Counterfactuals are an important explanatory tool to understand a models' cause for decisions. Generating counter factuals is main focus of this thesis. Hence, we will establish the most important chateristics of counterfactuals in this section.

0.0.1 What are Counterfactuals?

Counterfactuals have various definitions. However, their semantic meaning refers to "a conditional whose antecedent is false"[1]. A simpler definition from Starr states, counterfactual modality concerns itself with *what is not, but could or would have been.* Both definitions are related to linguistics and philosophy. Within AI and the mathematical framework various formal definitions can be found within causal inference[2]. Here, citeauthor describes a counterfactual as Causal inference definition. What binds all of these definitions is the notion of causality within "what if" scenarios.

However, for this paper, we will use the understanding established within the eXplanable AI (XAI) context. Within XAI, counterfactuals act as a prediction which "describes the smallest change to the feature values that changes the prediction to a predefined output"[3]. Note that XAI mainly concerns the explanation of models, which are always subject to inductive biases of the model itself and therefore inherently subjective. The idea behind counterfactuals as explanations[1] is that we understand the output of a model, if we know what change caused would cause a different outcome. For instance, lets denote a sequence 1 as $ABCDE\boldsymbol{FG}$, then a counterfactual $ABCDE\boldsymbol{XZ}$ would tell us that **F** (probably) caused **G** in sequence 1. As counterfactuals only address explanations of single model instances and not the model as a whole, they are called *local* explanation.

Valid counterfactuals satisfy four criteria. First, a counterfactual should be minimally different from the true instance. If the counterfactual to sequence 1 was $A\boldsymbol{A}CDE\boldsymbol{XZ}$ we would already have difficulties to discern whether B or F or both caused G at the end of sequence 1. Second, a counterfactual should produce a predefined outcome as closely as possible. This characteristic is ingrained in Molnars definition. If the counterfactual $ABCDE\boldsymbol{XZ}$ ends with Z but this sequence is highly unrealistic, then cannot be certain of our conclusion for sequence 1. Third, we typically desire multiple diverse counterfactuals. One counterfactual might not be enough to understand the causal relationships in a sequence. In the example above we might have a clue that F causes G but what if G is not only caused by F? If we

[1]There are other explanatory techniques in XAI like *feature importances* but counterfactuals are considered the most human-understandable

are able to find counterfactuals ***V****BCDEF****H*** and *ABCDE****XZ*** but all other configurations lead to G, then we know positions 1 and 6 cause G. As last criterion, each counterfactual should be possible. A sequence *ABCDE****1G*** would not be possible if numericals are not allowed. All four criteria allow us to assess the validity of each generated counterfactual and thus, help us to define an evaluation metric.

0.0.2 The Challenges of Counterfactual Sequence Generation

The current literature surounding counterfactuals expose a number of challenges when dealing with counterfactuals.

The most important disadvantage of counterfactuals is the Rashommon Effect[3, ch. 9.3]. If all of the counterfactuals are valid, but contradict eachother, we have to decide which of the *truths* are worth considering.

This decision reveals the next challenge of evaluation CITE . Although, the criteria can support us with the decision, it remains a question *how* to evaluate counterfactuals. Every automated measure comes with implicit assumptions and often do not guarantee a realistic explanation. We still need domain experts to assess their validity.

The generation of counterfactual sequences contribute to both former challenges, due to the combinatorial expansion of the solution space. This problem is common for counterfactual sentence generation and has been adressed within the Natural Language Processing (NLP) CITE . However, as process mining data not only consist of discrete objects like *words*, but also event and case features, the problem remains a daunting task. So far, little work has gone into the generation of multivariate counterfactual sequences like Process Instances CITE .