Semi-factual Explanations in AI

Saugat Aryal

School of Computer Science, University College Dublin, Dublin, Ireland
Insight Centre for Data Analytics, Dublin, Ireland
saugat.aryal@ucdconnect.ie

Abstract

Most of the recent works on post-hoc example-based eXplainable AI (XAI) methods revolves around employing counterfactual explanations to provide justification of the predictions made by AI systems. Counterfactuals show what changes to the 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. Semi-factuals are potentially as useful as counterfactuals but have received little attention in the XAI literature. My doctoral research aims to establish a comprehensive framework for the use of semi-factuals in XAI by developing novel methods for their computation, supported by user tests.

Introduction

Counterfactual explanations tell people about the input-features of an AI system that might change in order for to change an outcome decision (usually to the desired option; see e.g., (Smyth and Keane 2022)). For example, when a customer is refused a loan, the counterfactual explanation might say "*if* you asked for a loan with shorter term, it would have been approved". Semi-factuals, on the contrary, tell people about the feature changes that *do not change* the outcome decision. So, in the banking recourse example, the semi-facutal might say, "*even if* you doubled your income, you would still be refused the loan". Counterfactuals and semi-factuals have been shown to have different cognitive effects on users, the former tends to inform about enabling conditions, whereas the latter can weaken causal support (McCloy and Byrne 2002). Though semi-factuals have been studied in the Cognitive Sciences, they are only beginning to be appreciated in XAI, even though they have as much potential as counterfactuals (albeit in different contexts).

In my PhD research I am developing novel methods for generating semi-factual explanations which would help people better understand AI systems. Furthermore, I will also explore how people perceive semi-factuals in user studies and how their understanding of the causality can be affected. My dissertation focuses on 3 main research questions:

- **RQ1:** What does the prior literature on semi-factuals in the field of Cognitive Science and AI tell us about the desiderata for their use?
- **RQ2:** What novel methods can be devised to effectively generate and interpret semi-factual explanations for AI systems?
- **RQ3:** How do people comprehend and interpret semi-factuals and how do these explanations impact their trust and understanding of AI systems?

Progress To Date

My recent work (Aryal and Keane 2023) has focused on surveying the historical and recent works on semi-factuals and on formalising computational and cognitive desiderata for semi-factuals in XAI (**RQ1**). I have also performed benchmark evaluations of historical methods and propose a novel, baseline algorithm – the Most Distant Neighbor (MDN) method – to support benchmarking.

Semi-factuals have been extensively studied in Philosophy (Bennett 1982) and Psychology (McCloy and Byrne 2002) from different perspectives. The psychological research shows that 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.

The origin of semi-factuals in the context of XAI can be attributed to the research on post-hoc example-based explanations in the field of Case-Based Reasoning (CBR). In this research, semi-factual explanations have been characterized as *a fortiori* arguments which can provide more convincing explanations for a given case than a "standard" Nearest Neighbor (Doyle et al. 2004). Other work has used similarity to a Nearest Unlike Neighbor (NUN) (Cummins and Bridge 2006) or surrogate models (like LIME) (Nugent, Doyle, and Cunningham 2009) to compute the semi-factuals. More recently, Kenny & Keane (2021) advanced a generative method for computing both semi-factuals and counterfacutals in a unified framework, work that has significantly fueled interest in new uses of semi-factuals (e.g., (Artelt and Hammer 2022; Lu et al. 2022; Vats et al. 2022; Zhao et al. 2022; Kenny and Huang 2023)).

Stepping on from this previous work, we have intro-

duced a novel naïve benchmark, the Most Distant Neighbors (MDNs), that finds 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, 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. We have shown that they meet many of the desiderata for semi-factuals though they may not be an optimal solution. Furthermore, we experimentally compared four historical methods 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 four of the seven metrics including distance from the query, although MDN scored less well on sparsity.

Future Work and Timeline

- **Phase 1: Optimized MDNs**

The proposed MDN algorithm could be further optimized to find instances with fewer feature differences from the query while also not compromising its overall efficiency (to improve sparcity). This involves augmenting the original MDN method with a regularizer which penalizes the method for finding semi-factuals with more feature-differences, to improve its sparsity without compromising on other metrics.

- **Phase 2: Generative Model for MDNs**

All the historical methods including MDN, find semi-factuals that are actual instances in the dataset. This opens a promising avenue to explore the use of a generative model such as Variational Autoencoder (VAE) to produce semi-factual for the query, given its success in counterfactual explanations (Pawelczyk, Broelemann, and Kasneci 2020). Specifically, we propose to employ a variant of VAE, called Conditional VAE (CVAE) which takes a conditional information as additional input to both the encoder and decoder. This allows the latent space to not only learn the representations of the data but also the intrinsic relationship between the data and the provided information. In similar vein, we can train a CVAE with Query-MDN pairs where the model would learn to generate the MDN conditioned on its Query.

Both **Phases 1** & **2** aim to answer **RQ2** and are a work-in-progress. I anticipate results by the workshop date (February 21, 2024).

- **Phase 3: User Study**

To assess how semi-factual explanations impact people's understanding about the AI system and domain, we propose to conduct a carefully designed user study to psychologically validate these explanation methods (**RQ3**).

- **Phase 4: MDNs for Images**

The existing semi-factual methodologies have been specifically curated for tabular datasets to be able to compare with the historical benchmarks. However, their application within the image domain has received less attention. We propose to expand MDNs for images which primarily emphasize on generating semi-factual explanations (**RQ2**).

Objective	Timeline
Phase 1 & 2 (RQ2)	Sept 2023 - Nov 2023
Phase 3 (RQ3)	Dec 2024 - Feb 2024
Phase 4 (RQ2)	March 2024 - June 2024

Table 1: Research Timeline

References

Artelt, A.; and Hammer, B. 2022. "Even if..."–Diverse Semifactual Explanations of Reject. *arXiv preprint arXiv:2207.01898*.

Aryal, S.; and Keane, M. T. 2023. Even If Explanations: Prior Work, Desiderata & Benchmarks for Semi-Factual XAI. In Elkind, E., ed., *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI-23*, 6526–6535.

Bennett, J. 1982. Even if. *Linguistics and Philosophy*, 5(3): 403–418.

Cummins, L.; and Bridge, D. 2006. Kleor: A knowledge lite approach to explanation oriented retrieval. *Computing and Informatics*, 25(2-3): 173–193.

Doyle, D.; Cunningham, P.; Bridge, D.; and Rahman, Y. 2004. Explanation oriented retrieval. In *European Conference on Case-Based Reasoning*, 157–168. Springer.

Kenny, E. M.; and Huang, W. F. 2023. The Utility of "Even if" semifactual explanation to optimise positive outcomes. In *Thirty-seventh Conference on Neural Information Processing Systems (NeurIPS-23)*.

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

Lu, J.; Yang, L.; Mac Namee, B.; and Zhang, Y. 2022. A Rationale-Centric Framework for Human-in-the-loop Machine Learning. *arXiv preprint arXiv:2203.12918*.

McCloy, R.; and Byrne, R. M. 2002. Semifactual "even if" thinking. *Thinking & Reasoning*, 8(1): 41–67.

Nugent, C.; Doyle, D.; and Cunningham, P. 2009. Gaining insight through case-based explanation. *Journal of Intelligent Information Systems*, 32(3): 267–295.

Pawelczyk, M.; Broelemann, K.; and Kasneci, G. 2020. Learning model-agnostic counterfactual explanations for tabular data. In *Proceedings of the web conference 2020*, 3126–3132.

Smyth, B.; and Keane, M. T. 2022. A few good counterfactuals: generating interpretable, plausible and diverse counterfactual explanations. In *International Conference on Case-Based Reasoning*, 18–32. Springer.

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

Zhao, Z.; Leake, D.; Ye, X.; and Crandall, D. 2022. Generating Counterfactual Images: Towards a C2C-VAE Approach. In *4th Workshop on XCBR: Case-Based Reasoning for the Explanation of Intelligent Systems*.