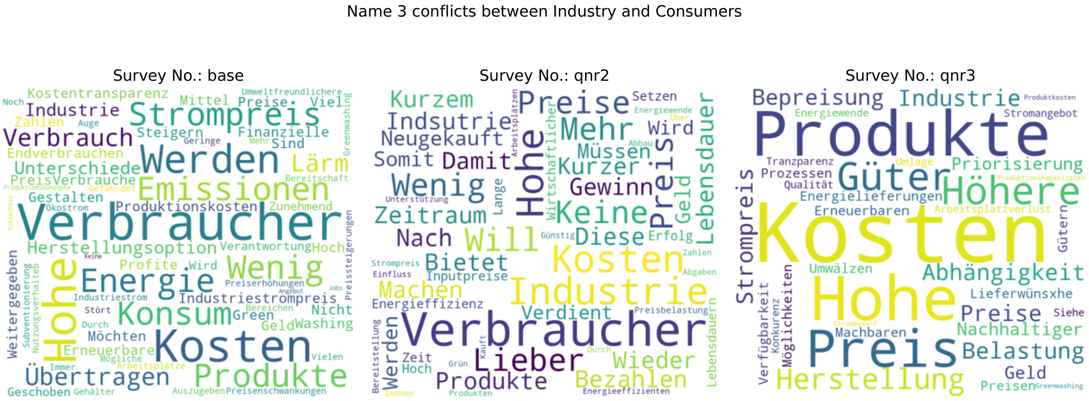


Figure C.13.: Word cloud of named conflicts of interests between private households and industry ($G03$).