

# Theoretical Basis and Computational Complexity of Semifactual Explanations

Gianvincenzo Alfano<sup>1</sup>, Sergio Greco<sup>1</sup>, Domenico Mandaglio<sup>1</sup>, Francesco Parisi<sup>1</sup>,  
Reza Shahbazian<sup>2</sup> and Irina Trubitsyna<sup>1</sup>

<sup>1</sup>Department of Informatics, Modeling, Electronics and System Engineering, University of Calabria, Italy

<sup>2</sup>Department of Humanities, University of Palermo, Italy

## Abstract

Explainable AI has received significant attention in recent years. Machine learning models often operate as black boxes, lacking explainability and transparency while supporting decision-making processes. Local post-hoc explainability queries attempt to answer why individual inputs are classified in a certain way by a given model. While there has been important work on counterfactual explanations, less attention has been devoted to semifactual ones. In this paper, we discuss the local post-hoc explainability queries for semifactual reasoning recently proposed in [1], analyze their computational complexity across different classification models, and examine the associated preference-based framework for semifactual and counterfactual explanations.

## Keywords

Explainable AI, Semifactual Explanations, Machine Learning

## 1. Introduction

The extensive study of counterfactual ‘if only’ thinking, exploring how things might have been different, has been a focal point for social and cognitive psychologists [2]. Consider a negative event, such as taking a taxi and due to traffic arriving late to a party. By analyzing this situation, an individual (e.g. Alice) might engage in counterfactual thinking by imagining how things could have unfolded differently, such as, ‘if only Alice had not taken the taxi, she would not have arrived late at the party’. This type of counterfactual thinking, where an alternative scenario is imagined, is a common aspect of daily life. In such a case the counterfactual scenario negates both the event’s cause (antecedent) and its outcome, presenting a false cause and a false outcome that are temporarily considered as true (e.g., Alice took the taxi and arrived late).

Counterfactual thinking forms the basis for crafting counterfactual explanations, which are crucial in automated decision-making processes. These explanations leverage alternative scenarios, aiding users in understanding why certain outcomes occurred and how different situations might have influenced decisions. Counterfactual explanations empower users to grasp the rationale behind decisions, fostering transparency and user trust in these systems. Several definitions of counterfactual explanations exist in the literature [3]. According to most of the literature, counterfactuals are defined as the minimum changes to apply to a given instance to let the prediction of the model be different [4].

While counterfactual explanations have received much attention in AI, semifactuals—based on “even if” reasoning—remain underexplored, though well-studied in cognitive science [5]. Whereas counterfactuals identify input changes that alter an AI system’s decision, semifactuals highlight changes that leave the outcome unchanged. Returning to the earlier example, a semifactual scenario would state: “Even if Alice had not taken the taxi, she would still have arrived late to the party.”

Sharing the same underlying idea of counterfactuals, we define semifactuals as the maximum changes to be applied to a given instance while keeping the same prediction. Indeed, the larger the feature differences asserted in the semifactual, the better (more convincing) the explanation [5]. Semifactual

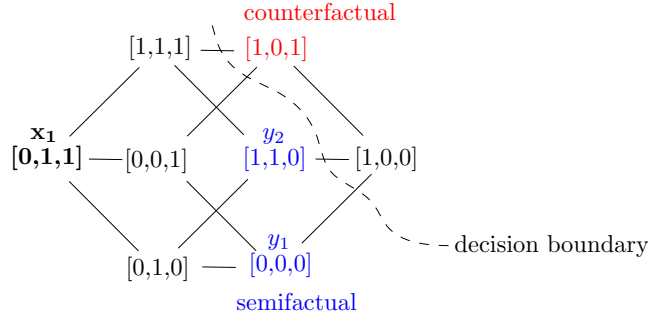
---

*Ital-IA 2025: 5th National Conference on Artificial Intelligence, organized by CINI, June 23-24, 2025, Trieste, Italy*

✉ g.alfano@dimes.unical.it (G. Alfano); greco@dimes.unical.it (S. Greco); d.mandaglio@dimes.unical.it (D. Mandaglio); fparisi@dimes.unical.it (F. Parisi); reza.shahbazian@unipa.it (R. Shahbazian); i.trubitsyna@dimes.unical.it (I. Trubitsyna)



© 2025 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).



**Figure 1:** Binary classification model  $\mathcal{M}$ :  $\text{step}(\mathbf{x} \cdot [-2, 2, 0] + 1)$  of Example 1. The binary feature  $f_1$  (resp.,  $f_2$  and  $f_3$ ) represents part-time employment contract (resp., salary lower than 5K\$, and on site-working) [1].

explanations incorporating more feature changes offer several benefits, including improved decision-making and enhanced interpretability. For decision-makers, understanding the extent of changes that do not affect the outcome can aid in optimizing processes. For instance, in resource allocation, knowing the maximum allowable changes helps in making adjustments without compromising results. Such explanations provide a comprehensive understanding of a model’s decision boundary by revealing how the model processes information and identifying which input aspects are critical for maintaining the decision. Semifactuals indicate which feature-sets are not relevant for classification, as they can be changed without altering the outcome. Also, considering a large number of feature changes in the semifactual intuitively captures the desire of an agent to have more flexibility and favorable conditions (represented by features changed), while keeping the (positive) status assigned to it by the model. Consider the following hiring scenario.

**Example 1.** Consider the binary and linear classification model  $\mathcal{M} : \{0, 1\}^3 \rightarrow \{0, 1\}$  shown in Figure 1, where  $\mathcal{M}$  is defined as  $\text{step}(\mathbf{x} \cdot [-2, 2, 0] + 1)$  and the input  $\mathbf{x} = [x_1, x_2, x_3]$  denotes an applicant (also called user) defined by means of the following three features: (i)  $f_1$  = “part-time job”; (ii)  $f_2$  = “requested (monthly) salary < 5K\$”; (iii)  $f_3$  = “on-site job”. For any instance  $\mathbf{x} \in \{0, 1\}^3$  we have that  $\mathcal{M}(\mathbf{x}) = 0$  if  $\mathbf{x} = [1, 0, 1]$  or  $\mathbf{x} = [1, 0, 0]$ , and  $\mathcal{M}(\mathbf{x}) = 1$  otherwise. Intuitively, this means that the company’s AI model does not approve the application only when the user applies for a part-time job and the requested salary is no less than 5K\$.

Consider a user  $\mathbf{x}_1$  that applies for a full-time and on-site job, and the requested salary is lower than 5K\$ (i.e.,  $\mathbf{x}_1 = [0, 1, 1]$ ), we have that  $\mathbf{y}_1 = [0, 0, 0]$  and  $\mathbf{y}_2 = [1, 1, 0]$  are semifactual of  $\mathbf{x}_1$  w.r.t.  $\mathcal{M}$  at maximum distance (i.e., 2) from  $\mathbf{x}_1$  in terms of number of features changed. Intuitively,  $\mathbf{y}_1$  represents the fact that ‘the user  $\mathbf{x}_1$  will be hired *even if* (s)he had requested for a remote job and the requested salary was greater than or equal to 5K\$’, while  $\mathbf{y}_2$  represents ‘the user  $\mathbf{x}_1$  will be hired *even if* (s)he had applied for a remote and part-time job’.  $\square$

Counterfactuals and semifactuals are strongly connected and they should be considered together in eXplainable AI (XAI) as they describe which changes to feature-inputs of a black-box AI system result in changes to or confirmation to a decision-outcome, that is both contribute in understanding the presence of a decision boundary in the classification process. Taking for instance our running example, whose feature-values are shown in Figure 1, where edges represent changes of a unique feature value, the decision boundary can be described by considering both counterfactuals and semifactuals.

As highlighted in the previous example, multiple semifactuals can exist for each given instance. In these situations, a user may prefer one semifactual to another, by expressing preferences over features so that the *best* semifactuals will be selected, as shown in the following example.

**Example 2.** Continuing with Example 1, suppose that the user  $\mathbf{x}_1$  looks for another opportunity and prefers to change feature  $f_2$  rather than  $f_1$  (irrespective of any other change), that is (s)he prefers semifactuals with  $f_2 = 0$  rather than those with  $f_1 = 1$ . Thus, (s)he would prefer to still get hired by changing the salary to be greater than or equal to 5K\$ (obtaining  $\mathbf{y}_1$ ); if this cannot be accomplished, then (s)he prefers to get it by changing the job to part-time (i.e.  $\mathbf{y}_2$ ).  $\square$

**Contributions.** In this paper, we discuss our recent in semifactual reasoning for XAI [1]. Our main contributions are as follows. We define semifactual explanations for three model classes—perceptrons, free binary decision diagrams (FBDDs), and multi-layer perceptrons (MLPs)—as local post-hoc queries under even-if reasoning. We analyze the computational complexity of related interpretability problems, showing they are no harder than those for counterfactuals. We also introduce a preference-based framework allowing users to prioritize certain features in explanations, covering both semifactuals and counterfactuals. Finally, we study the complexity of finding best explanations under preferences and identify tractable cases with corresponding algorithms.

## 2. Preliminaries

A (binary classification) model is a function  $\mathcal{M} : \{0, 1\}^n \rightarrow \{0, 1\}$ , specifically focusing on instances whose features are represented by binary values. Constraining inputs and outputs to booleans simplifies our context while encompassing numerous relevant practical scenarios. A class of models is just a way of grouping models together. An instance  $\mathbf{x}$  is a vector in  $\{0, 1\}^n$  and represents a possible input for a model. We recall 3 significant categories of ML models that will be the ones we will focus on.

A *Binary Decision Diagram (BDD)*  $\mathcal{M} = (V, E, \lambda_V, \lambda_E)$  [6] is a rooted directed acyclic graph  $(V, E)$  where leaves are labeled 0 or 1, and internal nodes are labeled via  $\lambda_V$  with values in  $\{1, \dots, n\}$ . Each internal node has two outgoing edges labeled by  $\lambda_E$  as 0 and 1. An input  $\mathbf{x} = [x_1, \dots, x_n] \in \{0, 1\}^n$  defines a unique path  $p_{\mathbf{x}}$  in  $\mathcal{M}$ , following the edge matching  $x_i$  at each node labeled  $i$ . The size  $|\mathcal{M}|$  is the number of edges. A BDD is *free* (FBDD) if no path visits the same variable twice. A *decision tree* is an FBDD with a tree structure.

A *multilayer perceptron (MLP)*  $\mathcal{M}$  with  $k$  layers is defined by weight matrices  $\mathbf{W}^{(1)}, \dots, \mathbf{W}^{(k)}$ , biases  $\mathbf{b}^{(1)}, \dots, \mathbf{b}^{(k)}$ , and activation functions  $a^{(1)}, \dots, a^{(k)}$ . For input  $\mathbf{x}$ , let  $\mathbf{h}^{(0)} = \mathbf{x}$  and compute recursively  $\mathbf{h}^{(i)} = a^{(i)}(\mathbf{h}^{(i-1)}\mathbf{W}^{(i)} + \mathbf{b}^{(i)})$ . The output is  $\mathcal{M}(\mathbf{x}) = \mathbf{h}^{(k)}$ . We assume rational weights and biases. Hidden layers are those with  $i < k$ , and we focus on ReLU activations for  $a^{(1)}, \dots, a^{(k-1)}$ , and the step function for  $a^{(k)}$ , defined as  $\text{step}(x) = 1$  if  $x \geq 0$ , else 0.

A *perceptron* is an MLP with no hidden layers (i.e.,  $k = 1$ ). That is, a perceptron  $\mathcal{M}$  is defined by a pair  $(\mathbf{W}, b)$  such that  $\mathbf{W} \in \mathbb{Q}^{n \times 1}$  and  $b \in \mathbb{Q}$ , and the output is  $\mathcal{M}(\mathbf{x}) = \text{step}(\mathbf{x}\mathbf{W} + b)$ . Because of its particular structure, a perceptron is usually defined as a pair  $(\mathbf{w}, b)$  with  $\mathbf{w} = \mathbf{W}^T$  a rational vector and  $b$  a rational number. The output of  $\mathcal{M}(\mathbf{x})$  is then 1 iff  $\mathbf{x} \cdot \mathbf{w} + b \geq 0$ , where  $\mathbf{x} \cdot \mathbf{w}$  denotes the dot product between  $\mathbf{x}$  and  $\mathbf{w}$ .

Boolean functions  $\mathcal{F}$  mapping strings to strings whose output is a single bit are called decision problems. We identify the computational problem of computing  $\mathcal{F}$  (i.e., given an input string  $x$  compute  $\mathcal{F}(x)$ ) with the problem of deciding whether  $\mathcal{F}(x) = 1$ .

## 3. Even-if Explanations

We illustrate our framework on three classes of boolean models and associated explainability queries, aiming to compare their interpretability. We begin by revisiting counterfactuals—central to the ‘if only’ reasoning—and their known complexity results [4]. To formalize the comparison between instances  $\mathbf{x}, \mathbf{y} \in \{0, 1\}^n$ , we use the Hamming distance, defined as  $d(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^n |x_i - y_i|$ , which counts the number of differing features.

**Definition 1 (Counterfactual).** *Given a pre-trained model  $\mathcal{M}$  and an instance  $\mathbf{x}$ , an instance  $\mathbf{y}$  is said to be a counterfactual of  $\mathbf{x}$  iff i)  $\mathcal{M}(\mathbf{x}) \neq \mathcal{M}(\mathbf{y})$ , and ii) there exists no other instance  $\mathbf{z} \neq \mathbf{y}$  s.t.  $\mathcal{M}(\mathbf{x}) \neq \mathcal{M}(\mathbf{z})$  and  $d(\mathbf{x}, \mathbf{z}) < d(\mathbf{x}, \mathbf{y})$ .*

**Example 3.** Continuing with our running example illustrated in Figure 1, for  $\mathbf{y}_3 = [1, 0, 1]$  we have that  $\mathbf{x}_2 = [0, 0, 1]$  and  $\mathbf{x}_3 = [1, 1, 1]$  are the only counterfactuals of  $\mathbf{y}_3$  w.r.t.  $\mathcal{M}$  (herein,  $d(\mathbf{y}_3, \mathbf{x}_2) = d(\mathbf{y}_3, \mathbf{x}_3) = 1$ ). Intuitively, this encodes the fact that user  $\mathbf{y}_3$  (that applied for a part-time and remote job, and a salary greater than or equal to 5K\$) will be hired *if only* (s)he would change the employment contract to be full time (obtaining  $\mathbf{x}_2$ ) or the requested salary to be lower than 5K\$ (obtaining  $\mathbf{x}_3$ ).  $\square$

The natural decision version of the problem of finding a counterfactual for  $\mathbf{x}$  is the following.

**Problem 1 ([4]).** [*MINIMUM CHANGE REQUIRED (MCR)*] Given a model  $\mathcal{M}$ , instance  $\mathbf{x}$ , and  $k \in \mathbb{N}$ , check whether there exists an instance  $\mathbf{y}$  with  $d(\mathbf{x}, \mathbf{y}) \leq k$  and  $\mathcal{M}(\mathbf{x}) \neq \mathcal{M}(\mathbf{y})$ .

**Theorem 1 ([4]).** *MCR is i) in PTIME for FBDDs and perceptrons, and ii) NP-complete for MLPs.*

We adopt a standard view linking interpretability to computational complexity [4]: *a model class is more interpretable if post-hoc queries can be answered more efficiently*. Under this view, Theorem 1 shows that perceptrons and FBDDs are strictly more interpretable than MLPs, as their related queries are computationally easier. This formally supports the common belief that linear models are more interpretable than deep networks in the context of counterfactual explanations.

An open question, which we consider in this discussion, is whether similar results extend to post-hoc queries grounded in the ‘even-if’ thinking setting, i.e., to semifactual explanations. To this end, we first recall the formal definition of semifactuals.

**Definition 2 (Semifactual).** Given a pre-trained model  $\mathcal{M}$  and an instance  $\mathbf{x}$ , an instance  $\mathbf{y}$  is said to be a semifactual of  $\mathbf{x}$  iff i)  $\mathcal{M}(\mathbf{x}) = \mathcal{M}(\mathbf{y})$ , and ii) there exists no other instance  $\mathbf{z} \neq \mathbf{y}$  s.t.  $\mathcal{M}(\mathbf{x}) = \mathcal{M}(\mathbf{z})$  and  $d(\mathbf{x}, \mathbf{z}) > d(\mathbf{x}, \mathbf{y})$ .

Similar to counterfactuals, the following problem is the decision version of the problem of finding a semifactual of an instance  $\mathbf{x}$  with a model  $\mathcal{M}$ .

**Problem 2.** [*MAXIMUM CHANGE ALLOWED (MCA)*] Given a model  $\mathcal{M}$ , instance  $\mathbf{x}$ , and  $k \in \mathbb{N}$ , check whether there is an instance  $\mathbf{y}$  with  $d(\mathbf{x}, \mathbf{y}) \geq k$  and  $\mathcal{M}(\mathbf{x}) = \mathcal{M}(\mathbf{y})$ .

Although semifactuals and counterfactuals appear to be similar, their mathematical definitions are different. Indeed, while counterfactuals minimize the changes in order to have a different outcome, semifactuals maximize the changes while keeping the same outcome. Notably, the two problems are not interchangeable - we do not see how to naturally reduce one to the other; however, a (possibly complex) reduction may exist as our complexity results presented in Theorem 2 below do not rule this out. For instance, considering Example 1 and the two semifactuals  $\mathbf{y}_1$  and  $\mathbf{y}_2$  of  $\mathbf{x}_1$ , they do not correspond to the counterfactuals of the counterfactuals of  $\mathbf{x}_1$ , that are  $[0, 0, 1]$  and  $[1, 1, 1]$ .

**Theorem 2.** *MCA is i) in PTIME for FBDDs and perceptrons, and ii) NP-complete for MLPs.*

It turns out that, under standard complexity assumptions, computing semifactuals under perceptrons and FBDDs is easier than under multi-layer perceptrons. Moreover, independently of the type of the model, computing semifactuals is as hard as computing counterfactuals. Thus, perceptrons and FBDDs are strictly more interpretable than MLPs, in the sense that the complexity of answering post-hoc queries for models in the first two classes is lower than for those in the latter.

**Preferences over Explanations.** The problem of preference handling has been extensively studied in AI. Several formalisms have been proposed to express and reason with different kinds of preferences [7, 8, 9, 10, 11]. In our work [1], we propose a framework to express and reason about user preferences over counterfactual and semifactual explanations, drawing inspiration from Brewka et al.’s work [12] on preference logic. We define preferences as rules over input features, where a rule ranks feature literals (e.g.,  $f_1 \succ \neg f_2$ ) either unconditionally or conditioned on other literals. A model with preferences (termed BCMP) is then a pair consisting of a binary classifier and a set of such preference rules. The semantics is based on a degree function  $\delta(\mathbf{y}, \kappa)$  that quantifies how well an explanation  $\mathbf{y}$  satisfies a preference rule  $\kappa$ . This allows us to define a partial order  $\sqsubseteq$  over explanations and identify the best ones according to the user-specified preferences. We formally define what it means for an explanation to be the best under this ordering, and then analyze the complexity of verifying whether a given explanation

is optimal (denoted CB-MCR for counterfactuals and CB-MCA for semifactuals). We prove that this problem is in coNP for perceptrons and FBDDs, and coNP-complete for MLPs. We also consider a simpler setting with linear preferences, where only one rule with an empty body is used. In this case, we show that best explanations can be computed in polynomial time for perceptrons and FBDDs, and provide a dedicated algorithm for this purpose [1].

## 4. Related Work

The pursuit of transparency and interpretability in AI has led to several explanation paradigms in XAI [13]. Factual explanations clarify why a model produced a certain prediction [14, 15, 16, 17], while counterfactual explanations explore how minimal changes to inputs could lead to different outcomes [18, 19, 3, 20]. Semifactual explanations [21, 5, 22, 23] sit between these two, identifying input changes that do not alter the decision.

Despite significant progress in methods, the computational complexity of these explanations has received limited attention [4, 24, 25]. Notably, prior work on semifactuals [21] proposes a heuristic solution to an NP-hard problem. Our approach differs in three key ways: (i) it adopts a different notion of semifactual (not based on maximal distance), (ii) incorporates user preferences, and (iii) provides exact rather than heuristic methods.

## 5. Conclusions and Future Work

We discussed a recent advancement in formal XAI [1], which analyzes the complexity of local post-hoc interpretability queries related to semifactuals across three model classes, and introduces a preference-based framework for personalizing semifactual and counterfactual explanations. Future directions include extending to counting problems, non-binary features, constraints [26], other classes of models (e.g. graph neural networks [27]) and preference structures.

## Acknowledgments

We acknowledge financial support from PNRR MUR projects FAIR (PE0000013) and SERICS (PE0000014), project Tech4You (ECS0000009), and MUR project PRIN 2022 EPICA (H53D23003660006).

## Declaration on Generative AI

The author(s) have not employed any Generative AI tools.

## References

- [1] G. Alfano, S. Greco, D. Mandaglio, F. Parisi, R. Shahbazian, I. Trubitsyna, Even-if explanations: Formal foundations, priorities and complexity, in: *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 2025, pp. 15347–15355.
- [2] R. McCloy, R. M. Byrne, Semifactual “even if” thinking, *Thinking & reasoning* 8 (2002) 41–67.
- [3] R. Guidotti, Counterfactual explanations and how to find them: literature review and benchmarking, *Data Mining and Knowledge Discovery* (2022) 1–55.
- [4] P. Barceló, M. Monet, J. Pérez, B. Subercaseaux, Model interpretability through the lens of computational complexity, in: *Proceedings of Advances in Neural Information Processing Systems*, 2020.
- [5] S. Aryal, M. T. Keane, Even if explanations: Prior work, desiderata & benchmarks for semi-factual xai, in: *Proceedings of International Joint Conference on Artificial Intelligence (IJCAI)*, 2023, pp. 6526–6535.

- [6] I. Wegener, Bdds—design, analysis, complexity, and applications, *Discrete Applied Mathematics* 138 (2004) 229–251.
- [7] F. Rossi, K. B. Venable, T. Walsh, *A Short Introduction to Preferences: Between Artificial Intelligence and Social Choice*, Synthesis Lectures on Artificial Intelligence and Machine Learning, Morgan & Claypool Publishers, 2011.
- [8] G. Alfano, S. Greco, F. Parisi, I. Trubitsyna, On preferences and priority rules in abstract argumentation, in: *Proceedings of International Joint Conference on Artificial Intelligence (IJCAI)*, 2022, pp. 2517–2524.
- [9] G. Alfano, S. Greco, D. Mandaglio, F. Parisi, I. Trubitsyna, Complexity of verification and existence problems in epistemic argumentation framework, in: *ECAI 2023*, IOS Press, 2023, pp. 77–84.
- [10] G. Alfano, S. Greco, F. Parisi, I. Trubitsyna, Abstract argumentation framework with conditional preferences, in: *Proceedings of AAAI Conference on Artificial Intelligence*, 2023, pp. 6218–6227.
- [11] G. Alfano, S. Greco, F. Parisi, I. Trubitsyna, Preferences and constraints in abstract argumentation, in: *Proceedings of International Joint Conference on Artificial Intelligence (IJCAI)*, 2023, pp. 3095–3103.
- [12] G. Brewka, I. Niemelä, M. Truszczynski, Answer set optimization, in: *Proceedings of International Joint Conference on Artificial Intelligence (IJCAI)*, 2003, pp. 867–872.
- [13] J. Marques-Silva, A. Ignatiev, Delivering trustworthy AI through formal XAI, in: *Proceedings of AAAI Conference on Artificial Intelligence*, 2022, pp. 12342–12350.
- [14] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, D. Pedreschi, A survey of methods for explaining black box models, *ACM computing surveys (CSUR)* 51 (2018) 1–42.
- [15] F. Bodria, F. Giannotti, R. Guidotti, F. Naretto, D. Pedreschi, S. Rinzivillo, Benchmarking and survey of explanation methods for black box models, *Data Mining and Knowledge Discovery (2023)* 1–60.
- [16] G. Ciravegna, P. Barbiero, F. Giannini, M. Gori, P. Lió, M. Maggini, S. Melacci, Logic explained networks, *Artificial Intelligence* 314 (2023) 103822.
- [17] M. C. Cooper, J. Marques-Silva, Tractability of explaining classifier decisions, *Artif. Intell.* 316 (2023) 103841.
- [18] E. Dervakos, K. Thomas, G. Filandrianos, G. Stamou, Choose your data wisely: A framework for semantic counterfactuals, *arXiv preprint arXiv:2305.17667* (2023).
- [19] E. Albini, A. Rago, P. Baroni, F. Toni, Relation-based counterfactual explanations for bayesian network classifiers., in: *Proceedings of International Joint Conference on Artificial Intelligence (IJCAI)*, 2020, pp. 451–457.
- [20] G. Audemard, J.-M. Lagniez, P. Marquis, N. Szczepanski, Deriving explanations for decision trees: The impact of domain theories, in: *Proceedings of International Joint Conference on Artificial Intelligence (IJCAI)*, 2024, pp. 3688–3696.
- [21] S. Dandl, G. Casalicchio, B. Bischl, L. Bothmann, Interpretable regional descriptors: Hyperbox-based local explanations, in: *Proceedings of Machine Learning and Knowledge Discovery in Databases*, volume 14171, Springer, 2023, pp. 479–495.
- [22] E. M. Kenny, M. T. Keane, On generating plausible counterfactual and semi-factual explanations for deep learning, in: *Proceedings of AAAI Conference on Artificial Intelligence*, 2021, pp. 11575–11585.
- [23] G. Alfano, S. Greco, F. Parisi, I. Trubitsyna, Counterfactual and Semifactual Explanations in Abstract Argumentation: Formal Foundations, Complexity and Computation, in: *Proceedings of the International Conference on Principles of Knowledge Representation and Reasoning*, 2024, pp. 14–26.
- [24] M. Arenas, P. Barceló, M. Romero Orth, B. Subercaseaux, On computing probabilistic explanations for decision trees, *Proceedings of Advances in Neural Information Processing Systems 35* (2022) 28695–28707.
- [25] O. El Harzli, B. C. Grau, I. Horrocks, Cardinality-minimal explanations for monotonic neural networks, in: *Proceedings of International Joint Conference on Artificial Intelligence (IJCAI)*, 2023, pp. 3677–3685.
- [26] G. Alfano, S. Greco, D. Mandaglio, F. Parisi, I. Trubitsyna, Abstract argumentation frameworks

with strong and weak constraints, *Artificial Intelligence* 336 (2024) 104205.

- [27] L. Zangari, D. Mandaglio, A. Tagarelli, Link prediction on multilayer networks through learning of within-layer and across-layer node-pair structural features and node embedding similarity, in: *Proceedings of the ACM Web Conference 2024*, 2024, pp. 924–935.