Generation of Counterfactual Explanations in Smart Environments

Master Thesis

Author: Anna Trapp (Student ID: 7354713)
Supervisor: Prof. Dr. Andreas Vogelsang
Co-Supervisor: Dr. Mersedeh Sadeghi

Chair of Software and Systems Engineering
Faculty of Mathematics and Natural Sciences
University of Cologne

December 23, 2024

Contents

1 Introduction

Smart environments consist of sensor-enabled devices that work together to support users through decision-making, monitoring and controlling systems, and managing abnormal situations. Prominent examples include smart homes, offices, and buildings (El-Din et al., 2021; Ahmed et al., 2016). The adoption of smart environments is rapidly increasing, driven by advancements in the Internet of Things (IoT) and Artificial Intelligence (AI), decreasing costs of smart devices, and the growing availability of integration systems (Li et al., 2021).

In this thesis, we focus on rule-based smart environments, as they are one of the most common approaches to implementing such systems (Nandi & Ernst, 2016). Rule-based smart environments operate by executing predefined rules whenever their preconditions are satisfied (Sadeghi et al., 2024). However, these systems are often perceived as black boxes with users expressing the desire to understand their behavior (Jakobi et al., 2018). Furthermore, Bunt et al. (2012) found that while most users understand the general concept of such systems, they often lack knowledge about their details and how they interact.

Providing explanations can address this issue, as explanations significantly enhance user understanding (Chazette & Schneider, 2020). Lombrozo (2006) also highlights the significance of explanations, as "they are central to our sense of understanding, and the currency in which we exchange beliefs". Furthermore, the provision of explanations can enhance performance, user perception, and learning, empowering users to predict and control future outcomes by offering insights about the past and highlighting key information (Lombrozo, 2006; Gregor & Benbasat, 1999). Additionally, integrating an explanation layer significantly improves users' trust in a system. This is particularly crucial when there is a mismatch between user expectation and system behavior, which is often the case in smart environments (Lim et al., 2009; Sadeghi et al., 2021). Winikoff (2018) even argues that explainability is a prerequisite for trust. In contrast, Kästner et al. (2021) cautions that explainability should not be used as a substitute for trust but also argues that to make a system more trustworthy, explainability is still important. Moreover, Balasubramaniam et al. (2023) identify a strong connection between explainability, trust, and the transparency of AI systems. Finally, explainability is increasingly recognized as a non-functional requirement of its own (Köhl et al., 2019).

Counterfactual explanations represent one type of explanation that focuses on counterfactual ("contrary-to-fact") events. They provide insights into how an outcome A could have been achieved by analyzing what would be different from the current situation if the counterfactual events had happened ("A would have happened if. . . ") (Stepin et al., 2021). The concept of thinking about counterfactual events is deeply rooted in humans as it plays a crucial role in how children learn (Guidotti, 2022). Research even suggests that humans derive stronger causal conclusions from counterfactual than from factual reasoning (Mandel et al., 2005). Hence, counterfactual explanations amplify causal judgments (Byrne, 2019). Mandel et al. (2005) even argue that counterfactual thinking is a defining characteristic of humanity. Furthermore, Wachter et al. (2017) examine the EU General Data Protection Regulation (GDPR) and the concept of a *Right to Explanation* from the perspective of the individual whose data is processed. They propose three aims that explanations should achieve in this context: (1) informing and helping users understand system decisions, (2) providing a basis for reversing incorrect decisions, and (3) clarifying what changes could lead to a desired outcome. Since counterfactual explanations meet these objectives without having to open the black box of the decision-making process, they provide a powerful tool in the context of machine learning and in explaining its decisions to users.

In addition, counterfactual explanations can not only be helpful in the context of machine learning but have the potential to show great results when applied to smart environments. Firstly, a key purpose of explanations is to enhance system usability by aiding user understanding and teaching users how to operate the system more effectively (Chazette & Schneider, 2020). This is precisely what counterfactual explanations in smart environments accomplish when they are designed in a way that enables users to implement the proposed changes. Additionally, Roese (1997) argues that counterfactual thoughts are most prevalent when there is a need for correction and are most effective in situations that are controllable and likely to repeat in the future. This aligns seamlessly with the use case of explanations in smart environments, making counterfactual explanations particularly well-suited for such contexts. Moreover, Lim et al. (2009) speculate that proactive systems such as smart environments may benefit more from explanations answering *how to* and *what if* rather than *why* questions. Counterfactual explanations are an example of explanations answering exactly these questions (Woodward, 2003) and, as such, have the potential to show excellent results when applied in smart environments.

Despite these promising propositions, there is currently no consensus on a formal definition of counterfactual explanations in the context of rule-based smart environments, nor are there any methods for their generation. To address this gap, we propose a formal definition of counterfactual explanations in smart environments grounded in existing literature. Furthermore, we present a framework for generating counterfactual explanations in rule-based smart environments and provide an implementation to evaluate its feasibility. Additionally, we conduct a user-centric evaluation, addressing a significant gap in the field as these types of evaluations remain rare and pose an open research challenge (Guidotti, 2022; Verma et al., 2024).

This work is organized as follows: Section 2 provides background by introducing relevant concepts related to counterfactual explanations and smart environments. Additionally, we give an overview of related work. In Section 3, we define counterfactual explanations in the context of smart environments, propose a framework for their generation, and describe the implementation. To assess the practical value of our framework, we conduct a user study, described in Section 4. The results are discussed in Section 5. Finally, Section 6 concludes our work with a summary of our contributions and a discussion of future work.

2 Background and Related Work

Before we propose our framework for generating counterfactual explanations in rule-based smart environments, we introduce the relevant concepts and provide an overview of the related work.

2.1 Background

To begin with, the term *smart environment system* refers to a connected system where sensor-enabled devices work together to improve the comfort of their users (Ahmed et al., 2016). These systems typically possess autonomous perception, reasoning, and action capabilities within their environment (Alam et al., 2012). So-called *rule-based systems* are commonly used in smart environments (Nandi & Ernst, 2016). They contain two primary components: a knowledge base and an inference engine. The knowledge base contains all necessary information, such as the rules and the state of each device within the smart environment (Masri et al., 2019; Hayes-Roth, 1985). Each state of a device, like the "color of the desk lamp", can either be manipulated by the user directly or via the action of a rule. Typically, a *rule* consists of two components: *preconditions* and *actions*. Preconditions are logical expressions that determine the truth value of device states in the smart environment, such as "the desk lamp is green" (Herbold et al., 2024). Actions, for example "turn on the fan", describe what a rule should do once all of its preconditions are met and it *fires* (Nandi & Ernst, 2016). The rules are controlled by the inference engine of the rule-based system. It checks whether preconditions are true and fires the rules. Furthermore, conflicting situations may arise when multiple rules with contradictory actions simultaneously have true preconditions. For such situations, conflict resolution techniques determine what rule should be fired. Common examples include minimum specificity (choosing the rule with minimal preconditions), regency (determining the changes that occurred most recently and taking the rule that uses them), breadth-first (choosing the rule that has true preconditions the longest), or a random strategy (Ali et al., 2018). However, we apply priority-based scheduling to our framework as it is especially suitable for IoT systems with overriding preconditions (Shah et al., 2019). Here, each rule is associated with a *priority*, and when a conflicting situation arises, only the rule with the highest one is fired (Hayes-Roth, 1985). Notably, no rules have the same priority.

Rule-based smart environment systems can be equipped with an explanation layer that enables them to explain their behavior (Masri et al., 2019). Formally, a system is considered *explainable* if, and only if, it can provide information (the explanation) to the explainee, such that the explainee understands the explanandum

(Köhl et al., 2019; Chazette et al., 2021). The *explanandum* refers to the event to be explained, and the *explainee* is the user that requests an explanation (Madumal et al., 2020). Typically, the provided explanation describes how the internal logic of an algorithm led to a decision. Counterfactual explanations, on the other hand, clarify how external facts influenced the outcome (Wachter et al., 2017) and contain instructions on how it could have been changed. In natural language, they consist of: (1) a condition C describing an alternative event to an actual one, (2) an outcome A that would have occurred had the condition been true, and (3) a relation between them, expressed as: A would (not) have happened if C had (not) happened. But definitions of counterfactual explanations and their relation to causation still compete (Stepin et al., 2021). Therefore, we provide an overview before utilizing them to propose our definition of counterfactual explanations in rule-based smart environments.

Theories on counterfactual explanations presupposing a causal nature (as is the case in rule-based systems) can be divided into four milestones (Stepin et al., 2021): Firstly, Stalnaker (1968) and Lewis (1973) consider *possible worlds*, which are worlds that coincide with the real one except for a specified difference. Then, the counterfactual statement "if A was true, then B was true" holds in the real world if, and only if, there is a possible world where both A and B are true and that is more similar to the real world than any possible world where A is true but B is not. Secondly, Pearl (2000) proposes the *Structural Causal Model* centered around the notion of sustenance. It states that A causally sustains B (ensures B remains unchanged during an intervention) if A is both necessary and sufficient to sustain B. Furthermore, Woodward (2003) proposes an *Interventionist Account of Explanation*: "It is only when one has identified conditions relevant to the manipulation of the explanandum that one has provided an explanation." Finally, the *Neyman-Rubin Causal Model* interprets counterfactuals in terms of potential outcomes of a dependent causal variable given some intervention (Stepin et al., 2021).

In practice, these theories have been used to develop several definitions of counterfactual explanations. Especially in the context of eXplainable Artificial Intelligence (XAI), numerous definitions have emerged. Wachter et al. (2017) implicitly include the notion of possible worlds by defining counterfactual explanations in machine learning as statements of the form: score p was returned because variables V had values $(v_1, v_2, \ldots)$ associated with them. If V instead had values $(v_1', v_2', \ldots)$, and all other variables had remained constant, score p' would have been returned. In contrast, Bertossi (2020) uses the interventionist account of explanation by Woodward (2003) to define an explanation for classification as a set of original feature values that are altered by a minimal counterfactual intervention. Guidotti (2022) also includes minimality in his definition of counterfactual explanations for classification by stating that they consist of an instance that differs minimally from the original input but receives a different classification. All in all, the most commonly accepted definition of counterfactual explanations in the context of XAI includes a minimal set of changes to features such that the model changes its prediction (Stepin et al., 2021).

When determining the minimal change, we employ a *Multi-Criteria Decision Making* (MCDM) method. MCDM methods determine the best choice among a finite set of alternatives available to the decision maker. The choice is determined by considering multiple *decision criteria*, which are desirable properties of the best solution. These criteria may be conflicting, and the MCDM method must weigh them against each other. Additionally, the criteria may be associated with weights that influence their importance in the decision making process (Triantaphyllou, 2000). One of the most widely used MCDM methods is TOPSIS (Technique for Order Preference by Similarity to Ideal Solution) (Taherdoost & Madanchian, 2023). It was first proposed by Hwang and Yoon (1981) and is based on the concept that the selected alternative should have the shortest distance from the ideal and the longest distance from the worst solution. TOPSIS assumes that each criterion is either beneficial or non-beneficial. It starts by determining the performance measure $x_{i,j}$ of alternative A_i in terms of the j-th criterion for all $i \in \{1, \dots, m\}$ and $j \in \{1, \dots, n\}$. Then, all $x_{i,j}$ are normalized using (usually) the Euclidean distance and, if available, multiplied by their according weights w_j:

$$v_{i,j} := \frac{x_{i,j}}{\sqrt{\sum_{k=1}^{m} x_{k,j}^2}} \cdot w_j. \tag{1}$$

Afterward, the ideal and worst solution are determined. For the ideal solution A_*, among the alternatives, the maximum for each beneficial criterion and the minimum for each non-beneficial criterion is taken. Contrarily, for the worst solution A_-, the minimum for each beneficial and the maximum for each non-beneficial criterion is taken. Subsequently, for each alternative, the Euclidean distance to the optimal and worst solution is determined:

$$S_{i,*} = \sqrt{\sum_{j=1}^{n} (v_{i,j} - v_{*,j})^2}, \quad \text{for } i = 1, \dots, m \tag{2}$$

$$S_{i,-} = \sqrt{\sum_{j=1}^{n} (v_{i,j} - v_{-,j})^2}, \quad \text{for } i = 1, \dots, m, \tag{3}$$

where $v_{*,j}$ is the best and $v_{-,j}$ the worst result for each criterion. Next, the relative closeness of each alternative A_i is determined:

$$C_{i,*} = \frac{S_{i,-}}{S_{i,*} + S_{i,-}}. \tag{4}$$

Finally, the best alternative is determined by choosing the alternative with the lowest $C_{i,*}$, which corresponds to the shortest distance to the optimal solution. From Equation (4), it is apparent that the best alternative also has the longest distance to the worst solution (Triantaphyllou, 2000).

2.2 Related Work

There are several works associated with either explainable smart environments or counterfactual explanations in the context of XAI.

Smart Environments Blumreiter et al. (2019) propose the framework *MAB-EX* (Monitor, Analyze, Build, Explain) intended for cyber-physical systems, such as smart environments. It sketches how to build self-explainable systems that evaluate requirements and explainability models at run-time. Furthermore, Houzé et al. (2022) deliver a modular XAI reference architecture for explainable smart homes. It contains *LECs* (Local Explanatory Components) that each can explain the behavior of one device in the smart home. They are extendable at run-time and communicate with each other. Additionally, a central component to coordinate the *LECs* and to generate system-wide explanations is included. Dobrovolskis et al. (2023) introduce an agent-based approach to develop explainable IoT systems. The framework incorporates various types of agents, including those responsible for sensing, data collection, decision-making, and executing physical activities. Their proposed method was applied to create an explainable, rule-based smart home system, which was subsequently evaluated through a one-year study conducted in three laboratory rooms. In addition, Das et al. (2023) propose an explainable Human Activity Recognition (HAR) framework that leverages state-of-the-art XAI methods to generate natural language explanations for why a specific activity was detected. The framework is evaluated in the context of smart homes, where caregivers remotely monitor individuals who live alone or require assistance. Sadeghi et al. (2024) propose *SmartEx*, a framework for generating user-centric explanations in rule-based smart environments by including different contexts in the explanation. Moreover, Herbold et al. (2024) extend *SmartEx* to include contrastive explanations, which explain why an event occurred instead of another one.

Counterfactual Explanations Counterfactual explanations are commonly used in the context of XAI. For example, Madumal et al. (2020) offer an approach that learns a Structural Causal Model as introduced by Pearl (2000) through reinforcement learning. Then, counterfactual explanations are generated through an analysis of the model. Furthermore, Poyiadzi et al. (2020) argue that a generated counterfactual may not represent the underlying data distribution and prescribe unrealistic goals. Therefore, they introduce *FACE* (Feasible and Actionable Counterfactual Explanations), a method to generate counterfactual explanations in the context of machine learning. *FACE* uncovers *feasible paths* between the current and desired state of the object by considering the shortest path distances defined via density-weighted metrics. Del Ser et al. (2024) use a GAN (Generative Adversarial Network) and multi-objective optimization weighing plausibility, the intensity of changes, and adversarial power to model the distribution of an input into a black box model and to generate a counterfactual explanation from it. In contrast, Mothilal et al. (2020) focus on diversity and feasibility. They propose *DiCE* (Diverse Counterfactual Explanations), a framework to explain machine learning classifiers through the generation and evaluation of a diverse set of counterfactual explanations. They incorporate Determinantal Point Processes (DDP) to promote diversity and include proximity and sparsity for feasibility. *DiCE*, however, assumes the underlying classification model to be differentiable and, therefore, excludes tree ensembles. Lucic et al. (2022) resolve this problem by presenting *FOCUS* (Flexible Optimizable CounterfactUal Explanations for Tree EnsembleS), which formulates the problem of finding counterfactuals as a gradient-based optimization task. *FOCUS* provides an optimal counterfactual example by determining a minimal change to the input that results in an alternative prediction. Bertossi (2020) suggest an ASP (Answer-Set Program) that determines counterfactual interventions to explain decisions made by classification models. Their approach focuses on determining the features most responsible for the classification and can be applied to black box models as well as rule-based classifiers. Van der Waa et al. (2018), in contrast, argue that this approach becomes infeasible in a high-dimensional feature space. Therefore, they deliver a method using one-versus-all decision trees to identify the set of rules that led the classifier to identify it as the foil and not the fact. Thus, they provide contrastive instead of counterfactual explanations. Finally, Ranjbar et al. (2024) present three methods to explain recommendation systems using counterfactual textual explanations.

But to the best of our knowledge, no work on counterfactual explanations in rule-based smart environments has been done. We try to close this gap by introducing our framework for generating counterfactual explanations.

3 Approach

In the following, we define counterfactual explanations in rule-based smart environments by adopting the notion of minimal change utilized by several related works mentioned in Section 2. Subsequently, we provide a framework for their generation and its implementation.

Definition (Counterfactual explanations in smart environments)**.** *A counterfactual explanation in a smart environment is an explanation containing the minimal change to explanation constructs, such that a specific foil would have occurred instead of the fact.*

Here, we follow Sadeghi et al. (2024) and define *explanation constructs* as a set of specifications, facts, propositions, and events related to both the internal elements and the external world. The notions of fact and foil stem from contrastive explanations that ask the question, "Why did the fact occur rather than the foil?" (Lipton, 1990). The *fact* refers to the event or piece of information that caused the need for an explanation, whereas the *foil* refers to the event the user expected to happen (T. Miller, 2021). Definitions of counterfactual explanations differ in their dealing with the foil. While Guidotti (2022) and Bertossi (2020) define the foil as any event different from the fact, Wachter et al. (2017) fix the foil to a specific incident. We argue that in the context of smart environments, the user is not only interested in undoing the fact but also wants to achieve a particular foil. Thus, we follow Wachter et al. (2017) by selecting a particular event as the foil. Furthermore, we adopt the concept of minimal change, as research shows that humans typically consider only one or two causes when explaining an event and prefer to avoid unnecessary information (T. Miller, 2019; Chazette & Schneider, 2020). Therefore, providing as little information as possible while still enabling users to resolve the confusing situation leads to the most effective outcomes.

Moreover, counterfactual explanations can be distinguished by their structure; there are additive and subtractive ones. *Additive* counterfactual explanations add new information to the situation while *subtractive* counterfactual explanations remove them (Roese & Epstude, 2017). We adapt this definition to rule-based systems and define additive counterfactual explanations as explanations concerned with the firing of rules. Conversely, we define subtractive counterfactual explanations in rule-based smart environments as explanations concerned with the removal of rules, where the removal refers to changing the system so that the rule no longer has true preconditions. Markman et al. (2007) showed that additive counterfactual explanations evoke an expansive processing style and, therefore, favor creative problem-solving, whereas subtractive ones evoke a relational processing style that facilitates analytic task performance. Together, they allow users to apply cognitive processes triggered in one task to another.

After formalizing the definition, we now present our framework to generate these counterfactual explanations in rule-based smart environments. The framework utilizes the foil determination method by Herbold et al. (2024) and determines the minimal change mentioned in the definition of counterfactual explanations. Finally, the framework uses a natural language pattern to transform the determined minimal change into a natural language explanation. Our framework assumes that the need for an explanation has already been identified, either by a user or through a dedicated system. We assume that a mismatch between reality (the fact) and the user's expectation (the foil) has occurred and define this as a *confusing situation*. Furthermore, we assume the inference engine of the underlying rule-based system works correctly, and the confusing situation was not caused by an error. In such cases, we redirect to the work by Herbold et al. (2024).

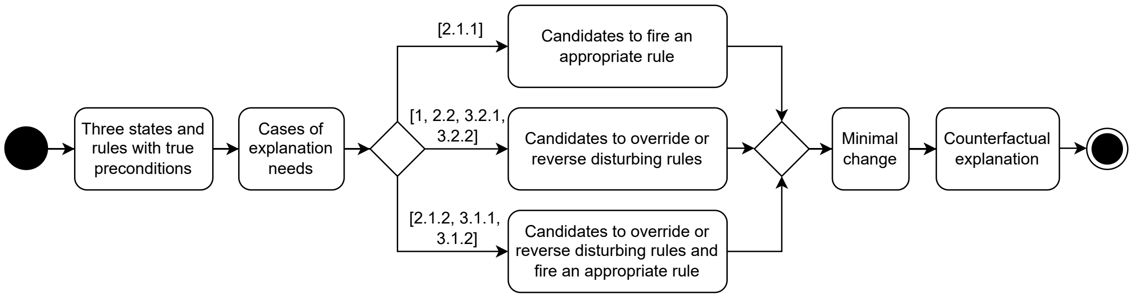

Three states and rules with true preconditions
Cases of explanation needs
[2.1.1]
Candidates to fire an appropriate rule
[1, 2.2, 3.2.1, 3.2.2]
Candidates to override or reverse disturbing rules
[2.1.2, 3.1.1, 3.1.2]
Candidates to override or reverse disturbing rules and fire an appropriate rule
Minimal change
Counterfactual explanation

Figure 1: Procedure for developing a counterfactual explanation

In the following chapter, we describe the procedure for developing a counterfactual explanation, as illustrated in Figure 1. In Section 3.1, we begin by identifying the current, previous, and expected states of the device that caused the confusing situation and the rules with true preconditions whose actions result in the system to assume these states. The states and rules are then used in Section 3.2 to determine the appropriate case of the explanation need. Based on the identified case, the framework calculates multiple candidates for the minimal change required to achieve the foil, as detailed in Section 3.3. Each candidate consists of a set of explanation constructs requiring either general or specific state changes so that the system would result in the foil instead of the fact. To reduce the number of candidates, several minima are computed during the collection phase, and once all candidates are collected, the minimal one is determined. All minima are computed the same way using the desirable properties outlined in Section 3.4. The specific computation process for determining the minimal change is detailed in Section 3.5. The selected minimal candidate is then transformed into a natural language counterfactual explanation using the pattern described in Section 3.6. Finally, in Section 3.7, we provide an implementation of our framework.

3.1 Different States and Rules with True Preconditions

In the following, we refer to the device responsible for the confusing situation as the *device of interest*. To determine the minimal change, three states of this device are relevant:

- *The current state:* The state at the moment the need for an explanation arises.
- *The previous state:* The state before the current state if the state change caused the need for an explanation. If this is not the case, the previous state is set equal to the current one.
- *The expected state:* The state the user expected.

While the current and previous states are among the explanation constructs, the expected one is derived from the foil, which is determined using the framework by Herbold et al. (2024). It is important to note that the previous state may coincide with either the expected or current one. However, the current and expected state must differ as we assume the existence of a need for an explanation. Using these three states, we can identify the sets of rules leading to each state. A rule is said to *lead to* a state if its actions result in the device of interest to assume that specific state. The rules with true preconditions leading to the current, previous, and expected states are then used to determine the candidates for the minimal change. Notably, although there may be multiple rules with true preconditions, only the rule with the highest priority is actually fired. Additionally, the fired rule must lead to the current state as it is the state the device of interest is in at the time the need for an explanation arises.

3.2 Cases of Explanation Needs

Similarly to Herbold et al. (2024) and based on the current, previous, and expected states, we present three different cases for explanation needs, which are displayed in Table 1. We further categorize them into sub-cases, determined by the rules with true preconditions leading to the three states. Each sub-case resolves the confusing situation differently, as displayed in Figure 1. For each case, there are *disturbing rules*, which are the rules with true preconditions that caused the confusing situation and prevent us from achieving the foil. To resolve the confusing situation, the disturbing rules must be overridden or removed. A rule is *overridden* by firing another rule with higher priority, which executes an action that changes the state of

Table 1: Three cases of explanation needs

Case	Fact	Foil	State equal to the previous one
1	Event y occurred	No event occurred	Expected state
2	No event occurred	Event z occurred	Current state
3	Event y occurred	Event z occurred	None

the device specified in the actions of the original rule to a different one. Then, the original rule does not affect the system anymore as it is nullified by the rule with the higher priority. To fire the higher-priority rule, the states associated with its preconditions are changed to the states required to make the preconditions true. A rule is *removed* by making one of its preconditions false. This is done by changing a state mentioned in a precondition to any other state.

Case 1 In the first case, the user encounters a confusing situation because an unexpected event occurs. Therefore, the expected state of the device of interest is identical to the previous one, as mentioned in Table 1. We assume that it is too complex for a user to know that a rule has been overridden and to, therefore, expect nothing to happen. Hence, we exclude this case and suppose that there are no rules with true preconditions leading to the expected state. However, there has to be at least one rule with true preconditions leading to the current state, as a rule must have caused the event to occur. To achieve the foil, the system must return to the previous state. Consequently, the disturbing rules are the rules with true preconditions leading to the current state. As shown in Figure 1, removing all rules, without overriding any, is possible as we end up in the previous state, which is equal to the expected one.

Case 2 In the second case, the user is confused because they expected an event that did not happen. Hence, the previous state is equal to the current one, as displayed in Table 1. There can be rules with true preconditions leading to the expected and current states. Firstly, we call the case where there are no such rules case 2.1.1. As there are no rules with true preconditions, there are also no disturbing rules. In this case, achieving the foil requires firing one rule that leads to the expected state.

In case 2.1.2, there are still no rules with true preconditions leading to the expected state, but some leading to the current one. These are disturbing rules as they must be removed or overridden to achieve the foil. Removing all rules without firing any is not possible, as we would remain in the current state and not reach the expected one.

Finally, if there are rules with true preconditions leading to the expected state, they must be overridden by at least one rule with true preconditions leading to the current state. Otherwise, there would be no need for an explanation as the device of interest would be in the expected state. We refer to the situation where there are rules with true preconditions leading to both the current and expected states as case 2.2. The disturbing rules are those leading to the current state again. If any of them are already overridden by a rule leading to the expected state, they require no additional change and, therefore, do not influence the minimal change. Hence, for simplicity, such rules can still be included in the disturbing rules. As outlined in Figure 1, removing all disturbing rules and overriding none is possible, as the rules with true preconditions leading to the expected state ensure that we end up at the foil.

Case 3 In the third case, the user undergoes a confusing situation because something other than what they expected happened. In this case, the current, previous, and expected states are all different from each other, as showcased in Table 1. Because the current state is not equal to the previous one, a rule must have fired. Thus, there are rules with true preconditions leading to the current state. However, rules leading to both the expected and previous states can exist but do not have to. Therefore, there are four sub-cases:

Firstly, case 3.1.1 refers to the situation where there are no rules with true preconditions other than the ones leading to the current state. Here, a rule was fired, but the user expected a different action to happen. The disturbing rules are the rules with true preconditions leading to the current state as they must be removed or overridden. Additionally, a rule leading to the expected state must be fired because just removing all disturbing rules results in the previous and not the expected state.

Next, in case 3.1.2, there are rules with true preconditions leading to the current and previous states but none leading to the expected one. Notably, as the current state is not equal to the previous one, the rules with true preconditions leading to the previous state are overridden by at least one rule leading to the current one. In contrast to case 3.1.1, the disturbing rules also include the ones leading to the previous state. The reason for this is that when the user applies the changes, the rules with true preconditions leading to the current state may get removed. Then, a rule leading to the previous state may have the highest priority and impact the state of the device of interest and, hence, the minimal change. Similarly to case 2.2, we can add these rules to the disturbing rules as they require no additional change if they are already overridden. Again, firing is necessary as removing all rules would result in the previous and not the expected state.

Moreover, case 3.2.1 is concerned with rules with true preconditions leading to the current and expected states but none leading to the previous one. This case covers the situation where a user knew of a rule that would fire in this situation but is caught off guard because the rule is overridden by another one. The disturbing rules are the ones leading to the current state. Firing is not necessary as there are rules with true preconditions leading to the expected state, which ensure that we end up at the foil when all disturbing rules are removed.

The final case 3.2.2 has rules with true preconditions for all three states. To the user, the situation looks similar to case 3.2.1, but, as in case 3.1.2, the rules with true preconditions leading to the previous state may impact the computation of the minimal change. Therefore, the rules with true preconditions leading to the previous and current states are disturbing. Again, firing is not necessary as there are rules with true preconditions leading to the expected state.

3.3 Collection of Candidates

After the disturbing rules have been determined, they must be removed or overridden. Overriding a rule results in an additive counterfactual explanation, as we fire a new rule that introduces changes to the system ("A had happened"). In contrast, removing a rule results in a subtractive counterfactual explanation, as it removes changes from the system ("A had not happened"). We further also refer to the changes as additive and subtractive if they add or remove events. As outlined at the beginning of this section, additive and subtractive counterfactual explanations have different strengths. To maximize their benefits and not restrict ourselves unnecessarily, we incorporate both structures into our counterfactual explanation. Therefore, we can override or remove each disturbing rule independently. If there are multiple disturbing rules, the possible combinations of overriding some rules and removing others increase exponentially. Thus, we reduce them using the following approach:

If there are two rules, they must have different priorities. Then, if the rule with the higher priority is overridden by a third rule with even higher priority, the rule with lower priority is automatically also overridden. Hence, we sort all disturbing rules by their priority and determine the changes to the system if, for each rule r:

1. All rules with lower or equal priority than r (including r) are overridden by a rule.

2. All rules with higher priority than r are removed.

For each rule r, we derive a set of changes, where each set represents a candidate for the minimal change. As displayed in Figure 1, in cases 1, 2.2, 3.2.1, and 3.2.2, it is possible to remove all disturbing rules and not fire any. Therefore, we add this as an additional option to the set of candidates. In contrast, in cases 2.1.2, 3.1.1, and 3.1.2, it is not possible to remove all disturbing rules and not fire any. Hence, we add the option of removing all disturbing rules and firing a rule leading to the expected state with arbitrary priority to the set of candidates. Finally, in case 2.1.1, there are no disturbing rules, we just have to fire any rule leading to the expected state. Notably, just like overriding, the firing of rules introduces an additive component to our counterfactual explanation. The next step is to determine the specific changes required to fire, override, or remove these rules.

Additive Counterfactual Explanations To achieve the expected state through the firing of a rule, all rules leading to the expected state are collected. If another rule must be overridden, only rules with a higher priority are considered. Then, for each of these rules, the minimal change to fire them is determined. To fire a rule, all of its preconditions must be true. Therefore, all false preconditions of the rule are collected. In addition to directly changing the state of a device mentioned in a precondition to the desired one, rules whose actions result in the altering of the state can also be fired if they exist. For these new rules, the minimal change to fire them is determined in the same way. Then, for each precondition, the minimum between directly changing the precondition or firing any of the rules with according actions is determined. This process is repeated for all false preconditions, and the changes to the system are collected in a set containing the minimal changes required to make the rule fire. When the minimal change for each rule is determined, the rule requiring the minimal change among them is chosen. Then, the changes required to fire the chosen rule are the minimal changes needed to achieve the expected state through the firing or overriding of a rule.

Subtractive Counterfactual Explanations To remove a rule, that is, change something in the system such that the rule does not have true preconditions anymore, it is enough to change one precondition as they all need to be fulfilled for a rule to be able to fire. Therefore, the minimal change to make a precondition false is determined for each precondition, and a minimum is computed among them. When determining the changes required to make a precondition false, two additional aspects must be taken into account. Firstly, as in the additive case, once we determine the change to the system that would make the precondition false, this change can also be implemented through the actions of a rule that fired instead of a direct manipulation. We implement this concept just as in the additive case. Secondly, imagine the precondition we want to make false is "device d has state s_1". Then, there may be a rule r_1 with true preconditions that has an action "change device

d to state s_1". If we change device d to any other state, r_1 changes d back to s_1 instantly, making it impossible to turn the precondition false. To solve this issue, r_1 must be removed as well. Therefore, we check for these rules and remove them before removing the original precondition.

3.4 Desirable Properties

We identify five desirable properties that the minimal change that is contained in the counterfactual explanation should have. These properties majorly influence the determination of the minimal change during the collection process and the minimality computation.

Controllability Firstly, we consider the property of *controllability.* It refers to the ability of a user to implement the changes to the explanation constructs described in the explanation themselves (Byrne, 2019). We distinguish between three different levels of controllability, as suggested by Karimi et al. (2021) and adapt it to rule-based smart environments: A change to an explanation construct is called *actionable* if the user can directly change it, such as the temperature of a heater. Secondly, we call the change to an explanation construct *mutable but non-actionable* if it cannot be manipulated by the explainee directly but through the action of a rule. For example, consider a lab whose door is usually closed for all staff members. In this scenario, the explainee is a staff member and, therefore, cannot open the lab door. Hence, the opening of the lab door is not actionable. However, there is a rule that opens the lab door if the manager is on the same level as the lab. Thus, the lab door can be opened, though not by the explainee themselves. They have to manipulate the situation, such as asking the manager to come to the floor, to make the rule fire, and to enter the lab. Finally, a change to an explanation construct is *immutable* if the explainee cannot change them in any way, such as the weather.

Controllability is often considered the most important property of counterfactual explanations, as people, when thinking about alternatives, mentally alter events within their control over ones that are not and want to actually implement the minimal change contained in the explanation (Guidotti et al., 2018; Poyiadzi et al., 2020; Karimi et al., 2021). Therefore, we do not include controllability in the computation of the minimal change but directly in the collection of the candidates: Immutable changes to explanation constructs are only included if there are no actionable ones. For mutable but non-actionable changes to explanation constructs, we search for rules to change them and then consider the controllability of the preconditions of these rules.

Sparsity The first property that is directly included in the computation of the minimal change is the *sparsity* of a set of changes to explanation constructs. It refers to the number of changes to the system required to achieve the foil (Mothilal et al., 2020). For example, consider the candidate "the door should be closed, the light should be on, and the temperature should not be over 20°C". Then, the sparsity of this candidate is three as there are three features to change. Effectively, sparsity is determined by the size of the set. It is considered a non-beneficial property as the user wants to change as few features as possible and receive a short explanation (Verma et al., 2024). Dai et al. (2023) found that users prefer short and simple explanations in smart homes, further underlining the importance of sparsity in the context of smart environments. In addition to the use as a property for the computation of the minimum, sparsity is also included as a constraint: If a rule requires more than three changes to its preconditions to fire, it is excluded, as the explanation would get too long.

Temporality The notion of *temporality* considers the fact that people tend to mentally undo more recent events over ones that happened longer ago, as shown in two studies by D. T. Miller and Gunasegaram (1990). Therefore, the more recent an event is, the more likely it should be chosen to be manipulated. Temporality is directly included in the computation of the minimum and determined by considering each change to an explanation construct separately and then taking the average across all determined temporality scores. The temporality of a change is computed by taking the difference in seconds from the point where the explanandum occurred to the point where the change to the explanation construct last happened. If no such point exists, the change is assigned the maximum integer value to ensure that it will not be contained in the minimal change. Temporality is not a beneficial property as the longer ago the state contained in the change was last true, the less likely it should be chosen to be manipulated.

Proximity When determining a minimum, we want the change to the system to be as minimal as possible. To account for this, the non-beneficial *proximity* score is added to the computation of the minimum. It counts how many resulting changes to the system there are if all changes in a candidate set were applied (Mothilal et al., 2020). Our notion of proximity also considers what rules would be fired if the explanation constructs were changed and how these rules would override or make other rules fire that have actions themselves. Additionally, the proximity score is determined differently depending on whether the change corresponds to an additive or subtractive explanation. Changes relating to additive explanations contain instructions, such as "the lamp was turned on". Therefore, we consider the state of the device after this change is implemented and what effects it would have. For changes related to subtractive explanations, we consider the state the device of interest had before the current state and its effects. This is due to changes relating to subtractive explanations containing the removal of actions, such as "the lamp was not turned on".

Abnormality As argued by Kahneman and Tversky (1982), people do not just undo the most unlikely event out of all necessary conditions for an event but the most exceptional one. This finding motivates the concept of *abnormality* as a measure of how unusual an event is (Byrne, 2019). The abnormality of a change is determined by the percentage of how often the state mentioned in the change was true in the past compared to the other states of the device. For additive changes, abnormality is a non-beneficial property, as the more abnormal an event, the less likely we want to make it happen. In contrast, abnormality is a beneficial property for subtractive changes as we remove events. To make abnormality beneficial for additive changes, we alter the computation. Instead of determining the abnormality of the state we want to achieve, the abnormality of all other states of the device is determined and summed up. The determined abnormality then represents how abnormal it is for the device not to be in the according state. But this coincides with determining how normal it is for the device to be in the according state. Hence, the score is beneficial and comparable to subtractive changes. Finally, the average across all changes in a candidate set is taken to determine the abnormality of the candidate.

3.5 Minimal Change Computation

Once all candidates are collected, the optimal candidate is determined and selected as the final minimal change that is included in the explanation. Additionally, several minima are determined during the collection process, as outlined in the previous sections. To determine the minimum, firstly, all duplicates within and between the candidate sets are removed. Then, for each candidate, we calculate its sparsity, temporality, proximity, and abnormality scores, as outlined in Section 3.4. These properties serve as the decision criteria for selecting the optimal candidate. To compute the minimum, we employ TOPSIS, an MCDM method outlined in Section 2. TOPSIS was chosen due to its widespread use (Taherdoost & Madanchian, 2023) and its integration in the framework by Herbold et al. (2024), which we also use for the foil determination. During the TOPSIS computation, sparsity, temporality, and proximity are treated as non-beneficial criteria, while abnormality is considered beneficial. Furthermore, users can assign weights to the criteria based on their preferences or if they consider certain criteria to be more important than others.

3.6 Generation of the Counterfactual Explanation

The returned minimum contains a set of state changes to explanation constructs where the change is either specific ("device d should have had state s") or general ("device d should not have had state s"), depending on if it is an additive or subtractive change. We use a natural language pattern to transform the changes into a counterfactual explanation, which is provided to the explainee. The pattern references (1) the device of interest, (2) the expected state, followed by (3) all additive, and (4) all subtractive changes that must be implemented to achieve the foil. To convey that the explanation is concerned with the minimal change that should have happened, the tense is adjusted. The resulting pattern is as follows:

$$\begin{gathered}\text{The } \textit{device of interest} \text{ would be } \textit{expected state} \text{ if } \textit{additive changes} \\ \text{had happened and } \textit{subtractive changes} \text{ had not happened.}\end{gathered} \tag{5}$$

For example, the user is at home and watching TV when the TV suddenly turns off. The user asks for an explanation and receives the following: "The TV would be on if it was not after 11 pm." Here, the device of interest is the TV, and the expected state is the TV being on. The framework determines that the minimal change to achieve the foil is to remove the rule: "If it is after 11 pm, the TV turns off." The required subtractive change is "not after 11 pm", which is added to the natural language pattern.

3.7 Implementation

We implement our proposed framework as a plugin[1] for *SmartEx* by Sadeghi et al. (2024). *SmartEx* is a RESTful web service that can generate causal and context-aware explanations in rule-based smart environments and is implemented in Java using MongoDB as a database. It can be integrated into existing smart environments and provides an explanation layer for them while remaining decoupled from the core intelligent system. *SmartEx* utilizes Home Assistant[2] to take advantage of the inference engine and to fetch runtime data such as rules, device states, and a log of all past activities and states through a RESTful API. Additionally, *SmartEx* can provide contrastive explanations through a plugin designed by Herbold et al. (2024). We present a reference architecture of our implemented *Counterfactual Explanation Service* and the components from *SmartEx* it interacts with in Figure 2.

When a counterfactual explanation is requested, the *Case Distinction* component identifies the appropriate case of explanation need. This process involves determin-

[1] `https://github.com/ExmartLab/SmartEx-Engine/tree/counterfactual`

[2] `https://www.home-assistant.io/`

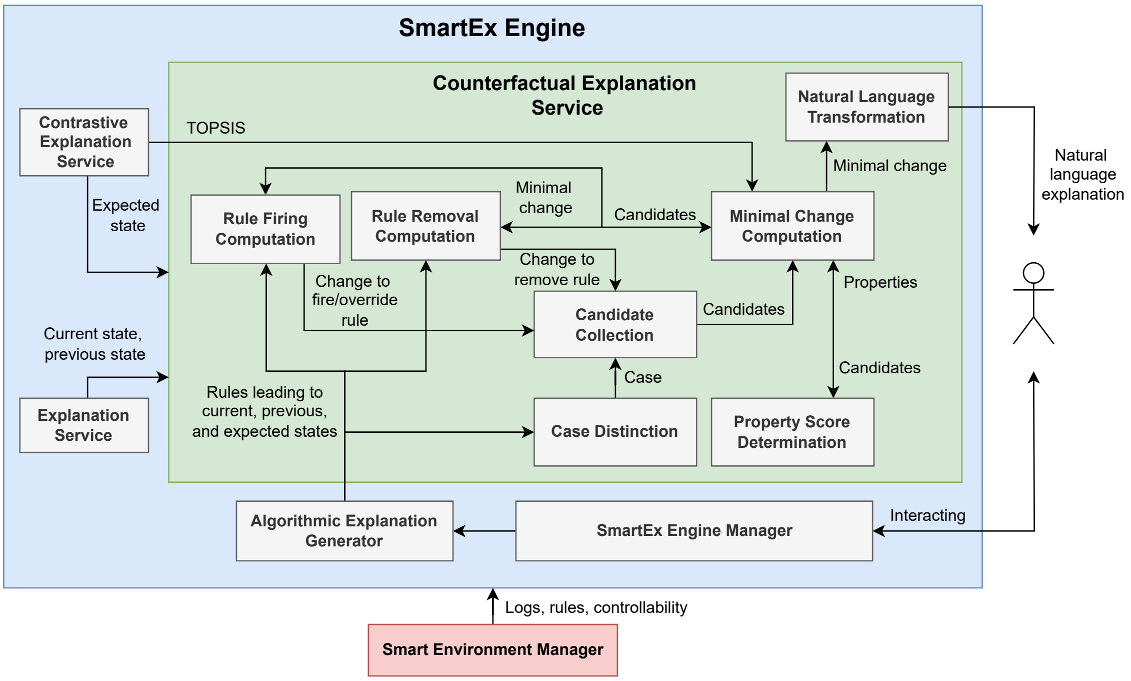

SmartEx Engine
Counterfactual Explanation Service
Contrastive Explanation Service
TOPSIS
Natural Language Transformation
Natural language explanation
Minimal change
Expected state
Rule Firing Computation
Rule Removal Computation
Minimal change
Candidates
Minimal Change Computation
Change to remove rule
Change to fire/override rule
Candidate Collection
Candidates
Properties
Current state, previous state
Case
Candidates
Explanation Service
Rules leading to current, previous, and expected states
Case Distinction
Property Score Determination
Algorithmic Explanation Generator
SmartEx Engine Manager
Interacting
Logs, rules, controllability
Smart Environment Manager

Figure 2: Reference architecture of *SmartEx* and the *Counterfactual Explanation Service*

ing the three relevant states (current, previous, and expected) and identifying the rules with true preconditions that lead to these states. The *Contrastive Explanation Service* is utilized to determine the foil from which the expected state is extracted. The current and previous states are provided by the *Explanation Service*, which retrieves them from the Home Assistant Logs provided by the *Smart Environment Manager*. The three states are further used in almost all components of the *Counterfactual Explanation Service*. The *Algorithmic Explanation Generator* is capable of determining the rules that lead to a particular state. It is equipped with this capability, as it allows the determination of all potential rules that may have fired, which is necessary for generating causal explanations as they include the preconditions of the fired rule if such a rule exists. Therefore, the *Algorithmic Explanation Generator* is used to determine the rules leading to the three states and passes them to the *Case Distinction* component. Subsequently, the rules without true preconditions are excluded. Finally, the remaining rules and the three identified states are used by the *Case Distinction* component to determine the appropriate case of explanation need, as described in Section 3.2.

The appropriate case is provided to the *Candidate Collection* component. Based on the case, it determines which rules need to be overridden or removed and whether another rule must be fired, as shown in Figure 1. The computation of the change to fire or remove a rule is done separately from the *Candidate Collection* component.

To fire a rule, the *Rule Firing Computation* component uses the *Algorithmic Explanation Generator* to determine all rules leading to the expected state. If another rule must be overridden, only the rules with higher priority are considered. Then, for each rule, the minimal change required to make the rule fire is computed, as described in the subsection regarding additive counterfactual explanations in Section 3.3. Additionally, the *Rule Firing Computation* component considers whether it is more minimal to fire another rule that causes a precondition of the current rule to change its state rather than changing the state directly. Finally, the rule requiring the minimal change to fire is determined. This is achieved by passing the necessary changes for each rule to the *Minimal Change Computation* component. The minimal candidate is then sent back to the *Rule Firing Computation* component, which forwards it to the *Candidate Collection* component.

To remove a rule, the *Rule Removal Computation* component identifies all preconditions of the rule that must be removed, as outlined in the subsection regarding subtractive counterfactual explanations in Section 3.3. Then, the *Smart Environment Manager* is used to assess the controllability of the change required to make each precondition false. If any change is actionable, all immutable ones are excluded. For each remaining precondition of the rule that must be removed, the component evaluates whether there are rules with true preconditions that enforce the state of the precondition. Such rules make it impossible to alter the precondition directly, as their actions would immediately restore the original state. These rules are subsequently also removed. Next, the component determines whether firing an alternative rule would result in a smaller change than directly manipulating the precondition. Finally, a minimal change computation across all preconditions of the rule is performed. Again, the *Minimal Change Computation* component is used, and the resulting change is provided to the *Candidate Collection* component.

When the *Candidate Collection* component receives the appropriate changes to fire/override or remove the rules, it creates combinations of how rules can be overridden or removed while reducing the number of combinations by using the approach outlined at the beginning of Section 3.3. Each combination forms one candidate for the final minimal change computation and is provided to the *Minimal Change Computation* component.

The *Minimal Change Computation* component removes any duplicates and excludes all candidates that are not actionable if there are fully actionable ones. The controllability of each candidate is determined by using the *Smart Environment*

Manager. The remaining candidates are provided to the *Property Score Determination* component, where the scores for sparsity, temporality, proximity, and abnormality are determined.

The *Property Score Determination* component computes the scores separately for each candidate and for each property. The sparsity of a candidate is determined by the size of the candidate set. The temporality of the candidate is calculated as the average across the temporality scores of the changes in the candidate set. For each change, the temporality score is determined as the difference in seconds from the point where the explanandum occurred to the point where the state mentioned in the change last turned true. The proximity of a candidate is determined as the number of resulting changes to the system if all additive changes are implemented and all subtractive changes are removed. Here, rules that may fire if the changes are implemented and how their actions influence the system further are considered. The abnormality of a candidate is determined as the average across the abnormality scores of the changes in the candidate set. For additive changes, the abnormality scores of all states of the device other than the one contained in the change are determined and summed up. For subtractive changes, only the state mentioned in the change is considered. The abnormality score of a change is calculated as the percentage of occurrences in which the state matched the current state relative to all states of the device of interest. After all property scores are determined, they are provided to the *Minimal Change Computation* component.

The determined scores are used as input into TOPSIS, which is provided by the *Contrastive Explanation Service*. The *Minimal Change Computation* component determines the minimal candidate and the corresponding minimal change is sent to the *Natural Language Transformation* component. It utilizes a pattern to create the explanation, which is then issued to the user.

4 Evaluation

To review how well our generated counterfactual explanations are received in practice, we conducted a quantitative human-centered evaluation as suggested by Vilone and Longo (2021). In the following, we discuss the study design and present the results, which form the basis of a further discussion in Section 5. To clearly define the scope of our evaluation, we present two research questions:

RQ1: Do users prefer counterfactual or causal explanations in smart environments?

RQ2: In which contexts do users prefer counterfactual or causal explanations in smart environments?

Here, causal explanations refer to the most commonly used explanations in smart environments (Sadeghi et al., 2024). A further definition is given in Section 4.1.3.

4.1 Study Design

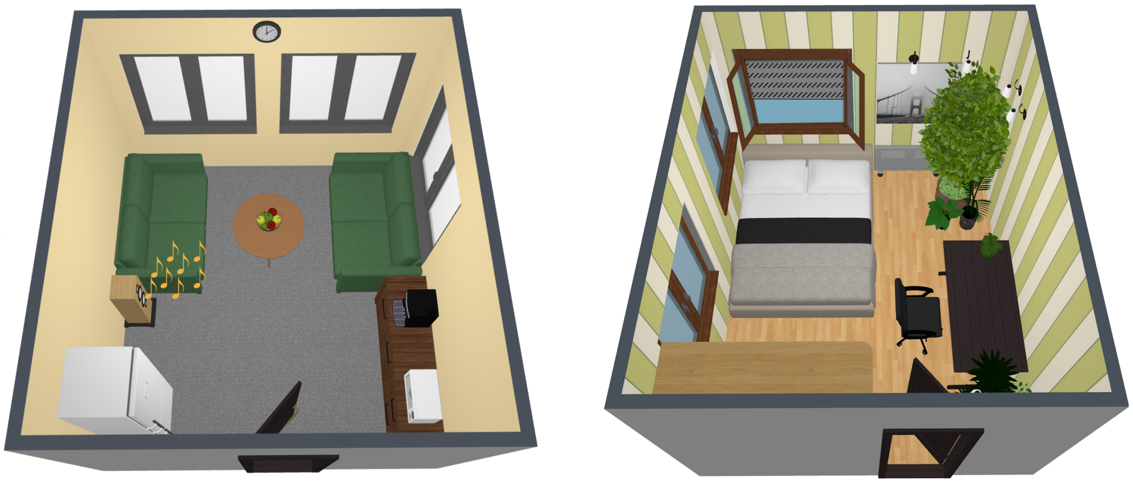
The study was conducted as an in-person interview and followed a within-subject design. Participants were presented with six scenes depicting confusing situations caused by automation in smart environments and received multiple explanations for them. As we followed a within-subject design, all participants were exposed to the same explanations. The scenes were experienced through a series of slides shown during the interview. Furthermore, all participants went through the same set of scenes in the same order.

4.1.1 Sampling and Participants

A total of 17 participants were recruited through personal contacts, though no demographic data was collected. In addition, no exclusion criterion was applied.

4.1.2 Study Format

Subjects did not experience a real smart environment but were presented with slides containing confusing situations in smart environments. Due to time and cost restraints, we refrained from placing subjects in real smart environments. Additionally, unlike in a long-term study where participants live in a smart environment, we could ensure that all participants experienced the same situations, making comparison easier. Moreover, we decided on an interview format as they are commonly used



(a) Social room with music playing (b) Bob's room with blinds down halfway

Figure 3: Reference images for participants

in XAI system evaluations (Lopes et al., 2022), and allowed us to clarify uncertainties during the study. To strengthen the mental model of the smart environment, the slides were enhanced via reference images, as shown in Figure 3. Additionally, animations and sounds were added to some images to emphasize the confusing situation.

4.1.3 Experiment Design

As we employed a within-subject design, participants received both causal and counterfactual explanations during the study. The causal explanations served as the control group as they are commonly used in smart environments (Sadeghi et al., 2024). They were determined by the *Algorithmic Explanation Generator* from *SmartEx* and are structured as follows:

$$X \text{ happened because } \textit{preconditions of the fired rule} \text{ are true.} \tag{6}$$

$$X \text{ remains } \textit{current state} \text{ because no rule was executed.} \tag{7}$$

Explanation (7) was shown in the case where no rule was fired, which corresponds to explanation need case 2.1.1 (Sadeghi et al., 2024). In contrast, the counterfactual explanations were generated using the plugin for *SmartEx* presented in Section 3.7. A complete list of all provided explanations that were provided during the study is displayed in Table 2. Finally, before experiencing each scenario, participants were asserted that all explanations were correct to ensure no influence on the rating of the explanation.

The counterfactual plugin utilizes the foil determination capabilities developed

Table 2: Explanations provided to participants

Scene	Exp. Type	Explanation
1	Causal	The speaker is on because no meeting is going on in a meeting room, and the social room is not empty.
	Counterfactual	The speaker would be off if there was a meeting going on in a meeting room.
2	Causal	The meeting room door is locked because it is before 8:30 am.
	Counterfactual	The meeting room door would be open if it was not before 8:30 am.
3	Causal	The brightness is at 70 % because there is only a single person in the room.
	Counterfactual	The brightness would be at 100 % if a device was connected to the beamer.
4	Causal	The speaker remains off because no rule was executed.
	Counterfactual	The speaker would be on if there was no meeting going on.
5	Causal	The air conditioning is on because it is sunny and all windows are closed.
	Counterfactual	The air conditioning would be off if the door was open longer than 10 min and not all windows were closed.
6	Causal	The blinds are rolled down halfway because the blind's controller down button was pressed twice, and the plant lights are off.
	Counterfactual	The blinds would be rolled down completely if the plant lights were not off.

by Herbold et al. (2024), but to isolate the impact of our proposed method for generating counterfactual explanations, we excluded it from our evaluation. To do so, we limited our evaluation to devices with binary states, such as lamps that can be either *on* or *off*. In these cases, the foil was defined as the device state opposite to the current one, making the foil determination component unnecessary for this study.

Across the six scenes, each of the three cases of explanation needs was represented twice, and sub-cases 1, 2.1.1, 2.2, 3.1.1, and 3.2.1 were incorporated. To address varying levels of urgency, for each case, one scene where participants were under time pressure and one where they were not was included. To account for multiple types of smart environments, the scenes were spread out over two scenarios: a smart office and a smart home. Furthermore, the causal and counterfactual explanations differed to various extents across the scenes. In scene 2, the differences were purely linguistic, while in scenes 1 and 6, the counterfactual explanations included fewer details than the causal ones. Moreover, in scenes 3 and 5, the counterfactual explanations were actionable, whereas the causal ones were not. Finally, in scene 4, the causal explanation issued explanation (7), while the counterfactual one provided a solution on how the confusing situation could be resolved. Further details regarding the explanations for each scene are presented in Section 4.1.5.

To evaluate RQ1, participants ranked their preferences for the two explanation types (causal and counterfactual) and the option of receiving no explanation after experiencing the confusing situation in each scene. Furthermore, a final questionnaire contained four 5-point Likert-scale questions to determine participants' agreement with statements about the content and linguistics of each of the two explanation types. This approach isolated participants' opinions of the actual content of counterfactual explanations generated by our framework. The decision to separate content from linguistic evaluation stemmed from feedback in a pilot study, where multiple participants highlighted that the linguistic phrasing of counterfactual explanations was difficult to understand.

Furthermore, we collected additional contexts to answer RQ2. Before starting the main study involving the six scenes, participants were asked about their general preferences regarding explanation length (shorter with fewer details or longer with more details) and objective (reason or solution). The question regarding explanation length was motivated by the design of counterfactual explanations, which often excludes details irrelevant to changing the situation. Similarly, preferences regarding explanation objectives were measured, as counterfactual explanations aim to provide solutions to resolve the confusing situation (Wachter et al., 2017), while causal explanations provide reasons by listing the preconditions of the fired rule. Moreover, after each confusing situation and before receiving the possible explanations, participants were asked whether they wanted to receive an explanation. This question allows us to analyze the impact of the need for an explanation on the specific rankings participants assigned to the provided explanation types.

4.1.4 Procedure

The interview starts with a welcoming of the participant, followed by an introduction and an overview of the study. Then, participants are given information regarding data privacy and are asked two preliminary questions:

1. What do you prefer?

    (i) Shorter but less detailed explanations.

(ii) Longer but more detailed explanations.

2. What is more important in an explanation?

    (i) Providing a reason for something that happened.

    (ii) Providing a solution for changing something that happened.

Subsequently, the main section of the study is started. Subjects experience two scenarios containing four and two scenes, respectively. In each scene, participants are introduced to the setting and undergo a confusing situation. Afterward, they are asked whether they would like to receive an explanation for the confusing situation and can choose to answer (1) *yes*, (2) *I don't care*, or (3) *no.* Regardless of their answer, they are provided with three paper snippets containing a causal and a counterfactual explanation as well as a snippet saying *no explanation.* Using the snippets, participants are asked to provide a ranking based on their preferences. All provided explanations are listed in Table 2.

After completing all six scenes, participants are introduced to the concepts of causal and counterfactual explanations. Causal explanations are framed as explanations of the form *X happened because...*, while counterfactual explanations as explanations of the form *Y would have happened if....* Participants are then provided with two reference lists: one containing all causal and another containing all counterfactual explanations participants received during the study. These lists can then be used for reference during the final questionnaire. There, they are asked to rank causal and counterfactual explanations separately on a 5-point Likert scale regarding their agreement (1 = strongly disagree, 5 = strongly agree) to the following two sentences:

(i) I liked the explanations linguistically.

(ii) I liked the explanations content-wise.

This concludes the study, and participants are thanked for their participation. In summary, the study takes, on average, 15 minutes.

4.1.5 Scenes

In the following, we provide an overview of the six scenes participants experienced, identify the according case of explanation need as described in Section 3.2, and elaborate on how the explanations were determined.

First Scenario The first scenario is set in an office. It is the participant's first day working there, and they must hold a presentation at 9:00 am. This setting enables us to introduce new rules to the participant during the following scenes as they are placed in an unfamiliar environment. Additionally, the impending presentation puts the participant under time pressure in the first three scenes. In the beginning, the participant receives an introduction to the smart office's capabilities, including the provision of explanations. Furthermore, they are asserted that all provided explanations are correct and are informed about one rule in the smart office:

r_1: If it is one hour before a meeting starts, open the social and meeting room doors.

As the participant is told that they arrive at 8:00 am, they assume the social and meeting room doors to be open.

The first scene takes place in the social room, where the participant wants to get a coffee. The confusing situation occurs when, as soon as they enter, the speaker turns on and starts playing music. Along with a reference image of the social room, the confusing situation is accentuated by music and animation of music notes, as shown in Figure 3a. The confusing situation corresponds to the explanation need case 1, as something unexpected occurs. It was caused by the following rule, whose existence is unknown to the participant:

r_2: If no meetings are going on, and the room is not empty, turn on the speaker.

The plugin described in Section 3.7 identifies that the minimal change required to turn off the music is to remove rule r_2 as no rule can override it. Consequently, the counterfactual and causal explanations share similar content, as both reference the preconditions of r_2. However, the counterfactual explanation contains only the precondition relevant to the minimal change, while the causal explanation lists all of them, as shown in Table 2.

Scene 2 takes place when the participant is going to the meeting room. Because of rule r_1, they expect the meeting room door to be open. This assumption is reinforced by scene 1, where the social room door was opened by r_1. However, in scene 2, the meeting room is locked, creating a confusing situation. The locked door is emphasized by a rattling sound and an animation of a moving door handle. The confusing situation was caused by the following rule, which overrides r_1 as it has higher priority and true preconditions:

r_3: If it is before 8:30 am, close the meeting room door.

The need for an explanation in this scene corresponds to case 2.2 as there are rules with true preconditions leading to the current (r_3) and expected state (r_1). The determined counterfactual explanation removes r_3 as no rule can override it. As this rule only has one precondition, the causal and counterfactual explanations contain the same content and only differ linguistically, as shown in Table 2.

Next, the third scene takes place after 8:30 am in the meeting room. The participant must hold the presentation shortly but finds the room to be too dark. Using a remote, they turn the brightness up from 50 % to 100 %. However, the brightness is automatically set back to 70 %. An image of the remote is provided to the user, and an animation setting back the number on the remote is added to reinforce the mental model. The scene covers case 3.1.1 as the confusing situation arises due to something other than expected happening and there only being a single rule with true preconditions leading to the current state:

r_4: If there is only a single person in the room, keep the brightness of the light below 70 %.

Therefore, the causal explanation contains its preconditions, whereas the counterfactual one considers another rule with higher priority that could be fired:

r_5: If there is a device connected to the beamer, turn the brightness to 100 %.

Unlike removing r_4, firing r_5 is actionable. Hence, it is determined as the minimal change. Thus, the counterfactual explanation contains the false precondition of rule r_5, and the causal and counterfactual explanations differ starkly, as shown in Table 2. Additionally, the counterfactual explanation is, as opposed to the causal one, actionable.

Scene 4 takes place in the social room again, but the participant is not under time pressure anymore since their presentation is over. The confusing situation arises because, unlike in the first scene, the speaker does not turn on as expected. Participants thus experience case 2.1.1 where unexpectedly no event occurs. This is due to the preconditions of r_2 no longer being met as a meeting is taking place in another meeting room that the participant does not know of. Therefore, the counterfactual explanation identifies the minimal change required to make r_1 fire. In contrast, as no rule was fired, the causal explanation issues explanation (7), which is its standard explanation for this case.

Second Scenario The second scenario takes place at home after the participant's first day at the office. They share their flat with a roommate called Bob, who is on vacation and asked the participant to look after his plants. Again, an introduction to smart devices, rules, and explanations is given, as the flat is a smart home. Once more, participants are told that all explanations are correct and are informed about two rules in the smart home:

r_6: If a room has been empty for more than 12 hours, turn off the air conditioning.

r_7: If the blind's controller down button is pressed twice, roll down the blinds.

Scene 5 takes place in Bob's room as he has been on vacation for two days, and the participant wants to water his plants. When entering the room, the participant is informed that the air conditioning is on. A reference image of Bob's room and an animation portraying wind coming out of the air conditioning are provided to enhance immersion. Furthermore, the participant is told to feel annoyed due to concerns about increasing electricity bills. The confusing situation arises as the participant knows that rule r_6 fired. Therefore, they expect the air conditioning to still be turned off. The confusing situation arises due to the following rule that has true preconditions and a higher priority than rule r_6:

r_8: If it is sunny and all windows are closed, turn on the air conditioning.

This scene aligns with case 1 again as something unexpected happened, though without time pressure, as the participant is at home. The causal explanation contains the preconditions of r_8, whereas the counterfactual one considers a third rule:

r_9: If the door is open for more than 10 min, turn off the air conditioning.

Since r_9 has a lower priority than r_8, it is not enough to just fire r_9. The framework determines the minimal change to resolve the confusing situation as firing r_9 and removing r_8, thereby providing an actionable counterfactual explanation.

In the final scene, the participant is still in Bob's room. The instructions from the issued counterfactual explanation were followed as the door, and a window were opened to turn off the air conditioning. Now, the participant is told to be worried that Bob's room may get too hot. Therefore, they want to roll down the blinds by utilizing rule r_7 presented to them in the introduction of the second scenario. However, when pressing the blind's controller down button twice, as suggested by r_7, the blinds only roll down halfway. This surprising situation is again enhanced by a reference image of Bob's room as well as animation and the sound of blinds rolling down, as displayed in Figure 3b. Here, the rule that overrode r_7 to ensure that the plants get enough light is:

r_{10}: If the blind's controller down button is pressed twice and the plant lights are turned off, roll down the blinds halfway.

Thus, the scene covers case 3.2.1. The causal explanation contains all preconditions of r_{10}, whereas the counterfactual one wants to remove the rule, therefore only containing one of its preconditions.

4.2 Results

In this section, we present the results of our conducted study, which are the basis of a further discussion in Section 5. As the number of participants was relatively low, we performed no statistical analysis.

As shown in Figure 4, participants generally preferred to receive an explanation. This preference was strongly pronounced in all but the first and fourth scenes, as at least 65 % of the participants expressed a desire for an explanation. Notably, in scene 5, all 17 participants, without exceptions, expressed the need for an explanation. In contrast, participants' opinions were more divided in scene 4. While a majority still preferred to receive an explanation, a notable proportion showed no preference. Furthermore, in scene 1, participants predominantly preferred not to receive an explanation, although this preference was not as strong compared to the other scenes.

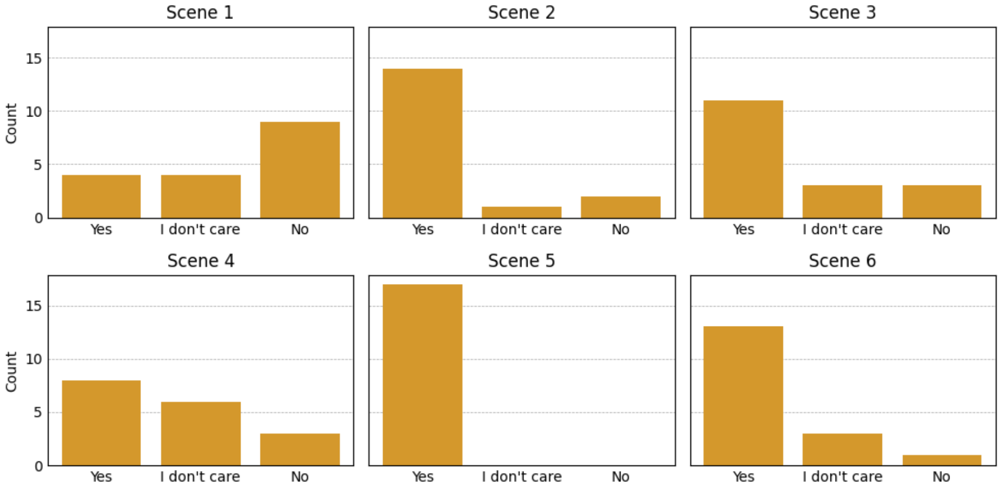


Figure 4: Explanation needs per scene

As shown in Table 3, participants expressed a strong preference for causal explanations over counterfactual ones linguistically. While counterfactual explanations were rated averagely on a 5-point Likert scale, causal explanations were rated significantly higher. In contrast, participants rated counterfactual explanations more favorably content-wise. However, the ratings concerned with the content of the explanations were not as extreme, and the difference between the ratings was significantly smaller. Finally, standard deviations were relatively high, indicating variability in participants' responses.

Table 3: Explanation types rated linguistically and content-wise

Explanation Type	Criterion	Mean	Std
Causal	Linguistically	4.235	0.752
Counterfactual	Linguistically	2.941	1.029
Causal	Content-wise	3.529	1.179
Counterfactual	Content-wise	3.647	0.996

During the main part of the study, causal explanations were slightly favored over counterfactual ones, as they were ranked first more often, as shown in Figure 5. However, causal explanations were also ranked last more often than counterfactual ones, though the difference in rankings was relatively small. Receiving no explanations was rarely chosen first and most often ranked last, indicating an overall desire for some form of explanation. Causal explanations were predominantly preferred in the first two scenes, where they were almost always ranked first. In contrast, counterfactual explanations were more frequently ranked first in the subsequent four scenes. Notably, a majority of the causal explanations that were ranked last came from the fourth scene, whereas for the counterfactual explanations, a majority belonged to the first scene. Finally, receiving no explanation was mostly ranked last across all scenes except for scenes 1 and 4.

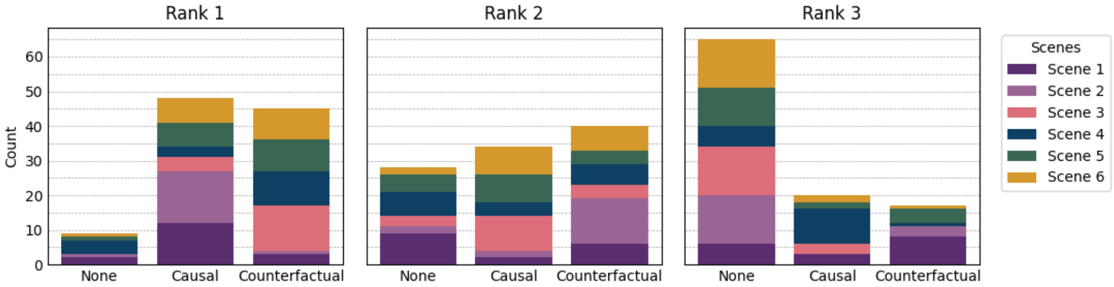
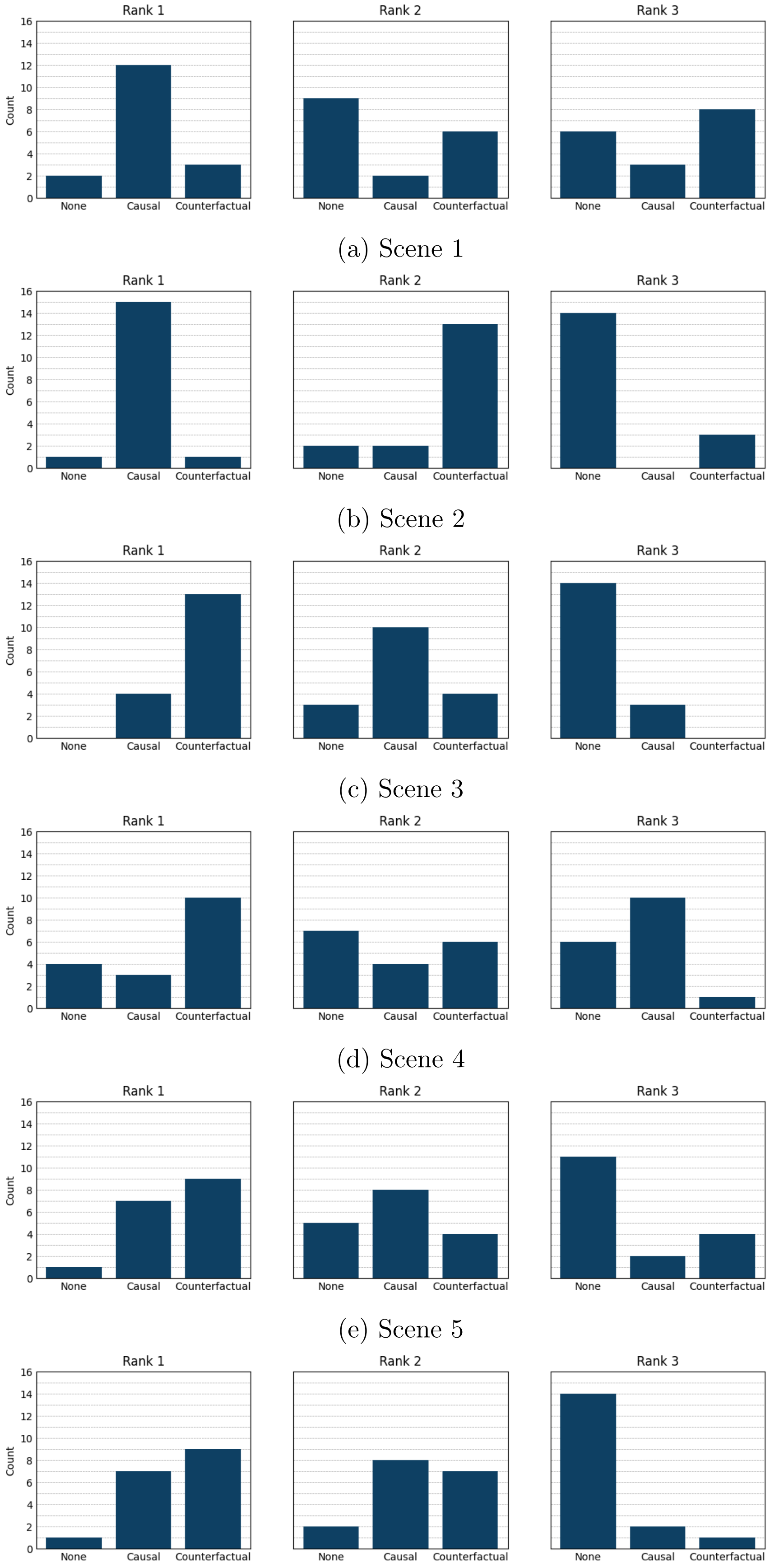


Figure 5: Combined rankings of no, causal and counterfactual explanations

When considering a distinction of the scenes, as done in Figure 6, further insides can be gained. In scene 1, causal explanations were preferred over counterfactual ones and over receiving no explanation, as they were ranked first most often. Receiving no explanation was selected second most often, while counterfactual explanations were predominantly ranked last. However, while causal explanations were strongly preferred over the other two options, the difference in rankings between counterfactual explanations and receiving no explanation was not as significant.

In scene 2, the ranking was very strongly pronounced. Causal explanations



(a) Scene 1

(b) Scene 2

(c) Scene 3

(d) Scene 4

(e) Scene 5

(f) Scene 6

Figure 6: Rankings of no, causal, and counterfactual explanations per scene

were ranked first by all but two participants, while counterfactual explanations were ranked second, except for four cases. Apart from three exceptions, receiving no explanation was always ranked last.

Contrarily, in the third scene, counterfactual explanations were strongly preferred over causal ones as they were ranked first by the majority of participants. However, causal explanations were strongly favored over not receiving any explanation, as they were most frequently ranked second. Receiving no explanation was ranked third by most participants.

In the fourth scene, counterfactual explanations were again strongly favored, as they were ranked first significantly more often than receiving no explanation or a causal one. Additionally, getting no explanation was chosen first and second marginally more frequently than getting a causal explanation. Causal explanations were ranked last most frequently. Hence, receiving no explanation was slightly favored over receiving a causal one.

In the fifth scene, subjects again favored counterfactual explanations, ranking them first more often than causal ones. However, the difference in rankings between causal and counterfactual explanations was less distinct compared to previous scenes. Receiving no explanation was preferred the least, as it was ranked last most often. Interestingly, although counterfactual explanations were ranked first more frequently than causal ones, they were also ranked last more often, suggesting a divisive opinion of counterfactual explanations among participants.

Finally, in scene 6, counterfactual explanations were again slightly preferred over causal ones, as they were ranked first marginally more often. Contrarily, causal explanations were ranked second and third minimally more often. Not receiving any explanation was the least preferred option, as it was predominantly ranked last.

In summary, causal explanations were primarily favored in the first two scenes. The difference was very pronounced as receiving no or a counterfactual explanation was rarely ranked first. In contrast, counterfactual explanations were preferred in the subsequent four scenes, with particularly strong preferences in scenes 3 and 4. The strongest consensus for a distinct rating occurred in scene 2, where at least than 12 out of 17 participants had the same preference for each rank. Across all scenes, receiving no explanation was generally not preferred, as it was scarcely ranked first and chosen last in four out of six scenes. In the first scene, receiving a counterfactual explanation was least preferred, while in the fourth scene, a causal explanation was most frequently ranked last.

Apart from analyzing preferences on a scene-by-scene basis, we also investigate the participants' rankings based on their preferences regarding explanation length and objective. We divide participants into two groups for each criterion and analyze the rankings for each group. The results are presented in Figure 7. Overall, 12 out of 17 participants expressed a preference for shorter over longer explanations, and 9 out of 17 participants indicated a preference for explanations providing a solution over a reason.

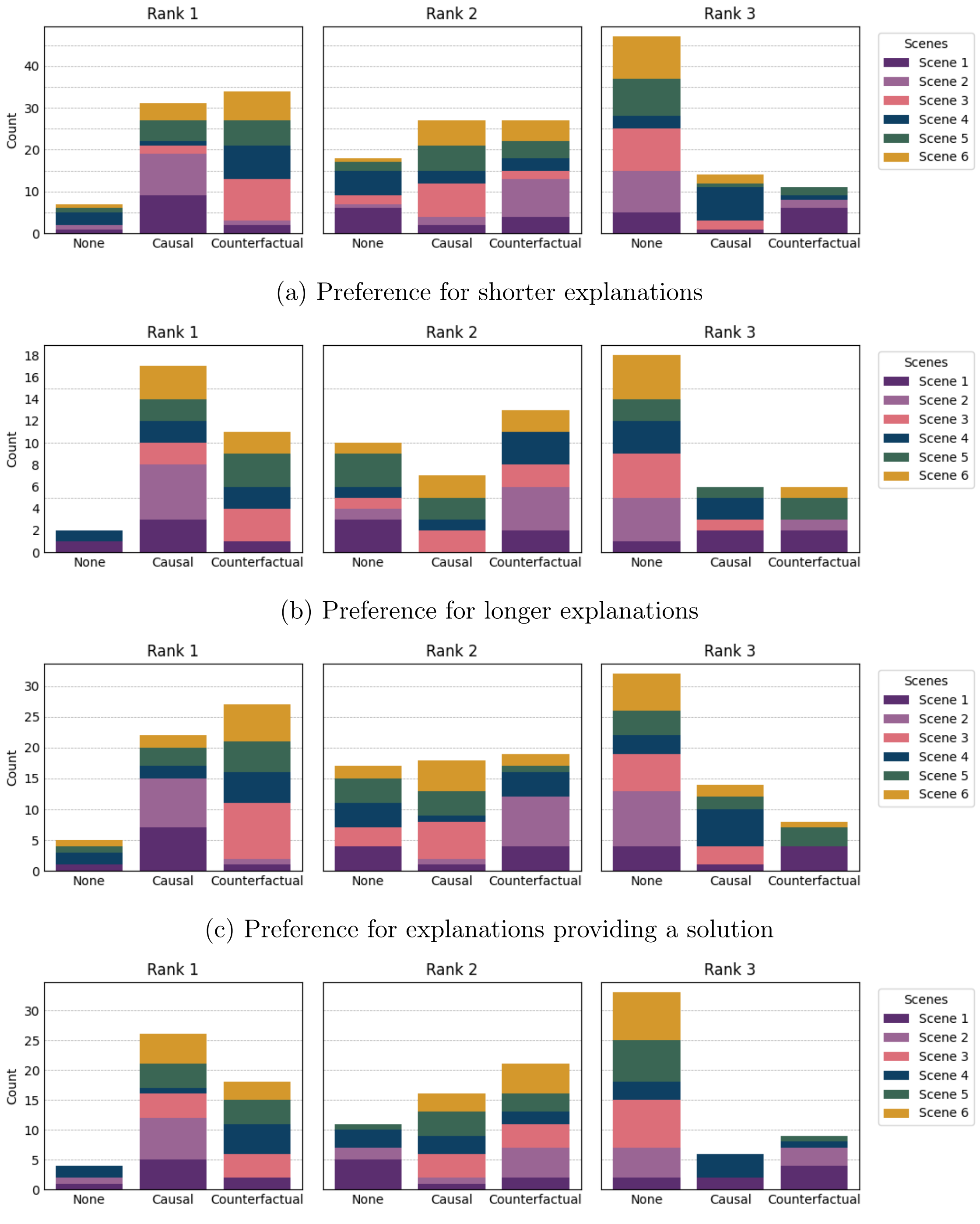
We begin by investigating the rankings of participants preferring shorter explanations, as displayed in Figure 7a. Among this group, counterfactual explanations were ranked first slightly more often than causal ones. Receiving no explanation was generally not preferred, as it was ranked third most often. When analyzing individual scenes, causal explanations were strongly favored over counterfactual ones in the first two scenes. However, in the subsequent four scenes, counterfactual explanations were preferred over causal ones, with the preference being particularly pronounced in scenes 3 and 4.

In contrast, the five participants preferring longer explanations generally strongly favored causal explanations over counterfactual ones, as they were ranked first most often, as shown in Figure 7b. Again, receiving no explanation was the least preferred option, as it was ranked last most frequently. When analyzing the scenes, causal explanations were favored over counterfactual ones in the first two scenes. However, in the remaining four scenes, no clear preference emerged as, due to the small number of participants in this group, no tendency of more than one participant was observed.

Among participants with a preference for explanations providing a solution rather than a reason, counterfactual explanations were slightly favored over causal ones as they were ranked first more often, as illustrated in Figure 7c. Receiving no explanation was the least preferred option, as it was ranked last by a majority of participants. Causal explanations were strongly preferred in the first two scenes, as most participants in this group ranked them first. In contrast, counterfactual explanations were favored in the subsequent four scenes, particularly in scene 3, where all participants ranked counterfactual explanations first. But counterfactual explanations were also strongly favored in scenes 4 and 6.

On the other hand, among participants with a preference for explanations providing reasons, causal explanations were strongly preferred as they were ranked first most often, as displayed in Figure 7d. Receiving no explanation was not favored as it was ranked last by most participants. When analyzing the scenes, causal explanations were preferred in scenes 1, 2, and 6, while counterfactual ones were favored in scene 4. In scenes 3 and 5, first-place rankings for the two explanation types were relatively evenly distributed, indicating no clear preference for either explanation type.

In summary, counterfactual explanations were preferred among participants, with a preference for shorter explanations and explanations providing a solution. Contrarily, causal explanations were preferred among participants, with a prefer-



(a) Preference for shorter explanations

(b) Preference for longer explanations

(c) Preference for explanations providing a solution

(d) Preference for explanations providing a reason

Figure 7: Combined rankings of no, causal, and counterfactual explanations by participants with different preferences

ence for longer explanations and explanations providing a reason. Again, receiving no explanation was preferred by neither of the groups. When analyzing each scene, causal explanations were preferred by all groups in the first two scenes. In scene 3, counterfactual explanations were favored by participants, preferring shorter explanations and explanations providing a solution. No tendency could be observed in participants favoring longer explanations and explanations providing a reason. In the fourth scene, counterfactual explanations were favored by all groups except for the group containing participants who preferred longer explanations. In this group, no tendency could be observed. In addition, the preference for counterfactual explanations was more pronounced in the groups favoring shorter explanations and explanations providing a solution. In scene 5, counterfactual explanations were marginally preferred by participants, favoring shorter explanations and explanations providing solutions. In contrast, participants preferring longer explanations or explanations providing reasons favored neither of the two explanation types. In the last scene, counterfactual explanations were preferred among subjects, with a preference for shorter explanations and explanations providing a solution. However, the difference was more significant in participants favoring explanations that provide a solution. Causal explanations were preferred by participants, favoring explanations that provide a reason, while no tendency could be observed for participants preferring longer explanations.

Finally, to conduct an analysis per participant, we present Figure 8, which displays the amount of no, causal, and counterfactual explanations ranked first by each participant. Among the 17 participants, five showed a preference for causal explanations, while four favored counterfactual ones. Receiving no explanation was not favored by any participant, as the remaining eight participants showed no clear preference towards any option.

Participants indicating a preference for shorter explanations with fewer details over longer ones with more details preferred counterfactual explanations in three cases, while causal ones were favored by two participants. The remaining six participants showed no clear preference. Conversely, among participants indicating a preference for longer explanations, causal explanations were preferred by three out of five participants. One participant preferred counterfactual ones, while another showed no preference.

Participants preferring explanations that provide a solution over a reason favored both causal and counterfactual explanations in two cases. The remaining five participants showed no preference for any option. In contrast, causal explanations were favored by three participants who indicated a preference for explanations providing a reason. Counterfactual explanations were preferred by two participants, while

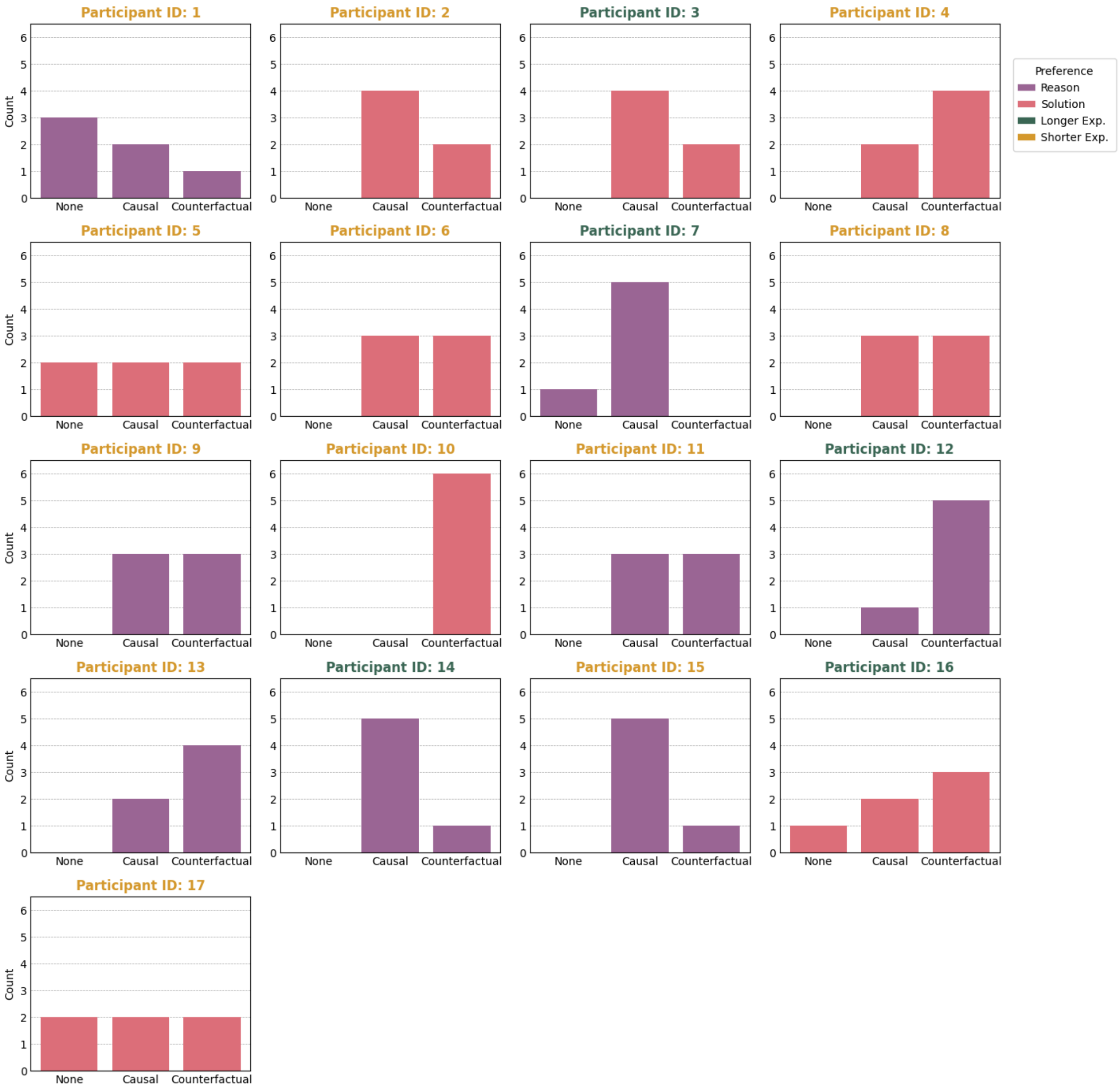


Figure 8: No, causal, and counterfactual explanations ranked first per participant

three participants had no preference.

In conclusion, the preference for no, causal, or counterfactual explanations varied significantly between the participants. Five participants rated one explanation type first in at least five out of six scenes, indicating a strong preference for one explanation type regardless of the scene. In contrast, about half of the participants (8 out of 17) had no preference. Additionally, among participants with either preference for an explanation objective, no explanation type was strongly preferred. Participants with a preference for longer explanations favored causal explanations, though no preference could be observed among participants favoring shorter explanations.

5 Discussion

In the following, we discuss the results of our evaluation by answering the research questions proposed in Section 4 and debate the threats to validity.

5.1 General Preferences

Regarding RQ1, we found that users, in general, do not prefer one explanation type over the other. Summarized across all scenes, causal explanations were preferred over counterfactual ones because they were ranked first slightly more often, as displayed in Figure 5. However, causal explanations were also ranked last somewhat more often, indicating a divisive opinion. Most first-place rankings for causal explanations stemmed from the first two scenes. There, causal explanations were strongly favored over counterfactual ones. However, counterfactual explanations were preferred in the remaining four scenes, even though in scenes 5 and 6, the difference among preferences was not as distinct. Receiving no explanation was usually the least preferred option except for scenes 1 and 4, where counterfactual and causal explanations were preferred the least, respectively.

When analyzing individual participants' preferences, no general preference for one explanation type could be determined, as shown in Figure 8. While approximately the same amount of participants displayed a strong preference for either causal or counterfactual explanations, about half of the participants indicated no distinct preference for one explanation type.

Additionally, participants were asked to rank the explanation types based on their agreement to liking them both linguistically and content-wise. Causal explanations were strongly preferred linguistically over counterfactual ones, while counterfactual explanations were rated slightly better than causal ones in terms of content. These results indicate a need for linguistic improvements to counterfactual explanations. Although the use of complex tenses in counterfactual explanations is necessary to convey the minimal change required to change the outcome, we propose the use of a large language model to improve their linguistic clarity. By doing so, counterfactual explanations may show the potential to be consistently rated higher than causal explanations, as the difference in linguistic ratings was significant, while the difference in rankings during the main study was relatively small.

In conclusion, no generalization regarding the overall preference for either explanation type can be made. Therefore, we proceed with a discussion of RQ2, where we analyze the contexts influencing users' preferences for an explanation type.

5.2 Contexts Influencing the Preferences

To answer RQ2, we analyze the impact of several user-centric, situational, and explanation-specific contexts on users' preferences for causal and counterfactual explanations. Each context is analyzed separately before conclusions about which contexts influence the preference for each explanation type are drawn. Finally, based on the contexts, we offer speculations to explain the rankings observed in each scene of our study.

5.2.1 User-Centric Contexts

Firstly, we analyze the effect of user preferences on the preference for an explanation type to provide insight into which users may benefit most from which type of explanation. Before the main study, participants were asked about their preferences regarding explanation length and objective. A majority preferred shorter explanations with fewer details over longer ones with more details. In contrast, participants' preferences on the explanation objective (solution vs. reason) were evenly divided. As shown in Figure 6, counterfactual explanations were slightly preferred over causal ones among participants with a preference for shorter explanations and explanations providing a solution. In contrast, participants with a preference for longer explanations and explanations providing a reason favored causal explanations over counterfactual ones.

However, when looking at each participant individually, as shown in Figure 8, preferences for any explanation objective and short explanations did not influence the preference for an explanation type. In contrast, participants preferring longer explanations favored causal explanations over counterfactual ones. Though, only five participants expressed a preference for longer explanations, indicating the need for further research.

Additionally, we consider participants' preferences for each scene and each group (shorter vs. longer and solution vs. reason) separately. In the first two scenes, causal explanations were preferred by all groups. However, in the remaining four scenes, preferences were more divided. While counterfactual explanations were preferred by the groups favoring shorter explanations and explanations providing solutions, the other two groups usually had no tendency. Only participants favoring explanations that provide reasons preferred causal explanations in scene 4 and counterfactual ones in scene 6. These findings are consistent with the previous results, as participants who preferred shorter explanations or explanations offering a solution generally tended to favor counterfactual explanations slightly more in each scene than the average participant, whereas the other two groups showed a slight preference for causal explanations in each scene when compared to all participants.

In summary, counterfactual explanations are preferred among users favoring shorter explanations and explanations providing a solution, while causal ones are favored by users preferring longer explanations and explanations providing a reason. This result coincides with our expectation outlined in Section 4.1.3. By design, counterfactual explanations frequently omit details unnecessary for changing the situation. Additionally, they aim to provide a solution for a confusing situation (Wachter et al., 2017), which makes them suitable for users liking shorter explanations that provide a solution. In contrast, as causal explanations contain the rule that fired and caused the confusion, they provide a reason for the current situation. Furthermore, they mention all preconditions of this rule and, hence, include more details than counterfactual explanations, making them suitable for users liking longer explanations that provide a reason.

Additionally, we identify further situational and explanation-specific contexts that we suspect influence the preference of users for an explanation type. For each scene, the contexts are presented in Table 4. We identify the effect of each context on the preference for counterfactual and causal explanations. To do so, we determine an integer for each explanation type and context, where the larger the integer, the larger the impact of the context on the preference for the explanation type. The integer is determined as a sum over all six scenes. For each scene, we add $+1$ if the explanation type was preferred and the context applied or if the explanation type was not preferred and the context did not apply. Conversely, we add -1 if the explanation type was preferred, but the context did not apply, or the explanation type was not preferred, but the context applied. Finally, we provide comments and speculations on the reason for the determined effect.

5.2.2 Situational Contexts

We begin by examining four situational contexts: the smart environment the user is in, whether they are under time pressure, prefer to receive an explanation, or want to alter the situation. These contexts are chosen as the relevance of an explanation is highly dependent on the specific situation in which they are provided (Hanson, 1972). Additionally, we aim to provide further insights into which types of explanations should be issued in which situations.

Setting Firstly, we discuss the setting or type of smart environment the user is in. The study included two types of smart environments: a smart office and a smart

Table 4: Influencing contexts in scenes
The x symbols: The context does not apply. The ✓symbols: The context does apply. The o symbols: The context neither does nor does not apply.

Scene	Setting	Time Pressure	Explanation Need	Desire to Change	Explanation Type	Negated	Longer	Actionable	Rank
1	Office	✓	x	x	Causal	✓	✓	x	↑
					Counterfactual	x	x	x	↓
2	Office	✓	✓	✓	Causal	x	o	x	↑
					Counterfactual	✓	o	x	↓
3	Office	✓	✓	✓	Causal	x	o	x	↓
					Counterfactual	x	o	✓	↑
4	Office	x	✓	x	Causal	✓	o	x	↓
					Counterfactual	✓	o	x	↑
5	Home	x	✓	✓	Causal	x	x	x	↓
					Counterfactual	✓	✓	✓	↑
6	Home	x	✓	✓	Causal	x	✓	✓	↓
					Counterfactual	✓	x	✓	↑

home. As shown in Table 4, scenes 1 to 4 were situated in a smart office, whereas scenes 5 and 6 were set in a smart home. In the four smart office scenes, causal and counterfactual explanations were preferred on two occasions each. Conversely, in the two smart home scenes, counterfactual explanations were always preferred. Consequently, the effect of the smart office setting on the preference for causal explanations is +2, while its effect on the preference for counterfactual explanations is −2. In contrast, the smart home setting has an effect of −2 on the preference for causal explanations and an effect of +2 on the preference for counterfactual ones. In summary, the setting does impact users' preferences for an explanation type. Causal explanations are more often favored in smart offices, while counterfactual explanations are preferred in smart home settings. One participant noted that they did not care what happened in the office environment as they were not in their own space. This suggests that the preference for causal explanations in smart offices may be attributed to their straightforward nature. Meanwhile, the preference for counterfactual explanations in smart homes may reflect users' greater willingness to invest cognitive effort in comprehending more complex explanations within their personal space. This distinction in cognitive effort requirements was also pointed out by several participants and is further evidenced by the linguistic ratings, as shown in Table 3.

Time Pressure Next, we analyze if users prefer different explanations when they are under time pressure. As displayed in Table 4, we identified that participants were under time pressure in the first three scenes because they had to hold a presentation soon. In both cases where participants preferred causal explanations, they were under time pressure. In contrast, counterfactual explanations were preferred in one scene where participants were under pressure and in three scenes where they were not. Therefore, the effect of the existence of a time pressure component on the preference for causal explanations is +4. Contrarily, its effect on the preference for counterfactual explanations is −4. Hence, counterfactual explanations are less preferred in scenes where users are under time pressure, while causal ones are more preferred. As in the setting context, we hypothesize that users prefer causal explanations when they are under time pressure because they require less time and effort to comprehend than counterfactual explanations.

Explanation Need Additionally, we examine the effect of the need for an explanation on the preference for an explanation type. After experiencing the confusing situation and before receiving any explanation, participants were asked if they desired one. As shown in Figure 4, participants generally preferred to receive an explanation, except in scene 1, where most participants did not express a desire for one. Out of the five scenes where receiving an explanation was favored, causal explanations were preferred once, while counterfactual explanations were favored four times. In scene 1, where not receiving an explanation was preferred, causal explanations were favored over counterfactual ones. Therefore, the existence of an explanation need has a strongly negative effect of -4 on the preference for causal explanations and a strongly positive effect of $+4$ on the preference for counterfactual ones. We deduce that counterfactual explanations are preferred when there is a need for an explanation, while causal ones are not. However, as there is only one scene where participants generally did not want to receive an explanation, further research is required to come to a definitive conclusion.

Desire to Change Moreover, we analyze the impact of the user's desire to change the situation on the preference for an explanation type. During our study, participants encountered four scenes where they had the desire to change the situation and two where they did not. As outlined in Table 4, in scenes 1 and 4, participants did not care if the speaker was on or off and, therefore, did not have the desire to change the situation. In contrast, when the meeting room door did not open in scene 2, participants could not brighten up the room in scene 3, the air conditioning was on in scene 5, and the blinds only rolled down halfway in scene 6, participants wanted to manipulate the confusing situation to achieve a goal. Notably, in scene 6, multiple participants asked whether the blinds that were rolled down halfway were sufficient for cooling down the room. Additionally, in scene 3, some participants wanted to know whether a brightness level of 70 % was enough, suggesting that a desire to change the situation would influence their responses. In the four scenes where users desired to change the situation, counterfactual explanations were preferred on three occasions, while causal explanations were preferred once. In the two scenes where no change was desired, causal and counterfactual explanations were each favored once. Consequently, the effect of a change being desired on the preference for causal explanations is -2, while it is $+2$ for counterfactual ones. Therefore, if there is a desire to change the situation, counterfactual explanations are usually preferred, while causal explanations are not. We suspect that this is due to counterfactual explanations answering *how to* questions (Woodward, 2003), and hence enabling users to change the situation. In contrast, causal explanations provide reasons for the confusing situation, as opposed to solutions to resolve them, as they mention all preconditions of the rule that fired.

5.2.3 Explanation-Specific Contexts

Next, we analyze multiple contexts regarding the explanations themselves. We consider their length, whether users can implement their proposed changes, and if they contain negations. Additionally, we consider the structure of the counterfactual explanations (additive vs. subtractive) and whether they omit any details.

Length We investigate the effect of explanation length as the literature suggests that users generally prefer to receive short explanations (Dai et al., 2023). We consider an explanation to be longer than another one if it contains at least three more words. Hence, we identified the causal explanations in scene 1 and 6 and the counterfactual explanation in scene 5 as being longer. As the explanation length is approximately the same in the remaining scenes, they are excluded from our analysis, as outlined in Table 4. The longer causal explanations were preferred in one instance and not preferred in another, while the longer counterfactual explanation was preferred. Hence, the effect of the explanation being longer on the preference for both causal and counterfactual explanations is determined as +1, suggesting that they are preferred if they are longer. However, this contradicts current literature (Dai et al., 2023) and participants' answers regarding preferences for explanation length in the preliminary questionnaire. Moreover, the effect is only marginal, and just three scenes were included in its determination, indicating the need for further research.

Actionability Our framework excludes, wherever possible, changes that users cannot implement themselves, i.e., changes that are not actionable, as outlined in Section 3.4. Consequently, actionability plays a crucial role. To evaluate this design decision, we analyze the effect of completely actionable explanations on participants' preferences for causal and counterfactual explanations. Causal explanations were fully actionable only in scene 6, whereas counterfactual explanations were completely actionable in scenes 3, 5, and 6, as shown in Table 2. Preferences for causal explanations were determined in two instances when they were non-actionable but never when they were actionable. Out of the four instances where counterfactual explanations were preferred, they were actionable in three instances and not actionable once. Hence, the effect of actionability on the preference for causal explanations is 0, while for counterfactual explanations, the effect is +4. From this, we infer that the actionability of an explanation has no impact on the preference for causal explanations. However, as they were only actionable in one scene, further research to confirm this finding is required. In contrast, the actionability of an explanation has a significant effect on the preference for counterfactual explanations. This observation aligns with the argument by Roese (1997), who claims that counterfactual functionality is maximized in actionable situations. Additionally, it aligns with the arguments by Poyiadzi et al. (2020), who emphasize the importance of controllability in counterfactual explanations. Finally, since counterfactual explanations are designed to offer solutions for changing the current situation (Wachter et al., 2017), their actionability becomes particularly crucial, as it constitutes the core purpose of those explanations. These findings reinforce our design decision to prioritize the actionability of the generated counterfactual explanation.

Negations As we received feedback from multiple participants that the counterfactual explanations were difficult to understand due to single and double negations, we analyze their impact on the preference for each explanation type. First, we determine the effect of any negation. The causal explanations contained negations in scene 1 and 4 while the counterfactual ones contained them in scenes 2, 4, 5, and 6, as displayed in Table 2. Causal explanations were preferred once when they contained a negation and once when they did not. In contrast, counterfactual explanations were preferred in three cases when they contained a negation and once when they did not. Hence, the effect of a single negation on the preference for causal and counterfactual explanations is +2, suggesting that both explanation types are preferred if they contain negations. However, this contradicts our preliminary hypothesis.

We hypothesize that instead of single negations, users mainly dislike double negations or unnecessarily complex expressions that can be easily simplified, as a second negation considerably adds to the comprehension time (Sherman, 1976). While the causal explanations issued in our study did not contain these types of expressions, the counterfactual explanations contained them in scenes 2, 5, and 6. In scene 2, *not before 8:30 am* could be simplified to *after 8:30 am*, in scene 5, *not all windows were closed* could be simplified to *a window was opened*, and in scene 6, *the plant lights were not off*, could be simplified to *the plant lights were on.* Counterfactual explanations were preferred twice when they contained these types of expressions and twice when they did not. Additionally, they were not preferred once when they did not contain them and once when they contained them. Hence, the determined effect of these phrases on the preference for counterfactual explanations is 0, indicating no impact. However, in scene 2, we determined the biggest difference between the rankings of causal and counterfactual explanations, though the explanations only differed linguistically. Here, the double negation in the counterfactual explanation was one of the only differences, suggesting that it significantly impacted the ranking. Moreover, while counterfactual explanations were preferred in scenes 5 and 6, the preference was not as strongly pronounced, which could be explained by the difficult phrases contained in the counterfactual explanations. However, these observations remain speculative, and further research is required.

Finally, we discuss two additional contexts specific to counterfactual explanations: the explanation structure (additive vs. subtractive) and whether the counterfactual explanation omitted any details, such as additional preconditions.

Explanation Structure Instead of considering negations as a directly influencing factor, we hypothesize that the structure of the explanation affects its preference, as subtractive explanations inherently include a negation, whereas additive explanations do not. Moreover, additive and subtractive counterfactual explanations possess different benefits and, hence, should be used in different contexts (Byrne, 2019; Markman et al., 2007). Among the six provided counterfactual explanations, two were additive, three were subtractive, and one explanation contained additive and subtractive components, as explained in Section 4.1.5. One additive and one subtractive counterfactual explanation were not preferred over the causal one, whereas two subtractive ones, an additive one and the explanation containing both, were preferred. Therefore, the effect of an additive component in the explanation on the preference for counterfactual explanations is 0, while for subtractive ones, it is +2. Hence, counterfactual explanations are preferred if they contain subtractive elements. We speculate that this is due to subtractive explanations usually being the more obvious ones. They contain information that is directly related to the confusing situation, while additive explanations include new rules and, therefore, add new and not directly related information. Additionally, subtractive counterfactual explanations evoke a relational processing style (Markman et al., 2007), which may be more suitable to grasp the interactions of rules in rule-based systems.

Omission of Details As the omission of unnecessary details is an integral part of our framework, we analyze its effect on the preference for counterfactual explanations. We define an explanation as omitting information when it does not include all preconditions of the referenced rule. Notably, causal explanations did not omit details. As described in Section 4.1.5, counterfactual explanations excluded details in scenes 1, 4, 5, and 6. Among the four instances where counterfactual explanations were preferred, three excluded details, while one contained all information. Conversely, when counterfactual explanations were not preferred, they excluded details in one instance and contained all information in another. Thus, the effect of an information omission on the preference for a counterfactual explanation is +2. We conclude that users generally prefer explanations that omit unnecessary information, a finding consistent with current literature (Chazette & Schneider, 2020).

In summary, we could not determine a general preference for counterfactual or causal explanations. While causal explanations were overall ranked first more frequently, counterfactual explanations were favored in more scenes. Counterfactual explanations were criticized for their complex linguistics, though their content was slightly preferred. Additionally, participants' preferences varied considerably.

We analyzed several contexts contributing to a preference for each explanation type. Causal explanations were preferred by users, favoring longer explanations and explanations providing a reason. Moreover, they were preferred when users were in an office environment or under time pressure. In contrast, counterfactual explanations were favored by users, preferring shorter explanations and explanations providing a solution. In addition, they were preferred when users were at home, wanted to change the situation, or when the explanation was actionable. Furthermore, subtractive explanations were favored over additive ones, and counterfactual explanations that excluded details over ones that did not.

We conclude that counterfactual explanations, due to their linguistic complexity, should not be issued in scenarios where users lack the time or willingness to comprehend them. In these situations, such as when the user is under time pressure or at the office, causal explanations should be used as they are more straightforward. However, when the user is willing to invest the effort to comprehend the counterfactual explanations, such as when they are at home or want to change the situation, they should preferably be issued to the user, though user preferences should be taken into account. The provided counterfactual explanations should be concise and, where possible, actionable and subtractive rather than additive.

Considering this analysis, we suspect the high ranking of causal explanations in the first two scenes to be due to the office setting and the user being under time pressure. Additionally, the counterfactual explanations were not actionable, and in scene 1, there was no need for an explanation, while in scene 2, the only difference between the explanations was linguistically where the counterfactual explanation contained the unnecessarily complex phrase *not before 8:30 am*. Moreover, we suspect that the low ranking of counterfactual explanations in the first two scenes is due to the participants needing time to familiarize themselves with the linguistic structure of these explanations. Counterfactual explanations contain complex tenses, making them challenging to comprehend, as pointed out by several participants. However, since the order of scenes was not randomized, definite conclusions cannot be drawn.

In scenes 3 and 4, counterfactual explanations were strongly preferred. We hypothesize that this is due to the counterfactual explanation being actionable in scene 3 and change being desired, while in scene 4, the issued causal explanation was explanation (7), which provided no real insight into the situation.

We speculate that the slight preference for counterfactual explanations over causal ones in scenes 5 and 6 was due to change being desired and the counterfactual explanations being actionable. Additionally, participants were at home, not under time pressure and used to the counterfactual explanations. We suspect that they were only slightly preferred because they contained double negations.

5.3 Threats to Validity

Finally, we discuss the threats to validity of our evaluation. We differentiate between internal, external, and construct threats (Cook et al., 2002).

Internal Threats Firstly, participants were not selected independently, as they were recruited through personal contacts, potentially introducing selection bias. Additionally, a within-subject design was employed, meaning that the explanation types and the option of receiving no explanation could not be evaluated independently. Participants were asked to rank the three options, requiring that they receive all of them simultaneously. This could have influenced their ranking, as they noticed, for example, when the counterfactual explanations omitted details, leading them to believe they were wrong or incomplete.

Moreover, during the study, participants were instructed on when to feel surprised and what outcome to expect. However, as we only included devices with binary states, the expected states could easily be inferred. In addition, since measuring explanation needs was not a primary objective of our study, this guidance did not significantly impact the results.

Furthermore, participants were only informed about the distinction between causal and counterfactual explanations after the main study and were asked to rank the two types separately. It is unclear whether participants were really able to differentiate between them, though they were not surprised to learn that they always received the same two types of explanations.

In addition, some participants observed that, by the end of the study, their preferences differed from the expectations they initially reported regarding explanation length and objective. This discrepancy suggests that self-reported preferences before experiencing any explanations may not be an appropriate measure.

Furthermore, the lack of randomization in the order of scenes introduced the risk of sequencing effects. In addition, participants were exposed to multiple scenes, potentially introducing a maturation effect. Together, these factors may have significantly impacted the rankings. This is especially likely as causal explanations were ranked first in the first two scenes, suggesting that participants needed time to comprehend the complex linguistic structure of counterfactual explanations.

Additionally, the analyzed contexts were not isolated within the scenes, complicating the interpretation of results since multiple contexts likely influenced participants' ratings simultaneously. For instance, setting and time pressure strongly correlated in the scenes as they both applied in the first three scenes. Finally, the distribution of contexts was uneven. For example, two scenes were set in a smart home, while four took place in a smart office, resulting in an overrepresentation of the latter. Finally, while for some contexts, an objective measure for when they applied could be used, other measures were more subjective. For example, the desire to change a situation may be subjective, as indicated by participants asking if a brightness of 70 % or the blinds rolling down halfway was enough.

External Threats Our findings may lack generalizability due to several limiting factors. First, the number of participants was relatively low, making the results vulnerable to sampling bias. Additionally, the study was conducted in an interview format with slides rather than in a real-life smart environment, which could impact the ecological validity of the findings. However, the slides used in the interview were enhanced with images and animations to facilitate the participants' development of a mental model of the smart environment.

To isolate the effect of our counterfactual explanation generation, we limited the study to devices with binary states, thereby excluding the foil determination method proposed by Herbold et al. (2024). While this approach ensured that our study only evaluated the counterfactual explanation framework, it also reduced the generalizability of the results to smart environments involving more complex devices.

Moreover, only six scenes were included in the study, and these were restricted to two settings: a smart home and a smart office. While we aimed to design the scenes to be as diverse as possible, the limited number may still restrict the applicability of our findings. Each scene referred to a distinct sub-case of explanation need, and all three primary cases of explanation needs were covered twice. Nevertheless, not all sub-cases described in Section 3.2 were included, further limiting the study's generalizability across all possible explanation needs.

Construct Threats Since our study was conducted as an in-person interview, the presence of an interviewer aware of the desired responses was necessary but may have introduced an observer bias. However, the use of standardized slides minimized variability between interviews and, therefore, reduced its potential impact.

Additionally, we analyzed several contexts that we hypothesize influenced participants' ratings of the explanation types. While these contexts were carefully motivated and supported by participant feedback, we cannot guarantee that all relevant factors were identified.

Furthermore, the decision on these contexts was made after the study, which allowed us to incorporate valuable participant insights. However, this approach also introduced the risk of confirmation bias, as the analysis could have unknowingly been influenced by preexisting expectations or interpretations of the results.

6 Conclusion

Counterfactual explanations offer insights into how an outcome could have been changed by examining what would have happened if an alternative event had occurred in the past (Stepin et al., 2021). They are especially valuable in rule-based smart environments as they enhance causal understanding and are most effective in situations that are controllable and likely to repeat in the future (Byrne, 2019; Roese, 1997).

Therefore, in this thesis, we proposed a framework for generating counterfactual explanations in rule-based smart environments. The framework determines the minimal change required for the system to align with the user's expectation. It begins by identifying the previous, current, and expected states of the device that caused the confusion and collects all rules with true preconditions whose actions result in the changing of the device to either of the three states. The rules are subsequently used to determine the appropriate case of explanation need. Depending on the case, the rules that must be reversed or overridden are determined. The framework then calculates the minimal change to override or remove each of these rules separately, prioritizing changes the user can implement themselves. Then, all possible combinations of overriding some rules and removing others are considered, and the minimal option is chosen. All minima are calculated using TOPSIS, which considers the properties sparsity, temporality, proximity, and abnormality. Finally, using a natural language pattern, the minimal change is transformed into an explanation that is issued to the user. This approach ensures that explanations are concise, actionable, and aligned with human reasoning, acknowledging that humans prefer to avoid unnecessary information and tend to select only one or two causes that they can manipulate as the explanation (Chazette & Schneider, 2020; T. Miller, 2019; Girotto et al., 1991).

To test the feasibility of our proposed framework, we implemented it as a plugin to *SmartEx*, a RESTful web service by Sadeghi et al. (2024) that can be integrated into pre-existing smart environments and offers an explanation layer for them while staying separate from the core intelligent system. Moreover, we conducted a user study to evaluate our framework in practice, thereby addressing a significant gap in the research of counterfactual explanations (Guidotti, 2022). The study followed a within-subject design and was conducted as an interview of 17 participants who experienced six confusing scenes in smart environments. Due to time and cost restraints, participants were not placed in real-life smart environments but were presented with slides that were enhanced with images and animations of smart environments. After each scene, participants were asked to rank their preference among receiving no, a causal, or a counterfactual explanation.

We found out that users prefer to receive an explanation but do not generally favor one explanation type over the other. Counterfactual explanations were criticized for their complex linguistics, as participants found them difficult to comprehend. However, the content of counterfactual explanations was marginally preferred.

We could identify several contexts in which one explanation type was preferred over the other. Causal explanations were favored when users were under time pressure, in an office setting, or preferred longer explanations providing a reason. In contrast, counterfactual explanations were preferred when users were at home, wanted to change the situation, preferred shorter explanations that offered a solution, and when the explanation was actionable. Furthermore, counterfactual explanations that omitted details were preferred over those that did not, and subtractive explanations were preferred over additive ones.

In the future, further research in real-life smart environments is needed. Participants should experience more scenes in a random order, where the contexts are isolated and evenly distributed. Furthermore, a between-subject design should be adopted to allow for independent measuring of the explanation types.

Finally, we suggest the development of a system where counterfactual explanations are only issued when users are willing to invest time and effort into understanding them, such as when they desire to change the situation. A large language model should be employed to improve the complex tenses and negations of counterfactual explanations, and user preferences should be considered, as suggested by Liao et al. (2020). The provided counterfactual explanations should be concise and, where possible, actionable and subtractive rather than additive.

References

Ahmed, E., Yaqoob, I., Gani, A., Imran, M., & Guizani, M. (2016). Internet-of-things-based smart environments: State of the art, taxonomy, and open research challenges. *IEEE Wireless Communications*, *23*(5), 10–16.

Alam, M. R., Reaz, M. B. I., & Ali, M. A. M. (2012). A review of smart homes-Past, present, and future. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, *42*(6), 1190–1203.

Ali, R., Afzal, M., Sadiq, M., Hussain, M., Ali, T., Lee, S., & Khattak, A. M. (2018). Knowledge-based reasoning and recommendation framework for intelligent decision making. *Expert Systems*, *35*(2), e12242.

Balasubramaniam, N., Kauppinen, M., Rannisto, A., Hiekkanen, K., & Kujala, S. (2023). Transparency and explainability of AI systems: From ethical guidelines to requirements. *Information and Software Technology*, *159*, 107197.

Bertossi, L. (2020). An ASP-based approach to counterfactual explanations for classification. In *Rules and Reasoning: 4th International Joint Conference* (pp. 70–81).

Blumreiter, M., Greenyer, J., Garcia, F. J. C., Klös, V., Schwammberger, M., Sommer, C., … Wortmann, A. (2019). Towards self-explainable cyber-physical systems. In *2019 ACM/IEEE 22nd International Conference on Model Driven Engineering Languages and Systems Companion (MODELS-C)* (pp. 543–548).

Bunt, A., Lount, M., & Lauzon, C. (2012). Are explanations always important? A study of deployed, low-cost intelligent interactive systems. In *Proceedings of the 2012 ACM international conference on Intelligent User Interfaces* (pp. 169–178).

Byrne, R. M. (2019). Counterfactuals in explainable artificial intelligence (XAI): Evidence from human reasoning. In *Proceedings of the 28th international joint conference on artificial intelligence (ijcai)* (pp. 6276–6282).

Chazette, L., Brunotte, W., & Speith, T. (2021). Exploring explainability: A definition, a model, and a knowledge catalogue. In *2021 IEEE 29th international requirements engineering conference (RE)* (pp. 197–208).

Chazette, L., & Schneider, K. (2020). Explainability as a non-functional requirement: Challenges and recommendations. *Requirements Engineering*, *25*(4), 493–514.

Cook, T. D., Campbell, D. T., & Shadish, W. (2002). *Experimental and quasi-experimental designs for generalized causal inference.* Houghton Mifflin.

Dai, J., Zhang, C., Aliakseyeu, D., Peeters, S., & Ijsselsteijn, W. A. (2023). The effect of explanation design on user perception of smart home lighting systems: A mixed-method investigation. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (pp. 1–14).

Das, D., Nishimura, Y., Vivek, R. P., Takeda, N., Fish, S. T., Ploetz, T., & Chernova, S. (2023). Explainable activity recognition for smart home systems. *ACM Transactions on Interactive Intelligent Systems*, *13*(2), 1–39.

Del Ser, J., Barredo-Arrieta, A., Díaz-Rodríguez, N., Herrera, F., Saranti, A., & Holzinger, A. (2024). On generating trustworthy counterfactual explanations. *Information Sciences*, *655*, 119898.

Dobrovolskis, A., Kazanavičius, E., & Kižauskienė, L. (2023). Building XAI-based agents for IoT systems. *Applied Sciences*, *13*(6), 4040.

El-Din, D. M., Hassanein, A. E., & Hassanien, E. E. (2021). Smart environments concepts, applications, and challenges. *Machine Learning and Big Data Analytics Paradigms: Analysis, Applications and Challenges*, 493–519.

Girotto, V., Legrenzi, P., & Rizzo, A. (1991). Event controllability in counterfactual thinking. *Acta Psychologica*, *78*(1-3), 111–133.

Gregor, S., & Benbasat, I. (1999). Explanations from intelligent systems: Theoretical foundations and implications for practice. *MIS quarterly*, 497–530.

Guidotti, R. (2022). Counterfactual explanations and how to find them: Literature review and benchmarking. *Data Mining and Knowledge Discovery*, 1–55.

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*, *51*(5), 1–42.

Hanson, N. R. (1972). *Patterns of discovery: An inquiry into the conceptual foundations of science.* Cambridge University Press.

Hayes-Roth, F. (1985). Rule-based systems. *Communications of the ACM*, *28*(9), 921–932.

Herbold, L., Sadeghi, M., & Vogelsang, A. (2024). Generating context-aware contrastive explanations in rule-based systems. In *Proceedings of the 2024 workshop on explainability engineering* (pp. 8–14).

Houzé, E., Diaconescu, A., Dessalles, J.-L., & Menga, D. (2022). A generic and modular reference architecture for self-explainable smart homes. In *2022 IEEE International Conference on Autonomic Computing and Self-Organizing Systems (ACSOS)* (pp. 101–110).

Hwang, C.-L., & Yoon, K. (1981). Methods for multiple attribute decision making. *Multiple attribute decision making: Methods and applications*, 58–191.

Jakobi, T., Stevens, G., Castelli, N., Ogonowski, C., Schaub, F., Vindice, N., ... Wulf, V. (2018). Evolving needs in IoT control and accountability: A longitudinal study on smart home intelligibility. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, *2*(4), 1–28.

Kahneman, D., & Tversky, A. (1982). The simulation heuristic. In *Judgment under uncertainty: Heuristics and biases* (p. 201-208). Cambridge University Press.

Karimi, A.-H., Schölkopf, B., & Valera, I. (2021). Algorithmic recourse: From counterfactual explanations to interventions. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency* (pp. 353–362).

Kästner, L., Langer, M., Lazar, V., Schomäcker, A., Speith, T., & Sterz, S. (2021). On the relation of trust and explainability: Why to engineer for trustworthiness. In *2021 IEEE 29th International Requirements Engineering Conference Workshops (REW)* (pp. 169–175).

Köhl, M. A., Baum, K., Langer, M., Oster, D., Speith, T., & Bohlender, D. (2019). Explainability as a non-functional requirement. In *2019 IEEE 27th International Requirements Engineering Conference (RE)* (pp. 363–368).

Lewis, D. (1973). Counterfactuals and comparative possibility. In *IFS: Conditionals, Belief, Decision, Chance and Time* (pp. 57–85). Springer.

Li, W., Yigitcanlar, T., Erol, I., & Liu, A. (2021). Motivations, barriers and risks of smart home adoption: From systematic literature review to conceptual framework. *Energy Research & Social Science*, *80*, 102211.

Liao, Q. V., Gruen, D., & Miller, S. (2020). Questioning the AI: Informing design practices for explainable AI user experiences. In *Proceedings of the 2020 CHI conference on human factors in computing systems* (pp. 1–15).

Lim, B. Y., Dey, A. K., & Avrahami, D. (2009). Why and why not explanations improve the intelligibility of context-aware intelligent systems. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 2119–2128).

Lipton, P. (1990). Contrastive explanation. *Royal Institute of Philosophy Supplements*, *27*, 247–266.

Lombrozo, T. (2006). The structure and function of explanations. *Trends in cognitive sciences*, *10*(10), 464–470.

Lopes, P., Silva, E., Braga, C., Oliveira, T., & Rosado, L. (2022). XAI systems evaluation: A review of human and computer-centred methods. *Applied Sciences*, *12*(19), 9423.

Lucic, A., Oosterhuis, H., Haned, H., & de Rijke, M. (2022). FOCUS: Flexible optimizable counterfactual explanations for tree ensembles. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 36, pp. 5313–5322).

Madumal, P., Miller, T., Sonenberg, L., & Vetere, F. (2020). Explainable reinforcement learning through a causal lens. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 34, pp. 2493–2500).

Mandel, D. R., Hilton, D. J., & Catellani, P. (2005). *The psychology of counterfactual thinking.* Routledge.

Markman, K. D., Lindberg, M. J., Kray, L. J., & Galinsky, A. D. (2007). Implications of counterfactual structure for creative generation and analytical problem solving. *Personality and Social Psychology Bulletin*, *33*(3), 312–324.

Masri, N., Sultan, Y. A., Akkila, A. N., Almasri, A., Ahmed, A., Mahmoud, A. Y., ... Abu-Naser, S. S. (2019). Survey of rule-based systems. *International Journal of Academic Information Systems Research (IJAISR)*, *3*(7), 1–23.

Miller, D. T., & Gunasegaram, S. (1990). Temporal order and the perceived mutability of events: Implications for blame assignment. *Journal of personality and social psychology*, *59*(6), 1111.

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, *267*, 1–38.

Miller, T. (2021). Contrastive explanation: A structural-model approach. *The Knowledge Engineering Review*, *36*.

Mothilal, R. K., Sharma, A., & Tan, C. (2020). Explaining machine learning classifiers through diverse counterfactual explanations. In *Proceedings of the 2020 conference on fairness, accountability, and transparency* (pp. 607–617).

Nandi, C., & Ernst, M. D. (2016). Automatic trigger generation for rule-based smart homes. In *Proceedings of the 2016 ACM Workshop on Programming Languages and Analysis for Security* (pp. 97–102).

Pearl, J. (2000). Models, reasoning and inference. *Cambridge University Press*, *19*(2), 3.

Poyiadzi, R., Sokol, K., Santos-Rodriguez, R., De Bie, T., & Flach, P. (2020). FACE: Feasible and actionable counterfactual explanations. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* (pp. 344–350).

Ranjbar, N., Momtazi, S., & Homayoonpour, M. (2024). Explaining recommendation system using counterfactual textual explanations. *Machine Learning*, *113*(4), 1989–2012.

Roese, N. J. (1997). Counterfactual thinking. *Psychological Bulletin*, *121*(1), 133.

Roese, N. J., & Epstude, K. (2017). The functional theory of counterfactual thinking: New evidence, new challenges, new insights. In *Advances in experimental social psychology* (Vol. 56, pp. 1–79). Academic Press.

Sadeghi, M., Herbold, L., Unterbusch, M., & Vogelsang, A. (2024). SmartEx: A framework for generating user-centric explanations in smart environments. In *2024 IEEE International Conference on Pervasive Computing and Communications (PerCom)* (pp. 106–113).

Sadeghi, M., Klös, V., & Vogelsang, A. (2021). Cases for explainable software systems: Characteristics and examples. In *2021 IEEE 29th International Requirements Engineering Conference Workshops (REW)* (pp. 181–187).

Shah, T., Venkatesan, S., Ngo, T., & Neelamegam, K. (2019). Conflict detection in rule based IoT systems. In *2019 IEEE 10th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)* (pp. 0276–0284).

Sherman, M. A. (1976). Adjectival negation and the comprehension of multiply negated sentences. *Journal of Verbal Learning and Verbal Behavior*, *15*(2), 143–157.

Stalnaker, R. C. (1968). A theory of conditionals. In *Ifs: Conditionals, belief, decision, chance and time* (pp. 41–55). Springer.

Stepin, I., Alonso, J. M., Catala, A., & Pereira-Fariña, M. (2021). A survey of contrastive and counterfactual explanation generation methods for explainable artificial intelligence. *IEEE*, *9*, 11974–12001.

Taherdoost, H., & Madanchian, M. (2023). Multi-criteria decision making (MCDM) methods and concepts. *Encyclopedia*, *3*(1), 77–87.

Triantaphyllou, E. (2000). *Multi-criteria decision making methods.* Kluwer.

Van der Waa, J., Robeer, M., van Diggelen, J., Brinkhuis, M., & Neerincx, M. (2018). Contrastive explanations with local foil trees. *Workshop on Human Interpretability in Machine Learning (WHI)*, 41–46.

Verma, S., Boonsanong, V., Hoang, M., Hines, K., Dickerson, J., & Shah, C. (2024). Counterfactual explanations and algorithmic recourses for machine learning: A review. *ACM Computing Surveys*, *56*(12), 1–42.

Vilone, G., & Longo, L. (2021). Notions of explainability and evaluation approaches for explainable artificial intelligence. *Information Fusion*, *76*, 89–106.

Wachter, S., Mittelstadt, B., & Russell, C. (2017). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harv. JL & Tech.*, *31*, 841.

Winikoff, M. (2018). Towards trusting autonomous systems. In *Engineering Multi-Agent Systems: 5th International Workshop, EMAS* (pp. 3–20).

Woodward, J. (2003). *Making things happen: A theory of causal explanation.* Oxford University Press.