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. A prominent figure within the causal inference discipline is Pearl et al., who postulates that a "*kind of statement – an 'if' statement in which the 'if' portion is untrue or unrealized – is known as a counterfactual*"[4]. What binds all of these definitions is the notion of causality within *what-if* scenarios.

For this paper, we 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, p. 212]. 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.

Let us assume, a student is approaching an important deadline, which she desires to meet. Every day, she has a multitude of options to choose from. Either, continue with the report (option A), focus on learning more about the topic (option B), pursue her hobby as a break (option C), meet up with friends (option D), or procrastinate (option E). Furthermore, we assume, there are 7 days left and she can either miss (0) the deadline or meet it (1). The approach she follwos is $ABABD\boldsymbol{E}A$ and she misses the deadline. Let us refer to this sequence of actions as the factual *sequence 1.* Then, a counterfactual $ABABD\boldsymbol{B}A$ that meets the deadline tells us that **E** (probably) caused missing the deadline. In other words, if the student had not procastinated two days before the deadline she could have make it on time.

As counterfactuals only address explanations of one model result and not the model as a whole, they are *local* explanations[3, p. 212]. According to Molnar *Valid* counterfactuals satisfy **four** criteria[3, p. 212]:

Similarity: A counterfactual should be similar to the original instance. If a successful counterfactual to sequence 1 was $ABAB\boldsymbol{EE}A$, we would already have difficulties to discern whether meeting with friends D, procrastinating E or both caused the outcome of missing the deadline 0. Hence,

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

we want to be able to easily compare the counterfactual with the original. We can archive this by minimizing their mutual distance.

Sparcity: In line with the notion of similarity, we want to change the original instance only minimally. If the sequence had many changes, it would similarly impede the understanding of causal relationships in sequence 1.

Feasibility: Each counterfactual should be feasible. In other words, impossible values are not allowed. As an example, if the student followed a strict $A\boldsymbol{A}A\boldsymbol{A}A\boldsymbol{A}EA$ would not be feasible if we consider students could burnout. Typically, we can use data to ensure this property. However, the *open-world assumption* impedes this solution. With *open-world*, we mean that processes may change and introduce behaviour that has not been measured before. A student might only attempt a Bachelor's thesis once. Especially, for long and cyclical sequences, we have to expect previously unseen sequences.

Likelihood: A counterfactual should produce the desired outcome if possible. This characteristic is ingrained in Molnar's definition. However, as the model might not be persuaded to change its prediction, we relax this condition. We say that we want to increase the likelihood of the outcome as much as possbile. If the counterfactual $ABABDXA$ hinges on X as in an earthquake occuring that postpones the deadline, the sequence would be highly unrealistic. Hence, we cannot be certain of our conclusion for sequence 1. Therefore, we want the counterfactual's likelihood to be at least more likely than the factual outcome.

All four criteria allow us to assess the viability of each generated counterfactual and thus, help us to define an evaluation metric for each individual counterfactual. However, we also seek to optimise certain qualities on the population level of the counterfactual candidates.

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 E causes an outcome of 0, but what ifoutcome 0 is by more than E? If we are able to find counterfactuals all counterfactuals that involve E and that lead to missing the deadline, we get a better understanding of what caused outcome 0.

Realism: For a real world application, we still have to evaluate their *reasonability* within the applied domain. This is a characteristic that can only be evaluated by a domain expert.

We refer to both sets of viability criterions as *individual viability* and *population viability*. However, to remains concise, we use *viability* to refer to the individual criterions only. We explicitly mention *population viability* if we refer to criterions that concern the population.