Towards feasible counterfactual explanations: a taxonomy guided template-based NLG method.

SALIMI, P., WIRATUNGA, N., CORSAR, D. and WIJEKOON, A.

2023

ECAI 2023
K. Gal et al. (Eds.)

doi:10.3233/FAIA230499

Towards Feasible Counterfactual Explanations: A Taxonomy Guided Template-Based NLG Method

Pedram Salimi[a], Nirmalie Wiratunga[a], David Corsar[a] and Anjana Wijekoon[a]

[a]Robert Gordon University, Aberdeen, UK
ORCiD ID: Nirmalie Wiratunga https://orcid.org/0000-0003-4040-2496,
David Corsar https://orcid.org/0000-0001-7059-4594, Anjana Wijekoon https://orcid.org/0000-0003-3848-3100

Abstract. Counterfactual Explanations (cf-XAI) describe the smallest changes in feature values necessary to change an outcome from one class to another. However, many cf-XAI methods neglect the feasibility of those changes. In this paper, we introduce a novel approach for presenting cf-XAI in natural language (Natural-XAI), giving careful consideration to actionable and comprehensible aspects while remaining cognizant of immutability and ethical concerns. We present three contributions to this endeavor. Firstly, through a user study, we identify two types of themes present in cf-XAI composed by humans: content-related, focusing on how features and their values are included from both the counterfactual and the query perspectives; and structure-related, focusing on the structure and terminology used for describing necessary value changes. Secondly, we introduce a feature actionability taxonomy with four clearly defined categories, to streamline the explanation presentation process. Using insights from the user study and our taxonomy, we created a generalisable template-based natural language generation (NLG) method compatible with existing explainers like DICE, NICE, and DisCERN, to produce counterfactuals that address the aforementioned limitations of existing approaches. Finally, we conducted a second user study to assess the performance of our taxonomy-guided NLG templates on three domains. Our findings show that the taxonomy-guided Natural-XAI approach (n-XAI^T) received higher user ratings across all dimensions, with significantly improved results in the majority of the domains assessed for articulation, acceptability, feasibility, and sensitivity dimensions.

1 Introduction

A counterfactual explanation (cf-XAI) shows how to get a different outcome from a black-box AI model by changing only a few input features. This aligns with human intuition by offering the black-box model's underlying rationale in the form of a counter-argument [4]. It serves three primary goals [25]: 1) elucidate the reasoning behind decisions; 2) supply adequate information to critique decisions with negative outcomes; and 3) enable a better understanding of the necessary changes to achieve desired outcomes in the future. There is an abundance of techniques to generate cf-XAI in the literature that achieve some subsets of these three goals [2, 9, 18, 24, 27]. The focus of this paper instead is to achieve the third goal as a post-processing step taking into account the user perspective.

The literature identifies many properties of good counterfactuals, such as sparsity, proximity, validity, diversity, feasibility, and plausibility [12]. To achieve the third goal, feasibility, must be integrated into the explanation generation process, such that the resulting cf-XAI provides a complete understanding of suggested changes. The challenge in attaining feasibility lies in meticulously evaluating the suggested alterations in the context of the user, taking into account factors such as appropriateness and ethics. For instance, proposing a change in an individual's weight might be acceptable in the medical domain; however, within the social sphere, such recommendations could be perceived as disrespectful and potentially offensive. The question of how to guide the user through the recommended changes while being sensitive to the types of features when presenting a cf-XAI remains unanswered.

To the best of our knowledge, no formal approach exists for managing user-specific feasibility considerations when presenting the counterfactual's recommended changes. Certain methods [18] generate a diverse set of counterfactuals in the expectation that users will identify one or more as feasible; others [2] entrust individuals with specifying requirements that can be integrated into the counterfactual generation algorithm. Both impose considerable cognitive load on the user. In this paper, we formalise feasibility requirements using a taxonomy to enable a natural language presentation of explanations (Natural-XAI), using a template-based natural language generation (NLG) approach to effectively address and handle feasibility-related criteria when presenting counterfactual recommendations. Accordingly we make the following contributions:

- propose a set of common natural language constructs, identified from a user study, that enables us to convey cf-XAI in a better textual presentation format;
- introduce a taxonomy that captures the knowledge of *feature actionability* and categorises features based on their mutability;
- present a template-based NLG method (n-XAI^T) that utilises the feature actionability taxonomy to generate counterfactual Natural-XAI;
- conduct a user study analysis of the proposed n-XAI^T method across three application domains, demonstrating that it improves counterfactual understandability with respect to sensitivity, acceptability, feasibility, and articulation; and
- provide useful guidelines and insights for XAI platform development, derived from a thematic analysis of user responses.

In Section 2 we present related work, while Section 3 describes an initial user study conducted to gather insights for improving counterfactual Natural-XAI. Section 4.1 formalises actionable recommendations for cf-XAI systems, based on our proposed taxonomy of fea-

ture actionability, and Section 4.2 describes a mapping from taxonomic categories to language generation templates for Natural-XAI. This section also discusses in detail our proposed n-XAI^T method, and the effectiveness of n-XAI^T, which incorporates a three-stage NLG pipeline integrating actionability knowledge, and evaluated in a second user study with results presented in Section 5.

2 Related Work

A cf-XAI is distinct from a factual explanation, as it aims to answer "What-If" and "Why-Not" user queries that relate to the input-output relationship of a black-box model, while factual explanations typically address "Why" questions. Given a user query and the AI's prediction, a cf-XAI defines the smallest change in feature values required to shift a prediction to the desired outcome. For example, in response to an AI loan application system's prediction, a cf-XAI may propose, "A smaller loan amount would have resulted in your application being accepted", where the action of decreasing by a small amount is the proposed action recommended by the cf-XAI system.

Proponents of counterfactual theories argue that they offer significant computational, psychological, and legal benefits [12]. Effective and interpretable cf-XAI must also satisfy several key requirements. Sparsity calls for minimising the number of modified features, while proximity ensures that the counterfactual instance is as close as possible to the original instance in the feature space, thereby seeking the minimal change necessary to achieve the desired outcome [12]. Both can be addressed either by case-based instance learning [2, 5, 27] or as parameters within optimisation minimisation techniques [18, 25]. Feasibility ensures suggested changes are achievable [20], and plausibility maintains realistic distributions [28]. This paper concentrates on feasibility and recourse, which involve users taking actionable steps based on provided explanations to achieve desired outcomes [23]. Feasibility, refers to whether a proposed change can realistically occur, and actionability, concerns the user's capacity to implement the change. Recourse emphasises the importance of suggesting feasible changes that users can implement realistically, to maintain user trust in AI systems. Specifically, we investigate post-processing generated explanations to refine their presentation format while carefully considering the feasibility of suggested actions, enabling our method to integrate with any cf-XAI algorithm.

Numerous studies have explored feasibility of counterfactuals, such as the FACE algorithm [20], which generates feasible counterfactuals by considering proximity in high-density regions. However, generalising to all individuals may be challenging due to diverse backgrounds and situations, rendering certain feasible counterfactuals ineffective for some users [1]. While feasibility knowledge aids in post-processing generated explanations by guiding the selection of suitable presentation styles for each recommended action, causal knowledge helps with grouping interrelated actions and effectively presenting actionable groups. However, most counterfactual explainers do not consider causal relationships. For example, while [18] highlights the importance of causal constraints for feasibility, they do not offer methods for generating them. Similarly, [11] emphasises causality and human intervention in feasible cf-XAI, but their approach necessitates deep causal model understanding. Although [17] presents a variational autoencoder-based learning method for generating feasible counterfactuals, it lacks scalability across domains and adequate data for learning causal constraints. Here, we focus on feasibility aspects in post-processing cf-XAI outputs, and argue that if causal knowledge is available, the presentation would involve combining presentations of individual features, which is less challenging compared to addressing the presentation of feasibility aspects.

Natural-XAI is more human-friendly and can be tailored to the user's specific context, beliefs, and preferences [13]. For instance, textual data representation has been shown to outperform visual graphs in clinical decision-making [14, 19], enhancing trust, transparency, acceptability, and usability. Effectively presenting counterfactuals requires managing actionable changes, which can be difficult to absorb when presented in tabular form. A natural language format is likely to be more accessible and capable of clearly describing the recommended actionable changes. In domains such as finance and health, controlled text generation, an advancement in numerical reasoning for language models, is critical for accurately conveying cf-XAI where inaccurate suggestions for numerical values or attributes can have significant consequences, such as rejected loan applications and financial harm to the applicant [22]. Despite advancements in large language models like GPT-3, such inaccuracies persist [10] due to hallucinations in generated text. The alternative template-based approach is a more reliable solution to integrate feasibility and ethical considerations for Natural-XAI. In this paper, we use a taxonomy to generate feature-based templates that inform the NLG process in Natural-XAI sentence planning, with surface realisation focused on choosing comparative adjectives, action verbs, and other forms of language constructs to convey actionable changes based on the taxonomy node type. Discourse planning involving the ordering of these sentences is typically influenced by the importance of the recommended action, based on feature attribution explainer weights [16].

3 Understanding How to Compose Counterfactuals

To understand how counterfactuals are authored, we carried out a user study examining naturally expressed counterfactuals in the widely-used loan-approval dataset, with the aim of identifying reusable linguistic constructs without requiring domain expertise.

3.1 User Study Setup

Using DICE [18], counterfactuals with 4-5 actionable feature changes for seven *rejected* loan applications was selected. Here the choice in the number of changes was based on the cf-XAI's actionable changes distribution observed on the loan dataset. A query to the cf-XAI system provides feature value pairs, and the corresponding counterfactual recommends alternative values for a subset of these features to achieve a desired outcome from the black-box model. For each such {query & counterfactual} pair, we generate a tabular form of the counterfactual that includes recommended changes to the query feature values ordered by SHAP [16] local feature importances. Additionally, we create a zero-centered visual chart (maintaining SHAP orderings) and a basic Natural-XAI format as alternative presentations of the same counterfactual (see Figure 1). For the Natural-XAI presentation, two basic NLG templates were used: n-XAI^{B1}, where feature changes are presented in order of SHAP feature value importance (e.g., increase feature F1 to V1 then decrease feature F2 to V2 ...); or n-XAI^{B2}, where feature changes are ordered by SHAP but grouped by action (verb) type order (e.g., increase features F1 & F3 to values V1 & V3, thereafter decrease features ...).

We had two independent cohorts, with alternative explanation formats allocated as follows: Cohort1 was shown a tabular form of the counterfactual and a visual chart highlighting differences between query and counterfactual, while Cohort2 was presented with

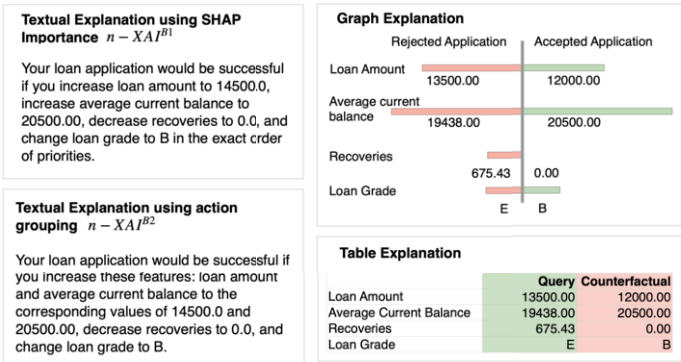


Figure 1: Examples of cf-XAI presentation formats

the Natural-XAI forms. In setting up the study we wanted to: 1) minimise the impact on users' natural expressions of counterfactuals by presenting multiple alternatives to address potential biases, such as cognitive, visual, and familiarity biases; 2) examine if access to basic natural XAI would influence the quality of users' responses; 3) determine if any correlation exists between the authored text and either n-XAI^{B1} or n-XAI^{B2} concerning preferences for sentence ordering; and 4) assess the extent to which access to alternative formats might influence the quality of user responses.

3.2 User Study Protocol

The study was conducted in October 2022 with 33 participants aged 18-24 from an undergraduate AI course but prior to learning about XAI and with no prior domain knowledge, although it is reasonable to assume that some familiarity with "student loans" is to be expected amongst a majority of our participants. After providing details about the dataset, including possible prediction outcomes and descriptions of loan features, participants were divided into cohorts. They received a query and corresponding counterfactual in alternative formats depending on their cohort. Participants were asked to use these counterfactuals to complete the tasks listed below:

1. Task1: Compose a piece of natural language guidance for the loan applicant to help them achieve a better outcome in the future.
2. Task2: Evaluate and rank the alternative explanation presentations based on their usefulness for composing the recommendation in Task1, along with justifications for the preferences.
3. Repeat Tasks 1 and 2, for 4 query instances from a set of 7 randomly selected queries.

In Task2, each cohort had access to different formats of the counterfactual: Cohort1 had table and chart; and Cohort2 had n-XAI^{B1} and n-XAI^{B2}. We expect that responses to Task1 will enable better understanding of the sentence planning requirements for Natural-XAI and provide insights into discourse planning, such as determining the most effective order in which to present content. Task2 aims to assess the degree to which the SHAP ordering of features in the table and chart influences the order of authored text for Cohort1. For Cohort2, the task seeks to identify any preferences or influences resulting from the ordering presented in both the SHAP-based n-XAI^{B1} and action group-based n-XAI^{B2}. Accordingly, Task2 is expected to provide valuable insights for discourse planning.

3.3 Quantitative Evaluation

The quantitative text analysis of the 108 responses used 7 criteria (see Table 1). Here the "Alternatives Preferred" criteria relate to Task2 whilst the rest relate to Task1.

Criteria	Analysis	Cohort1	Cohort2
Response Statistics	# of samples	55	53
	Min. Length	2	4
	Max. Length	63	75
	Ave. Length	22.86	31.35
	Std.Dev.	14.48	14.29
Readability	Flesch Score	52.79	41.93
	Style Description	Complex	V.Complex
Grammatical	Mean Error	1.85	2.05
	Error-free frequency	16	18
	Max Errors	8	7
Avg Similarity	Token-wise	7%	23%
	Semantic	55.4%	70.3%
Ordering Correlation Analysis	SHAP	17.28	34.75
	Action group	–	78.03
Alternatives Preferred	n-XAI^{B2} > n-XAI^{B1}	–	55%
	Chart > Tabular	84.61%	–

Table 1: Response analysis by cohort

Response Statistics for 108 responses were analysed after excluding brief or non-compliant submissions. Length statistics were calculated without 4 max length outliers. A minimum length was 2 to accommodate responses like "Higher total_rec_prncp"

Readability is assessed using the Flesch score [8] which calculates a value between 0-100, where lower values indicate lower readability due to complex constructs. We find that Cohort2 participants responded with more complex explanations i.e. Style Description = Difficult, where text is likely to contain, longer sentences, technical vocabulary, and more specialised ideas.

Grammatical error analysis was conducted using the Python LanguageTool (for spelling and grammar). Cohort2 outperformed Cohort1 suggesting that basic Natural-XAI formats of n-XAI^{B1} and n-XAI^{B2} had influenced the cohort to write more grammatically correct explanations, compared to Cohort1 who only saw tabular and graphical formats. Token-wise average similarity between Cohort2's text compositions and n-XAI^{B1} and n-XAI^{B2} showed a 23% match, indicating that participants did not merely copy-paste the content.

Semantic similarity uses the all-MiniLM-L6-v2 model from the sentence-transformers library, to assess the semantic similarity between both cohorts' authored text to the n-XAI^{B1} and n-XAI^{B2} Natural-XAI forms. As expected results showed an average resemblance of 55.4% for Cohort1 and 70.28% for Cohort2. Since n-XAI^{B1} and n-XAI^{B2} were factually correct, a greater resemblance to these baselines can be used as a reliable measure of the plausibility of the participant-generated explanations.

Correlation between the feature ordering methods of n-XAI^{B1} and n-XAI^{B2}, which are based on the SHAP and Action Group models, and the ordering of features in the text generated by Cohort2 were compared using the Spearman's rank order coefficient. The analysis indicated that Cohort2 preferred grouping their text by actionability verbs before ordering features by SHAP order. However, Cohort1 did not exhibit a significant correlation in the ordering of features, despite being presented with SHAP ordering through the tabular and graphical alternatives.

Alternatives Preferred relate to Task2, and findings suggest that 55% of cohort 2 preferred n-XAI^{B2} and 84.61% of Cohort1 preferred chart over tabular explanation.

3.4 Qualitative Evaluation - Thematic Analysis

The qualitative evaluation consists of two steps: 1) analyse all authored explanations manually to identify common themes; and 2) conduct a clustering to compare automatically created clusters with the manually formed themes. The manual analysis was conducted using a thematic analysis approach, where the content was coded by two researchers. Common themes were then aggregated and defined. This process identified five content-related and two structure-related themes. Table 2 lists the themes, with examples and the frequency of each theme in the responses. The "Content" theme identified five variations in the text authored by the study participants, based on the presence or absence of feature values with reference to the query, counterfactual, or both. The "Vague" theme refers to responses where the recommended actions remained unclear (i.e. the exact amount by which to change was missing), while "Reduced" indicates responses where mentions of subsets of features and actions were missing. The "Structure" themes were examined using ordinal adverbs and ordering styles (e.g. bullet pointing), and the use of unusual/interesting "actionable words" (verbs).

Agglomerative clustering was used on responses to assess alignment with manually extracted themes. The average linkage method is used to merge the clusters based on their similarity. The textual responses were encoded using the pre-trained all-MiniLM-L6-v2 model from the sentence transformers library to create embeddings, which were used as input to the clustering algorithm. The resulting clusters were compared to the manually identified themes, and the similarity was assessed (see Table 2). We found that the clustering results were mostly consistent with the manually identified themes, with the "Vague" theme split into many smaller clusters separate from the other themes. Specifically, we found that themes 2, 3, and 4 were highly similar and mostly clustered together in clusters 0 and 2. These related to the differences as to whether or not feature names and their values were mentioned with reference to the counterfactual, query or both. For instance, consider these two alternative sentences to convey an actionable change: "s1: your loan amount of 13,5K needs to be reduced to 12K" and "s2: you must reduce your loan to 12K". Here s1 uses feature values from both the query and counterfactual whilst s2 only refers to the counterfactual. Results show that actionable changes are more frequently referenced with counterfactual values and feature names, rather than using query values.

3.5 Findings for Counterfactual Natural-XAI

Our analysis suggests that the use of Natural-XAI formats, such as n-XAI^{B1} and n-XAI^{B2}, may have had a positive impact on the quality and accuracy of the explanations generated by users, as evidenced by the higher level of grammatical correctness observed in Cohort2 and the preference for organising text by actionability verbs before ordering features by SHAP order. Accordingly, to help with discourse planning we can adopt such ordering strategies to organise sentences. For sentence planning, employing action verbs like "negotiate" and phrases like "strive to" highlights the need for actionability concepts that can capture varying degrees of actionability. The factual similarity and the variation in semantic similarity observed in Cohort2's text to n-XAI^{B1} and n-XAI^{B2} suggest value in further studying Cohort2's content and structural organisation to derive generalisable Natural XAI templates.

Our thematic analysis integrates findings from both content and structure analysis to inform two components. First, we develop a taxonomy in Section 4.1 that discerns different categories of feature actionability. Second, we create Natural-XAI templates in Section 4.2 that encapsulate common constructs based on ordinal adverbs, comparative adjectives, and action verbs. These templates are inspired by and take ideas from the human-authored text in the user study.

4 Counterfactual Natural-XAI Method

Our Natural-XAI template-based NLG method, n-XAI^{T}, uses a Feature Actionability Taxonomy (FAT) to guide template selection. This taxonomy, which categorises features based on their level of mutability enables the use of appropriate sentence constructs for sentence planning. FAT is informed by findings from the user study and is used to determine the template structure for each feature, which provides alternatives for slot filling. These options include using both query and counterfactual values, using only counterfactual values, employing ordinal adverbs or bullet points, and selecting alternative forms of action terms such as "increase" or "raise". Once templates are identified they are presented in order of taxonomic category using SHAP weights for within category ordering.

4.1 Feature Actionability Taxonomy (FAT)

FAT was defined using a data-driven methodology that relied on examining features extracted from six datasets [7, 15] related to Fair AI. They span three distinct domains, with each feature analysed to determine suitable actionability categories. The resulting categories and distributions are summarised in Table 3. We observe that, while current cf-XAI systems consider recipients can change all features directly, such features appear least, highlighting the need for n-XAI^{T}.

FAT category definitions appear in Figure 2, where features are categorised into two groups: those that the recipients of counterfactuals can change through their actions (i.e., *Mutable*) and those that cannot be changed (i.e., *Immutable*). This categorisation allows n-XAI^{T} to carefully consider how to present information for each category, even for immutable features. While recipients of explanations cannot change immutable features, classifying them in the taxonomy allows NLG systems to present these as “factual explanations” (in contrast to suggesting a counterfactual-driven change). Recognising likely ethical concerns about presenting certain features like race or ethnicity, we classify them as *Immutable Sensitive* in n-XAI^{T} enabling the system to consider them when generating explanation texts, and exposing bias with human-in-the-loop [6]. We further categorise those features that recipients can change into those that will be directly impacted by actions (i.e. *Mutable Directly)*, and those that the recipient can only change by acting on another feature (*Mutable Indirectly*), such as discretionary income, which changes by increasing salary or decreasing expenses. The latter category can also be useful should the XAI system provide causal knowledge.

4.2 FAT Template-based NLG

In n-XAI^{T} we adopt a three-staged template-based NLG approach of sentence planning, surface realisation and discourse planning [21]. The FAT is used for sentence planning where relevant Feature Sentence Templates (see Table 4) are identified based on the feature's categorisation in the FAT. Thereafter a mapping of content and structure themes to templates guides the surface realisation step of NLG. This is detailed in Table 4, where sentences explaining mutable features use content themes C2, C3, or C4, meaning they can opt

Content Theme	Example	Frequency	Cluster(s)
C1:**Vague** action	The customer should **lower** their **collection fee** and their **recoveries** as well as **increase** their **total payment** and their **principal**	23	1, 4, 5, 3
C2: **Counterfactual** values only	For your application to be improved, you would need to **increase total_pymnt to 12268.08** and **total_rec_prncp to 10925**.You would need to **decrease recoveries and collection_recovery_fee to 0**	39	0, 2
C3: **Counterfactual and Query** values	For a successful application you should increase total payment **to 12268.08 from 5040.00** while also increasing principal **to 10925.00 from 1492.93** and decrease recoveries and collection fee to 0	5	0, 2
C4: **Combined** T2 & T3	For your application to be successful you should increase your **principal to 14999.99** and your **fico to 669.00 from 585** while **decreasing recovery** and **collection recovery fees to 0** and **decreasing** your **last payment amount to 33.35**	3	0, 2
C5: **Reduced** explanation	**Reduce** the **recoveries** to **0** and **total_rec_prncp** needs to be **higher**	20	0, 5
Structure Theme	Example	Frequency	Cluster(s)
S1. Use of ordinal adverbs / ordering with Bullet-pointing	For your application to be accepted, you will need to prioritize the following in the **order** given: **1) increase** the total principal received (total_rec_prncp) to 14,999 **2) decrease** the recoveries (recoveries) to 0.00 **3) decrease** the collection recovery fee (collection_recovery_fee) to 0.00 **4) increase** the last payment amount (last_pymnt_amnt) to 11,448.66	12	2
S2. Creative Action Verbs	Cust1 should **pay off** their recoveries and **negotiate** to have their charge off **removed**	4	2

Table 2: Content and structure themes with examples and alignment to response clusters.

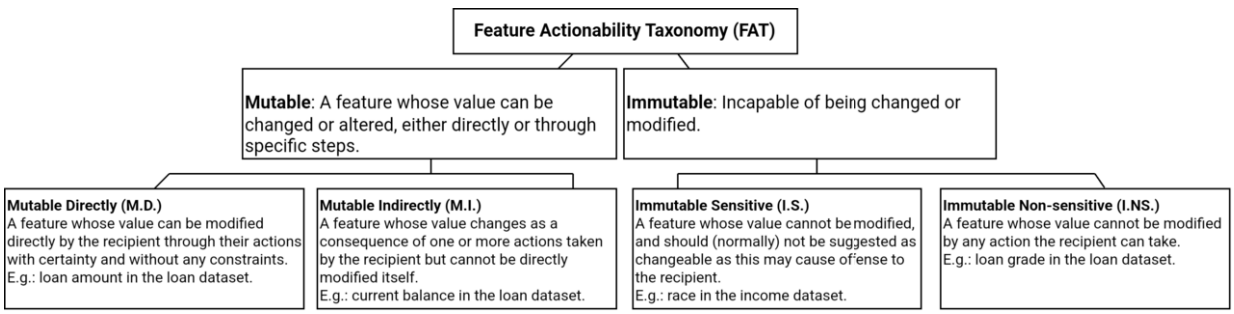


Figure 2: The Feature Actionability Taxonomy (FAT).

Domain	Dataset	#Features	#Features by Category			
			M.D.	M.I.	I.S.	I.NS.
Health	Diabetes	8	1	4	1	2
	Breast Cancer	9	0	2	0	7
Education	OULAD	8	0	2	4	2
	Student (UCI)	31	4	12	6	9
Finance	Loan Approval	67	4	50	0	13
	Income	8	0	2	1	5
Total		131	9	72	12	38

Table 3: Dataset overview with feature count and actionability category distribution based on the FAT knowledge in Figure 2. Categories: Mutable Directly (M.D.), Mutable Indirectly (M.I.), Immutable Sensitive (I.S.), and Immutable Non-sensitive (I.N.).

whether to include values from the query along with the counterfactual value. Sentences explaining immutable features use theme C1 or C5. Further, mutable features are presented using structure theme S1 to make it easier for users to focus on them, with immutable features generally using S2. Additionally, we incorporate positive reinforcement language, based on insights from psychology research [3], into the immutable non-sensitive feature template. As a result, the template generates positive explanations, such as "your loan has a high chance of approval", through the use of surface realisation techniques. This approach effectively conveys information while maintaining a supportive tone, as opposed to negative language such as "your loan has a less chance of rejection". Thereafter, for discourse planning, sentences are grouped by taxonomic category and sorted by a feature-based ordering (such as SHAP). For a given query, counterfactual pair, the pipeline in Figure 3 shows the input and output for each feature at each of the three stages. At deployment, the explanation is customised with an initial sentence specific to the dataset, including the number of actionable features, and a domain-specific epilogue, such as "Stay healthy!" for a health domain and "Good luck with your loan!" for a finance domain.

5 Evaluating Actionability in Natural-XAI

This study aims to evaluate the utility of the n-XAI^T approach to Natural-XAI using actionability knowledge. It was conducted across three new test domains: health (heart disease dataset), finance (credit risk - kaggle) and education (student performance - kaggle). Manual instantiation of the taxonomy for each dataset is summarised in Table 5. Using cf-XAI system (DICE), six scenarios were devised, two per domain, and a comparative analysis was conducted with two cohorts: Cohort1 using a baseline n-XAI^B and Cohort2 using n-XAI^T. We improved the formatting of n-XAI^{B1} based on previous findings, resulting in the creation of n-XAI^B. The key difference between the two is in the handling of mutable features. n-XAI^T generates factual explanations for features in categories I.N and I.NS, e.g., "... your parental level of education is a contributing factor to the risk of obtaining an overall credit risk score of below average". In contrast n-XAI^B treats all features as actionable e.g., " ... change your parental level of education to bachelor's degree". Further examples for each domain are provided in the section 6.

5.1 User Study Protocol

Our study consisted of a non-randomised between-subjects design with three domains, each with two scenarios. A total of 60 participants were prescreened and assigned to one of two cohorts based on

Template Variables with Synonym Examples	
`VERB={Take|Initiate|Undertake|Pursue|Negotiate}`	
`OBJECT={steps|measures|actions|}`	
`ACTION={Pos: (increase|improve|raise)| Neg:(decrease|reduce)}`	
`COMPARATIVE={Pos: (increase|higher|better) | Neg: (decrease|lower|worse)}`	
`OUTCOME={undesired:(rejected|fail) | desired:(accepted|pass)}`	
`FEATURE= feature name in dataset, QUERY_VALUE= feature value from query, CF_VALUE= feature value from counterfactual, POSSESSIVE={Your}`	
Actionability Category	*Feature Sentence Template*
Mutable Directly	1. `{ACTION} {FEATURE} from {QUERY_VALUE} value to {CF_VALUE}` 2. `{ACTION} {FEATURE} to {CF_VALUE}`
Mutable Indirectly	1.`{VERB} {OBJECT} to {ACTION} {FEATURE} from {QUERY_VALUE} to {CF_VALUE}` 2.`{VERB} {OBJECT} to {ACTION} {FEATURE} to {CF_VALUE}`
Immutable Non-sensitive	`Having a value of {CF_VALUE} for {FEATURE} would provide a {COMPARATIVE} chance of {DESIRED_OUTCOME} compared to a value of {QUERY_VALUE}`
Immutable Sensitive	`{POSSESSIVE} {FEATURE} has contributed to {OUTCOME}`

Table 4: Templates from mapping Content and Structure themes to Feature Actionability Taxonomy categories.

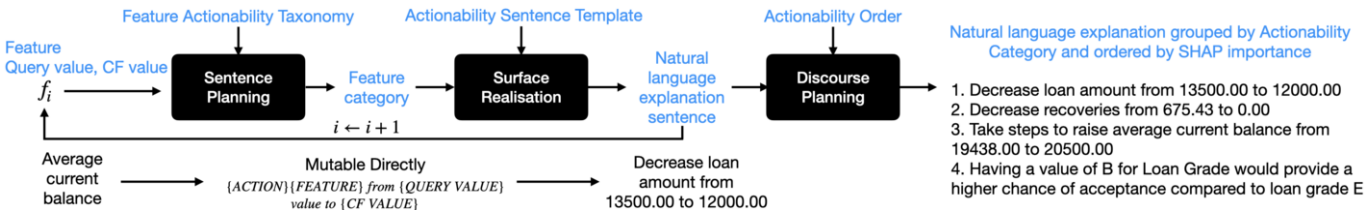


Figure 3: NLG pipeline for n-XAI^T, using FAT and Feature Sentence Templates.

Domain	Dataset	#Features	#Features by Category			
			M.D.	M.I.	I.S.	I.NS.
Health	Heart	13	0	8	2	3
Education	Student (Kaggle)	8	5	0	3	0
Finance	credit	11	2	1	7	1
Total		32	7	9	12	4

Table 5: Overview of test datasets in each domain, displaying the number of features and the actionability category distribution.

their domain knowledge alignment, with 20 participants allocated to each domain (and divided into 10 per cohort). Cohort1 was presented with the n-XAI^B explanations for the two scenarios in their respective domain, while Cohort2 was presented with explanations generated using the n-XAI^T pipeline. Each cohort completed the same two scenarios in their respective domain. The between-subjects design allowed us to compare the effectiveness of the two types of explanations across different domains and independent cohorts while controlling for individual differences between participants.

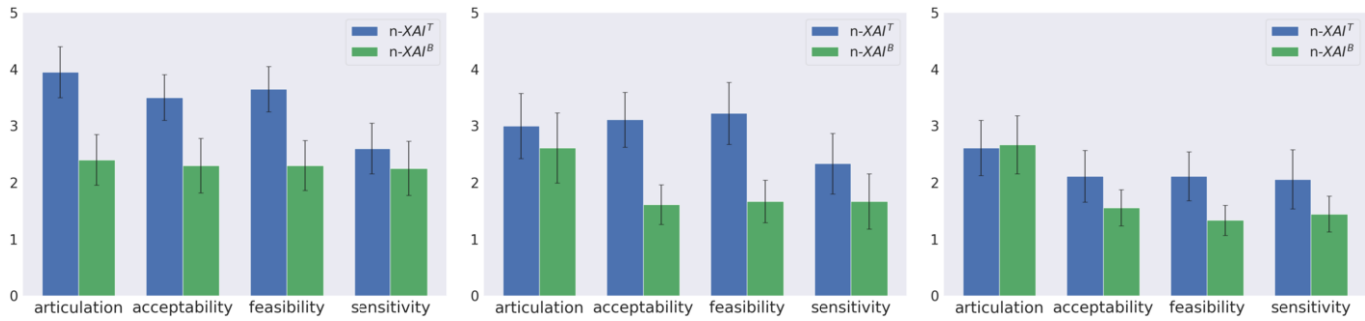
After recruiting participants for our study, we carefully screened their responses to ensure that only reliable and qualified participants were included. Upon close examination of the responses, we discarded six responses that failed the attention test, resulting in a final sample of 54 participants. Our participant pool consisted of native English speakers from the USA, UK, Ireland, Australia, Canada, and New Zealand, with an equal distribution of participants based on sex. Additionally, we prescreened participants on age, ranging from 18 to 74 years old for the health and education domains, and from 28 to 74 years old for the finance domain. We also prescreened them based on relevant domain expertise.

Given the unique nature of each domain, we conducted separate user studies for each domain, with different groups of participants completing the scenarios for each domain. After being presented with the two scenarios, participants were asked to answer 4 questions related to the explanations they received, evaluating them based on the following criteria: Articulation, Acceptability, Feasibility, and Sensitivity. Participants rated their response to each question using a 5-point Likert scale (1-5) and were also asked to provide their rationale for selecting a specific rating. The results from each domain were analysed separately, allowing us to draw distinct conclusions and insights for each domain. The user studies were conducted on the Prolific platform, a reputable source for obtaining high-quality data from diverse participant samples using the prolific platform.

5.2 User Study Outcomes

Figure 4 presents the results for each domain on the 4 criteria with 95% confidence bars. The Health domain received the highest ratings, followed by Education and Finance. We conducted a Shapiro-Wilk test to determine the normal distribution of both cohorts for each domain, followed by Levene's test to assess variance equality. Based on the outcomes of these tests, we used the Wilcoxen Signed Rank test to evaluate the significance of each domain's cohorts for the 4 criteria. Our analysis revealed significant improvements in the Feasibility ratings for explanations across all domains. Incorporating actionability knowledge has also resulted in significantly Acceptable recommendations in the Education and Health domains. Although n-XAI^T outperformed n-XAI^B in terms of Articulation and Sensitivity in all domains, the differences were not statistically significant.

All textual responses were thematically analysed to identify areas for improvement in the explanations provided. Table 6 shows the identified themes and provides example participant responses from both cohorts. Further analysis of these responses, indicates that users' subjective perceptions of their expectations significantly influenced their feedback, with this trend being particularly noticeable in the Fi-



(a) Healthcare Domain (b) Education Domain (c) Finance Domain

Figure 4: Comparison of participant responses between n-XAI^T and n-XAI^B across all domains.

Theme	*Domain(s)*	*Example of user suggestions*
Feasibility Knowledge	H, E, F	"The 4 **steps** to take are **better at indicating what can be done** to offer practical advice" - H2 "The patient should be given **suggestions** for how to decrease their resting BP " - H2
Handling Protected Features	H, E, F	"The suggestion to reduce her age and subtype of thalassemia **are essentially discriminatory** and **should not be included**." - H1 "It is **highly unethical** to suggest someone changes their sex" - E1 "Statements were **factual & not offensive** although they could be interpreted this way" - E2, "Not taking into account work history of 5 years & consistent employment is **unethical**" - F2
Hybrid Explanations	H, E	"...**reasons** for these need to be **clear**, and the **means with which to achieve** them " - F2 "Suggest **reasons** as to why he may be behind but these are not evidenced" - E2
Personalised Revision of Counterfactuals	F	"...it would be **hard to change** ownership to rent as you would need to sell your house" - F2 "Considering her age and income /risk again **not sure this is suitable** for subjects unless they work in finance and **understand these terms**" - F2
Feasibility Impact	F	"The potential **negative outcomes are not provided** to the user by the AI, meaning that harm may come to them by following advice that they do not understand" - F1 "...limit the suggestions to things that are **actually possible** for a person to do" - F1
Structure and Context	H, E, F	"Tense & use of **capitals** & upper case letters prevent advice from being well articulated" - F2 "Suggestion is too wordy - E2"

Table 6: User response analysis: explanation improvement themes (column 1), domain-wise theme coverage (column 2), and mix of user quotes from the domains, Healthcare (H), Education (E) & Finance (F) and Cohort (1 or 2) (column 3).

nance domain (acceptability criteria). For instance, those who saw n-XAI^T explanations wanted more detailed strategies and explanations on how to achieve the suggested changes, which would require extensive domain knowledge beyond the scope of this study (see quotes in themes Feasibility Knowledge and Hybrid Explanations for example). Conversely, those who saw baseline explanations found them to be less acceptable and feasible due to their perceived lack of rationality and real-world applicability (theme Handling Protected Features).

Regarding Sensitivity and articulation, in both n-XAI^T and n-XAI^B, concerns remained about the need for personalisation and ethically considerate explanations. For instance, recommending actions related to changing ownership to rent (from a mortgage) applies only to those that don't already own a house; but to support this level of inference requires the XAI system to have access to the user's background, which may not always be practical. These observations emphasise the user-dependent nature of applying taxonomy definitions. To address this in deployed systems, an interactive iterative process with the user is necessary, enabling adaptive feature actionability categories based on individual circumstances. This approach aligns with prior research advocating for interactive and personalised XAI systems [26]. Many responses about the structure emphasised the need for enhanced articulation. Suggestions included using intuitive names, and employing appropriate capitalisation, and adopting understandable units of measurement. Interviews to understand expectations were desirable but not possible due to the online evaluation. Overall we found common themes related to wanting more guidance on how to achieve suggested changes across all domains. This was especially true of Finance. The use of factual explanation style for sensitive features and counterfactual action recommendations for others in n-XAI^T was liked by Cohort2 participants, especially in Health and Education (see Hybrid theme).

6 Conclusion

This paper presents n-XAI^T, a Natural-XAI approach that enhances natural language counterfactual explanations with actionability knowledge, resulting in better results for articulation, sensitivity, feasibility, and actionability criteria. Guided by an actionability taxonomy, and feature attribution weights, n-XAI^T selects feature-based sentence-level templates to generate natural explanations. Results from a user study (n=60) provide useful guidelines for counterfactual XAI platforms to enhance the feasibility of recommended actions, including incorporating mixed XAI strategies (factual and counterfactual), use of domain knowledge to guide users on how to implement recommended changes and personalising actionability categories to individual preferences. The taxonomy is open-sourced for community contributions in new domains, and future work includes extending the taxonomy, simplifying structures for accessibility, and tailoring language templates to user personas.

Acknowledgements

This research was partially funded by the iSee project (https://isee4xai.com/). iSee is an EU CHIST-ERA project which received funding for the UK from EPSRC under grant number EP/V061755/1. We are grateful to the University Research Ethics Committee for approving the study protocol. Special thanks to all our participants, who were fully informed about their right to withdraw and the handling and storage of their data. Finally, we wish to thank all our anonymous reviewers for their valuable feedback that helped to improve the paper.

Resources

All code and materials from this paper are available from http://github.com/pedramsalimi/NLGXAI.

References

[1] Solon Barocas, Andrew D Selbst, and Manish Raghavan, 'The hidden assumptions behind counterfactual explanations and principal reasons', in *Proc. Conf. on fairness, accountability, and transparency*, pp. 80–89, (2020).

[2] Brughmans, Dieter and Leyman, Pieter and Martens, David, 'NICE : an algorithm for nearest instance counterfactual explanations', *Data mining and knowledge discovery*, (2023).

[3] Raykhona Burieva, 'The effectiveness of teaching writing to the students with the technique "rewards and positive reinforcement"', *Academic research in educational sciences*, (1), 229–232, (2020).

[4] Ruth M.J. Byrne, 'Counterfactual thought', *Annual Review of Psychology*, **67**(1), 135–157, (2016).

[5] Susan Craw, Stewart Massie, and Nirmalie Wiratunga, 'Informed case base maintenance: A complexity profiling approach', in *Proc. 22nd AAAI Conf. on AI*. AAAI Press, (2007).

[6] Jonathan Dodge, Q. Vera Liao, Yunfeng Zhang, Rachel K. E. Bellamy, and Casey Dugan, 'Explaining models: An empirical study of how explanations impact fairness judgment', in *Proc. 24th Int. Conf. on Intelligent User Interfaces*, p. 275–285. ACM, (2019).

[7] Dheeru Dua and Casey Graff. UCI machine learning repository, 2017.

[8] Rudolph Flesch, 'A new readability yardstick.', *Journal of applied psychology*, **32**(3), 221, (1948).

[9] Riccardo Guidotti, 'Counterfactual explanations and how to find them: literature review and benchmarking', *Data Mining and Knowledge Discovery*, 1–55, (2022).

[10] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Andrea Madotto, and Pascale Fung, 'Survey of hallucination in NLG', *ACM Computing Surveys*, (2022).

[11] Amir-Hossein Karimi, Bernhard Schölkopf, and Isabel Valera, 'Algorithmic recourse: from counterfactual explanations to interventions', in *Proc. ACM Conf. on fairness, accountability, and transparency*, pp. 353–362, (2021).

[12] Mark T Keane and Barry Smyth, 'Good counterfactuals and where to find them: A case-based technique for generating counterfactuals for XAI', in *Int. Conf. on CBR*, pp. 163–178. Springer, (2020).

[13] Sawan Kumar and Partha Talukdar, 'NILE : Natural language inference with faithful natural language explanations', in *Proc. 58th Annual Meeting of the ACL*, pp. 8730–8742, Online, (July 2020). ACL.

[14] Anna S Law, Yvonne Freer, Jim Hunter, Robert H Logie, Neil McIntosh, and John Quinn, 'A comparison of graphical and textual presentations of time series data to support medical decision making in the neonatal intensive care unit', *Journal of clinical monitoring and computing*, **19**(3), 183–194, (2005).

[15] Tai Le Quy, Arjun Roy, Vasileios Iosifidis, Wenbin Zhang, and Eirini Ntoutsi, 'A survey on datasets for fairness-aware machine learning', *WIREs Data Mining and Knowledge Discovery*, **12**(3), e1452, (2022).

[16] Scott M Lundberg and Su-In Lee, 'A unified approach to interpreting model predictions', *Advances in neural information processing systems*, **30**, (2017).

[17] Divyat Mahajan, Chenhao Tan, and Amit Sharma, 'Preserving causal constraints in counterfactual explanations for machine learning classifiers', in *Proc. CausalML workshop at NeurIPS*, (2019).

[18] Ramaravind K Mothilal, Amit Sharma, and Chenhao Tan, 'Explaining machine learning classifiers through diverse counterfactual explanations', in *Proc. Conf. on Fairness, Accountability, and Transparency*, pp. 607–617, (2020).

[19] Marian Petre, 'Why looking isn't always seeing: readership skills and graphical programming', *Communications ACM*, **38**(6), 33–44, (1995).

[20] Rafael Poyiadzi, Kacper Sokol, Raul Santos-Rodriguez, Tijl De Bie, and Peter Flach, 'Face: feasible and actionable counterfactual explanations', in *Proc. AAAI/ACM Conf. on AI, Ethics, and Society*, pp. 344–350, (2020).

[21] Ehud Reiter, 'An architecture for data-to-text systems', in *Proc. 11th European Workshop on NLG (ENLG 07)*, pp. 97–104, Saarbrücken, Germany, (June 2007). DFKI GmbH.

[22] Lya Hulliyyatus Suadaa, Hidetaka Kamigaito, Kotaro Funakoshi, Manabu Okumura, and Hiroya Takamura, 'Towards table-to-text generation with numerical reasoning', in *Proc. 59th Annual Meeting of the ACL and 11th Int. Joint Conf. on NLP (Volume 1)*, pp. 1451–1465, (2021).

[23] Berk Ustun, Alexander Spangher, and Yang Liu, 'Actionable recourse in linear classification', in *Proc. Conf. on fairness, accountability, and transparency*, pp. 10–19, (2019).

[24] Sahil Verma, Varich Boonsanong, Minh Hoang, Keegan E Hines, John P Dickerson, and Chirag Shah, 'Counterfactual explanations and algorithmic recourses for machine learning: A review', in *Proc. ML-RSA workshop at NeurIPS*, (2020).

[25] Sandra Wachter, Brent Mittelstadt, and Chris Russell, 'Counterfactual explanations without opening the black box: Automated decisions and the gdpr', *Harv. JL & Tech.*, **31**, 841, (2017).

[26] Anjana Wijekoon, Nirmalie Wiratunga, Kyle Martin, David Corsar, Ikechukwu Nkisi-Orji, Chamath Palihawadana, Derek Bridge, Preeja Pradeep, Belen Diaz-Agudo, and Marta Caro-Martínez, 'CBR driven interactive XAI', in *Proc. 31st Int. Conf. on CBR*. Springer, (2023).

[27] Nirmalie Wiratunga, Anjana Wijekoon, Ikechukwu Nkisi-Orji, Kyle Martin, Chamath Palihawadana, and David Corsar, 'Discern: Discovering counterfactual explanations using relevance features from neighbourhoods', in *2021 IEEE 33rd Int. Conf. on Tools with Artificial Intelligence (ICTAI)*, pp. 1466–1473. IEEE, (2021).

[28] Linyi Yang, Eoin Kenny, Tin Lok James Ng, Yi Yang, Barry Smyth, and Ruihai Dong, 'Generating plausible counterfactual explanations for deep transformers in financial text classification', in *Proc. 28th Int. Conf. on Computational Linguistics*, pp. 6150–6160, (2020).