Natural XAI: Generating Feasible, Actionable, and Causally-Aware Counterfactual Explanations in Natural Language

Pedram Salimi[1,2,*]

[1]Robert Gordon University, Garthdee Rd, Aberdeen, AB10 7GJ, United Kingdom

Abstract

Counterfactual explanations have become a significant component in eXplainable AI (XAI), offering intuitive "what if" scenarios. However, typical numeric or tabular outputs can be vague to non-technical audiences. Additionally, many counterfactual methods ignore causal relationships or suggest inactionable changes such as "be younger by five years," raising concerns over realism and ethics. To address these issues, we propose a holistic approach that integrates a Feature Actionability Taxonomy (FAT) and causal discovery into counterfactual generation, thereby ensuring realistic, ethically sound, and semantically transparent explanations in natural language. We further introduce an interactive, agentic workflow enabling users to iteratively refine constraints. Through extensive user studies, pilot evaluations, and synergy with Case-Based Reasoning (CBR), this approach yields explanations that are accessible, trust-enhancing, and practically useful in domains such as healthcare, finance, and education.

Keywords

Explainable AI, Counterfactual Explanations, Causality, Natural Language Generation

1. Introduction

Explainable AI (XAI) aims to mitigate the opacity of complex models such as deep neural networks and ensemble methods. Among the many post-hoc explanatory approaches, Counterfactual Explanations have emerged as particularly intuitive and user-friendly. By illustrating the minimal alterations required to flip an AI model's outcome (e.g. from "reject" to "approve"), counterfactuals enable stakeholders to understand *how* a model's decision might be changed in practice.

Despite these benefits, several key challenges persist:

- **Complexity in Presentation:** Many methods present counterfactuals numerically (e.g. in tables or lists of feature tweaks), which can be difficult for non-technical users to interpret and understand.
- **Inactionable or Unethical Suggestions:** Traditional approaches often recommend modifying immutable or sensitive features (such as gender or race), conflicting with ethical and practical constraints.
- **Causal Oversight:** Ignoring real-world causal relationships can generate implausible recommendations, such as simultaneously increasing one feature and reducing another in ways that are contradictory or infeasible.

Our research addresses these shortcomings by integrating three pivotal elements:

(i) **Feature Actionability Taxonomy (FAT)**, which systematically categorises features by their mutability and sensitivity;

(ii) **Causal Discovery and Integration**, ensuring that recommended changes respect underlying cause-effect relationships;

ICCBR DC'25: Doctoral Consortium at ICCBR-2025, July, 2025, Biarritz, France

*Corresponding author.

p.salimi@rgu.ac.uk (P. Salimi)
https://rgu-repository.worktribe.com/person/1607134/pedram-salimi (P. Salimi)
0000-0002-6572-8516 (P. Salimi)

(iii) **Natural Language Generation (NLG)**, presenting counterfactuals in user-friendly textual form. Finally, we propose an agentic framework, in which users can iteratively refine or reject certain suggestions, receiving updated counterfactuals each time. This paper outlines our progress towards a more transparent, practical, and interactive XAI, referred to here as Natural-XAI.

2. Motivation and Related Work

Modern decision-support systems increasingly rely on complex, "black-box" models. Trust and adoption highly rely on effective explanation. Counterfactual methods stand out for their simplicity and actionability: rather than simply stating "your loan was denied because of X," a counterfactual might say "if your monthly income increased by $200, your loan would have been approved."

However, user trust can degrade if suggested feature changes are not plausible or ethically-sound. For instance, suggestions like "be ten years younger" or "change your race" are not only inactionable but also unethical in many contexts. Prior works such as DiCE [1] address diversity in counterfactual explanations, while FACE [2] focuses on data manifold feasibility. Causality-oriented approaches (e.g. [3]) endeavour to align recommendations with real-world cause-and-effect. Meanwhile, template-based NLG [4] has shown promise for generating explanations that resonate more strongly with end-users compared to raw numeric or tabular outputs.

Despite these advancements, few solutions comprehensively unify actionability, causality, and natural language. Our approach bridges these dimensions, while also employing Case-Based Reasoning (CBR), which complements counterfactuals by providing exemplars of similar cases and how they differ from the current instance.

3. Proposed Approach

Building on the limitations of purely numeric or tabular counterfactuals, our proposed framework introduces a holistic pipeline that progressively addresses actionability, presentation, causal coherence, and iterative user engagement. We begin by specifying how features can and cannot be changed in the Feature Actionability Taxonomy (FAT). Next, we incorporate these constraints into a template-based Natural Language Generation (NLG) module to communicate potential changes in a user-friendly manner. We then extend the counterfactual generation process to handle causal dependencies so that recommended interventions remain realistic. Finally, we embed all these components into an agentic workflow, allowing iterative dialogue and refinement of constraints. The sections that follow describe these four core pillars in turn.

3.1. Feature Actionability Taxonomy (FAT)

A foundational element of our approach is the **Feature Actionability Taxonomy (FAT)**, which classifies features into three categories based on how realistically they can be modified:

- **Directly Mutable (DM):** Easily adjustable features such as requested loan amount.
- **Indirectly Mutable (IM):** Features that are changeable but only through more extended or intricate actions (e.g. educational level, occupation type).
- **Non-sensitive or Sensitive Immutable:** Characteristics like age or gender, which typically cannot or should not be changed for ethical or practical reasons.

FAT was defined using a data-driven methodology that relied on examining features extracted from six datasets [5, 6] related to Fair AI. They span three distinct domains, with each feature analysed to determine suitable actionability categories [4].

By explicitly encoding each feature's mutability, FAT automatically filters out suggestions that violate real-world or ethical constraints, thereby increasing user trust in the system [7]. Figure 1 illustrates this taxonomy, showing how each feature is funnelled into a relevant branch, thus ensuring that impossible or unethical recommendations. for example, *"reduce your age by 10 years"* are never generated [4].

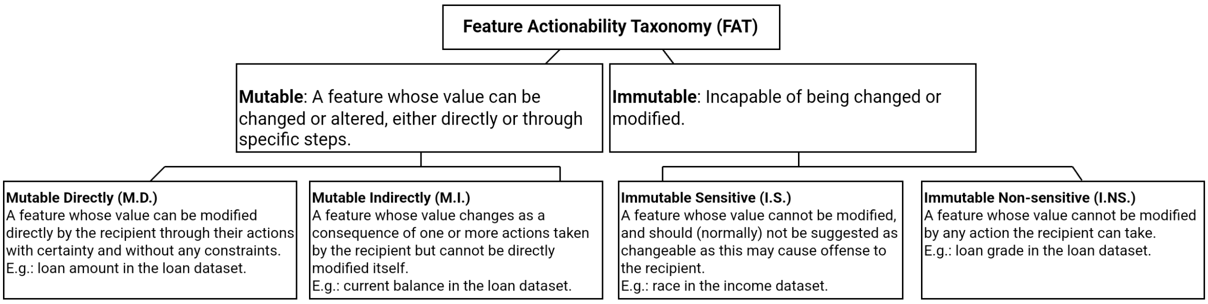
Feature Actionability Taxonomy (FAT)
Mutable: A feature whose value can be changed or altered, either directly or through specific steps.
Immutable: Incapable of being changed or modified.
Mutable Directly (M.D.)
A feature whose value can be modified directly by the recipient through their actions with certainty and without any constraints.
E.g.: loan amount in the loan dataset.
Mutable Indirectly (M.I.)
A feature whose value changes as a consequence of one or more actions taken by the recipient but cannot be directly modified itself.
E.g.: current balance in the loan dataset.
Immutable Sensitive (I.S.)
A feature whose value cannot be modified, and should (normally) not be suggested as changeable as this may cause offense to the recipient.
E.g.: race in the income dataset.
Immutable Non-sensitive (I.NS.)
A feature whose value cannot be modified by any action the recipient can take.
E.g.: loan grade in the loan dataset.

Figure 1: An illustrative overview of the Feature Actionability Taxonomy, detailing different types of features and their permissible actions.

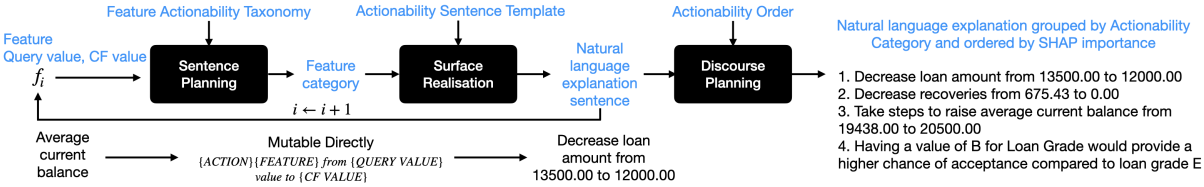
Feature Actionability Taxonomy
Actionability Sentence Template
Actionability Order
Feature
Query value, CF value
f_i
Sentence Planning
Feature category
Surface Realisation
Natural language explanation sentence
Discourse Planning
$i \leftarrow i+1$
Natural language explanation grouped by Actionability Category and ordered by SHAP importance
1. Decrease loan amount from 13500.00 to 12000.00
2. Decrease recoveries from 675.43 to 0.00
3. Take steps to raise average current balance from 19438.00 to 20500.00
4. Having a value of B for Loan Grade would provide a higher chance of acceptance compared to loan grade E
Average current balance
Mutable Directly
{ACTION}{FEATURE} from {QUERY VALUE} value to {CF VALUE}
Decrease loan amount from 13500.00 to 12000.00

Figure 2: NLG pipeline for Natural-XAI, using FAT and Feature Sentence Templates.

3.2. Natural Language Generation (NLG)

We employ a three-stage, template-based approach to generate user-friendly counterfactual explanations in natural language [8]. First, *sentence planning* leverages the Feature Actionability Taxonomy (FAT) to identify appropriate templates based on whether a feature is directly mutable, indirectly mutable, or immutable. Second, *surface realisation* populates these templates with user-specific details while taking into account thematic preferences (e.g. emphasising feasibility or positive language). For instance, immutable features are described using templates that convey encouragement rather than negative warnings, following recommendations from psychology research [9]. Finally, *discourse planning* groups and orders the individual sentences by category (e.g. mutable vs. immutable) and by feature-based priority (e.g. a SHAP ranking), generating a coherent paragraph-level explanation.

Figure 2 (referenced in the text) illustrates the full pipeline, from matching features to templates through to sentence sorting and domain-specific epilogues. For example, a finance application might conclude with "Good luck with your loan!" while a healthcare scenario might close with "Stay healthy!" This layered approach ensures that explanations are both contextually accurate and reassuring in tone.

3.3. Causally-Aware Counterfactual Explanations

While FAT addresses the actionability of individual features, many real-world scenarios involve intricate causal dependencies. Changing one feature (e.g. "Education Level") may, in fact, affect others (e.g. "Occupation," "Income"), which in turn could impact the ultimate prediction (e.g. "Loan Approval"). To capture these relationships, we integrate causal discovery, for instance, the DECI framework [10] or domain-informed causal graphs. In this work the causal graph is learned with DECI, a deep end-to-end causal discovery method which use a bayesian approach to learn causal relationships using observational data. Also, when partial causal knowledge is available, domain experts can directly adjust or constrain edges in the causal graph. These human-in-the-loop edits ensure that only plausible relationships feed into the counterfactual discovery.

Once such a causal graph is in place, our counterfactual discovery engine intervenes on chosen features while propagating changes through the causal network, ensuring each recommended scenario respects real cause-and-effect pathways. This stands in contrast to naive methods that treat all features as mutually independent, potentially yielding contradictory suggestions (e.g. recommending both fewer working hours and a higher income). By linking causal modelling with FAT, we ensure that only permissible features are altered, and do so in a way that maintains internal consistency across all features.

3.4. Agentic and Iterative Workflow

Finally, we embed these components in an agentic workflow, transitioning from static explanations to dynamic, user-driven conversations. Our agentic workflow is simultaneously interactive, which means the user can add or relax constraints in natural language as well as being iterative so that the system regenerates counterfactuals until the user is satisfied. Basically, after the system presents a counterfactual recommendations in natural language, the user can refine constraints or discard unfeasible changes by updating Feature Actionability Taxonomy which is embeded in this workflow. Concretely:

- *Rejecting a Recommendation:* Users may specify, *"I cannot reduce my monthly outgoings below $1,000 due to fixed costs."*
- *Prioritising Feasibility:* The system re-executes the counterfactual search, guided by both FAT and the user's updated constraints, and then re-runs the NLG to produce revised textual explanations.

This interactive loop ensures each subsequent iteration of recommendations is increasingly tailored to the user's personal limitations and priorities. Consequently, Natural Language Counterfactual Explanations evolve from a one-off advisory statement into an iterative conversation, fostering transparency, trust, and practical usability.

4. Progress and Preliminary Results

Thus far, our work has centred around three key pillars ensuring actionability, incorporating causal insights, and enabling interactive user refinement and has yielded promising preliminary outcomes.

First, we developed a Feature Actionability Taxonomy (FAT) to categorise features as directly mutable, indirectly mutable, or immutable/sensitive. Using insights from a focused user study, we then crafted a template-based NLG module that transforms raw numeric deltas into readable, context-rich sentences (e.g. emphasising time frames and feasibility cues). Early feedback showed that these text-based explanations were perceived as significantly more transparent and actionable than purely tabular representations.

Next, to address the risk of contradictory or unrealistic suggestions, we integrated causal discovery (e.g. via DECI) into the counterfactual search process. This ensures that interventions on one feature (e.g. increasing "Education Level") properly cascade to related variables (e.g. "JobType" or "Income"), thus reflecting genuine real-world cause-and-effect. Pilot experiments demonstrated a measurable reduction in contradictory recommendations, particularly in financial scenarios, when compared to correlation-only methods.

Finally, we introduced an agentic workflow that treats users as active participants rather than passive recipients of one-shot explanations. After receiving an initial set of textual counterfactuals, users can impose additional constraints (*"I cannot reduce my expenses below $1,000"*) or request alternative actions. The system promptly regenerates revised solutions, again expressed via the NLG module. Preliminary user testing suggests that iterative refinement can boosts clarity and trust.

5. Conclusion and Future Directions

In moving towards truly Natural XAI, our framework unifies a Feature Actionability Taxonomy (FAT), causal integration, and a template-based NLG approach within an agentic workflow. This design ensures that counterfactual explanations remain feasible, ethically grounded, and responsive to user feedback. As the next major step, we plan a user study that deploys this interactive system in finance, healthcare, and education scenarios each featuring its own domain-specific features and constraints. Participants will provide iterative feedback on the system's recommendations (e.g. indicating infeasible changes) and then observe how the workflow adapts the generated counterfactuals in real time. We will evaluate how iterative refinement influences trust, clarity, and overall satisfaction, testing whether our method significantly outperforms traditional single-shot explanations. Ultimately, we aim to demonstrate that combining actionability, causality, and user engagement not only enhances the transparency of AI decisions but also offers a meaningful path towards ethically and practically sound recourse.

Acknowledgements. The author would like to thank the supervisory team, domain experts who provided guidance on ethical considerations, and participants in pilot studies whose feedback has substantially shaped the direction of this research.

Declaration on Generative AI

During preparation of this work, the authors used ChatGPT for the purpose of: grammar and spelling check, paraphrase and reword. After using this tool, the authors reviewed and edited the content and take full responsibility for the publication's content.

References

[1] R. K. Mothilal, A. Sharma, C. Tan, Explaining machine learning classifiers through diverse counterfactual explanations, in: Proc. Conf. on Fairness, Accountability, and Transparency, 2020, pp. 607–617.

[2] R. Poyiadzi, K. Sokol, R. Santos-Rodriguez, T. De Bie, P. Flach, Face: feasible and actionable counterfactual explanations, in: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, 2020, pp. 344–350.

[3] A.-H. Karimi, B. Schölkopf, I. Valera, Algorithmic recourse: from counterfactual explanations to interventions, in: Proceedings of the 2021 ACM conference on fairness, accountability, and transparency, 2021, pp. 353–362.

[4] P. Salimi, N. Wiratunga, D. Corsar, A. Wijekoon, Towards feasible counterfactual explanations: A taxonomy guided template-based nlg method, in: ECAI 2023, IOS Press, 2023, pp. 2057–2064.

[5] D. Dua, C. Graff, UCI machine learning repository, 2017. URL: http://archive.ics.uci.edu/ml.

[6] T. Le Quy, A. Roy, V. Iosifidis, W. Zhang, E. Ntoutsi, A survey on datasets for fairness-aware machine learning, WIREs Data Mining and Knowledge Discovery 12 (2022) e1452. URL: https://wires.onlinelibrary.wiley.com/doi/abs/10.1002/widm.1452. doi:https://doi.org/10.1002/widm.1452. arXiv:https://wires.onlinelibrary.wiley.com/doi/pdf/10.1002/widm.1452.

[7] P. Salimi, Addressing trust and mutability issues in xai utilising case based reasoning., ICCBR Doctoral Consortium 1613 (2022) 0073.

[8] E. Reiter, An architecture for data-to-text systems, in: Proc. 11th European Workshop on NLG (ENLG 07), DFKI GmbH, Saarbrücken, Germany, 2007, pp. 97–104. URL: https://aclanthology.org/W07-2315.

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

[10] T. Geffner, J. Antoran, A. Foster, W. Gong, C. Ma, E. Kiciman, A. Sharma, A. Lamb, M. Kukla, A. Hilmkil, et al., Deep end-to-end causal inference, in: NeurIPS 2022 Workshop on Causality for Real-world Impact, 2022.