Semi-Factual Explanations in AI

Saugat Aryal[1,*]

[1]University College Dublin, Belfield, Dublin, Ireland

Abstract

Recent works on post-hoc example-based eXplainable AI (XAI) methods have focused on counterfactual explanations to provide justifications for predictions made by AI systems. Counterfactuals explain by showing what changes to input-features change the output decision. However, a lesser-known, special-case of the counterfacual is the *semi-factual*, which provide explanations about what changes to the input-features *do not change* the output decision. Despite their significant potential, semi-factuals have largely been unexplored in the XAI literature. My PhD research aims to address this gap by establishing a comprehensive framework for the use of semi-factuals in XAI. This includes development of novel methods for their computation, validated through user studies.

Keywords

XAI, XCBR, Semi-Factual

1. Introduction

In recent years, research on eXplainable AI (XAI) have garnered significant attention as it aims to improve the transparency and interpretability of the black-box AI models. Among different XAI strategies, post-hoc example-based explanation methods which provides after-the-fact justification have been very popular. Within this landscape, significant effort have been expended on counterfactual explanations [1, 2, 3, 4]. However, a special-case of the counterfactual, called *semi-factual explanations* have been largely ignored even though they have as much potential as counterfactuals (albeit in different contexts).

Counterfactual explanations inform users about how the output-decision can be altered by changing key input-features in the form of *"if only"* reasoning. For example, when a customer is refused a loan, the counterfacutal might say "*if only* you asked for a loan with shorter term, it would have been approved". Semi-factual explanations, on the contrary, inform users about how the output-decision remains the same when the key input-features change using *"even if"* reasoning. So, in the banking recourse example, the semi-factual might say "*even if* you doubled your income, your loan would still be refused". In cognitive science, counterfactuals and semi-factuals have been shown to have different psychological impacts on users where the former tends to enable strong causal relation whereas the latter weakens the causal support [5].

The origin of semi-factual as explanations can be traced back to early works on post-hoc explanatory case-based reasoning (XCBR)[6, 7, 8]. In these studies, semi-factuals were used in the form of *"a fortiori"* arguments to provide strong convincing explanation for a proposition. The scholars observed that in some scenarios, a case-based neighbor which was farther away from the query and closer to the decision boundary (semi-factual) could provide better explanation than the actual nearest neighbor. However, they lacked to formally define the behaviour of semi-factual as explanations. In recent works, semi-factuals have emerged mostly in association with counterfactuals in that they are obtained as a consequence of generating counterfactual explanations [9]. As such, semi-factuals as explanations have not been well-defined and studied independently in the XAI domain which drives the motivation of my research.

The aim of my PhD thesis primarily focus on defining the characterstics of semi-factual explanations and establishing a comprehensive framework for their use in XAI.

ICCBR DC'24: Doctoral Consortium at ICCBR2024, July 1, 2024, Mérida, Mexico

*Corresponding author.

saugat.aryal@ucdconnect.ie (S. Aryal)

0000-0001-6357-3904 (S. Aryal)

2. Research Plan

My PhD research seek to establish a formal definition for semi-factuals in XAI unifying their computational (i.e, "what needs to be computed") and cognitive requirements (i.e, "the response to be elicited in users"). Based on the desiderata, novel methods will be developed to generate these explanations. This work also intend to propose key evaluation measures to asses their quality and hence provide a strong foundation for future advancements in this area. Furthermore, user studies will also be conducted to analyze how people perceive such explanations.

2.1. Research Questions

My dissertation focuses on 3 main research questions:

- **RQ1:** What does the prior literature on semi-factuals in the field of Cognitive Science, CBR and AI tell us about the fundamental characteristics and desiderata for their use? The aim is to systematically identify and articulate a set of desirable attributes for the utilization of semi-factual explanations in the context of AI.
- **RQ2:** What novel methods can be devised to effectively generate, interpret and evaluate semi-factual explanations in AI systems? This research question addresses the core objective of my PhD thesis which is to explore innovative methodologies for the generation of semi-factual explanations that meet the established desiderata.
- **RQ3:** How do people comprehend and interpret semi-factuals and how do these explanations impact their trust and understanding of AI systems? The objective is to conduct comprehensive user tests to analyze their impact and effect on people.

3. Progress Summary

The progress so far can be divided into three phases.

3.1. Literature Review, MDN & Benchmarking Study

In my first phase of work [10], a systematic literature review was conducted surveying the historical and recent works on semi-factuals across several domains including Philosophy, Psychology, CBR and AI. Following the literature, a formalised computational and cognitive desiderata for semi-factuals in XAI was introduced (**RQ1**). In the same work, a benchmark evaluation of historical methods was performed along with the proposal of a novel, baseline algorithm, the Most Distant Neighbor (MDN) method to support benchmarking.

In cognitive science, semi-factuals have been extensively studied in Philosophy [11] and Psychology [5] under different guises. Philosophers have argued if semi-factuals are fundamentally different from counterfactuals and psychological research show that they have different cognitive impact on people. Specifically, semi-factuals tend to weaken the causal dependencies between the input and the outcome. When someone is told that "even doubling your income will not lead to a loan approval" they are more likely to think that income is really not causally important in the domain.

Similarly, a number of methods were proposed in early CBR research where semi-factuals were characterized as *a fortiori* arguments. Doyle et. al [12] was the seminal paper which first proposed the use of such reasoning and used utility functions to obtain them. Other works used similarity to Nearest Unlike Neighbor (NUN) [7] and surrogate models (similar to LIME) [8] to compute them. Recently, Kenny & Keane [9] advanced a generative method for computing both semi-factuals and counterfacutals in a unified framework which instigated, what could be called, the modern-era of semi-factual XAI research. Consequently, several uses of semi-factuals have been proposed in several areas such as healthcare diagnosis [13], decision space analysis [14, 15] and data augmentation [16].

The proposed novel MDN method finds the furthest neighbor of the query along some feature-dimension while being in the same class as query (this is analogous to the use of NUNs in counterfactuals,

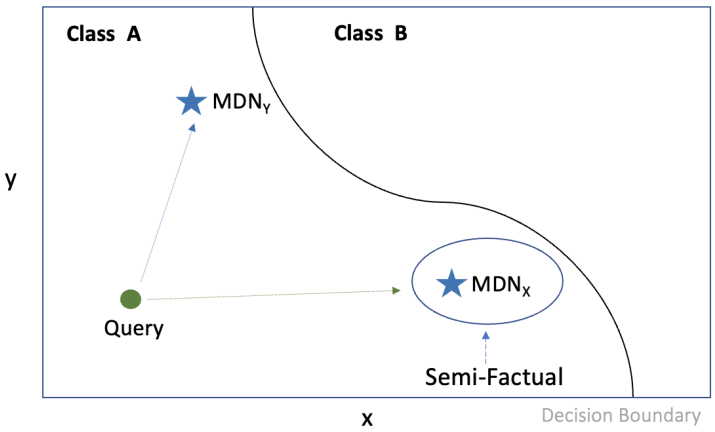

Figure 1: Consider a 2-dimensional feature space with features, **x** and **y**. MDN_X is the Most Distant Neighbor along **x** for Query while MDN_Y is the MDN along **y**. MDN_X achieves high *sfs()* score than MDN_Y as it is farther from the query along **x** while also being most similar to query along **y**. Hence, MDN_X is selected as the best MDN and hence, Semi-Factual for the Query and **x** is chosen as the final key-feature.

where an existing datapoint is used as an explanation). MDNs are known data-points in the dataset that share some common features with the query but are far from it on some key-feature.

To compute MDNs, for a given feature of q, its neighbours along that dimension are partitioned into instance-sets that have higher values (i.e., HighSet) or lower values (i.e., LowSet) than the query. Each of these sets are ranked-ordered separately using a *Semi-Factual Scoring* (*sfs*) function, which is a distance measure that prioritises instances that are sparse (few feature differences) while also having the highest value-differences on the selected dimension, as follows:

$$sfs(q, S, F) = \frac{same(q, x)}{F} + \frac{diff(q_f, x_f)}{diff_{max}(q_f, S_f)} \tag{1}$$

where S is High or Low Set and $x \in S$, *same()* counts the features that are equal between q and x, F is the total number of features, *diff()* gives the difference-value of key-feature, f, and $diff_{max}()$ is the maximum difference-value for that key-feature in the HighSet/LowSet. Basically, the instance with the highest overall *sfs* value from the HighSet/LowSet is the best MDN for that feature. This computation is done for each feature of q, independently, with the best of the best instance (i.e., with the highest *sfs* value across all features) is selected to be the semi-factual for the query as shown in Figure 1.

$$\mathrm{SF}_{\mathrm{MDN}}(q, S) = \arg\max_{x \in S} sfs(x) \tag{2}$$

We show that MDNs meet many of the desiderata for semi-factuals though they may not be an optimal solution. Furthermore, we experimentally compared four historical CBR methods (three KLEOR-variants [7] and Local-Region [8]) against the MDN algorithm to provide a solid baseline for future works. The algorithms were evaluated on key distance metrics for assessing good semi-factuals. The results show that MDN performed best in finding semi-factuals that are farthest away from the query in both feature and instance space. However, it falls behind on three measures: distance to the query's class distribution, distance to the NUN and sparsity.

3.2. Optimized MDNs

In the second phase, a series of experiments were performed to propose two new MDN variants to overcome their initial limitations. The custom *sfs()* function in MDN scores each candidate instance based on their relative feature-value distance as well as closeness to the query. Along this line, we modify the scoring function to optimize their behaviour.

The two components in the scoring function are equally weighted to find the best semi-factual. Since original MDN (MDNv1) performed relatively poorly in the sparsity metric we propose Sparse-MDNs (MDNv2), which prioritises the similarity between non-key features. Essentially, we introduce a regularizer in the original *sfs()* function, which penalizes the algorithm for finding semi-factuals with higher feature-differences, thus promoting sparse explanations. We modify the scoring function by weighing it with the proportion of features that are "not same" between the query and the instance. Hence, instances with higher number of similar features will be assigned high scores to obtain sparse MDNs.

$$sfs_{v2}(q, S, F) = \frac{1}{F - same(q, x)} * (\frac{diff(q_f, x_f)}{diff_{max}(q_f, S_f)} + \frac{same(q, x)}{F}) \tag{3}$$

In both MDNv1 and MDNv2, the similarity between non-key features between query and instances using *same()* involves a direct comparison of their values. Specifically, the function checks if the values are identical in case of categorical features, while the continuous features are considered same if they fall within a predefined threshold range. However, it is not always straight-forward to determine the optimal threshold and it may vary across different features. Hence, we propose Dist-MDNs (MDNv3) where we modify the scoring function to compute similarity directly in the feature space as:

$$sfs_{v3}(q, S) = \frac{diff(q_f, x_f)}{diff_{max}(q_f, S_f)} * \frac{1}{dist(q_{nf}, x_{nf})} \tag{4}$$

where *dist()* computes the L_2-norm distance and q_{nf} and x_{nf} represents the query and instance with only non-key features (i.e excluding the key-feature in consideration) respectively.

We also performed comprehensive tests to evaluate the proposed methods. The results show that the proposed variants could improve on the initial MDN limitations, however, the historical CBR methods still perform better on some measures.

3.3. Analysis of Counterfactuals for Semi-Factuals

Finally, in my most recent work, comprehensive tests were conducted to determine if counterfactuals are needed to obtain the best semi-factuals.

We divide the literature on semi-factual methods into two groups: Counterfactual-Guided and Counterfactual-Free. The Counterfactual-Guided group uses counterfactuals as guides to find the semi-factuals, whereas, Counterfactual-Free methods considers exploration within the query-class without explicitly relying on counterfactuals. Specifically, we consider the key question "Are the best semi-factuals found by using counterfactuals as guides?". We evaluated 8 semi-factuals methods (4 from each group) on five evaluation metrics that attempt to capture key aspects of the desiderata for "good" semi-factuals. The results indicate that counterfactual-guidance is "not" necessary to find best semi-factuals. Each method do well on one or two metrics but then poorly on others. However, there is no single method that is consistently good across all measures. This work has been accepted in the main proceedings of ICCBR'24.

4. Conclusion and Future Work

Overall, **[RQ1]** have been successfully addressed and completed, and **[RQ2]** is well underway with an in-depth exploration. Based on the findings so far, the current focus is to devise novel generative methods to obtain semi-factual explanations. In the future, a primary emphasis will be on conducting user tests. The objective will be to study human perception of these explanations and gather insights about their impact, particularly within the context of AI systems. The collected user feedback will contribute to a comprehensive understanding of the nature and utility of semi-factual explanations in AI, thereby facilitating further advancements in this field. The results of these studies will address the objectives of **[RQ3]**.

References

[1] B. Smyth, M. T. Keane, A few good counterfactuals: generating interpretable, plausible and diverse counterfactual explanations, in: International Conference on Case-Based Reasoning, Springer, 2022, pp. 18–32.

[2] R. M. Byrne, Counterfactuals in explainable artificial intelligence (xai): evidence from human reasoning, in: Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19, 2019, pp. 6276–6282.

[3] T. Miller, Explanation in artificial intelligence: Insights from the social sciences, Artificial Intelligence 267 (2019) 1–38.

[4] M. T. Keane, E. M. Kenny, E. Delaney, B. Smyth, If only we had better counterfactual explanations, in: Proceedings of the 30th International Joint Conference on Artificial Intelligence (IJCAI-21), 2021.

[5] R. McCloy, R. M. Byrne, Semifactual "even if" thinking, Thinking & Reasoning 8 (2002) 41–67.

[6] C. Nugent, P. Cunningham, D. Doyle, The best way to instil confidence is by being right, in: International Conference on Case-Based Reasoning, Springer, 2005, pp. 368–381.

[7] L. Cummins, D. Bridge, Kleor: A knowledge lite approach to explanation oriented retrieval, Computing and Informatics 25 (2006) 173–193.

[8] C. Nugent, D. Doyle, P. Cunningham, Gaining insight through case-based explanation, Journal of Intelligent Information Systems 32 (2009) 267–295.

[9] E. M. Kenny, M. T. Keane, On generating plausible counterfactual and semi-factual explanations for deep learning, in: Proceedings of the 35th AAAI Conference on Artificial Intelligence (AAAI-21), 2021, pp. 11575–11585.

[10] S. Aryal, M. T. Keane, Even if explanations: Prior work, desiderata & benchmarks for semi-factual xai, in: E. Elkind (Ed.), Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI-23, International Joint Conferences on Artificial Intelligence Organization, 2023, pp. 6526–6535. URL: https://doi.org/10.24963/ijcai.2023/732. doi:10.24963/ijcai.2023/732, survey Track.

[11] J. Bennett, Even if, Linguistics and Philosophy 5 (1982) 403–418.

[12] D. Doyle, P. Cunningham, D. Bridge, Y. Rahman, Explanation oriented retrieval, in: European Conference on Case-Based Reasoning, Springer, 2004, pp. 157–168.

[13] A. Vats, A. Mohammed, M. Pedersen, N. Wiratunga, This changes to that: Combining causal and non-causal explanations to generate disease progression in capsule endoscopy, arXiv preprint arXiv:2212.02506 (2022).

[14] A. Artelt, B. Hammer, "even if..."–diverse semifactual explanations of reject, arXiv preprint arXiv:2207.01898 (2022).

[15] S. Mertes, C. Karle, T. Huber, K. Weitz, R. Schlagowski, E. André, Alterfactual explanations–the relevance of irrelevance for explaining ai systems, arXiv preprint arXiv:2207.09374 (2022).

[16] J. Lu, L. Yang, B. Mac Namee, Y. Zhang, A rationale-centric framework for human-in-the-loop machine learning, arXiv preprint arXiv:2203.12918 (2022).