### 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 that 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 in the causal inference[2] literature. Here, CITE 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 eXplainable 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" according to Molnar[3, ch.9.3]. Note that XAI mainly concerns itself with the explanation of *models*, which are always subject to inductive biases and therefore, inherently subjective. The idea behind counterfactuals as explanatory tool[1]is simple. We understand the outcome of a model, if we know *what* outcome would occur *if* we changed its input. For instance, lets declare a sequence 1 as *ABCDE****FG***. Then a counterfactual *ABCDE****XZ*** would tell us that **F** (probably) caused **G** in sequence 1. As counterfactuals only address explanations of one model result and not the model as a whole, they are called *local* explanations[3, ch.9.3]. According to Molnar *Valid* counterfactuals satisfy **four** criteria[3, ch.9.3]: **[FIX Indentation!]**

Similarity: A counterfactual should be minimally different from the original instance. If the counterfactual to sequence 1 was *A****A****CDE****XZ*** we would already have difficulties to discern whether B or F or both caused G at the end of sequence 1. This criterion is closely linked to the similarity of the original instance and the counterfactual instance.

Precision: A counterfactual should produce a predefined outcome as closely as possible. This characteristic is ingrained in Molnar's definition. If the counterfactual *ABCDE****XZ*** ends with Z but this sequence is highly unrealistic, we cannot be certain of our conclusion for sequence 1. Therefore, we want to have its likelihood being the highest of all other potential outcomes among *ABCDE[?]Z***[Needs a better description]**. We simplify this criterion by just focusing on options that are at least more likely than **the original sequence**.

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

Diversity: 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 are able to find counterfactuals $\boldsymbol{V}BCDEF\boldsymbol{H}$ and $ABCDE\boldsymbol{XZ}$ but all other configurations lead to G, then we know positions 1 and 6 cause G. **Ensuring this criterion often heavily relies on the generative process and its variability of producing counterfactual candidates.**

Sparsity: Each counterfactual should be feasible. As an example, a sequence $ABCDE\boldsymbol{1G}$ would not be feasible if numericals are not allowed. **The *open-world assumption* makes this criterion quite hard to fulfill. However, we can use the evidence given by the data to help us restrict the search space.**

All four criteria allow us to assess the validity of each generated counterfactual and thus, help us to define an evaluation metric. It is important to note however, that they only constitute the basis for an domain-agnostic evaluation of counterfactuals. For a real world application, we still have to evaluate their *reasonability* within the applied domain.**[Need to establish model evaluation and realisticness evaluation groups... The first constitute the core criterions (internal validity) and the second the auxiliary (external validity).]**

### 0.0.2 The Challenges of Counterfactual Sequence Generation

The current literature surounding counterfactuals exposes 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***[Rather realisticness like the 4+1 criterions above?]**.

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] .