On Counterfactual and Semifactual Explanations in Abstract Argumentation

(Discussion Paper)

Gianvincenzo Alfano[1,*,†], Sergio Greco[1,†], Francesco Parisi[1,†] and Irina Trubitsyna[1,†]

[1]Department of Informatics, Modeling, Electronics and System Engineering (DIMES), University of Calabria, Rende, Italy

Abstract

Explainable Artificial Intelligence and Formal Argumentation have received significant attention in recent years. Argumentation frameworks are useful for representing knowledge and reasoning on it. Counterfactual and semifactual explanations are interpretability techniques that provide insights into the outcome of a model by generating alternative hypothetical instances. While there has been important work on counterfactual and semifactual explanations for Machine Learning (ML) models, less attention has been devoted to these kinds of problems in argumentation. In this paper, we discuss counterfactual and semifactual reasoning in abstract Argumentation Framework recently proposed in [1].

Keywords

Formal Argumentation Theory, Explainable AI, Counterfactual and Semifactual Reasoning.

1. Introduction

In the last decades, Formal Argumentation has become an important research field in the area of knowledge representation and reasoning [2]. Argumentation has potential applications in several contexts, including e.g. modeling dialogues, negotiation [3, 4], and persuasion [5]. Dung's Argumentation Framework (AF) is a simple yet powerful formalism for modeling disputes between two or more agents [6]. An AF consists of a set of *arguments* and a binary *attack* relation over the set of arguments that specifies the interactions between arguments: intuitively, if argument a attacks argument b, then b is acceptable only if a is not. Hence, arguments are abstract entities whose status is entirely determined by the attack relation. An AF can be seen as a directed graph, whose nodes represent arguments and edges represent attacks. Several argumentation semantics—e.g. *grounded* (gr), *complete* (co), *stable* (st), *preferred* (pr), and *semi-stable* (sst) [6, 7]—have been defined for AF, leading to the characterization of *σ-extensions*, that intuitively consist of the sets of arguments that can be collectively accepted under semantics $\sigma \in \{\texttt{gr}, \texttt{co}, \texttt{st}, \texttt{pr}, \texttt{sst}\}$.

Example 1. Consider the AF Λ in Figure 1, describing tasting menus proposed by a chef. Intuitively, (s)he proposes to have either `fish`, `meat`, or `pasta` and to drink either `white` wine or `red` wine. However, if serving `meat` or `pasta` then `white` wine is not paired with. AF Λ has four stable extensions (that are also preferred and semi-stable extensions) representing alternative menus: E_1=$\{\texttt{fish}, \texttt{white}\}$, E_2=$\{\texttt{fish}, \texttt{red}\}$, E_3=$\{\texttt{meat}, \texttt{red}\}$, and E_4=$\{\texttt{pasta}, \texttt{red}\}$. □

AIxIA 2024 Discussion Papers - 23rd International Conference of the Italian Association for Artificial Intelligence, Bolzano, Italy, November 25–28, 2024

*Corresponding author.

†These authors contributed equally.

g.alfano@dimes.unical.it (G. Alfano); greco@dimes.unical.it (S. Greco); fparisi@dimes.unical.it (F. Parisi); i.trubitsyna@dimes.unical.it (I. Trubitsyna)

https://gianvincenzoalfano.net/ (G. Alfano); https://people.dimes.unical.it/sergiogreco/ (S. Greco); http://wwwinfo.deis.unical.it/~parisi/ (F. Parisi); https://sites.google.com/dimes.unical.it/itrubitsyna/home (I. Trubitsyna)

0000-0002-7280-4759 (G. Alfano); 0000-0003-2966-3484 (S. Greco); 0000-0001-9977-1355 (F. Parisi); 0000-0002-9031-0672 (I. Trubitsyna)

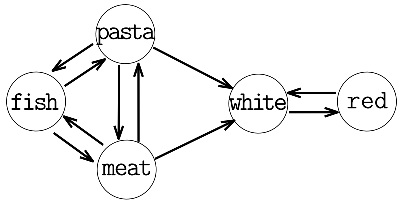

Figure 1: AF Λ of Example 1.

Argumentation semantics can be also defined in terms of labelling [8]. Intuitively, a σ-labelling for an AF is a total function $\mathcal{L}$ assigning to each argument the label **in** if its status is accepted, **out** if its status is rejected, and **und** if its status is undecided under semantics σ. For instance, the σ-labellings for AF Λ of Example 1, with $\sigma \in \{\mathtt{st}, \mathtt{pr}, \mathtt{sst}\}$, are as follows:

$\mathcal{L}_1 = \{\mathbf{in}(\mathtt{fish}), \mathbf{out}(\mathtt{meat}), \mathbf{out}(\mathtt{pasta}), \mathbf{in}(\mathtt{white}), \mathbf{out}(\mathtt{red})\}$,
$\mathcal{L}_2 = \{\mathbf{in}(\mathtt{fish}), \mathbf{out}(\mathtt{meat}), \mathbf{out}(\mathtt{pasta}), \mathbf{out}(\mathtt{white}), \mathbf{in}(\mathtt{red})\}$,
$\mathcal{L}_3 = \{\mathbf{out}(\mathtt{fish}), \mathbf{in}(\mathtt{meat}), \mathbf{out}(\mathtt{pasta}), \mathbf{out}(\mathtt{white}), \mathbf{in}(\mathtt{red})\}$,
$\mathcal{L}_4 = \{\mathbf{out}(\mathtt{fish}), \mathbf{out}(\mathtt{meat}), \mathbf{in}(\mathtt{pasta}), \mathbf{out}(\mathtt{white})\ , \mathbf{in}(\mathtt{red})\}$,

where $\mathcal{L}_i$ corresponds to extension E_i, with $i \in [1..4]$, respectively.

Integrating explanations in argumentation-based reasoners is important for enhancing argumentation and persuasion capabilities of software agents [9, 10, 11, 12]. For this reasons, several researchers explored how to deal with explanations in formal argumentation. Counterfactual and semifactual explanations are types of interpretability techniques that provide insights into the outcome of a model by generating hypothetical instances, known as counterfactuals and semifactual, respectively [13, 14]. On one hand, a counterfactual explanation reveals what should have been different in an instance to obtain a diverse outcome [15]—minimum changes w.r.t. the given instance are usually considered [16]. On the other hand, a semifactual explanation provides a maximally-changed instance yielding the same outcome of that considered [17].

While there has been interesting work on counterfactual and semifactual explanations for ML models, e.g. [18, 19, 20, 21, 22, 23], less attention has been devoted to these problems in argumentation.

In this paper, we discuss counterfactual and semifactual reasoning in AF [1]. Analogously to counterfactual explanations in ML that reveal what should have been minimally different in an instance to obtain a different outcome, our counterfactuals tell what should have been minimally different in a solution, i.e. a σ-labeling with a given acceptance status for a goal argument, to obtain an alternative solution where the goal has a different status.

Example 2. Continuing with Example 1, assume that the chef suggests the menu $\mathcal{L}_3 = \{\mathbf{out}(\mathtt{fish}), \mathbf{in}(\mathtt{meat}), \mathbf{out}(\mathtt{pasta}), \mathbf{out}(\mathtt{white}), \mathbf{in}(\mathtt{red})\}$ and the customer replies that (s)he likes everything except meat (as (s)he is vegetarian). Therefore, the chef looks for the closest menus not containing meat, that are $\mathcal{L}_2 = \{\mathbf{in}(\mathtt{fish}), \mathbf{out}(\mathtt{meat}), \mathbf{out}(\mathtt{pasta}), \mathbf{out}(\mathtt{white}), \mathbf{in}(\mathtt{red})\}$ and $\mathcal{L}_4 = \{\mathbf{out}(\mathtt{fish}), \mathbf{out}(\mathtt{meat}), \mathbf{in}(\mathtt{pasta}), \mathbf{out}(\mathtt{white}), \mathbf{in}(\mathtt{red})\}$. In this context, we say that $\mathcal{L}_2$ and $\mathcal{L}_4$ are *counterfactuals* for $\mathcal{L}_3$ w.r.t. the goal argument meat. □

Given a σ-labelling $\mathcal{L}$ of an AF Λ, and a goal argument g, a *counterfactual* of $\mathcal{L}$ w.r.t. g is a closest σ-labelling $\mathcal{L}'$ of Λ that changes the acceptance status of g. Hence, counterfactuals explain how to minimally change a solution to avoid a given acceptance status of a goal argument.

In contrast, semifactuals give the maximal changes to the considered solution in order to keep the status of a goal argument. That is, a *semifactual* of $\mathcal{L}$ w.r.t. goal g is a farthest σ-labelling $\mathcal{L}'$ of Λ that keeps the acceptance status of argument g.

Example 3. Continuing with Example 1, suppose now that a customer has tasted menu $\mathcal{L}_3 = \{\mathbf{out}(\mathtt{fish}), \mathbf{in}(\mathtt{meat}), \mathbf{out}(\mathtt{pasta}), \mathbf{out}(\mathtt{white}), \mathbf{in}(\mathtt{red})\}$, and asks to try completely new flavors while still maintaining the previous choice of wine as (s)he liked it a lot. Here the chef is interested in the farthest menus containing red wine. These menus are $\mathcal{L}_2 = \{\mathbf{in}(\mathtt{fish}), \mathbf{out}(\mathtt{meat}), \mathbf{out}(\mathtt{pasta}),$

$\mathbf{out}(\texttt{white})$, $\mathbf{in}(\texttt{red})\}$ and $\mathcal{L}_4 = \{\mathbf{out}(\texttt{fish}), \mathbf{out}(\texttt{meat}), \mathbf{in}(\texttt{pasta}), \mathbf{out}(\texttt{white}), \mathbf{in}(\texttt{red})\}$. We say that the labellings $\mathcal{L}_2$ and $\mathcal{L}_4$ are *semifactuals* for the labelling $\mathcal{L}_3$ w.r.t. the goal argument $\texttt{red}$. □

2. Counterfactual and Semifactual Reasoning

Intuitively, a counterfactual of a given σ-labelling w.r.t. a given goal argument g is a minimum-distance σ-labelling altering the acceptance status of g. More in detail, let $\langle \mathrm{A}, \mathrm{R}\rangle$ be an AF, $\sigma \in \{\texttt{gr}, \texttt{co}, \texttt{st}, \texttt{pr}, \texttt{sst}\}$ a semantics, $g \in \mathrm{A}$ a goal argument, and $\mathcal{L}$ a σ-labelling for $\langle \mathrm{A}, \mathrm{R}\rangle$. Then, a labelling $\mathcal{L}' \in \sigma(\langle \mathrm{A}, \mathrm{R}\rangle)$ is a *counterfactual* of $\mathcal{L}$ w.r.t. g if:

(i) $\mathcal{L}(g) \neq \mathcal{L}'(g)$, and

(ii) there exists no $\mathcal{L}'' \in \sigma(\langle \mathrm{A}, \mathrm{R}\rangle)$ such that $\mathcal{L}(g) \neq \mathcal{L}''(g)$ and $\delta(\mathcal{L}, \mathcal{L}'') < \delta(\mathcal{L}, \mathcal{L}')$.

We use $\mathcal{CF}^\sigma(g, \mathcal{L})$ to denote the set of counterfactuals of $\mathcal{L}$ w.r.t. g.

Example 4. Continuing with Example 2, under stable semantics, for the labelling $\mathcal{L}_3 = \{\mathbf{out}(\texttt{fish}), \mathbf{in}(\texttt{meat}), \mathbf{out}(\texttt{pasta}), \mathbf{out}(\texttt{white}), \mathbf{in}(\texttt{red})\}$, we have that $\mathcal{L}_2 = \{\mathbf{in}(\texttt{fish}), \mathbf{out}(\texttt{meat}), \mathbf{out}(\texttt{pasta}), \mathbf{out}(\texttt{white}), \mathbf{in}(\texttt{red})\}$ and $\mathcal{L}_4 = \{\mathbf{out}(\texttt{fish}), \mathbf{out}(\texttt{meat}), \mathbf{in}(\texttt{pasta}), \mathbf{out}(\texttt{white}), \mathbf{in}(\texttt{red})\}$ are its only counterfactuals w.r.t. argument $\texttt{meat}$, as their distance, $\delta(\mathcal{L}_3, \mathcal{L}_2) = \delta(\mathcal{L}_3, \mathcal{L}_4) = 2$, is minimal. The other labelling $\mathcal{L}_1 = \{\mathbf{in}(\texttt{fish}), \mathbf{out}(\texttt{meat}), \mathbf{out}(\texttt{pasta}), \mathbf{in}(\texttt{white}), \mathbf{out}(\texttt{red})\}$, such that $\mathcal{L}_3(\texttt{meat}) \neq \mathcal{L}_1(\texttt{meat})$ is not at minimum distance as $\delta(\mathcal{L}_3, \mathcal{L}_1) = 4 > \delta(\mathcal{L}_3, \mathcal{L}_2)$. Therefore, $\mathcal{CF}^{\texttt{st}}(\texttt{meat}, \mathcal{L}_3) = \{\mathcal{L}_2, \mathcal{L}_4\}$. □

The concept of semifactual is, in a sense, symmetrical and complementary to that of a counterfactual. Indeed, let $\langle \mathrm{A}, \mathrm{R}\rangle$ be an AF, $\sigma \in \{\texttt{gr}, \texttt{co}, \texttt{st}, \texttt{pr}, \texttt{sst}\}$ a semantics, $g \in \mathrm{A}$ a goal argument, and $\mathcal{L}$ a σ-labelling for $\langle \mathrm{A}, \mathrm{R}\rangle$. Then, $\mathcal{L}' \in \sigma(\langle \mathrm{A}, \mathrm{R}\rangle)$ is a semifactual of $\mathcal{L}$ w.r.t. g if:

(i) $\mathcal{L}(g) = \mathcal{L}'(g)$, and

(ii) there exists no $\mathcal{L}'' \in \sigma(\langle \mathrm{A}, \mathrm{R}\rangle)$ such that $\mathcal{L}(g) = \mathcal{L}''(g)$ and $\delta(\mathcal{L}, \mathcal{L}'') > \delta(\mathcal{L}, \mathcal{L}')$.

We use $\mathcal{SF}^\sigma(g, \mathcal{L})$ to denote the set of semifactuals of $\mathcal{L}$ w.r.t. g.

Example 5. Consider the stable labelling $\mathcal{L}_3 = \{\mathbf{out}(\texttt{fish}), \mathbf{in}(\texttt{meat}), \mathbf{out}(\texttt{pasta}), \mathbf{out}(\texttt{white}), \mathbf{in}(\texttt{red})\}$ for the AF of Example 3. We have that $\mathcal{L}_2 = \{\mathbf{in}(\texttt{fish}), \mathbf{out}(\texttt{meat}), \mathbf{out}(\texttt{pasta}), \mathbf{out}(\texttt{white}), \mathbf{in}(\texttt{red})\}$ and $\mathcal{L}_4 = \{\mathbf{out}(\texttt{fish}), \mathbf{out}(\texttt{meat}), \mathbf{in}(\texttt{pasta}), \mathbf{out}(\texttt{white}), \mathbf{in}(\texttt{red})\}$ are the only semifactuals of $\mathcal{L}_3$ w.r.t. the argument $\texttt{red}$ as there is no other $\texttt{st}$-labelling agreeing on $\texttt{red}$ and having distance greater than $\delta(\mathcal{L}_3, \mathcal{L}_2) = \delta(\mathcal{L}_3, \mathcal{L}_4) = 2$. In fact, $\mathcal{L}_1 = \{\mathbf{in}(\texttt{fish}), \mathbf{out}(\texttt{meat}), \mathbf{out}(\texttt{pasta}), \mathbf{in}(\texttt{white}), \mathbf{out}(\texttt{red})\}$, having distance $\delta(\mathcal{L}_3, \mathcal{L}_1) = 4$, is not a semifactual for $\mathcal{L}_3$ w.r.t. $\texttt{red}$ as $\mathcal{L}_1(\texttt{red}) \neq \mathcal{L}_3(\texttt{red})$. Thus, $\mathcal{SF}^{\texttt{st}}(\texttt{red}, \mathcal{L}_3) = \{\mathcal{L}_2, \mathcal{L}_4\}$. □

2.1. Existence and Verification Problems

Finding a counterfactual (resp., semifactual) means looking for a minimum (resp., maximum) distance labelling. The first problem we consider is a natural decision version of that problem.

Given as input an AF $\Lambda = \langle \mathrm{A}, \mathrm{R}\rangle$, a semantics $\sigma \in \{\texttt{co}, \texttt{st}, \texttt{pr}, \texttt{sst}\}$, a goal argument $g \in \mathrm{A}$, an integer $k \in \mathbb{N}$, and a σ-labelling $\mathcal{L} \in \sigma(\Lambda)$, CF-EX$^\sigma$ (resp., SF-EX$^\sigma$) is the problem of deciding whether there exists a labelling $\mathcal{L}' \in \sigma(\Lambda)$ s.t. $\mathcal{L}(g) \neq \mathcal{L}'(g)$ (resp., $\mathcal{L}(g) = \mathcal{L}'(g)$) and $\delta(\mathcal{L}, \mathcal{L}') \leq k$ (resp., $\delta(\mathcal{L}, \mathcal{L}') \geq k$).

The complexity of the existence problem under counterfactual and semifactual reasoning (i.e., CF-EX$^\sigma$ and SF-EX$^\sigma$) has been recently proved to be $i)$ NP-complete for $\sigma \in \{\texttt{co}, \texttt{st}\}$; and $ii)$ Σ_2^p-complete for $\sigma \in \{\texttt{pr}, \texttt{sst}\}$ [1].

A problem related to CF-EX$^\sigma$ and SF-EX$^\sigma$ is that of verifying whether a given labelling $\mathcal{L}'$ is a counterfactual/semifactual for $\mathcal{L}$ and g, and thus that the distance between the two labelling is minimum/maximum.

Given as input an AF $\Lambda = \langle \mathrm{A}, \mathrm{R} \rangle$, a semantics $\sigma \in \{\mathtt{co}, \mathtt{st}, \mathtt{pr}, \mathtt{sst}\}$, a goal argument $g \in \mathrm{A}$, a σ-labelling $\mathcal{L} \in \sigma(\Lambda)$, and a labelling $\mathcal{L}'$, $\mathsf{CF\text{-}VE}^\sigma$ (resp., $\mathsf{SF\text{-}VE}^\sigma$) is the problem of deciding whether $\mathcal{L}'$ belongs to $\mathcal{CF}^\sigma(g, \mathcal{L})$ (resp., $\mathcal{SF}^\sigma(g, \mathcal{L})$).

The problems CF-VE$^\sigma$ and CF-EX$^\sigma$ (resp., SF-VE$^\sigma$ and SF-EX$^\sigma$) are on the same level of the polynomial hierarchy. In fact $\mathsf{CF\text{-}VE}^\sigma$ and $\mathsf{SF\text{-}VE}^\sigma$ are $i)$ co$\mathbb{NP}$-complete for $\sigma \in \{\mathtt{co}, \mathtt{st}\}$; and $ii)$ Π_2^p-complete for $\sigma \in \{\mathtt{pr}, \mathtt{sst}\}$ [1].

3. Conclusions

Several researchers explored how to deal with explanations with in formal argumentation [24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38]. Counterfactual reasoning in AF has been firstly introduced in [39], where considering sentences of the form "if a were rejected, then b would be accepted", an AF Λ is modified to another AF Λ' such that (*i*) argument a which is accepted in Λ is rejected in Λ' (*ii*) and the Λ' is as close as possible to Λ.

However, none of the above-mentioned approaches deals with semifactual reasoning and most of them manipulate the AF by adding arguments or meta-knowledge. In contrast, in our approach, focusing on a given AF, novel definitions of counterfactual and semifactual are introduced to help understand what should be different in a solution (not in the AF) to accommodate a user requirement concerning a given goal. It turns out that the complexity of the considered problems is not lower than those of corresponding classical problems in AF, and is provably higher for fundamental problems such as the verification problem.

Although counterfactual- and semifactual-based reasoning suffers from high computational complexity (as many other computational problems in argumentation [40, 41, 42, 43, 44, 45, 46, 47]), several tools and techniques emerged in the last few years that can tackle such kinds of computational issues, including ASP- and SAT-based solvers. This is witnessed by the several efficient approaches presented at the ICCMA competition,[1] which aims at nurturing research and development of implementations for computational models of argumentation.

Acknowledgements

We acknowledge the support from project Tech4You (ECS0000009), and PNRR MUR projects FAIR (PE0000013) and SERICS (PE00000014).

Declaration on Generative AI

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

References

[1] G. Alfano, S. Greco, F. Parisi, I. Trubitsyna, Counterfactual and Semifactual Explanations in Abstract Argumentation: Formal Foundations, Complexity and Computation, in: Proc. of International Conference on Principles of Knowledge Representation and Reasoning (KR), 2024, pp. 14–26.
[2] D. Gabbay, M. Giacomin, G. R. Simari, M. Thimm (Eds.), Handbook of Formal Argumentation, volume 2, College Publications, 2021.
[3] L. Amgoud, Y. Dimopoulos, P. Moraitis, A unified and general framework for argumentation-based negotiation, in: Proc. of International Joint Conference on Autonomous Agents and Multiagent Systems, 2007, p. 158.

[1]https://argumentationcompetition.org

[4] Y. Dimopoulos, J. Mailly, P. Moraitis, Argumentation-based negotiation with incomplete opponent profiles, in: Proc. of International Joint Conference on Autonomous Agents and Multiagent Systems, 2019, pp. 1252–1260.
[5] H. Prakken, Models of persuasion dialogue, in: Argumentation in Artificial Intelligence, 2009, pp. 281–300.
[6] P. M. Dung, On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games, Artif. Intell. 77 (1995) 321–358.
[7] M. Caminada, Semi-stable semantics, in: Proc. of COMMA, 2006, pp. 121–130.
[8] P. Baroni, M. Caminada, M. Giacomin, An introduction to argumentation semantics, Knowl. Eng. Rev. 26 (2011) 365–410.
[9] B. Moulin, H. Irandoust, M. Bélanger, G. Desbordes, Explanation and argumentation capabilities: Towards the creation of more persuasive agents, Artificial Intelligence Review 17 (2002) 169–222.
[10] F. Bex, D. Walton, Combining explanation and argumentation in dialogue, Argument & Computation 7 (2016) 55–68.
[11] K. Cyras, D. Birch, Y. Guo, F. Toni, R. Dulay, S. Turvey, D. Greenberg, T. Hapuarachchi, Explanations by arbitrated argumentative dispute, Expert Systems with Applications 127 (2019) 141–156.
[12] T. Miller, Explanation in artificial intelligence: Insights from the social sciences, Artif. Intell. 267 (2019) 1–38.
[13] D. Kahneman, A. Tversky, The simulation heuristic, National Technical Information Service, 1981.
[14] R. McCloy, R. M. Byrne, Semifactual "even if" thinking, Thinking & reasoning 8 (2002) 41–67.
[15] R. Guidotti, Counterfactual explanations and how to find them: literature review and benchmarking, Data Mining and Knowledge Discovery (2022) 1–55.
[16] P. Barceló, M. Monet, J. Pérez, B. Subercaseaux, Model interpretability through the lens of computational complexity, in: Proc. of Advances in Neural Information Processing Systems, 2020.
[17] E. M. Kenny, M. T. Keane, On generating plausible counterfactual and semi-factual explanations for deep learning, in: Proc. of AAAI Conference on Artificial Intelligence, 2021, pp. 11575–11585.
[18] Y. Wu, L. Zhang, X. Wu, Counterfactual fairness: Unidentification, bound and algorithm, in: Proc. of International Joint Conference on Artificial Intelligence (IJCAI), 2019, pp. 1438–1444.
[19] E. Albini, A. Rago, P. Baroni, F. Toni, Relation-based counterfactual explanations for bayesian network classifiers., in: Proc. of International Joint Conference on Artificial Intelligence (IJCAI), 2020, pp. 451–457.
[20] G. Alfano, S. Greco, D. Mandaglio, F. Parisi, R. Shahbazian, I. Trubitsyna, Even-if explanations: Formal foundations, priorities and complexity, in: In Proc. of AAAI Conference on Artificial Intelligence, 2025, p. (to appear).
[21] P. Romashov, M. Gjoreski, K. Sokol, M. V. Martinez, M. Langheinrich, Baycon: Model-agnostic bayesian counterfactual generator, in: Proc. of International Joint Conference on Artificial Intelligence (IJCAI), 2022, pp. 23–29.
[22] S. Dandl, G. Casalicchio, B. Bischl, L. Bothmann, Interpretable regional descriptors: Hyperbox-based local explanations, in: Proc. of Machine Learning and Knowledge Discovery in Databases, volume 14171, 2023, pp. 479–495.
[23] S. Aryal, M. T. Keane, Even if explanations: Prior work, desiderata & benchmarks for semi-factual xai, in: Proc. of International Joint Conference on Artificial Intelligence (IJCAI), 2023, pp. 6526–6535.
[24] K. Cyras, A. Rago, E. Albini, P. Baroni, F. Toni, Argumentative XAI: A survey, in: Proc. of Proc. of International Joint Conference on Artificial Intelligence (IJCAI), 2021, pp. 4392–4399.
[25] A. Vassiliades, N. Bassiliades, T. Patkos, Argumentation and explainable artificial intelligence: a survey, Knowl. Eng. Rev. 36 (2021) e5.
[26] R. Craven, F. Toni, Argument graphs and assumption-based argumentation, Artif. Intell. 233 (2016) 1–59.
[27] P. M. Dung, R. A. Kowalski, F. Toni, Assumption-based argumentation, in: Argumentation in Artificial Intelligence, Springer, 2009, pp. 199–218.
[28] N. D. Hung, Computing probabilistic assumption-based argumentation, in: Proc. of Pacific Rim

International Conference on Artificial Intelligence (PRICAI), 2016, pp. 152–166.
[29] P. Dung, P. Mancarella, F. Toni, Computing ideal sceptical argumentation, Artif. Intell. 171 (2007) 642–674.
[30] P. M. Thang, P. M. Dung, N. D. Hung, Towards a common framework for dialectical proof procedures in abstract argumentation, Journal of Logic and Computation 19 (2009) 1071–1109.
[31] G. Alfano, M. Calautti, S. Greco, F. Parisi, I. Trubitsyna, Explainable acceptance in probabilistic abstract argumentation: Complexity and approximation, in: Proc. of International Conference on Principles of Knowledge Representation and Reasoning (KR), 2020, pp. 33–43.
[32] G. Alfano, M. Calautti, S. Greco, F. Parisi, I. Trubitsyna, Explainable acceptance in probabilistic and incomplete abstract argumentation frameworks, Artif. Intell. 323 (2023) 103967.
[33] R. Baumann, M. Ulbricht, Choices and their consequences - explaining acceptable sets in abstract argumentation frameworks, in: Proc. of International Conference on Principles of Knowledge Representation and Reasoning (KR), 2021, pp. 110–119.
[34] M. Ulbricht, J. P. Wallner, Strong explanations in abstract argumentation, in: Proc. of AAAI Conference on Artificial Intelligence, 2021, pp. 6496–6504.
[35] G. Brewka, M. Ulbricht, Strong explanations for nonmonotonic reasoning, in: Description Logic, Theory Combination, and All That, volume 11560 of *Lecture Notes in Computer Science*, 2019, pp. 135–146.
[36] G. Brewka, M. Thimm, M. Ulbricht, Strong inconsistency, Artif. Intell. 267 (2019) 78–117.
[37] Z. G. Saribatur, J. P. Wallner, S. Woltran, Explaining non-acceptability in abstract argumentation, in: Proc. of European Conference on Artificial Intelligence (ECAI), 2020, pp. 881–888.
[38] O. Cocarascu, A. Rago, F. Toni, Extracting dialogical explanations for review aggregations with argumentative dialogical agents, in: Proc. of AAMAS, 2019, pp. 1261–1269.
[39] C. Sakama, Counterfactual reasoning in argumentation frameworks., in: COMMA, 2014, pp. 385–396.
[40] G. Alfano, S. Greco, F. Parisi, On scaling the enumeration of the preferred extensions of abstract argumentation frameworks, in: Proceedings of ACM/SIGAPP Symposium on Applied Computing (SAC), 2019, pp. 1147–1153.
[41] G. Alfano, S. Greco, F. Parisi, Incremental computation in dynamic argumentation frameworks, IEEE Intell. Syst. 36 (2021) 80–86.
[42] G. Alfano, S. Greco, F. Parisi, I. Trubitsyna, Preferences and constraints in abstract argumentation, in: Proc. of International Joint Conference on Artificial Intelligence (IJCAI), ijcai.org, 2023, pp. 3095–3103.
[43] G. Alfano, S. Greco, F. Parisi, I. Trubitsyna, On acceptance conditions in abstract argumentation frameworks, Inf. Sci. 625 (2023) 757–779.
[44] G. Alfano, S. Greco, F. Parisi, I. Trubitsyna, Epistemic abstract argumentation framework: Formal foundations, computation and complexity, in: Proc. of International Conference on Autonomous Agents and Multiagent Systems (AAMAS), ACM, 2023, pp. 409–417.
[45] G. Alfano, S. Greco, F. Parisi, I. Trubitsyna, Abstract argumentation framework with conditional preferences, in: In Proc. of AAAI Conference on Artificial Intelligence, 2023, pp. 6218–6227.
[46] G. Alfano, S. Greco, D. Mandaglio, F. Parisi, I. Trubitsyna, Abstract argumentation frameworks with strong and weak constraints, Artif. Intell. 336 (2024) 104205.
[47] G. Alfano, S. Greco, F. Parisi, I. Trubitsyna, Complexity of credulous and skeptical acceptance in epistemic argumentation framework, in: Proc. of AAAI Conference on Artificial Intelligence, 2024, pp. 10423–10432.