Capturing Users' Reality: A Novel Approach to Generate Coherent Counterfactual Explanations

Maximilian Förster
University of Ulm
maximilian.foerster@uni-ulm.de

Philipp Hühn
University of Ulm
philipp.huehn@uni-ulm.de

Mathias Klier
University of Ulm
mathias.klier@uni-ulm.de

Kilian Kluge
University of Ulm
kilian.kluge@uni-ulm.de

Abstract

The opacity of Artificial Intelligence (AI) systems is a major impediment to their deployment. Explainable AI (XAI) methods that automatically generate counterfactual explanations for AI decisions can increase users' trust in AI systems. Coherence is an essential property of explanations but is not yet addressed sufficiently by existing XAI methods. We design a novel optimization-based approach to generate coherent counterfactual explanations, which is applicable to numerical, categorical, and mixed data. We demonstrate the approach in a realistic setting and assess its efficacy in a human-grounded evaluation. Results suggest that our approach produces explanations that are perceived as coherent as well as suitable to explain the factual situation.

1. Introduction

An expert group on Artificial Intelligence (AI) appointed by the European Commission states: "Without AI systems [..] being demonstrably worthy of trust, unwanted consequences may ensue and their uptake might be hindered" [1, p. 4]. Many AI systems are "black boxes" in that the reasons for their decisions and recommendations remain hidden from their users [2]. Consequently, users blindly follow AI systems' recommendations, distrust their decisions, or do not use the systems at all [2, 3].

In light of these challenges, the emerging research field of Explainable AI (XAI) provides approaches to automatically generate explanations along with AI systems' outputs. In this context, explanations are defined as human-understandable lines of reasoning for why an AI system maps a given input to a specific output [4]. Whereas the primary aim of XAI is to enable users to scrutinize AI outputs [1], existing methods are often criticized for producing explanations that only their developers appreciate, rather than their users [5].

Insights from the social sciences into how humans perceive explanations might inform the design of XAI methods [6]. One key finding is that humans predominantly construct counterfactual explanations, which are thus seen as a promising path for XAI [6]. Counterfactual explanations expose why an AI system yielded a particular output instead of another, similarly perceivable one [7]. In the case of the rejection of a loan, for example, a counterfactual explanation points out the contrast between the fact (e.g., customer's income and savings) and a so-called foil (e.g., higher income) that would lead to an approval. One major requirement for counterfactual explanations found by social sciences and confirmed by XAI user studies is their coherence [5, 6]. Coherence demands that the counterfactual scenario appears realistic to the user [6]. Further, the scenario contrasted should be suitable to explain the factual situation and not differ too much [6, 8].

Existing XAI research underpins the relevance of generating coherent explanations and provides promising ideas to address specific aspects of coherence [9–14]. However, to date, no approach exists that considers coherence to its full extent in the generation of counterfactual explanations [15]. Against this background, we propose a novel approach to generate coherent counterfactual explanations, i.e., realistic scenarios suitable for explaining an AI system's output. In a nutshell, our optimization-based approach utilizes a density estimate to find foils that represent real and typical scenarios. Harmonized distance measures ensure that the scenario contrasted in the explanation is suitable to explain the factual situation. Finally, the approach enables the incorporation of external knowledge into the explanation generation, thus enabling the refinement of coherence in a specific application context. Besides addressing coher-

URI: https://hdl.handle.net/10125/70767
978-0-9981331-4-0

ence, our approach natively handles numerical and categorical variables, thus expanding the applicability of optimization-based XAI methods to mixed data.

The remainder of this paper, following the Design Science methodology [16], is structured as follows: In Section 2, we present the theoretical background. In Section 3, we propose a novel approach to generate coherent counterfactual explanations for AI systems' outputs. Subsequently, in Section 4, we demonstrate its applicability and efficacy in a realistic setting based on a real-world data set. We conclude the paper with a discussion of the implications of our research, a reflection on its limitations, and directions for further research.

2. Theoretical background

2.1. Coherence of counterfactual explanations

The research field of XAI aims to help users "appropriately trust" AI systems by providing automatically generated explanations along with their outputs [1]. These explanations need to be user-centric in that users find them helpful to scrutinize AI decisions [5]. There exists rich literature proposing XAI methods that automatically generate explanations. For an overview, see the recent review by Arrieta et al. [17]. Often, XAI methods are model-agnostic, i.e., they can be used for any kind of AI system while not influencing its performance [11]. Inspired by how humans construct explanations, user-centric XAI research focuses on counterfactual explanations [6, 8]. The central elements of a counterfactual explanation are the *fact* (the event resulting in the AI system's output) and the *foil* (the event resulting in an alternative output). The difference between the fact and the foil is the *contrast* [6, 7].

The design of methods generating counterfactual explanations is informed by insights from social sciences investigating how humans perceive explanations [6]. Empirical studies from cognitive and social psychology provide desired characteristics of explanations [18] (for an overview, see [6]). These studies find coherence to be a decisive characteristic of counterfactual explanations [18, 19]. In the context of XAI, the relevance of coherence has been underpinned by various researchers [6, 9, 12, 13] and recently confirmed by a user study [5].

In general, an explanation is coherent if it relates to its recipients' prior beliefs [18, 19]. Research in social sciences and XAI reveals two requirements for explanations to be perceived as coherent [6, 19]. First, the counterfactual explanation must represent a realistic scenario [6]. From a social science perspective, the counterfactual scenario should "describe the results of observation" [19, p. 435]. XAI literature translates this into a foil representing a realistic and typical data point [13]. In the case of a loan rejection, a non-realistic counterfactual scenario would be, for example, a situation where the customer is a teenager but has already held a full-time job for ten years. Second, the counterfactual explanation is required to point out a contrast that is suitable to explain the factual situation [6]. Social sciences propose that the counterfactual situation should relate to the factual situation [19]. XAI researchers translate this aspect to the contrast being sparse and small (i.e., the foil is close to the fact as determined by some distance measure) [8, 10, 12] as well as feasible (i.e., the foil can indeed be reached from the fact) [9]. In the example of a loan rejection, a non-suitable contrast would, e.g., demand major changes of all of the customer's attributes, even though a slightly higher income alone would also lead to an approval. In specific application scenarios, it might be required that beyond coherence, the contrast is actionable, i.e., bridging the contrast between the fact and the foil is achievable for the explanations' recipient [9, 14].

2.2. Methods for the generation of coherent counterfactual explanations

The choice of a foil is crucial for counterfactual explanations to be perceived as coherent by their recipients [6, 8]. Formally, an AI system is a model $f(x)$ that produces an output y (e.g., loan rejection) for a given fact x from the input space of $f(x)$ (e.g., possible combinations of customers' features). An alternative outcome y' (e.g., loan approval) can be determined automatically or provided by the user. On that basis, a method searches for a suitable foil x', such that $f(x')=y'$.

XAI literature proposes two main classes of methods to identify the foil. The first class locally approximates the AI system with a simpler model from which a foil is extracted [20]. For example, a decision tree is used to approximate the AI system's outputs in the vicinity of the fact to derive a foil that lies close to the fact with respect to the decision tree's structure [21]. The second class frames the search for a foil as an optimization problem, finding foils by directly utilizing the respective outputs of the AI system [8]. More concretely, the value of an objective function capturing the foil's desired characteristics is optimized [8]. In the absence of local approximations, these methods can reliably produce explanations faithful to AI systems, a regulatory requirement for many applications [11]. Indeed, counterfactual explanations are mainly generated using optimization-based methods [15]. Beyond, as we discuss in the following, they constitute a promising starting point to address the coherence of explanations.

The pioneering optimization-based XAI approach by Wachter et al. [8] minimizes a weighted Manhattan

distance between x' and x as the objective function, with the constraint that the foil is classified into the foil class by the AI system. The particular choice of distance measure leads to a foil with a small and sparse contrast [8], thereby capturing a critical requirement for coherence. However, the approach incorporates no built-in mechanism to ensure that the foil represents a realistic and typical data point or that the contrast between the fact and the foil is feasible. Hence, recent XAI literature expands on this seminal work, aiming to control and improve the explanations' properties, and underpins the need for novel XAI methods producing coherent counterfactual explanations [9, 10, 12–14, 22].

Seeking to increase the realism and typicality of foils, researchers propose to add a term to the objective function that contains the difference from a foil to its auto-encoded value [11, 22]. Other researchers expand on this idea and propose a term that contains the distance of the auto-encoded foil to an average auto-encoded data point of the foil class [13]. However, these approaches lack practical applicability and transparency, as the explanations' quality highly depends on that of the auto-encoders. These are computationally expensive to train and constitute complex and hardly interpretable black-box models [22].

Other recent work focuses on the requirement that the contrast between the fact and the foil is feasible. Several studies propose to incorporate expert knowledge in the explanation generation process [10, 12, 14, 23]. For instance, the search for a foil may be restricted to the adaption of certain features or specific ranges previously defined as feasible by experts [12, 14]. Others suggest expert knowledge to assist the selection of foils that yield feasible contrasts [10, 12], e.g., by using it to instantiate filters that exclude foils with non-feasible contrasts [10]. While this idea might substantially contribute to feasibility, it suffers from a high dependency on expert knowledge's availability and quality. Another major drawback is the inability to consider complex interrelations concerning feasible changes (e.g., a change in a feature that is feasible only for some specific value combination of other features) [10, 12, 14].

Apart from optimization-based approaches, researchers suggest finding foils with feasible contrasts by taking the density of the AI system's training data into account. To find realistic paths between fact and foil, Poyiadzi et al. construct a graph from this data, with node weights calculated from a k-nearest-neighbor algorithm or kernel density estimate [9]. While the idea of considering the density of the training data appears promising, the proposed method [9] exhibits two major drawbacks. First, foils used for explanations are selected from the training data. In most use cases, the number of possible data combinations greatly exceeds that of training data points. Hence, the training data might not include points with a sufficiently small and sparse contrast to a given fact, an important requirement for foils to be perceived as coherent. Second, the approach suffers from high computational complexity, as feasible paths have to be calculated separately for each AI system's output. Indeed, the approach's applicability has only been demonstrated on a small synthetic data set [9].

To sum up, prior research provides valuable ideas to incorporate specific aspects of coherence in generating explanations for AI systems [9–14, 22]. However, none of the existing approaches addresses all requirements for counterfactual explanations to be perceived as coherent. Moreover, aside from preliminary qualitative evaluation with domain experts [11, 22], none of the existing approaches has been evaluated with users – a crucial step in the development of user-centric XAI methods [5].

Beyond coherence, optimization-based counterfactual XAI methods to date are not capable of incorporating categorical data, which is often encountered in real-world applications [12, 13]. Indeed, existing approaches rely on gradient-based optimization. However, for categorical variables, no gradient can be defined [11, 13]. One would need to optimize the objective function for each possible configuration of categorical variables and select the one that yields the best value [8], resulting in a computationally intractable problem [12]. Existing workarounds are limited to linearly ordered categorical data [11] and cannot cope with the complex interrelationships of categorical variables [13]. Others (e.g., one-hot encoding) mitigate the computational problem at the expense of neglecting the complex relationships between and within categorical variables [10]. Recently, researchers have proposed to utilize genetic algorithms to generate counterfactual explanations [23, 24]. While capable of generating foils for mixed data, these approaches neither capture the full complexity of categorical variables nor effectively address desired characteristics of explanations.

3. A novel approach to generate coherent counterfactual explanations

We design a novel approach to automatically generate counterfactual explanations that are perceived as coherent by their human recipients. For this, the foil is required to be both realistic and typical, while the contrast must be small, sparse, and feasible. To achieve this goal, we frame the search for a foil as an optimization problem and design an objective function based on a density estimate over the AI systems' input space and harmonized distance measures for both numerical and categorical variables. Moreover, we provide an option to integrate external knowledge.

First, to ensure that the foil is both realistic and typical and leads to a feasible contrast, we utilize an estimated probability density function (PDF). This enables to identify foils in regions of the input space with a high density. Such foils are not only realistic and typical but less likely to contain unusual or contradictory combinations of features. Further, guiding the search for a foil towards and along regions of high density ensures that the foil can indeed be feasibly reached from the fact. Consequently, incorporation of an estimated PDF contributes significantly to several aspects of the explanations' perceived coherence.

Second, to ensure that the contrast between the fact and the foil is both small and sparse, we utilize a pair of distance measures, one each for numerical and categorical variables. Framing the distance measurement as a cost-of-change estimate, we address the yet unsolved problem of providing a distance measure for categorical variables that is both consistent with reality and compatible with the one for numerical variables. Ensuring that the contrast is equally small for numerical and categorical variables contributes to the foil's perceived suitability to explain the fact.

Finally, to further refine the coherence of explanations, we include an option to integrate external knowledge. Our approach makes it possible to constrain the values of numerical variables to feasible ranges and exclude or adjust the costs of specific transitions for categorical variables based on information obtained from third parties. This not only contributes to foils that are perceived as both realistic and typical but further enhances the contrast's feasibility.

The basis for our approach and starting point for its design is gradient-free optimization, which natively handles mixed data. We integrate the first two core design components into an objective function

$$O(x') = \alpha\, T_{density}(x') + \mu\, T_{distance}(x', x)$$

that is minimized with constraint $f(x') = y'$ to obtain a foil x' for the fact x. The third core design component, external knowledge, can be integrated through modifying the loss terms $T_{density}$ and $T_{distance}$, imposing further constraints on x', and by influencing the optimizer's search heuristic. In the following, we detail our design decisions, formalize the objective function's terms and constraints, and describe their integration to a novel approach that yields coherent counterfactual explanations.

3.1. Starting point: gradient-free optimization

Optimization-based approaches are well-suited to generate counterfactual explanations, as they are model-agnostic and can guarantee the foil's faithfulness to the AI system. Therefore, we design our novel approach as an optimization problem. However, existing approaches optimize their objective function with a gradient-based optimizer, which leads to two inherent drawbacks. First, these approaches cannot adequately handle categorical or mixed data, which is ubiquitous in real-world applications [13], because no distance measure and consequently no gradient can be defined for categorical variables [10, 12]. Second, existing optimization-based approaches de facto require that a model's gradients can be computed analytically [10–13]. However, this is not possible for many popular kinds of AI systems (e.g., random forests) or cases where only the model's input and output values are accessible to the XAI method. In those cases, one has to resort to the extremely inefficient numerical computation of gradients [13].

To avoid these drawbacks, as the foundation of our approach, we employ gradient-free optimization based on the class of evolutionary algorithms [25]. We minimize the objective function $O(x')$ by randomly modifying the data point currently known to yield its smallest value ("parent") and determine whether this new data point ("child") leads to an even smaller value [25, 26]. While for numerical variables, we draw the new values from a normal distribution, we model the mutation of categorical variables as Markov chains on their value space [25], which define the probability that a variable's value is changed from its current value v_i to v_j. These transition probabilities are determined based on an estimated probability density function (cf. Section 3.3) and external knowledge (cf. Section 3.4). This process (cf. Algorithm 1) is repeated until a data point that yields a minimal value is found.

With this optimization procedure as its foundation, our approach can be applied to mixed data and is truly model-agnostic. The bandwidths and transition matrices influence not only the efficiency of the search but also the feasibility of the contrast. The properties of the obtained explanations are further determined by the objective function and additional constraints imposed on x'. In the following, we describe the design of the objective function's terms and the constraints in detail.

Algorithm 1. Gradient-free optimization

```
parent ← fact

for each optimization step do
  child ← parent
  for variable in object do
    if random(0,1) < 1/length(object) then
      if variable is numeric then
        variable ← draw from normal distribution
      if variable is categorical then
        variable ← select from Markov chain

  if O(child) < O(parent) then parent ← child
```

3.2. Estimated probability density function

For an explanation to be perceived as coherent, the foil has not only to be realistic but typical, i.e., exhibit a combination of feature values that is common in reality [13, 22]. To this end, existing counterfactual explanation methods either focus on auto-encoders [11, 13, 22] or expert knowledge [10, 12, 14]. As these approaches fall short due to their lack of transparency or dependence on the availability and quality of expert knowledge, in our approach, we utilize an estimated PDF to ensure realistic and typical foils.

From a probabilistic perspective, a real and typical data point has a high likelihood of occurring in reality. The function that describes the likelihood of any data point to be part of a population is its probability density function (PDF). In many AI applications, the PDF of the input data's population is unknown. However, in most application scenarios, a representative sample P of the population can be obtained. For example, in many machine-learning-based AI applications, the training data set constitutes such a representative sample or can be turned into one by weighting its data points according to their labels' probability. While requiring that foils belong to P ensures their realism [9], this restriction does not guarantee typicality and is detrimental to both the sparseness and smallness of the contrast.

Hence, in our approach, we evaluate the realism and typicality of data points using an estimate for the PDF obtained from the representative sample P [27]. To this end, we employ a mixed-variable multivariate Kernel Density Estimate (KDE) [27], which converges to the PDF with increasing size of P and decreasing bandwidths [28]. Instantiating a KDE for a given P is equivalent to selecting appropriate kernel functions and setting their bandwidths. We use a multivariate generalized product-kernel, whose value at a data point is the product of each variable's univariate kernel value [27]. As it allows us to choose a kernel function and its respective bandwidth independently for each variable, its parametrization can be tuned and evaluated separately, fostering the robustness and reliability of the obtained KDE. The bandwidth defines the size of the region around a data point considered for the estimation. The kernel function prescribes its shape as well as the relative weight given to points within it. In contrast to the training of auto-encoders [11, 13, 22], this proceeding is transparent and involves only a few clearly interpretable and testable parameters.

We integrate the resulting density estimate into our approach by adding its inverse to the objective function:

$$T_{density}(x') = 1/\mathrm{KDE}_P(x')$$

This loss term guides the search for a foil towards dense regions of the representative sample P and thus towards typical and realistic data points.

Beyond ensuring realistic and typical foils, this term substantially contributes to reach foils with a feasible contrast. More concretely, every optimization step aims to find an (intermediate) foil with high density. Thus, not only the final foil but also all foil candidates are less likely to contain unusual or contradictory combinations of features. This leads to a path of realistic data points from the fact to the foil [9]. In other words, the foil can be reached from the fact, resulting in a feasible contrast.

Finally, the density-related loss term contributes to the perceived coherence of explanations in an additional way, as data points are preferred that are similar to those from a sample that is representative of the data for which the AI system was designed. Therefore, the foil is more likely to be a data point for which the underlying AI system's output is reliable and realistic [25], preventing incoherent explanations.

3.3. Harmonized distance measures

An explanation's perceived coherence depends on the foil's reality and typicality and the contrast's feasibility, smallness, and sparseness. In optimization-based approaches, the latter is generally achieved through employing a suitable distance measure [8]. However, established distance measures are only applicable to numerical variables [13]. Researchers have attempted to construct ad-hoc distance measures for categorical variables [10, 11]. However, these approaches fail to capture the complex relationships represented by categorical variables and can lead to highly inconsistent distances. For example, linear ordering based on frequencies [11] is not applicable to unordered and non-linearly ordered categorical variables. Moreover, it falls short, even for linear categorical variables [13]. Further, for mixed data, the distance measures for numerical and categorical variables need to be balanced to prevent either type from dominating the distance.

Against this background, for our approach, we propose harmonized distance measures. Wachter et al. [8] found the Manhattan distance weighted by the median absolute deviation from the median (MAD) to be appropriate to obtain a small and sparse contrast. The MAD puts the change in one variable's value in relation to changes in all other variables' values, which can be interpreted as a cost-of-change estimate. We extend this concept to categorical variables by considering the likelihood that a variable's value changes from v_i to v_j as equivalent to the distance between v_i and v_j. As categorical variables can be unordered (e.g., a person's profession) or (partly) ordered (e.g., a person's education), a distance measure should reflect this aspect. Hence, we base the distance measure for categorical variables on Markov chains over their value space (cf. Section 3.1),

which are represented as transition matrices M [25]. In detail, we determine the transition probability M_{ij} from a value v_i to a value v_j based on the influence this change has on the estimated PDF. For this, we first calculate the average PDF of all data points with the value v_i for the categorical variable c. Next, we calculate the average PDF of the same data points after swapping v_i for v_j. Then, the distance between v_i and v_j is the inverse probability of the most probable path from v_i to v_j through the Markov chain, i.e.,

$$d_c(v_i, v_j) = 1/\max(\{M_{ik}^c \dots M_{lj}^c\}) - 1,$$

where $\{M_{ik}^c \dots M_{lj}^c\}$ denotes the set of probabilities of all possible paths from v_i to v_j. Subtracting 1 ensures that $d_c(v_i, v_j) = 0$ if the likeliest path's probability is 1. We harmonize the distance measures by combining

$$T_{distance}(x', x) = \sum_n \frac{|x_n' - x_n|}{MAD_n} + \beta \sum_c d_c(x_c', x_c),$$

where the first (second) sum is over all numerical (categorical) variables. With β chosen such that both sums are of the same order of magnitude, this term constitutes a distance measure for mixed data that takes the complexity of categorical variables into account. It guides the optimizer along feasible paths towards foils that result in a small and sparse contrast and thus contributes to the perceived coherence of explanations.

3.4. Integration of external knowledge

Ensuring that the foil is a data point of high density and the contrast is small and sparse is critical for the perception of explanations as coherent. The perceived coherence can be enhanced by constraining foils (e.g., excluding specific values) or the contrast (e.g., limiting transitions of categorical variables). Unlike previous approaches resorting to post-hoc filtering of explanations [10], our approach incorporates such constraints directly into the generation of explanations. On the one hand, this is more efficient, as only one foil needs to be generated [10]. On the other hand, even the search for a foil is guided through regions associated with coherent data points.

Incubation of external knowledge (i.e., from sources other than the AI system or its underlying training data) is accomplished in three distinct ways: First, by setting constraints on the values that individual variables can take. Second, by modifying the Markov chains for categorical variables that guide the optimizer. Third, by adjusting the sampling bandwidths for numerical variables. External knowledge can be obtained from, e.g., domain experts, federal statistical offices, and other public or corporate data sources. Its integration, if available, might notably increase the coherence of explanations in a specific application.

4. Demonstration and evaluation

In the following, as an essential part of the Design Science research process [16], we demonstrate the applicability of our approach and evaluate its efficacy in a realistic setting [29]. Following the Framework for Evaluation in Design Science Research (FEDS), we conduct a series of summative evaluations [16, 29] utilizing established XAI evaluation concepts introduced by Doshi-Velez and Kim [30]. To verify that our artifact meets its design goal, we first perform an artificial evaluation [29]. To this end, following the concept of functionally-grounded evaluation of XAI systems, we assess a large number of explanations using proxy measures [16, 30]. Second, to determine whether XAI users indeed perceive the resulting explanations as coherent, we conduct a more naturalistic evaluation [29]. Following the established concept of human-grounded evaluation, we analyze users' perception of explanations [5, 30, 31].

4.1. Setting and data set

We select price prediction for houses as the use case for our demonstration and evaluation. A fully functional AI system [4] suggests a price range to users that plan to sell a house. This classification task is representative of typical AI applications [5]. Explanations are intended to justify the price suggestions and thereby increase users' trust in the AI system.

To ensure rigor, we base the use case on a functionally complex AI system as well as a real-world data set [4, 29]. The data set contains 44,957 houses in Germany offered for sale on a popular online platform. The variables are presented in Table 1. For a majority of houses in the data set, not all variables' values are known. We use 80% of the data set entries to train a multi-layer neural network with about 50,000 parameters that classifies a house into one of 8 price ranges. The remaining 20% of the data set serves as the test set throughout the demonstration and evaluation. The AI model achieves an accuracy of 82% on the test data set. Neither the choice of the AI model nor its performance influence the instantiation and performance of our approach.

4.2. Instantiation of the approach

To instantiate our approach in the given setting, we first prepare the multivariate KDE, parametrize the harmonized distance measures, and integrate external knowledge. Subsequently, we initialize and parametrize the gradient-free optimizer and tune the weights of the objective function.

Table 1. Description of the data set as well as exemplary fact and corresponding generated foil

Variable	*Values*	*Exemplary Fact*	*Corresponding Foil*
building type	12 categories	single-family house	single-family house
year of construction	[1058; 2023]	1958	1958
living space (in m^2)	[10; 38,500]	163	170
no. rooms	[1; 420]	6	6
no. floors	[1; 12]	2	2
lot area (in m^2)	[20; 600,000]	3000	3000
heating type	14 categories	central heating	floor heating
cellar	y/n	n	n
condition	11 categories	need of renovation	well kept
interior quality	5 categories	simple	simple
no. parking lots	[0; 95]	2	4
state	16 categories	Baden-Württemberg	Baden-Württemberg
county	418 categories	Alb-Donau-Kreis	Alb-Donau-Kreis
price	8 categories	200,000€ to 350,000€	600,000€ to 800,000€

Instantiating the multivariate KDE is equivalent to selecting appropriate kernel functions for the generalized product-kernel and setting the bandwidths. To obtain a smooth density estimate, we use a Gaussian Kernel for numerical variables. For ordered categorical variables, we select a Wang-Ryzin kernel, as it can model the relationship between different variable values. For unordered categorical variables, we resort to an Aitchison-Aitken kernel that considers all values equally distinct. All three kernel functions are established standard choices for the respective type of variable [27]. We instantiate the KDE on the training set and select the bandwidths by inspecting and adjusting each variable's univariate KDE. We note that the goal for the KDE is not to fit the noisy data in every detail but to approximate the overall distribution smoothly.

To parametrize the harmonized distance measures, we first compute the MAD of the numerical variables using the training set. Second, to instantiate the distance measure for categorical variables, we compute the Markov chains' transition probabilities. To this end, we first select all data points from the training set with a particular value v_i of a categorical variable. Then, in order to estimate the likelihood of a transition to another value v_j, we determine the average change in the KDE that results from swapping v_i with v_j. To obtain transition probabilities, we normalize and smoothen the resulting values such that they are of the same order of magnitude.

We further modify the Markov chains based on external knowledge. First, we exclude (i.e., set to 0) transitions to values that indicate a lack of information, which prevents explanations where the fact contains specific information about the house (e.g., "central heating"), but the foil does not ("unknown heating"). Second, we exclude transitions that represent changes perceived as very large (e.g., from "ripe for demolition" to "mint condition"). Note that these transitions are still possible via intermediate values. Third, we modify the Markov chain for the variable "condition." As the estimated transition probabilities from "no information" are high predominantly for target values indicating unfavorable conditions, we mitigate this bias by replacing these probabilities with the target values' frequencies in the training data set.

Based on the KDE and the harmonized distance measures, we instantiate the objective function. We use a 1-plus-1 optimizer with the "$1/n$" mutation rule [26], guided by the Markov chains, to find its minimum. To compute a foil, we initialize the optimization problem with the fact and minimize the objective function with a budget of 1,000 steps. The pre-factors μ and α have to be set such that the weight of $T_{density}$ is sufficient to guide the optimizer along paths of high density, while ensuring that the term does not dominate $T_{distance}$. We find $\alpha = 2.5\mu$ to be a suitable ratio in our setting.

As a benchmark, we instantiate an upper bound on the state of the art (*BENCHMARK*). The most prominent method to generate counterfactual explanations, already providing a small and sparse contrast, is the pioneering optimization-based approach proposed by Wachter et al. [8]. However, it cannot handle categorical variables [11, 12] and is thus not directly applicable to our evaluation use case. To incorporate categorical variables in the optimization, we adapt the approach by using our gradient-free optimization algorithm with the Hamming distance for categorical variables, favoring a small and sparse contrast. In a setting with only numerical variables, the resulting explanations would be identical to those produced by the original approach by Wachter et al. [8].

4.3. Analysis of explanations' properties

Throughout the evaluation, we require a fixed set of explanations generated both with our novel approach (*ARTIFACT*) and with *BENCHMARK*. For pairwise comparison, we generate explanations for 1,000 houses from the test set and randomly select 51 houses for further evaluation. The number of distinct explanations ensures that a diverse set of explanations is judged, fostering the generalizability of our evaluation. At the same time, internal validity is improved by having multiple participants judge each explanation.

Following FEDS, in a first step, we assess whether the instantiated artifact meets its design goal [29] to generate counterfactual explanations that simultaneously address all aspects of coherence, which we defined as foils being realistic and typical while yielding a small, sparse, and feasible contrast. To this end, we analyze explanations with proxy measures, a concept well-established in Design Science research [16], which is known as functionally-grounded evaluation in the context of XAI [30].

First, to analyze the effect of our design decision to include harmonized distance measures on the coherence of explanations, we assess the explanations' contrasts. To this end, we compute both the *simplified distance* [8] (with the Hamming distance for categorical variables) and the *harmonized distance* (cf. Section 3.3). In line with our design assumptions, *ARTIFACT* contrasts exhibit a larger simplified distance in 92.2% of pairs and are in the median 90.9% larger than *BENCHMARK* contrasts. For the harmonized distance, *ARTIFACT* contrasts are larger in 70.6% of pairs, with a median increase of 10.7%. To assess the *sparseness*, we count the number of changed features [8] and find this proxy measure to be larger for *ARTIFACT* contrasts in 92.2% of pairs than *BENCHMARK* contrasts (median of 6 vs. 3 changed features). This implies that *ARTIFACT*, on average, produces foils that yield a less small and sparse contrast. Following our design hypothesis, this indicates that our approach focuses not only on these two aspects of coherence but takes others into account.

Second, we analyze the explanations with respect to the realism and typicality of the foils and their feasibility. To address these requirements, we designed our approach to incorporate a density-component in the search process (cf. Section 3.2). To measure its effect on the feasibility of explanations, we compute the *mean density* of all foil candidates in the optimization process and the *minimum density* of any foil candidate [9], both according to the KDE instantiated in Section 4.2. The mean density for *ARTIFACT* is higher than for *BENCHMARK* in 94.1% of pairs with a median increase of 2,620%, as it is the case for the minimum density (92.2%, 4,480%). To measure the realism and typicality of foils, we compare the *density estimates* for the final foils and find that *ARTIFACT* results in higher density estimates than *BENCHMARK* (98.0%, 54,800%). These findings indicate that our approach's density component markedly contributes to the coherence of explanations, thus supporting our design hypothesis. All shares of larger pairs reported above are significant based on a Wilcoxon signed-rank test ($p<0.01$). To sum up, based on the proxy measures, we find that our artifact meets its design goal. However, this artificial evaluation does not necessarily translate into users' perception [29, 30].

4.4. Human-grounded evaluation

Evaluation in a realistic setting is a crucial step in evaluating design artifacts [29]. In the context of XAI, this translates to evaluation with users. Indeed, while XAI methods are often merely evaluated with respect to proxy measures [5], only human-grounded evaluation can ensure that design assumptions reflect users' perception [5, 30, 32]. Therefore, we verify that our approach yields explanations perceived as coherent by users [5, 30]. In a study implemented as an online survey using the oTree framework [5, 33], participants first judge foils in terms of perceived realism and typicality. Then, they assess if a foil is suitable to explain a given fact. We survey 46 students (25 males and 21 females between 19 and 29 years). To ensure that participants, similar to users of an online real-estate platform, are familiar with the houses they judge, we restrict the survey to houses in the vicinity of their place of study.

To evaluate if foils are perceived as realistic and typical, in a crossover design with multiple periods, each participant is asked to rate ten houses. These houses are randomly sampled from the explanations generated by *ARTIFACT* and *BENCHMARK* for the initially selected 51 facts, as well as from two control sets. First, as a reference level for users' perception of real houses, we draw a set of samples from the data set (*REAL*). Second, as the baseline for unrealistic houses, we generate houses by independently drawing each variable's value from the training set (*FAKE*). Participants are shown houses in a random order to mitigate carryover effects. They rate the perceived realism and typicality each on a four-point Likert-like scale ranging from 0 ("not real/typical") to 3 ("real/typical"). Table 2 shows the means of the ratings' distributions for each set of foils.

Table 2. Means for realism and typicality

	Realism	*Typicality*
ARTIFACT	1.78	1.32
BENCHMARK	1.50	1.12
REAL	1.70	1.44
FAKE	1.06	0.82

As the ratings are not normally distributed, we conduct Mann-Whitney U tests to determine whether the ratings between sets differ significantly. We find that participants rate foils generated by *ARTIFACT* as significantly more real ($p<0.01$, effect size=0.59) and typical ($p<0.05$, 0.57) than *BENCHMARK* foils. Foils generated by *ARTIFACT* are rated as (non-significantly) more realistic and less typical than *REAL* houses. Moreover, foils generated by *ARTIFACT* or *BENCHMARK* are perceived as significantly more real ($p<0.001$, 0.72, and $p<0.001$, 0.63) and typical ($p<0.001$, 0.69 and $p<0.01$, 0.60) than *FAKE* houses. In order to evaluate the suitability of foils to explain the fact, participants are presented with a given house (fact) paired with an alternative house that is classified into a different price category (foil). For each participant, we randomly sample eight foils generated by *ARTIFACT* or *BENCHMARK* for the same set of facts. Participants are asked to rate the foil's *suitability* ("The houses differ exactly in the variables that explain the difference in price.") as well as the contrast's *sparseness* ("The given and the alternative house differ in too many variables.") on a five-point Likert-like scale from 1 ("I don't agree at all") to 5 ("I fully agree"). Pairwise comparison of the mean ratings reveals that *ARTIFACT* foils are perceived as less sparse than *BENCHMARK* foils generated for the same fact, with a larger rating for *sparseness* in 63% of pairs. However, foils generated by *ARTIFACT* are perceived as more suitable, with a larger rating for *suitability* in again 63% of pairs. In both cases, $p<0.1$ according to a Wilcoxon signed-rank test. In sum, users perceive foils produced by our novel approach as more realistic, typical, and suitable to explain the fact, although they perceive the contrasts as less sparse.

5. Conclusion, limitations, and directions for further research

Automatically generated explanations promise to help users scrutinize AI decisions and reduce their distrust in AI systems. In this context, the coherence of explanations is essential. Counterfactual explanations are perceived as coherent if the counterfactual scenario is realistic and typical as well as suitable to explain the factual situation. Although prior research provides ideas to address specific aspects of coherence, none of the existing approaches incorporates all aspects.

Against this background, we designed a novel approach for the generation of coherent counterfactual explanations. Our artifact includes a density estimate contributing to the creation of realistic and typical foils with a feasible contrast. Further, harmonized distance measures ensure that the contrast is small and sparse. Finally, external knowledge can be included to refine the coherence of explanations. We demonstrated the approach by generating explanations for house price estimates. After instantiating our approach utilizing a real-world data set, we evaluated the explanations in a user study. Results suggest that our approach produces explanations with foils perceived as significantly more realistic and typical as well as more suitable to explain the factual situation than those in state-of-the-art counterfactual explanations. To the best of our knowledge, our approach is the first that addresses all aspects of coherence simultaneously. We were also first to verify that resulting explanations are perceived as coherent by users. Further, our approach is applicable to mixed data and takes the complex relationships encoded by categorical variables into account.

Although our work constitutes a substantial step towards the generation of coherent counterfactual explanations, it is subject to several limitations. First, we demonstrated the approach only for one single use case. Hence, we encourage application in other domains, especially to use cases with mixed data, to which prior approaches are not applicable. As the integration of external knowledge played only a secondary role in our use case and thus in the evaluation, we particularly encourage studies that more deeply investigate its effect on coherence. Second, our experiment was conducted with students. While non-experts are a primary target group of user-centric XAI systems, future research should seek a more diverse range of participants with respect to demographics and domain knowledge. Third, although we provided evidence for our approach's efficacy through human-grounded evaluation, we call for future application-grounded evaluations to verify that coherent counterfactual explanations indeed enable users to scrutinize AI decisions.

References

[1] Independent High-Level Expert Group on Artificial Intelligence set up by the European Commission, "Ethics Guidelines for Trustworthy AI", European Commission, Brussels, 2019, https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=60419.

[2] S. Herse, J. Vitale, M. Tonkin, D. Ebrahimian, S. Ojha, B. Johnston, W. Judge, and M. Williams, "Do You Trust Me, Blindly? Factors Influencing Trust Towards a Robot Recommender System", 27th IEEE International Symposium on Robot and Human Interactive Communication, IEEE, Nanjing, China, 2018, pp. 7–14.

[3] E. Rader and R. Gray, "Understanding User Beliefs About Algorithmic Curation in the Facebook News Feed", Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, ACM, Seoul, South Korea, 2015, pp. 173–182.

[4] A. Abdul, J. Vermeulen, D. Wang, B.Y. Lim, and M. Kankanhalli, "Trends and Trajectories for Explainable,

Accountable and Intelligible Systems: An HCI Research Agenda", Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, ACM, Montréal, Canada, 2018.
[5] M. Förster, M. Klier, K. Kluge, and I. Sigler, "Evaluating Explainable Artificial Intelligence – What Users Really Appreciate", Proceedings of the European Conference on Information Systems 2020, AIS, 2020.
[6] T. Miller, "Explanation in artificial intelligence: Insights from the social sciences", Artificial Intelligence, 267, 2019, pp. 1–38.
[7] P. Lipton, "Contrastive Explanation", Royal Institute of Philosophy Supplement, 27, 1990, pp. 247–266.
[8] S. Wachter, B. Mittelstadt, and C. Russell, "Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR", Harvard Journal of Law & Technology, 31, 2018, pp. 841–887.
[9] R. Poyiadzi, K. Sokol, R. Santos-Rodriguez, T. De Bie, and P. Flach, "FACE: Feasible and Actionable Counterfactual Explanations", Proceedings of the 2020 AAAI/ACM Conference on AI, Ethics, and Society, ACM, New York, NY, 2020.
[10] R.K. Mothilal, A. Sharma, and C. Tan, "Explaining machine learning classifiers through diverse counterfactual explanations", Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, ACM, Barcelona, Spain, 2020, pp. 607–617.
[11] A. Dhurandhar, T. Pedapati, A. Balakrishnan, P.-Y. Chen, K. Shanmugam, and R. Puri, "Model Agnostic Contrastive Explanations for Structured Data", arXiv:1906.00117, 2019.
[12] C. Russell, "Efficient Search for Diverse Coherent Explanations", Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency, ACM, Atlanta, GA, 2019, pp. 20–28.
[13] A. Van Looveren and J. Klaise, "Interpretable Counterfactual Explanations Guided by Prototypes", arXiv:1907.02584, 2019.
[14] B. Ustun, A. Spangher, and Y. Liu, "Actionable recourse in linear classification", Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency, ACM, Atlanta, GA, 2019, pp. 10–19.
[15] U. Bhatt, A. Xiang, S. Sharma, A. Weller, A. Taly, Y. Jia, J. Ghosh, R. Puri, J.M.F. Moura, and P. Eckersley, "Explainable machine learning in deployment", Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, ACM, Barcelona, Spain, 2020, pp. 648–657.
[16] A.R. Hevner, S.T. March, J. Park, and S. Ram, "Design Science in Information Systems Research", MIS Quarterly, 28, 2004, pp. 75–105.
[17] A. Barredo Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, and F. Herrera, "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI", Information Fusion, 58, 2020, pp. 82–115.
[18] T. Lombrozo, "Explanation and Abductive Inference", The Oxford Handbook of Thinking and Reasoning, Oxford University Press, Oxford, 2012.
[19] P. Thagard, "Explanatory coherence", Behavioral and Brain Sciences, 12, 1989, pp. 435–467.
[20] R. Guidotti, A. Monreale, F. Giannotti, D. Pedreschi, S. Ruggieri, and F. Turini, "Factual and Counterfactual Explanations for Black Box Decision Making", IEEE Intelligent Systems, 34, 2019, pp. 14–23.
[21] J. van der Waa, M. Robeer, J. van Diggelen, M. Brinkhuis, and M. Neerincx, "Contrastive Explanations with Local Foil Trees", Proceedings of the 2018 ICML Workshop on Human Interpretability in Machine Learning, Stockholm, Sweden, 2018, pp. 41–46.
[22] A. Dhurandhar, P. Chen, R. Luss, C. Tu, P. Ting, K. Shanmugam, and P. Das, "Explanations based on the Missing: Towards Contrastive Explanations with Pertinent Negatives", Advances in Neural Information Processing Systems 31, Curran Associates, Montréal, Canada, 2018, pp. 592–603.
[23] S. Sharma, J. Henderson, and J. Gosh, "CERTIFAI: A Common Framework to Provide Explanations and Analyse the Fairness and Robustness of Black-box Models", Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, New York, NY, 2020.
[24] S. Dandl, C. Molnar, M. Binder, and B. Bischl, "Multi-Objective Counterfactual Explanations", arXiv:2004.11165v2, 2020.
[25] Russel, S.J. and P. Norvig, Artificial Intelligence: A Modern Approach, Pearson Education London, 2010.
[26] B. Doerr, H.P. Le, R. Makhmara, and T.D. Nguyen, "Fast Genetic Algorithms", Proceedings of the Genetic and Evolutionary Computation Conference, ACM, Berlin, Germany, 2017, pp. 777–784.
[27] J.S. Racine, "Nonparametric Econometrics: A Primer", Foundations and Trends in Econometrics, 3, 2008, pp. 1–88.
[28] Silverman, B.W., Density Estimation for Statistics and Data Analysis, Chapman and Hall, London, 1986.
[29] J. Venable, J. Pries-Heje, and R. Baskerville, "FEDS: a Framework for Evaluation in Design Science Research", European Journal of Information Systems, 25, 2016, pp. 77–89.
[30] F. Doshi-Velez and B. Kim, "Considerations for Evaluation and Generalization in Interpretable Machine Learning", Explainable and Interpretable Models in Computer Vision and Machine Learning, Springer, Cham, 2018, pp. 3–17.
[31] H.J.P. Weerts, W. van Ipenburg, and M. Pechenizkiy, "A Human-Grounded Evaluation of SHAP for Alert Processing", Proceedings of the KDD Workshop on Explainable AI, Anchorage, AK, 2019.
[32] T. Miller, P. Howe, and L. Sonenberg, "Explainable AI: Beware of Inmates Running the Asylum Or: How I Learnt to Stop Worrying and Love the Social and Behavioural Sciences", IJCAI-17 Workshop on Explainable AI (XAI) Proceedings, Melbourne, Australia, 2017, pp. 36–42.
[33] D.L. Chen, M. Schonger, and C. Wickens, "oTree—An open-source platform for laboratory, online, and field experiments", Journal of Behavioral and Experimental Finance, 9, 2016, pp. 88–97.