Proceedings of the 21st International Conference on Principles of Knowledge Representation and Reasoning
Main Track

Counterfactual and Semifactual Explanations in Abstract Argumentation: Formal Foundations, Complexity and Computation

Gianvincenzo Alfano , Sergio Greco , Francesco Parisi , Irina Trubitsyna
Department of Informatics, Modeling, Electronics and System Engineering,
University of Calabria, Italy
{g.alfano, greco, fparisi, trubitsyna}@dimes.unical.it

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 explore counterfactual and semifactual reasoning in abstract Argumentation Framework. We investigate the computational complexity of counterfactual- and semifactual-based reasoning problems, showing that they are generally harder than classical argumentation problems such as credulous and skeptical acceptance. Finally, we show that counterfactual and semifactual queries can be encoded in weak-constrained Argumentation Framework, and provide a computational strategy through ASP solvers.

1 Introduction

In the last decades, Formal Argumentation has become an important research field in the area of knowledge representation and reasoning (Gabbay et al. 2021). Argumentation has potential applications in several contexts, including e.g. modeling dialogues, negotiation (Amgoud, Dimopoulos, and Moraitis 2007; Dimopoulos, Mailly, and Moraitis 2019), and persuasion (Prakken 2009). Dung's Argumentation Framework (AF) is a simple yet powerful formalism for modeling disputes between two or more agents (Dung 1995). 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* ($\mathtt{gr}$), *complete* ($\mathtt{co}$), *stable* ($\mathtt{st}$), *preferred* ($\mathtt{pr}$), and *semi-stable* ($\mathtt{sst}$) (Dung 1995; Caminada 2006)—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 $\sigma \in \{\mathtt{gr}, \mathtt{co}, \mathtt{st}, \mathtt{pr}, \mathtt{sst}\}$.

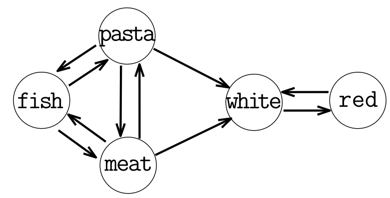

Figure 1: AF Λ of Example 1.

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

Argumentation semantics can be also defined in terms of labelling (Baroni, Caminada, and Giacomin 2011). Intuitively, a σ-labelling for an AF is a total function $\mathcal{L}$ assigning to each argument the label $\mathbf{in}$ if its status is accepted, $\mathbf{out}$ if its status is rejected, and $\mathbf{und}$ 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 (Moulin et al. 2002; Bex and Walton 2016; Cyras et al. 2019; Miller 2019). For this reasons, several researchers explored how to deal with explanations in formal argumentation (see related work in Section 7). 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 (Kahneman and Tversky 1981; McCloy and Byrne 2002). On one hand, a counterfactual explanation reveals what should have been different in an instance to obtain a diverse outcome (Guidotti 2022)—minimum changes w.r.t. the given instance are usually considered (Barceló et al. 2020). On the other hand, a semifactual explanation provides a maximally-changed instance yielding the same outcome of that considered (Kenny and Keane 2021).

While there has been interesting work on counterfactual and semifactual explanations for ML models, e.g. (Wu, Zhang, and Wu 2019; Albini et al. 2020; Romashov et al. 2022; Dandl et al. 2023; Aryal and Keane 2023), less attention has been devoted to these problems in argumentation.

In this paper, we explore counterfactual and semifactual reasoning in AF. 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 $\mathcal{L}_3 = \{\mathbf{out}(\texttt{fish}), \mathbf{in}(\texttt{meat}), \mathbf{out}(\texttt{pasta}), \mathbf{out}(\texttt{white}), \mathbf{in}(\texttt{red})\}$ 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 $\mathcal{L}_2 = \{\mathbf{in}(\texttt{fish}), \mathbf{out}(\texttt{meat}), \mathbf{out}(\texttt{pasta}), \mathbf{out}(\texttt{white}), \mathbf{in}(\texttt{red})\}$ and $\mathcal{L}_4 = \{\mathbf{out}(\texttt{fish}), \mathbf{out}(\texttt{meat}), \mathbf{in}(\texttt{pasta}), \mathbf{out}(\texttt{white}), \mathbf{in}(\texttt{red})\}$. 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 $\mathcal{L}_3 = \{\mathbf{out}(\texttt{fish}), \mathbf{in}(\texttt{meat}), \mathbf{out}(\texttt{pasta}), \mathbf{out}(\texttt{white}), \mathbf{in}(\texttt{red})\}$, 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 $\mathcal{L}_2 = \{\mathbf{in}(\texttt{fish}), \mathbf{out}(\texttt{meat}), \mathbf{out}(\texttt{pasta}), \mathbf{out}(\texttt{white}), \mathbf{in}(\texttt{red})\}$ and $\mathcal{L}_4 = \{\mathbf{out}(\texttt{fish}), \mathbf{out}(\texttt{meat}), \mathbf{in}(\texttt{pasta}), \mathbf{out}(\texttt{white}), \mathbf{in}(\texttt{red})\}$. 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 red. □

In this paper we introduce the concepts of counterfactual and semifactual explanations in AF. To the best of our knowledge, this is the first work addressing explainability queries in AF under both counterfactual and semifactual reasoning.

Contributions. Our main contributions are as follows.

- We introduce counterfactual-based and semifactual-based reasoning problems for existence, verification, and credulous and skeptical acceptance in AF.
- We investigate the complexity of above-mentioned problems showing that they are generally harder than classical ones; this particularly holds for verification and credulous acceptance problems. Notably, our results hold even for different generalizations of the concepts of counterfactual and semifactual, of measures for computing the distance between labellings, and for multiple goals (cf. Section 6).
- We show that the above-mentioned problems can be encoded through weak-constrained AF, that is a generalization of AF with both strong and weak constraints.
- Finally, we provide an algorithm for the computation of counterfactuals and semifactuals which makes use of well-known ASP encoding of AF semantics.

2 Preliminaries

In this section, after briefly recalling some complexity classes, we review the Dung's framework.

2.1 Complexity Classes

We recall the main complexity classes used in the paper. PTIME (or simply P) consists of the problems that can be decided in polynomial-time. Moreover, we have that
•$\Sigma_0^p = \Pi_0^p = \Delta_0^p = P$; •$\Sigma_1^p = NP$ and $\Pi_1^p = coNP$; and
•$\Delta_h^p = P^{\Sigma_{h-1}^p}$, $\Sigma_h^p = NP^{\Sigma_{h-1}^p}$, and $\Pi_h^p = co\Sigma_h^p$, $\forall h > 0$ (Papadimitriou 1994). Thus, P^C (resp., NP^C) denotes the class of problems that can be decided in polynomial time using an oracle in the class C by a deterministic (resp., non-deterministic) Turing machine. The class $\Theta_h^p = \Delta_h^p[log\ n]$ denotes the subclass of Δ_h^p consisting of the problems that can be decided in polynomial time by a deterministic Turing machine performing $O(log\ n)$ calls to an oracle in the class Σ_{h-1}^p. It is known that $\Sigma_h^p \subseteq \Theta_{h+1}^p \subseteq \Delta_{h+1}^p \subseteq \Sigma_{h+1}^p \subseteq PSPACE$ and $\Pi_h^p \subseteq \Theta_{h+1}^p \subseteq \Delta_{h+1}^p \subseteq \Pi_{h+1}^p \subseteq PSPACE$.

2.2 Argumentation Framework

An abstract *Argumentation Framework* (AF) is a pair $\langle \mathrm{A}, \mathrm{R} \rangle$, where A is a set of *arguments* and $\mathrm{R} \subseteq \mathrm{A} \times \mathrm{A}$ is a set of *attacks*. If $(a, b) \in \mathrm{R}$ then we say that a attacks b.

Given an AF $\Lambda = \langle \mathrm{A}, \mathrm{R} \rangle$ and a set $S \subseteq \mathrm{A}$ of arguments, an argument $a \in \mathrm{A}$ is said to be *i*) *defeated* w.r.t. S iff $\exists b \in S$ such that $(b, a) \in \mathrm{R}$, and *ii*) *acceptable* w.r.t. S iff for every argument $b \in \mathrm{A}$ with $(b, a) \in \mathrm{R}$, there is $c \in S$ such that $(c, b) \in \mathrm{R}$. The sets of defeated and acceptable arguments w.r.t. S are as follows (where Λ is fixed):

- $\mathit{Def}(S) = \{a \in \mathrm{A} \mid \exists (b, a) \in \mathrm{R} \,.\, b \in S\}$;
- $\mathit{Acc}(S) = \{a \in \mathrm{A} \mid \forall (b, a) \in \mathrm{R} \,.\, b \in \mathit{Def}(S)\}$.

Given an AF $\langle \mathrm{A}, \mathrm{R} \rangle$, a set $S \subseteq \mathrm{A}$ of arguments is said to be *i*) *conflict-free* iff $S \cap Def(S) = \emptyset$; *ii*) *admissible* iff it is conflict-free and $S \subseteq Acc(S)$.

Different argumentation semantics have been proposed to characterize collectively acceptable sets of arguments, called *extensions* (Dung 1995; Caminada 2006). Every extension is an admissible set satisfying additional conditions. Specifically, the complete, preferred, stable, semi-stable, and grounded extensions of an AF are defined as follows.

Given an AF $\langle \mathrm{A}, \mathrm{R} \rangle$, a set $S \subseteq \mathrm{A}$ is an *extension* called:

- *complete* (co) iff it is an admissible set and $S = Acc(S)$;
- *preferred* (pr) iff it is a $\subseteq$-maximal complete extension;
- *stable* (st) iff it is a total preferred extension, i.e. a preferred extension such that $S \cup Def(S) = \mathrm{A}$;
- *semi-stable* (sst) iff it is a preferred extension such that $S \cup Def(S)$ is maximal (w.r.t. $\subseteq$);
- *grounded* (gr) iff it is a $\subseteq$-minimal complete extension.

The argumentation semantics can be also defined in terms of *labelling* (Baroni, Caminada, and Giacomin 2011). A labelling for an AF $\langle \mathrm{A}, \mathrm{R}\rangle$ is a total function $\mathcal{L} : \mathrm{A} \rightarrow \{\mathbf{in}, \mathbf{out}, \mathbf{und}\}$ assigning to each argument a label: $\mathcal{L}(a) = \mathbf{in}$ means that a is accepted, $\mathcal{L}(a) = \mathbf{out}$ means that a is rejected, and $\mathcal{L}(a) = \mathbf{und}$ means that a is undecided.

Hereafter, with a little abuse of notation, a labelling will be also used to denote a set of labelled arguments, that is $\mathcal{L}$ also denotes the set $\{\ell(a) \mid \ell \in \{\mathbf{in}, \mathbf{out}, \mathbf{und}\} \wedge a \in A \wedge \mathcal{L}(a) = \ell\}$. Moreover, we also use the notations $\mathbf{in}(\mathcal{L}) = \{a \mid a \in \mathrm{A} \wedge \mathcal{L}(a) = \mathbf{in}\}$, $\mathbf{out}(\mathcal{L}) = \{a \mid a \in \mathrm{A} \wedge \mathcal{L}(a) = \mathbf{out}\}$, and $\mathbf{und}(\mathcal{L}) = \{a \mid a \in \mathrm{A} \wedge \mathcal{L}(a) = \mathbf{und}\}$, to denote the sets of arguments labelled as $\mathbf{in}$, $\mathbf{out}$, and $\mathbf{und}$ by $\mathcal{L}$, respectively. For any labelling $\mathcal{L}$ and argument a, $\mathcal{L}[a] = \ell(a)$, where $\ell \in \{\mathbf{in}, \mathbf{out}, \mathbf{und}\}$ and $\mathcal{L}(a) = \ell$, denotes the projection of $\mathcal{L}$ over a.

Given an AF $\Lambda = \langle \mathrm{A}, \mathrm{R}\rangle$, a labelling $\mathcal{L}$ for A is said to be *admissible (or legal)* if $\forall a \in \mathbf{in}(\mathcal{L}) \cup \mathbf{out}(\mathcal{L})$ it holds that:
(*i*) $\mathcal{L}(a) = \mathbf{out}$ iff $\exists (b, a) \in \mathrm{R}$ such that $\mathcal{L}(b) = \mathbf{in}$; and
(*ii*) $\mathcal{L}(a) = \mathbf{in}$ iff $\forall (b, a) \in \mathrm{R}$, $\mathcal{L}(b) = \mathbf{out}$ holds.

Moreover, $\mathcal{L}$ is a *complete* labelling (or co-labelling) iff conditions (*i*) and (*ii*) hold for all arguments $a \in \mathrm{A}$.[1]

Between complete extensions and complete labellings there is a bijective mapping defined as follows: for each extension E there is a unique labelling $\mathcal{L}(E) = \{\mathbf{in}(a) \mid a \in E\} \cup \{\mathbf{out}(a) \mid a \in Def(E)\} \cup \{\mathbf{und}(a) \mid a \in \mathrm{A} \setminus (E \cup Def(E))\}$, and for each labelling $\mathcal{L}$ there is a unique extension, that is $\mathbf{in}(\mathcal{L})$. We say that $\mathcal{L}(E)$ is the labelling *corresponding* to E. Moreover, we say that $\mathcal{L}(E)$ is a σ-labelling for a given AF Λ and semantics $\sigma \in \{\mathtt{co}, \mathtt{pr}, \mathtt{st}, \mathtt{sst}, \mathtt{gr}\}$ iff E is a σ-extension of Λ.

In the following, we say that the *status of an argument* a w.r.t. a labelling $\mathcal{L}$ (or its corresponding extension $\mathbf{in}(\mathcal{L})$) is $\mathbf{in}$ (resp., $\mathbf{out}$, $\mathbf{und}$) iff $\mathcal{L}(a) = \mathbf{in}$ (resp., $\mathcal{L}(a) = \mathbf{out}$, $\mathcal{L}(a) = \mathbf{und}$). We will avoid to mention explicitly the labelling (or the extension) whenever it is understood.

The set of complete (resp., preferred, stable, semi-stable, grounded) labellings of an AF Λ will be denoted by $\mathtt{co}(\Lambda)$ (resp., $\mathtt{pr}(\Lambda)$, $\mathtt{st}(\Lambda)$, $\mathtt{sst}(\Lambda)$, $\mathtt{gr}(\Lambda)$). All the above-mentioned semantics except the stable admit at least one labelling. The grounded semantics, that admits exactly one labelling, is said to be a *unique status* semantics, while the others are said to be *multiple status* semantics. With a little abuse of notation, in the following we also use $\mathtt{gr}(\Lambda)$ to denote the grounded labelling. For any AF Λ, it holds that: *i*) $\mathtt{st}(\Lambda) \subseteq \mathtt{sst}(\Lambda) \subseteq \mathtt{pr}(\Lambda) \subseteq \mathtt{co}(\Lambda)$, *ii*) $\mathtt{gr}(\Lambda) \in \mathtt{co}(\Lambda)$, and *iii*) $\mathtt{st}(\Lambda) \neq \emptyset$ implies that $\mathtt{st}(\Lambda) = \mathtt{sst}(\Lambda)$.

[1] Although $\mathbf{und}$ is not explicitly mentioned, since $\mathcal{L}$ is a total function, it suffices to characterize complete labelings (Caminada and Pigozzi 2011; Baroni, Caminada, and Giacomin 2011).

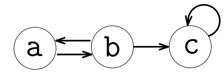

Figure 2: AF of Example 4.

For any pair ($\mathcal{L}$,$\mathcal{L}'$) of σ-labellings of AF $\Lambda = \langle \mathrm{A}, \mathrm{R}\rangle$ (with $\sigma \in \{\mathtt{gr}, \mathtt{co}, \mathtt{st}, \mathtt{pr}, \mathtt{sst}\}$), we use $\delta(\mathcal{L}, \mathcal{L}')$ to denote the *distance* $|\{a \in \mathrm{A} \mid \mathcal{L}(a) \neq \mathcal{L}'(a)\}|$ between $\mathcal{L}$ and $\mathcal{L}'$ in terms of the number of arguments having a different status.

Example 4. Let $\Lambda = \langle \mathrm{A}, \mathrm{R}\rangle$ be an AF where $\mathrm{A} = \{\mathtt{a}, \mathtt{b}, \mathtt{c}\}$ and $\mathrm{R} = \{(\mathtt{a}, \mathtt{b}), (\mathtt{b}, \mathtt{a}), (\mathtt{b}, \mathtt{c}), (\mathtt{c}, \mathtt{c})\}$ whose graph is shown in Figure 2. AF Λ has three complete labellings: $\mathcal{L}_1 = \{\mathbf{und}(\mathtt{a}), \mathbf{und}(\mathtt{b}), \mathbf{und}(\mathtt{c})\}$, $\mathcal{L}_2 = \{\mathbf{in}(\mathtt{a}), \mathbf{out}(\mathtt{b}), \mathbf{und}(\mathtt{c})\}$, and $\mathcal{L}_3 = \{\mathbf{out}(\mathtt{a}), \mathbf{in}(\mathtt{b}), \mathbf{out}(\mathtt{c})\}$, and we have that $\delta(\mathcal{L}_1, \mathcal{L}_2) = 2$ and $\delta(\mathcal{L}_1, \mathcal{L}_3) = \delta(\mathcal{L}_2, \mathcal{L}_3) = 3$. Moreover, the set of preferred labellings is $\mathtt{pr}(\Lambda) = \{\mathcal{L}_2, \mathcal{L}_3\}$, whereas the set of stable (and semi-stable) labellings is $\mathtt{st}(\Lambda) = \mathtt{sst}(\Lambda) = \{\mathcal{L}_3\}$, and the grounded labelling is $\mathcal{L}_1$. □

Four canonical argumentation problems are *existence*, *verification*, and *credulous* and *skeptical acceptance*. These problems can be formalized as follows. Given an AF $\Lambda = \langle \mathrm{A}, \mathrm{R}\rangle$, for any semantics $\sigma \in \{\mathtt{gr}, \mathtt{co}, \mathtt{st}, \mathtt{pr}, \mathtt{sst}\}$, (*i*) the existence problem (denoted EX^σ) consists in deciding whether there is at least one σ-labelling for Λ; (*ii*) the verification problem (denoted VE^σ) consists in deciding whether a given labelling is a σ-labelling for Λ; and (*iii*) given a (goal) argument $g \in \mathrm{A}$, the *credulous* (resp., *skeptical*) acceptance problem, denoted as CA^σ (resp., SA^σ), is the problem of deciding whether $\mathbf{in}(g)$ belongs to any (resp., all) σ-labellings of Λ. Clearly, for the grounded semantics, which admits exactly one labelling, credulous and skeptical acceptance problems become identical.

The complexity of the above-mentioned problems has been thoroughly investigated (see e.g. (Gabbay et al. 2021) for a survey), and the results are summarized in Table 1.

3 Counterfactual Reasoning

In this section, after formally defining the concept of counterfactual, we investigate the complexity of counterfactual-based argumentation problems.

As stated next, 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.

Definition 1 (Counterfactual (CF))**.** Let $\langle \mathrm{A}, \mathrm{R}\rangle$ be an AF, $\sigma \in \{\mathtt{gr}, \mathtt{co}, \mathtt{st}, \mathtt{pr}, \mathtt{sst}\}$ 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 5. Continuing with Example 2, under stable semantics, for the labelling $\mathcal{L}_3 = \{\mathbf{out}(\mathtt{fish}), \mathbf{in}(\mathtt{meat}),$

$\mathbf{out}(\texttt{pasta})$, $\mathbf{out}(\texttt{white})$, $\mathbf{in}(\texttt{red})\}$, 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 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, $\mathcal{CF}^{\texttt{st}}(\texttt{meat}, \mathcal{L}_3) = \{\mathcal{L}_2, \mathcal{L}_4\}$. □

Observe that, in general, the counterfactual relationship is not symmetric, in the sense that $\mathcal{L}' \in \mathcal{CF}^\sigma(g, \mathcal{L})$ does not entail that $\mathcal{L} \in \mathcal{CF}^\sigma(g, \mathcal{L}')$. For instance, in our running example, we have that $\mathcal{L}_3 \in \mathcal{CF}^{\texttt{st}}(\texttt{meat}, \mathcal{L}_1)$ while $\mathcal{L}_1 \notin \mathcal{CF}^{\texttt{st}}(\texttt{meat}, \mathcal{L}_3)$. Moreover, counterfactual reasoning makes sense for multiple status semantics only. In fact, for any AF $\langle \mathrm{A}, \mathrm{R}\rangle$ and goal $g \in \mathrm{A}$, it holds that $\mathcal{CF}^{\texttt{gr}}(g, \mathcal{L}) = \emptyset$. Thus, hereafter we focus on multiple semantics only, as for the grounded semantics all considered problems are trivial.

Finding a counterfactual means looking for a minimum distance labelling. The first problem we consider is a natural decision version of that problem.

Definition 2 (CF-Existence 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 $\mathcal{L} \in \sigma(\Lambda)$, $\mathsf{CF\text{-}EX}^\sigma$ is the problem of deciding whether there exists a labelling $\mathcal{L}' \in \sigma(\Lambda)$ s.t. $\mathcal{L}(g) \neq \mathcal{L}'(g)$ and $\delta(\mathcal{L}, \mathcal{L}') \leq k$.

In the following, we use $\text{CF-EX}^\sigma_\Lambda(g, k, \mathcal{L})$ (or simply $\text{CF-EX}^\sigma(g, k, \mathcal{L})$ whenever Λ is fixed) to denote the output of the $\mathsf{CF\text{-}EX}^\sigma$ problem with input Λ, g, k, and $\mathcal{L}$.

Example 6. Continuing with Example 2, assume the customer asks whether there is a menu not containing meat and differing from menu $\mathcal{L}_3$ by at most two items. Under stable semantics, the answer to the question is given by $\text{CF-EX}^{\texttt{st}}(\texttt{meat}, 2, \mathcal{L}_3)$: it is yes, as there is menu $\mathcal{L}_2 \in \texttt{st}(\Lambda)$, with $\mathcal{L}_2(\texttt{meat}) \neq \mathcal{L}_3(\texttt{meat})$ and $\delta(\mathcal{L}_3, \mathcal{L}_2) = 2$. □

The following theorem characterizes the complexity of the existence problem under counterfactual reasoning.

Theorem 1. $\mathsf{CF\text{-}EX}^\sigma$ *is:*

- NP*-complete for* $\sigma \in \{\texttt{co}, \texttt{st}\}$*; and*
- Σ_2^p*-complete for* $\sigma \in \{\texttt{pr}, \texttt{sst}\}$.

Proof. The membership result follows from the following guess-and-check strategy: guess a labelling $\mathcal{L}'$ with $\mathcal{L}(g) \neq \mathcal{L}'(g)$ and check in PTIME (resp., PTIME, coNP, and coNP) that $\mathcal{L}' \in \sigma(\Lambda)$ with $\sigma = \texttt{co}$ (resp., $\texttt{st}, \texttt{pr}, \texttt{sst}$) (Dvorák and Dunne 2017) and that $\delta(\mathcal{L}, \mathcal{L}') \leq k$.

(Sketch.) The hardness results derive from many-to-one reductions from known problems in AF. For $\sigma = \texttt{co}$ (resp., $\sigma = \texttt{st}$), we provide a reduction from the non-empty co-existence problem $\mathsf{EX}^{\texttt{co}}_{\neg\emptyset}$ (resp., st-existence problem $\mathsf{EX}^{\texttt{st}}$). We show that $\mathsf{EX}^{\texttt{co}}_{\neg\emptyset}(\langle \mathrm{A}, \mathrm{R}\rangle)$ (resp., $\mathsf{EX}^{\texttt{st}}(\langle \mathrm{A}, \mathrm{R}\rangle)$) is true iff $(\langle \mathrm{A}^*, \mathrm{R}^*\rangle, g, |\mathrm{A}|, \mathcal{L}^*)$ is a true instance of $\mathsf{CF\text{-}EX}^\sigma$, where: $\mathrm{A}^* = \mathrm{A} \cup \{g\}$, $\mathrm{R}^* = \mathrm{R} \cup \{(g, a), (a, g) \mid a \in \mathrm{A}\}$, and $\mathcal{L}^* = \{\mathbf{in}(g)\} \cup \{\mathbf{out}(a) \mid a \in \mathrm{A}\}$. For $\sigma \in \{\texttt{pr}, \texttt{sst}\}$ we provide a reduction from the complement of the skeptical σ-acceptance problem for AF $\overline{\mathsf{SA}}^\sigma$. We show that $\overline{\mathsf{SA}}^\sigma(\langle \mathrm{A}, \mathrm{R}\rangle, g)$ is true iff $(\langle \mathrm{A}^*, \mathrm{R}^*\rangle, g', |\mathrm{A}^*|, \mathcal{L}^*)$ is a true instance of $\mathsf{CF\text{-}EX}^\sigma$, where (*i*) for $\sigma = \texttt{pr}$, $\mathrm{A}^* = \mathrm{A} \cup \{x, g'\}$, $\mathrm{R}^* = \mathrm{R} \cup \{(x, a), (a, x), (x, g'), (g, g') \mid a \in \mathrm{A}\}$, and $\mathcal{L}^* = \{\mathbf{in}(x)\} \cup \{\mathbf{out}(a) \mid a \in \mathrm{A}^* \setminus \{x\}\}$; (*ii*) for $\sigma = \texttt{sst}$, $\mathrm{A}^* = \mathrm{A} \cup \{x, g', g''\}$, $\mathrm{R}^* = \mathrm{R} \cup \{(x, a), (a, x), (a, g''), (x, g'), (g', g''), (g'', g''), (g, g') \mid a \in \mathrm{A}\}$, and $\mathcal{L}^* = \{\mathbf{in}(x), \mathbf{out}(g'), \mathbf{und}(g'')\} \cup \{\mathbf{out}(a) \mid a \in \mathrm{A}\}$. □

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

Definition 3 (CF-Verification 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 $g \in \mathrm{A}$, a σ-labelling $\mathcal{L} \in \sigma(\Lambda)$, and a labelling $\mathcal{L}'$, $\mathsf{CF\text{-}VE}^\sigma$ is the problem of deciding whether $\mathcal{L}'$ belongs to $\mathcal{CF}^\sigma(g, \mathcal{L})$.

We use $\text{CF-VE}^\sigma_\Lambda(g, \mathcal{L}, \mathcal{L}')$ (or simply $\text{CF-VE}^\sigma(g, \mathcal{L}, \mathcal{L}')$) to denote the output of $\mathsf{CF\text{-}VE}^\sigma$ with input $\Lambda, g, \mathcal{L}$, and $\mathcal{L}'$.

Example 7. Consider again the situation of Example 2, and assume the customer is interested to know whether $\mathcal{L}_2$ is the closest menu to $\mathcal{L}_3$ not containing meat. This problem can be answered by deciding $\text{CF-VE}^{\texttt{st}}(\texttt{meat}, \mathcal{L}_3, \mathcal{L}_2)$, which is true as we have that $\mathcal{L}_2 \in \mathcal{CF}^{\texttt{st}}(\texttt{meat}, \mathcal{L}_3)$. □

Observe that $\text{CF-VE}^\sigma_\Lambda(g, \mathcal{L}, \mathcal{L}')$ is true iff *i*) $\mathcal{L}(g) \neq \mathcal{L}'(g)$ *ii*) $\mathcal{L}' \in \sigma(\Lambda)$ (i.e. the classical verification problem, VE^σ, is true for input $\mathcal{L}'$), and *iii*) $\text{CF-EX}^\sigma_\Lambda(g, \delta(\mathcal{L}, \mathcal{L}') - 1, \mathcal{L})$ is false. Since the complexity of VE^σ is lower than that of CF-EX^σ (cf. Table 1), and checking $\mathcal{L}(g) \neq \mathcal{L}'(g)$ is in PTIME, the above-mentioned three-steps strategy suggests that CF-VE^σ can be decided by an algorithm in the complement complexity class of CF-EX^σ. In fact, as entailed by the next result, the problems CF-VE^σ and CF-EX^σ are on the same level of the polynomial hierarchy.

Theorem 2. $\mathsf{CF\text{-}VE}^\sigma$ *is:*

- coNP*-complete for* $\sigma \in \{\texttt{co}, \texttt{st}\}$*; and*
- Π_2^p*-complete for* $\sigma \in \{\texttt{pr}, \texttt{sst}\}$.

Proof. We sketch the proofs of hardness results only, as the guess-and-check strategy for showing the membership results has been discussed earlier. For $\sigma = \texttt{co}$ (resp., $\sigma = \texttt{st}$) we can provide a reduction from the complement of the non-empty co-existence problem $\mathsf{EX}^{\texttt{co}}_{\neg\emptyset}$ (resp., the st-existence problem $\mathsf{EX}^{\texttt{st}}$) by showing that $\langle \mathrm{A}, \mathrm{R}\rangle$ is a false (resp., true) instance of $\mathsf{EX}^\sigma_{\neg\emptyset}$ (resp., $\mathsf{EX}^{\texttt{st}}$) iff $(\langle \mathrm{A}^*, \mathrm{R}^*\rangle, g, \mathcal{L}, \mathcal{L}')$ is a true instance of $\mathsf{CF\text{-}VE}^\sigma$, where (*i*) for $\sigma = \texttt{co}$, $\mathrm{A}^* = \mathrm{A} \cup \{g, g'\}$, $\mathrm{R}^* = \mathrm{R} \cup \{(g, g'), (a, g'), (g, a), (a, g) \mid a \in \mathrm{A}\}$, $\mathcal{L} = \{\mathbf{in}(g)\} \cup \{\mathbf{out}(x) \mid x \in \mathrm{A}^* \setminus \{g\}\}$, and $\mathcal{L}' = \{\mathbf{und}(x) \mid x \in \mathrm{A}^*\}$, (*ii*) for $\sigma = \texttt{st}$, $\mathrm{A}^* = \mathrm{A} \cup \{g, g', x_i \mid i \in [1, 2|\mathrm{A}|]\}$, $\mathrm{R}^* = \mathrm{R} \cup \{(g, g'), (g', g)\} \cup \{(g', a), (a, g')\ (g, a), (a, g) \mid a \in \mathrm{A}\} \cup \{(g', x_i) \mid i \in [1, 2|\mathrm{A}|]\}$, $\mathcal{L} = \{\mathbf{in}(g), \mathbf{out}(g'), \mathbf{out}(a) \mid a \in \mathrm{A}\} \cup \{\mathbf{in}(x_i) \mid i \in [1, 2|\mathrm{A}|]\}$, and $\mathcal{L}' = \{\mathbf{in}(g'), \mathbf{out}(x) \mid x \in \mathrm{A}^* \setminus \{g'\}\}$. Finally, for $\sigma \in \{\texttt{pr}, \texttt{sst}\}$, we provide a reduction from the coherence problem CO (Dunne and Bench-Capon 2002). □

Counterfactual-based acceptance problems extend credulous and skeptical reasoning w.r.t. a set $\mathcal{CF}^\sigma(g,\mathcal{L})$ of counterfactuals for labelling $\mathcal{L}$ and goal g (under σ). They consist in deciding whether an additional argument g' is accepted in any or all counterfactuals of $\mathcal{L}$ w.r.t. g, respectively.

Definition 4 (CF-Acceptance Problems)**.** Given as input an AF $\Lambda = \langle \mathrm{A}, \mathrm{R}\rangle$, a semantics $\sigma \in \{\mathtt{co}, \mathtt{st}, \mathtt{pr}, \mathtt{sst}\}$, an argument $g \in \mathrm{A}$, a σ-labelling $\mathcal{L} \in \sigma(\Lambda)$, and a goal argument $g' \in \mathrm{A}$, $\mathsf{CF\text{-}CA}^\sigma$ (resp., $\mathsf{CF\text{-}SA}^\sigma$) is the problem of checking whether g' is accepted in any (resp., every) counterfactual in $\mathcal{CF}^\sigma(g,\mathcal{L})$.

We use CF-CA$^\sigma_\Lambda(g,\mathcal{L},g')$ and CF-SA$^\sigma_\Lambda(g,\mathcal{L},g')$ (or simply CF-CA$^\sigma(g,\mathcal{L},g')$ and CF-SA$^\sigma(g,\mathcal{L},g')$) to denote the output of the $\mathsf{CF\text{-}CA}^\sigma$ and $\mathsf{CF\text{-}SA}^\sigma$ with input $\Lambda, g, \mathcal{L}$ and g', respectively.

Example 8. Consider Example 2 where the chef suggests menu $\mathcal{L}_3$ and the customer does not eat meat (but she likes the rest of the menu). Assume now that the customer is interested to know whether there exists some menu, among the ones in $\mathcal{CF}^{\mathtt{st}}(\mathtt{meat},\mathcal{L}_3)$, containing pasta. The answer to this question, computed by means of CF-CA$^\sigma_\Lambda(\mathtt{meat},\mathcal{L},\mathtt{pasta})$, is "yes" since $\mathcal{CF}^{\mathtt{st}}(\mathtt{meat},\mathcal{L}_3) = \{\mathcal{L}_2,\mathcal{L}_4\}$ and $\mathcal{L}_4(\mathtt{pasta}) = \mathbf{in}$.

Moreover, if we assume that the customer is interested to know whether all menus in $\mathcal{CF}^{\mathtt{st}}(\mathtt{meat},\mathcal{L}_3)$ contain pasta. The answer to this question, computed by means of CF-SA$^{\mathtt{st}}_\Lambda(\mathtt{meat},\mathcal{L},\mathtt{pasta})$, is "no" since $\mathcal{CF}^{\mathtt{st}}(\mathtt{meat},\mathcal{L}_3) = \{\mathcal{L}_2,\mathcal{L}_4\}$ and $\mathcal{L}_2(\mathtt{pasta}) = \mathbf{out}$. □

Our next results address the complexity of counterfactual-based credulous and skeptical acceptance.

Theorem 3. $\mathsf{CF\text{-}CA}^\sigma$ *is:*

- NP*-hard and in* Θ_2^p *for* $\sigma \in \{\mathtt{co}, \mathtt{st}\}$*; and*
- Σ_2^p*-hard and in* Θ_3^p *for* $\sigma \in \{\mathtt{pr}, \mathtt{sst}\}$.

Proof. Consider an instance $(\Lambda, g, \mathcal{L}, g')$ of $\mathsf{CF\text{-}CA}^\sigma$. For $\sigma \in \{\mathtt{co}, \mathtt{st}\}$ (resp., $\sigma \in \{\mathtt{pr}, \mathtt{sst}\}$) the following Θ_2^p (resp., Θ_3^p) algorithm suffices to prove the membership result. In the following, we also use $\mathsf{CF\text{-}\widetilde{EX}}^\sigma$, that is, a variant of $\mathsf{CF\text{-}EX}^\sigma$ where $\mathcal{L}'(g') = \mathbf{in}$, cf. Definition 2; clearly, $\mathsf{CF\text{-}\widetilde{EX}}^\sigma$ is in the same complexity class of $\mathsf{CF\text{-}EX}^\sigma$. The algorithm starts by computing, by $O(log_2(|\mathrm{A}|))$ calls to $\mathsf{CF\text{-}EX}^\sigma$, the minimum value $k' \in [0, |\mathrm{A}|]$ s.t. CF-EX$^\sigma(g, k', \mathcal{L}) = \mathtt{true}$. Then, with an additional call to (the above-introduced problem), $\mathsf{CF\text{-}\widetilde{EX}}^\sigma$ it checks the existence of a counterfactual $\mathcal{L}'$ of $\mathcal{L}$ w.r.t g such that (i) it is at distance k' and (ii) $\mathcal{L}'(g') = \mathbf{in}$. As the number of calls to the NP (resp., Σ_2^p) oracle is bounded by $O(log_2(|\mathrm{A}|) + 1)$ the result follows.

(Sketch.) The hardness results for $\sigma \in \{\mathtt{st}, \mathtt{co}, \mathtt{sst}\}$ (resp., $\sigma = \mathtt{pr}$) derive by reductions from σ-credulous acceptance (resp., complement of the preferred skeptical acceptance) problem in AF. We prove that, $\mathsf{CA}^\sigma(\langle \mathrm{A}, \mathrm{R}\rangle, g)$ is true iff $(\langle \mathrm{A}^*, \mathrm{R}^*\rangle,\ x,\ \mathcal{L}^*, g)$ is a true instance of $\mathsf{CF\text{-}CA}^\sigma$, where (i) for $\sigma \in \{\mathtt{co}, \mathtt{st}\}$, $\mathrm{A}^* = \mathrm{A} \cup \{x\}$, $\mathrm{R}^* = \mathrm{R} \cup \{(x, g), (g, x)\} \cup \{(x, a) \mid a \in \mathrm{A}\}$, and $\mathcal{L}^* = \{\mathbf{in}(x)\} \cup \{\mathbf{out}(a) \mid a \in \mathrm{A}\}$; (ii) for $\sigma = \mathtt{sst}$, $\mathrm{A}^* = \mathrm{A} \cup \{x, \bar{x}, x^u\}$, $\mathrm{R}^* = \mathrm{R} \cup \{(x, \bar{x}), (\bar{x}, x^u), (x^u, x^u)\ (x, g), (g, x)\} \cup \{(x, a) \mid a \in \mathrm{A}\}$, and $\mathcal{L}^* = \{\mathbf{in}(x), \mathbf{out}(\bar{x}), \mathbf{und}(x^u)\} \cup \{\mathbf{out}(a) \mid a \in \mathrm{A}\}$. Moreover, for $\sigma = \mathtt{pr}$ we prove that $\overline{\mathsf{SA}}^\sigma(\langle \mathrm{A}, \mathrm{R}\rangle, g)$ is true iff $(\langle \mathrm{A}^*, \mathrm{R}^*\rangle,\ g^u,\ \mathcal{L}^*, z)$ is a true instance of $\mathsf{CF\text{-}CA}^\sigma$, where $\mathrm{A}^* = \mathrm{A} \cup \{x, g^u, z\}$, $\mathrm{R}^* = \mathrm{R} \cup \{(x, g^u), (g, g^u), (g^u, g^u), (x, z)\} \cup \{(x, a), (a, x)\ \mid \in \mathrm{A}\}$, and $\mathcal{L}^* = \{\mathbf{in}(x)\} \cup \{\mathbf{out}(a) \mid a \in \mathrm{A}^* \setminus \{x\}\}$. □

Theorem 4. $\mathsf{CF\text{-}SA}^\sigma$ *is:*

- coNP*-hard and in* Θ_2^p *for* $\sigma \in \{\mathtt{co}, \mathtt{st}\}$*; and*
- Π_2^p*-hard and in* Θ_3^p *for* $\sigma \in \{\mathtt{pr}, \mathtt{sst}\}$.

Proof. Let $(\Lambda, g, \mathcal{L}, g')$ be an instance of $\mathsf{CF\text{-}SA}^\sigma$. For $\sigma \in \{\mathtt{co}, \mathtt{st}\}$ (resp., $\sigma \in \{\mathtt{pr}, \mathtt{sst}\}$) the following Θ_2^p (resp., Θ_3^p) algorithm suffices to prove the membership result for the complement $\overline{\mathsf{CF\text{-}SA}^\sigma}$ of our problem, that is the problem of checking whether g' is not skeptical accepted in any counterfactual $\mathcal{L}' \in \mathcal{CF}^\sigma(g,\mathcal{L})$. Consider the problem $\mathsf{CF\text{-}\bar{E}X}^\sigma$ that is a variant of $\mathsf{CF\text{-}EX}^\sigma$ where $\mathcal{L}'(g') \neq \mathbf{in}$; observe that $\mathsf{CF\text{-}\bar{E}X}^\sigma$ is in the same complexity class of $\mathsf{CF\text{-}EX}^\sigma$. The algorithm starts by computing, by $O(log_2(|\mathrm{A}|))$ calls to $\mathsf{CF\text{-}EX}^\sigma$, the minimum value $k' \in [0, |\mathrm{A}|]$ s.t. CF-EX$^\sigma(g, \mathcal{L}, k') = \mathtt{true}$. Then with an additional call to $\mathsf{CF\text{-}\bar{E}X}^\sigma$ we check the existence of a counterfactual $\mathcal{L}'$ of $\mathcal{L}$ w.r.t g such that (i) it is at distance k' and that (ii) $\mathcal{L}'(g') \neq \mathbf{in}$. As the number of calls to the NP (resp., Σ_2^p) oracles is bounded by $O(log_2(|\mathrm{A}|) + 1)$ we have that $\overline{\mathsf{CF\text{-}SA}^\sigma}$ is in Θ_2^p (resp., Θ_3^p). As Θ_i^p is closed under complement, the result follows.

(Sketch.) The hardness results derive for $\sigma \in \{\mathtt{st}, \mathtt{co}, \mathtt{sst}\}$ (resp., $\sigma = \mathtt{pr}$) by reductions from the complement of the σ-credulous acceptance (resp., preferred skeptical acceptance) problem in AF. By using constructions similar to those used in the hardness proof of Theorem 3, we can show that, for $\sigma \in \{\mathtt{st}, \mathtt{co}, \mathtt{sst}\}$ (resp., $\sigma = \mathtt{pr}$) $\overline{\mathsf{CA}^\sigma}(\langle \mathrm{A}, \mathrm{R}\rangle, g)$ is true iff $(\langle \mathrm{A}^*, \mathrm{R}^*\rangle,\ x,\ \mathcal{L}^*, x)$ (resp., $(\langle \mathrm{A}^*, \mathrm{R}^*\rangle,\ g^u,\ \mathcal{L}^*, g^u)$) is a true instance of $\mathsf{CF\text{-}SA}^\sigma$. □

4 Semifactual Reasoning

In this section, following what is done in the previous section, we investigate the complexity of semifactual-based argumentation problems. We first introduce the concept of semifactual that, in a sense, is symmetrical and complementary to that of a counterfactual.

Definition 5 (Semifactual (SF))**.** Let $\langle \mathrm{A}, \mathrm{R}\rangle$ be an AF, $\sigma \in \{\mathtt{gr}, \mathtt{co}, \mathtt{st}, \mathtt{pr}, \mathtt{sst}\}$ 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.

	Classical problems				Counterfactual-based problems				Semifactual-based problems			
σ	EX^σ	VE^σ	CA^σ	SA^σ	$\mathsf{CF\text{-}EX}^\sigma$	$\mathsf{CF\text{-}VE}^\sigma$	$\mathsf{CF\text{-}CA}^\sigma$	$\mathsf{CF\text{-}SA}^\sigma$	$\mathsf{SF\text{-}EX}^\sigma$	$\mathsf{SF\text{-}VE}^\sigma$	$\mathsf{SF\text{-}CA}^\sigma$	$\mathsf{SF\text{-}SA}^\sigma$
$\mathtt{co}$	T	P	NP-c	P	NP-c	coNP-c	NP-h, Θ_2^p	coNP-h, Θ_2^p	NP-c	coNP-c	NP-h, Θ_2^p	coNP-h, Θ_2^p
$\mathtt{st}$	NP-c	P	NP-c	coNP-c	NP-c	coNP-c	NP-h, Θ_2^p	coNP-h, Θ_2^p	NP-c	coNP-c	NP-h, Θ_2^p	coNP-h, Θ_2^p
$\mathtt{pr}$	T	coNP-c	NP-c	Π_2^p-c	Σ_2^p-c	Π_2^p-c	Σ_2^p-h, Θ_3^p	Π_2^p-h, Θ_3^p	Σ_2^p-c	Π_2^p-c	Σ_2^p-h, Θ_3^p	Π_2^p-h, Θ_3^p
$\mathtt{sst}$	T	coNP-c	Σ_2^p-c	Π_2^p-c	Σ_2^p-c	Π_2^p-c	Σ_2^p-h, Θ_3^p	Π_2^p-h, Θ_3^p	Σ_2^p-c	Π_2^p-c	Σ_2^p-h, Θ_3^p	Π_2^p-h, Θ_3^p

Table 1: Complexity of classical, counterfactual-based and semifactual-based problems in AF under complete ($\mathtt{co}$), stable ($\mathtt{st}$), preferred ($\mathtt{pr}$), and semi-stable ($\mathtt{sst}$) semantics. For any complexity class C, C-c (resp., C-h) means C-complete (resp., C-hard); an interval C-h, C' means C-hard and in C'. New results are highlighted in cyan. T means trivial (from the computational standpoint).

Example 9. Consider the stable labelling $\mathcal{L}_3 = \{\mathbf{out}(\mathtt{fish}), \mathbf{in}(\mathtt{meat}), \mathbf{out}(\mathtt{pasta}), \mathbf{out}(\mathtt{white}), \mathbf{in}(\mathtt{red})\}$ for the AF of Example 3. We have that $\mathcal{L}_2 = \{\mathbf{in}(\mathtt{fish}), \mathbf{out}(\mathtt{meat}), \mathbf{out}(\mathtt{pasta}), \mathbf{out}(\mathtt{white}), \mathbf{in}(\mathtt{red})\}$ and $\mathcal{L}_4 = \{\mathbf{out}(\mathtt{fish}), \mathbf{out}(\mathtt{meat}), \mathbf{in}(\mathtt{pasta}), \mathbf{out}(\mathtt{white}), \mathbf{in}(\mathtt{red})\}$ are the only semifactuals of $\mathcal{L}_3$ w.r.t. the argument $\mathtt{red}$ as there is no other $\mathtt{st}$-labelling agreeing on $\mathtt{red}$ and having distance greater than $\delta(\mathcal{L}_3, \mathcal{L}_2) = \delta(\mathcal{L}_3, \mathcal{L}_4) = 2$. In fact, $\mathcal{L}_1 = \{\mathbf{in}(\mathtt{fish}), \mathbf{out}(\mathtt{meat}), \mathbf{out}(\mathtt{pasta}), \mathbf{in}(\mathtt{white}), \mathbf{out}(\mathtt{red})\}$, having distance $\delta(\mathcal{L}_3, \mathcal{L}_1) = 4$, is not a semifactual for $\mathcal{L}_3$ w.r.t. $\mathtt{red}$ as $\mathcal{L}_1(\mathtt{red}) \neq \mathcal{L}_3(\mathtt{red})$. Thus, $\mathcal{SF}^{\mathtt{st}}(\mathtt{red}, \mathcal{L}_3) = \{\mathcal{L}_2, \mathcal{L}_4\}$. □

Similarly to the case of counterfactuals, the semifactual relationship is not symmetric, that is, $\mathcal{L}' \in \mathcal{SF}^\sigma(g, \mathcal{L})$ does not entail that $\mathcal{L} \in \mathcal{SF}^\sigma(g, \mathcal{L}')$.

As for the case of counterfactuals, semifactual reasoning makes sense only for multiple status semantics. Thus, hereafter, we do not consider the grounded semantics.

The semifactual-based existence problem is as follows.

Definition 6 (SF-Existence Problem)**.** 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 $g \in \mathrm{A}$, an integer $k \in \mathbb{N}$, and a σ-labelling $\mathcal{L} \in \sigma(\Lambda)$, $\mathsf{SF\text{-}EX}^\sigma$ is the problem of checking whether there exists a labelling $\mathcal{L}' \in \sigma(\Lambda)$ s.t. $\mathcal{L}(g) = \mathcal{L}'(g)$ and $\delta(\mathcal{L}, \mathcal{L}') \geq k$.

We use $\text{SF-EX}_\Lambda^\sigma(g, k, \mathcal{L})$ (or simply $\text{SF-EX}^\sigma(g, k, \mathcal{L})$ whenever Λ is fixed) to denote the output of $\mathsf{SF\text{-}EX}^\sigma$ with input Λ, g, k, and $\mathcal{L}$.

Example 10. Continuing with Example 3, assume the customer is interested to know whether there exists a menu containing $\mathtt{red}$ wine and differing from $\mathcal{L}_3$ by at least two items. Under stable semantics, the answer to this question is yes, as there exists menu $\mathcal{L}_2 \in \mathtt{st}(\Lambda)$, with $\mathcal{L}_2(\mathtt{red}) = \mathcal{L}_3(\mathtt{red})$ and $\delta(\mathcal{L}_3, \mathcal{L}_2) = 2$, i.e. $\text{SF-EX}^{\mathtt{st}}(\mathtt{red}, 2, \mathcal{L}_3)$ is true. □

The following theorem characterizes the complexity of the existence problem under semifactual reasoning.

Theorem 5. $\mathsf{SF\text{-}EX}^\sigma$ *is:*

- NP*-complete for* $\sigma \in \{\mathtt{co}, \mathtt{st}\}$*; and*
- Σ_2^p*-complete for* $\sigma \in \{\mathtt{pr}, \mathtt{sst}\}$.

Proof. For the membership result, guess a labelling $\mathcal{L}'$ with $\mathcal{L}(g) = \mathcal{L}'(g)$ and check in PTIME (resp., PTIME, coNP, and coNP) that $\mathcal{L}' \in \sigma(\Lambda)$ with $\sigma = \mathtt{co}$ (resp., $\mathtt{st}, \mathtt{pr}, \mathtt{sst}$) (Dvorák and Dunne 2017) and $\delta(\mathcal{L}, \mathcal{L}') \geq k$.

(Sketch.) For $\sigma = \mathtt{co}$ (resp., $\sigma = \mathtt{st}$), we provide a reduction from the non-empty $\mathtt{co}$-existence problem $\mathsf{EX}_{\neg\emptyset}^{\mathtt{co}}$ (resp., $\mathtt{st}$-existence problem $\mathsf{EX}^{\mathtt{st}}$), as done for the hardness results of $\mathsf{CF\text{-}EX}^\sigma$ but with different constructions. We show that $\mathsf{EX}_{\neg\emptyset}^{\mathtt{co}}(\langle \mathrm{A}, \mathrm{R} \rangle)$ (resp., $\mathsf{EX}^{\mathtt{st}}(\langle \mathrm{A}, \mathrm{R} \rangle)$) is true iff $(\langle \mathrm{A}^*, \mathrm{R}^* \rangle, \bar{g}, k = 1, \mathcal{L}^*)$ is a true instance of $\mathsf{SF\text{-}EX}^\sigma$, where: $\mathrm{A}^* = \mathrm{A} \cup \{g, \bar{g}\}$, $\mathrm{R}^* = \mathrm{R} \cup \{(g, a), (a, g), (a, \bar{g}), (g, \bar{g}) \mid a \in \mathrm{A}\}$, and $\mathcal{L}^* = \{\mathbf{in}(g)\} \cup \{\mathbf{out}(a) \mid a \in \mathrm{A}^* \setminus \{g\}\}$. For $\sigma \in \{\mathtt{pr}, \mathtt{sst}\}$ we provide a reduction from the complement of the skeptical σ-acceptance problem for AF, namely $\overline{\mathsf{SA}}^\sigma$. We we show that $\overline{\mathsf{SA}}^\sigma(\langle \mathrm{A}, \mathrm{R} \rangle, g)$ is true iff $(\langle \mathrm{A}^*, \mathrm{R}^* \rangle, \bar{g}, k = 1, \mathcal{L}^*)$ is a true instance of $\mathsf{SF\text{-}EX}^\sigma$, where $\mathrm{A}^* = \mathrm{A} \cup \{x, \bar{x}, x^u, \bar{g}, g^u\}$, $\mathrm{R}^* = \mathrm{R} \cup \{(g, \bar{g}), (\bar{g}, \bar{g}), (\bar{g}, g^u), (g^u, g^u), (x, \bar{x}), (\bar{x}, x^u), (x^u, x^u), (x, g^u) \cup \{(x, a), (a, x) \mid a \in \mathrm{A}\}$, and $\mathcal{L}^* = \{\mathbf{in}(x), \mathbf{out}(\bar{x}), \mathbf{und}(x^u), \mathbf{und}(\bar{g}), \mathbf{out}(g^u)\} \cup \{\mathbf{out}(a) \mid a \in \mathrm{A}\}$. □

For semifactuals, the verification problem is checking whether a given labelling $\mathcal{L}'$ is a semifactual for $\mathcal{L}$ and g (hence the distance between the two labelling is maximal).

Definition 7 (SF-Verification Problem)**.** 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 $\mathcal{L}'$, $\mathsf{SF\text{-}VE}^\sigma$ is the problem of checking whether $\mathcal{L}'$ belongs to $\mathcal{SF}^\sigma(g, \mathcal{L})$.

We use $\text{SF-VE}_\Lambda^\sigma(g, \mathcal{L}, \mathcal{L}')$ (or simply $\text{SF-VE}^\sigma(g, \mathcal{L}, \mathcal{L}')$) to denote the output of $\mathsf{SF\text{-}VE}^\sigma$ with input $\Lambda, g, \mathcal{L}$, and $\mathcal{L}'$.

Example 11. Consider the situation in Example 3, and assume the customer is interested to know whether $\mathcal{L}_2$ is the farthest menu w.r.t. $\mathcal{L}_3$ containing $\mathtt{red}$ wine. This problem can be answered by deciding $\text{SF-VE}^{\mathtt{st}}(\mathtt{meat}, \mathcal{L}_3, \mathcal{L}_2)$, which is true as we have that $\mathcal{L}_2 \in \mathcal{SF}^{\mathtt{st}}(\mathtt{red}, \mathcal{L}_3)$. □

Theorem 6. $\mathsf{SF\text{-}VE}^\sigma$ *is:*

- coNP*-complete for* $\sigma \in \{\mathtt{co}, \mathtt{st}\}$*; and*
- Π_2^p*-complete for* $\sigma \in \{\mathtt{pr}, \mathtt{sst}\}$.

Proof. The following guess-and-check algorithm provides the membership result for the complement of our problem. First, check whether $\mathcal{L}(g) = \mathcal{L}'(g)$, and then guess a labelling $\mathcal{L}''$ with $\mathcal{L}(g) = \mathcal{L}''(g)$ and check in PTIME (resp., PTIME, coNP, and coNP) that $\mathcal{L}'' \in \sigma(\Lambda)$ with $\sigma = \mathtt{co}$ (resp., $\mathtt{st}, \mathtt{pr}, \mathtt{sst}$) (Dvorák and Dunne 2017) and that $\delta(\mathcal{L}, \mathcal{L}'') > \delta(\mathcal{L}, \mathcal{L}')$ (in PTIME). If both condition holds then the complement of our problem is true.

(Sketch.) For $\sigma = \texttt{co}$ (resp., $\sigma = \texttt{st}$) we provide a reduction from the complement of the non-empty co-existence problem $\mathsf{EX}^{\texttt{co}}_{\neg\emptyset}$ (resp., st-existence problem $\mathsf{EX}^{\texttt{st}}$) by showing that $\langle \mathrm{A}, \mathrm{R}\rangle$ is a false (resp., true) instance of $\mathsf{EX}^{\sigma}_{\neg\emptyset}$ (resp., $\mathsf{EX}^{\texttt{st}}$) iff $(\langle \mathrm{A}^*, \mathrm{R}^*\rangle, g, \mathcal{L}, \mathcal{L})$ is a true instance of $\mathsf{SF\text{-}VE}^{\sigma}$, where (*i*) for $\sigma = \texttt{co}$, $\mathrm{A}^* = \mathrm{A} \cup \{g, g^u\}$, $\mathrm{R}^* = \mathrm{R} \cup \{(g, g), (g^u, g^u), (g^u, a), (a, g^u), (g^u, g) \mid a \in \mathrm{A}\}$, and $\mathcal{L} = \{\mathbf{und}(x) \mid x \in \mathrm{A}^*\}$; (*ii*) for $\sigma = \texttt{st}$ $\mathrm{A}^* = \mathrm{A} \cup \{g, \bar{g}\}$, $\mathrm{R}^* = \mathrm{R} \cup \{(\bar{g}, g), (a, g), (a, \bar{g}), (\bar{g}, a) \mid a \in \mathrm{A}\}$, and $\mathcal{L} = \{\mathbf{in}(\bar{g}), \mathbf{out}(g)\} \cup \{\mathbf{out}(a) \mid a \in \mathrm{A}\}$. For $\sigma \in \{\texttt{pr}, \texttt{sst}\}$, we provide a reduction from coherence problem CO by proving that $(\langle \mathrm{A}, \mathrm{R}\rangle)$ is a true instance of CO iff $(\langle \mathrm{A}^*, \mathrm{R}^*\rangle, b, \mathcal{L}, \mathcal{L})$ is a true instance of $\mathsf{SF\text{-}VE}_{\sigma}$ where $\mathrm{A}^* = \mathrm{A} \cup \{g, b, z, h, k, c\} \cup \{\bar{a}, \hat{a} \mid a \in \mathrm{A}\}$, $\mathrm{R}^* = \mathrm{R} \cup \{(g, h), (h, k), (k, k), (g, c), (c, c)\} \cup \{(g, a), (a, g), (g, z), (a, \bar{a}), (\bar{a}, \hat{a}), (a, \hat{a}), (a, h), (a, k), (\hat{a}, z), (z, b), (b, b) \mid a \in \mathrm{A}\}$, and $\mathcal{L} = \{\mathbf{in}(g), \mathbf{out}(c), \mathbf{out}(h), \mathbf{und}(k), \mathbf{out}(z), \mathbf{und}(b)\} \cup \{\mathbf{out}(a), \mathbf{in}(\bar{a}), \mathbf{out}(\hat{a}) \mid a \in \mathrm{A}\}$. $\square$

Finally, we investigate semifactual-based acceptance problems, that is credulous and skeptical reasoning w.r.t. a set $\mathcal{SF}^{\sigma}(g, \mathcal{L})$ of semifactual for labelling $\mathcal{L}$ and goal g.

Definition 8 (SF-Acceptance Problems)**.** Given as input an AF $\Lambda = \langle \mathrm{A}, \mathrm{R}\rangle$, a semantics $\sigma \in \{\texttt{co}, \texttt{st}, \texttt{pr}, \texttt{sst}\}$, an argument $g \in \mathrm{A}$, a σ-labelling $\mathcal{L} \in \sigma(\Lambda)$, and a goal argument $g' \in \mathrm{A}$, $\mathsf{SF\text{-}CA}^{\sigma}$ (resp., $\mathsf{SF\text{-}SA}^{\sigma}$) is the problem of deciding whether g' is accepted in any (resp., every) semifactual $\mathcal{L}' \in \mathcal{SF}^{\sigma}(g, \mathcal{L})$.

We use SF-CA$^{\sigma}_{\Lambda}(g, \mathcal{L}, g')$ and SF-SA$^{\sigma}_{\Lambda}(g, \mathcal{L}, g')$ (or SF-CA$^{\sigma}(g, \mathcal{L}, g')$ and SF-SA$^{\sigma}(g, \mathcal{L}, g')$) to denote the output of $\mathsf{SF\text{-}CA}^{\sigma}$ and $\mathsf{SF\text{-}SA}^{\sigma}$ with input $\Lambda, g, \mathcal{L}$ and g', respectively.

Example 12. In our running example, assume the customer is interested to know whether in any (resp., all) of the farthest menus w.r.t. $\mathcal{L}_3$ containing `red` wine, `pasta` is present. The answer to the first question is positive as SF-CA$^{\texttt{st}}(\texttt{red}, \mathcal{L}_3, \texttt{pasta})$ is true, while that to the second one is negative as SF-SA$^{\texttt{st}}(\texttt{red}, \mathcal{L}_3, \texttt{pasta})$ is false. $\square$

The next theorems state the complexity of semifactual-based credulous and skeptical acceptance problems.

Theorem 7. $\mathsf{SF\text{-}CA}^{\sigma}$ *is:*

- NP*-hard and in* Θ^p_2 *for* $\sigma \in \{\texttt{co}, \texttt{st}\}$*; and*
- Σ^p_2*-hard and in* Θ^p_3 *for* $\sigma \in \{\texttt{pr}, \texttt{sst}\}$.

Theorem 8. $\mathsf{SF\text{-}SA}^{\sigma}$ *is:*

- coNP*-hard and in* Θ^p_2 *for* $\sigma \in \{\texttt{co}, \texttt{st}\}$*; and*
- Π^p_2*-hard and in* Θ^p_3 *for* $\sigma \in \{\texttt{pr}, \texttt{sst}\}$.

Thus, semifactual- and counterfactual-based reasoning problems share the same complexity bounds. Moreover, the considered problems are generally harder than the corresponding classical argumentation problems; this particularly holds if we focus on the verification problem, credulous acceptance under preferred semantics, and skeptical acceptance under complete semantics.

5 WAF and ASP Mappings

We first show that counterfactual and semifactual explanations can be encoded through Weak constrained AF (WAF), which is a generalization of AF with strong and weak constraints, and then provide an algorithm for computing counterfactuals and semifactuals by making use of well-known ASP encoding of AF semantics as well as of constraints capturing counterfactuals and semifactuals' semantics.

5.1 Weak Constrained AF

Constraints in argumentation frameworks have been investigated in several works (Coste-Marquis, Devred, and Marquis 2006; Arieli 2015; Sakama and Son 2020; Alfano et al. 2024b; 2023c). They extend AF by considering a set of strong and weak constraints, that are sets of propositional formulae to be satisfied by extensions. Intuitively, constraints introduce subjective knowledge of agents, whereas the AF encodes objective knowledge. Strong constraints in AF allow restricting the set of feasible solutions, but do not help in finding "best" or preferable solutions. To express this kind of conditions, *weak* constraints have been introduced, that is, constraints that are required to be satisfied *if possible* (Alfano et al. 2021b). In the following, for the sake of presentation, we consider constraints as propositional formulae built over labelled arguments, as e.g. in (Sakama and Inoue 2000), instead of propositional formulae defined over argument literals. An AF with strong and weak constraints is said to be a *Weak constrained Argumentation Framework* (WAF).

Definition 9 (WAF)**.** *A* Weak Constrained AF (WAF) *is a quadruple* $\langle \mathrm{A}, \mathrm{R}, \mathcal{C}, \mathcal{W}\rangle$ *where* $\langle \mathrm{A}, \mathrm{R}\rangle$ *is an AF, and* $\mathcal{C}$ *and* $\mathcal{W}$ *are sets of propositional formulae called* strong *and* weak *constraints, respectively, both built from the set of labeled arguments* $\lambda_A = \{\mathbf{in}(a), \mathbf{out}(a), \mathbf{und}(a) \mid a \in \mathrm{A}\}$ *by using the connectives* $\neg$, $\vee$*, and* $\wedge$.

We say that a labelling $\mathcal{L}$ satisfies a constraint κ if and only if $\mathcal{L}$ is a (2-valued) model of κ, denoted as $\mathcal{L} \models \kappa$, that is, if the formula obtained from κ by replacing every atom occurring in $\mathcal{L}$ with `true`, and every atom not occurring in $\mathcal{L}$ with `false`, evaluates to true. Moreover, we say that $\mathcal{L}$ satisfies a set $\mathcal{K} = \{\kappa_1, \dots, \kappa_n\}$ of constraints, denoted as $\mathcal{L} \models \mathcal{K}$, whenever $\mathcal{L} \models \kappa_i\ \forall i \in [1, n]$. The set of constraints in $\mathcal{K}$ satisfied by a labelling $\mathcal{L}$ will be denoted by $\mathcal{K}_{\mathcal{L}}$.

Maximum-cardinality semantics for WAF prescribes as preferable extensions those satisfying the largest number of weak constraints (Alfano et al. 2021b). This is similar to the semantics of weak constraints in DLV (Alviano et al. 2017) where, in addition, each constraint has assigned a weight.

Definition 10. *Given a WAF* $\Upsilon = \langle A, R, \mathcal{C}, \mathcal{W}\rangle$*, a* σ*-labelling* $\mathcal{L}$ *for* $\langle A, R\rangle$ *is a maximum-cardinality* σ*-labelling for* Υ *if* $\mathcal{L} \models \mathcal{C}$ *and there is no* σ*-labelling* $\mathcal{L}'$ *for* $\langle A, R\rangle$ *with* $\mathcal{L}' \models \mathcal{C}$ *a such that* $|\mathcal{W}_{\mathcal{L}}| < |\mathcal{W}_{\mathcal{L}'}|$.

The set of maximum-cardinality σ-labellings of a WAF Υ will be denoted by $\texttt{mc-}\sigma(\Upsilon)$, with $\sigma \in \{\texttt{co}, \texttt{st}, \texttt{pr}, \texttt{sst}\}$.

As stated next, counterfactual and semifactual explanations one-to-one correspond to maximum-cardinality labellings of appropriate WAFs.

Proposition 1. *For any AF $\Lambda = \langle A, R\rangle$, semantics $\sigma \in \{\mathtt{co}, \mathtt{st}, \mathtt{pr}, \mathtt{sst}\}$, goal $g \in A$, and σ-labelling $\mathcal{L}$ of Λ,*
- $\mathcal{CF}^{\sigma}(g, \mathcal{L})$=$\mathtt{mc}$-$\sigma(\langle A, R, \{\neg\mathcal{L}[g]\}, \{\mathcal{L}[a] \mid a \in A \setminus \{g\}\}\rangle)$;
- $\mathcal{SF}^{\sigma}(g, \mathcal{L})$=$\mathtt{mc}$-$\sigma(\langle A, R, \{\mathcal{L}[g]\}, \{\neg\mathcal{L}[a] \mid a \in A \setminus \{g\}\}\rangle)$.

Example 13. Consider the AF $\langle \mathrm{A}, \mathrm{R}\rangle$ of Example 1 and recall from Example 5 that $\mathcal{CF}^{\mathtt{st}}(\mathtt{meat}, \mathcal{L}_3)$=$\{\mathcal{L}_2, \mathcal{L}_4\}$. We have that $\mathcal{L}_2 \in \mathtt{mc\text{-}st}(\langle \mathrm{A}, \mathrm{R}, \mathcal{C}, \mathcal{W}\rangle)$ where $\mathcal{C} = \{\neg\mathbf{in}(\mathtt{meat})\}$, and $\mathcal{W} = \{\mathbf{out}(\mathtt{fish}), \mathbf{out}(\mathtt{pasta}), \mathbf{out}(\mathtt{white}), \mathbf{in}(\mathtt{red})\}$ ($\mathcal{C}$ consists of a single strong constraint, while $\mathcal{W}$ consists of four weak constraints). Moreover, recalling from Example 9 that $\mathcal{SF}^{\mathtt{st}}(\mathtt{red}, \mathcal{L}_3) = \{\mathcal{L}_2, \mathcal{L}_4\}$, we have that $\mathcal{L}_2 \in \mathtt{mc\text{-}st}(\langle \mathrm{A}, \mathrm{R}, \mathcal{C}, \mathcal{W}\rangle)$ where $\mathcal{C} = \{\mathbf{in}(\mathtt{red})\}$, and $\mathcal{W} = \{\neg\mathbf{out}(\mathtt{fish}), \neg\mathbf{out}(\mathtt{pasta}), \neg\mathbf{out}(\mathtt{white}), \neg\mathbf{in}(\mathtt{red})\}$. □

asprin Encoding

Given the tight relationship between formal argumentation and Answer Set Programming (ASP), we introduce EXPLAIN in Algorithm 1 that computes the set of counterfactual and semifactual explanations by leveraging existing ASP-based solvers. In particular, we rely on the *asprin* framework, that is *ASP for preference handling* (Brewka et al. 2015a; 2015b). Intuitively, Algorithm 1 encodes the distance measure δ between labellings/extensions through (weighted) preferences in ASP to select best extensions, among extensions given by ASP encodings of AF semantics. We use P_σ to denote a set of rules corresponding to the encoding of semantics σ. As an example, an encoding for stable semantics is as follows[2] (Dvorák et al. 2020):

$$P_{\mathtt{st}} = \left\{ \begin{array}{l} \mathbf{in}(X) \text{ :- } \mathbf{not\ out}(X), \mathbf{arg}(X); \\ \mathbf{out}(X) \text{ :- } \mathbf{not\ in}(X), \mathbf{arg}(X); \\ \mathbf{defeated}(X) \text{ :- } \mathbf{in}(Y), \mathbf{att}(Y, X); \\ \text{ :- } \mathbf{in}(X), \mathbf{in}(Y), \mathbf{att}(X, Y); \\ \text{ :- } \mathbf{out}(X), \mathbf{not\ defeated}(X); \end{array} \right\}.$$

Algorithm 1 takes as input an AF $\Lambda = \langle \mathrm{A}, \mathrm{R}\rangle$, a goal argument $g \in \mathrm{A}$, a semantics $\sigma \in \{\mathtt{co}, \mathtt{st}, \mathtt{pr}, \mathtt{sst}\}$, a σ-labelling $\mathcal{L}$, and the task type $T \in \{CF, SF\}$ (either counterfactual or semifactual). After defining the set of ASP rules, P_σ, encoding AF semantics (Line 1), the ASP encoding P_Λ for AF Λ is computed (Line 2). Then, using the result of Proposition 1, a set P_S containing a single strong constraint (Line 4 for CF, Line 7 for SF) and a set of weak constraints P_W (Line 5 for CF, Line 8 for SF) are computed. As an example, if $T = CF$, to compute answer sets representing counterfactuals, the constraint ":- $\mathcal{L}[g]$", stating that $\mathcal{L}[g]$ must be false is added. In contrast, if $T = SF$, to compute answer sets representing semifactuals, the constraint ":- **not** $\mathcal{L}[g]$", stating that $\mathcal{L}[g]$ must be true, is added. Moreover, to ensure that an answer set corresponds to a counterfactual of $\mathcal{L}$ w.r.t. g, it should satisfy as less constraints of the form $\mathbf{w}$(a) :- **not** $\mathcal{L}[a]$ as possible, so that the distance w.r.t. $\mathcal{L}$ is minimized (a similar approach is used for semifactuals). To this end the following *asprin* preference statement and optimization directive are given when invoking *asprin* over $P = P_\sigma \cup P_\Lambda \cup P_S \cup P_W$ (Line 9).

[2] To denote sets of clauses in asprin we shall will the semicolon, as the comma is used to represent the "and" operator.

Algorithm 1 EXPLAIN($\Lambda, g, \sigma, \mathcal{L}, T$)

Input: AF $\Lambda = \langle \mathrm{A}, \mathrm{R}\rangle$, goal $g \in \mathrm{A}$, $\sigma \in \{\mathtt{co}, \mathtt{st}, \mathtt{pr}, \mathtt{sst}\}$, σ-labelling $\mathcal{L}$ of Λ, task $T \in \{CF, SF\}$.
Output: $\mathcal{CF}^{\sigma}(g, \mathcal{L})$ if $T = CF$, $\mathcal{SF}^{\sigma}(g, \mathcal{L})$ if $T = SF$.
1: Let P_σ be the ASP encoding for semantics σ;
2: $P_\Lambda = \{\mathbf{arg}(a) \mid a \in \mathrm{A}\} \cup \{\mathbf{att}(a, b) \mid (a, b) \in \mathrm{R}\}$;
3: **if** $T = CF$ **then**
4: $\quad P_S = \{$:- $\mathcal{L}[g]\}$;
5: $\quad P_W = \{\mathbf{w}(a)$:- **not** $\mathcal{L}[a] \mid a \in \mathrm{A} \setminus \{g\}\}$;
6: **else**
7: $\quad P_S = \{$:- **not** $\mathcal{L}[g]\}$;
8: $\quad P_W = \{\mathbf{w}(a)$:- $\mathcal{L}[a] \mid a \in \mathrm{A} \setminus \{g\}\}$;
9: **return** ASPRIN($P_\Lambda \cup P_\sigma \cup P_S \cup P_W$);

- #preference(p,less(cardinality))$\{\mathbf{w}(X) : \mathbf{arg}(X)\}$;
- #optimize(p).

As stated next, our algorithm is sound and complete.

Theorem 9. *Algorithm 1 is sound and complete.*

Example 14. Considering the AF Λ of Example 1, $\mathtt{st}$-labelling $\mathcal{L}_3 = \{\mathbf{out}(\mathtt{fish}), \mathbf{in}(\mathtt{meat}), \mathbf{out}(\mathtt{pasta}), \mathbf{out}(\mathtt{white}), \mathbf{in}(\mathtt{red})\}$, and goal argument $\mathtt{meat}$, the *asprin* program P built in Algorithm 1 with $T = CF$ is $P = P_{\mathtt{st}} \cup P_\Lambda \cup P_S \cup P_W$, where $P_{\mathtt{st}}$ is as shown earlier, and:

$$\mathrm{P}_\Lambda = \left\{ \begin{array}{l} \mathbf{arg}(\mathtt{fish}); \mathbf{att}(\mathtt{meat}, \mathtt{fish}); \mathbf{att}(\mathtt{fish}, \mathtt{meat}); \\ \mathbf{arg}(\mathtt{white}); \mathbf{att}(\mathtt{pasta}, \mathtt{white}); \mathbf{att}(\mathtt{meat}, \mathtt{white}); \\ \mathbf{arg}(\mathtt{red}); \mathbf{att}(\mathtt{red}, \mathtt{white}); \mathbf{att}(\mathtt{white}, \mathtt{red}); \\ \mathbf{arg}(\mathtt{meat}); \mathbf{att}(\mathtt{pasta}, \mathtt{meat}); \mathbf{att}(\mathtt{meat}, \mathtt{pasta}); \\ \mathbf{arg}(\mathtt{pasta}); \mathbf{att}(\mathtt{pasta}, \mathtt{fish}); \mathbf{att}(\mathtt{fish}, \mathtt{pasta}); \end{array} \right\}$$

$$P_S = \{ \text{ :- } \mathbf{in}(\mathtt{meat}); \}$$

$$P_W = \left\{ \begin{array}{ll} \mathbf{w}(\mathtt{fish}) & \text{:- } \mathbf{not\ out}(\mathtt{fish}); \\ \mathbf{w}(\mathtt{pasta}) & \text{:- } \mathbf{not\ out}(\mathtt{pasta}); \\ \mathbf{w}(\mathtt{white}) & \text{:- } \mathbf{not\ out}(\mathtt{white}); \\ \mathbf{w}(\mathtt{red}) & \text{:- } \mathbf{not\ in}(\mathtt{red}); \end{array} \right\}$$

The result of ASPRIN(P) is $\mathcal{CF}^{\mathtt{st}}(\mathtt{meat}, \mathcal{L}_3)$=$\{\mathcal{L}_2, \mathcal{L}_4\}$. □

6 Discussion

In this section, we consider definitions of counterfactual and semifactual more general than those in given Definitions 1 and 5, respectively. In particular, we focus on the following extensions: (*a*) more general distance measures between labellings; (*b*) more general criterion for changing the status of the goal argument; and *(c)* set of goal arguments (instead of a single argument). Notably, regardless of the above-mentioned generalization adopted for Definitions 1 and 5, the complexity bounds in Table 1 still hold.

More general distance measures. The measure δ used Definitions 1 and 5 does not distinguish among different labelling changes. For example, labelling $\mathbf{in}(\mathtt{fish})$ has the same distance, 1, from $\mathbf{out}(\mathtt{fish})$ or $\mathbf{und}(\mathtt{fish})$. Here, we redefine the concept of counterfactual and semifactual w.r.t. a more general distance measure η, that is, a (polynomial-time computable) function $\eta : \mathrm{A} \times \{\mathbf{in}, \mathbf{out}, \mathbf{und}\} \times \{\mathbf{in}, \mathbf{out}, \mathbf{und}\} \rightarrow \mathbb{N}$, which assigns a (positive) number to every possible labelling change for each argument. Then, we use $\eta(\mathcal{L}, \mathcal{L}') = \sum_{a \in \mathrm{A}} \eta(a, \mathcal{L}(a), \mathcal{L}'(a))$ to denote the (weighted) distance between σ-labellings $\mathcal{L}$ and $\mathcal{L}'$.

Observe that δ is a special case of η where, $\forall a \in \mathrm{A}$, $\eta(a, \mathcal{L}(a), \mathcal{L}'(a)) = 0$ if $\mathcal{L}(a) = \mathcal{L}'(a)$, 1 otherwise.

Given the new distance measure η, we can redefine counterfactuals and semifactuals by replacing item *(ii)* of Definitions 1 and 5 as follows.

- Minimality condition for Definition 1: there is no $\mathcal{L}'' \in \sigma(\Lambda)$ s.t. $\mathcal{L}(g) \neq \mathcal{L}''(g)$ and $\eta(\mathcal{L}, \mathcal{L}'') < \eta(\mathcal{L}, \mathcal{L}')$.
- Maximality condition for Definition 5: there is no $\mathcal{L}'' \in \sigma(\Lambda)$ s.t. $\mathcal{L}(g) = \mathcal{L}''(g)$ and $\eta(\mathcal{L}, \mathcal{L}'') > \eta(\mathcal{L}, \mathcal{L}')$.

As a proposal for η to capture the idea that the distance should distinguish between the changes $\mathbf{in}(a)$-to/from-$\mathbf{out}(a)$ from those $\mathbf{in}(a)$-to/from-$\mathbf{und}(a)$ or $\mathbf{out}(a)$-to/from-$\mathbf{und}(a)$, we could define it as follows: $\eta(a, \mathcal{L}(a), \mathcal{L}'(a)) = 0$ if $\mathcal{L}(a) = \mathcal{L}'(a)$; otherwise, $\eta(a, \mathcal{L}(a), \mathcal{L}'(a)) = 1$ (resp., 2) if $\mathbf{und}(a) \in \{\mathcal{L}(a), \mathcal{L}'(a)\}$ (resp., $\mathbf{und}(a) \notin \{\mathcal{L}(a), \mathcal{L}'(a)\}$). This way changes between decided values are weighted double of the others.

Example 15. Consider the AF obtained from that of Figure 1 by removing the argument pasta (and the attacks involving it), whose co-labellings are :
$\mathcal{L}_1 = \{\mathbf{in}(\mathtt{fish}), \mathbf{out}(\mathtt{meat}), \mathbf{in}(\mathtt{white}), \mathbf{out}(\mathtt{red})\}$,
$\mathcal{L}_2 = \{\mathbf{in}(\mathtt{fish}), \mathbf{out}(\mathtt{meat}), \mathbf{out}(\mathtt{white}), \mathbf{in}(\mathtt{red})\}$,
$\mathcal{L}_3 = \{\mathbf{out}(\mathtt{fish}), \mathbf{in}(\mathtt{meat}), \mathbf{out}(\mathtt{white}), \mathbf{in}(\mathtt{red})\}$,
$\mathcal{L}_4 = \{\mathbf{und}(\mathtt{fish}), \mathbf{und}(\mathtt{meat}), \mathbf{out}(\mathtt{white}), \mathbf{in}(\mathtt{red})\}$,
$\mathcal{L}_5 = \{\mathbf{in}(\mathtt{fish}), \mathbf{out}(\mathtt{meat}), \mathbf{und}(\mathtt{white}), \mathbf{und}(\mathtt{red})\}$,
$\mathcal{L}_6 = \{\mathbf{und}(\mathtt{fish}), \mathbf{und}(\mathtt{meat}), \mathbf{und}(\mathtt{white}), \mathbf{und}(\mathtt{red})\}$.

Using distance δ, we have that $\mathcal{CF}^{\mathtt{co}}(\mathtt{meat}, \mathcal{L}_3) = \{\mathcal{L}_2, \mathcal{L}_5\}$ and $\mathcal{SF}^{\mathtt{co}}(\mathtt{red}, \mathcal{L}_3) = \{\mathcal{L}_2, \mathcal{L}_4\}$, while for η defined as in the paragraph preceding this example, we have that $\mathcal{CF}^{\mathtt{co}}(\mathtt{meat}, \mathcal{L}_3) = \{\mathcal{L}_5\}$ and $\mathcal{SF}^{\mathtt{co}}(\mathtt{red}, \mathcal{L}_3) = \{\mathcal{L}_2\}$. □

As stated next, our complexity results holds irrespective of how η is instantiated.

Proposition 2. *The results of Theorems 1–8 still hold if measure η is used (instead of δ) in Definitions 1 and 5.*

Replacing criterion for goals' status change. A counterfactual $\mathcal{L}'$ for $\mathcal{L}$ and g could also be defined by requiring that the status of the goal g w.r.t. $\mathcal{L}'$ is not undecided, that is by changing condition (i) of Definition 1. For instance, we could replace condition (i) $\mathcal{L}(g) \neq \mathcal{L}'(g)$, equivalently rewritten as

$$\mathcal{L}'(g) \in \begin{cases} \{\mathbf{out}, \mathbf{und}\} & \text{if } \mathcal{L}(g) = \mathbf{in} \\ \{\mathbf{in}, \mathbf{und}\} & \text{if } \mathcal{L}(g) = \mathbf{out} \\ \{\mathbf{in}, \mathbf{out}\} & \text{if } \mathcal{L}(g) = \mathbf{und} \end{cases}$$

with the following one:

$$\mathcal{L}'(g) \in \begin{cases} \{\mathbf{out}\} & \text{if } \mathcal{L}(g) = \mathbf{in} \\ \{\mathbf{in}\} & \text{if } \mathcal{L}(g) = \mathbf{out} \\ \{\mathbf{in}, \mathbf{out}\} & \text{if } \mathcal{L}(g) = \mathbf{und} \end{cases}$$

Again, replacing condition (i) of Definition 1 with that above—or any other that can be checked in PTIME—does not alter the complexity results in Table 1.

Considering multiple goal arguments. For the sake of presentation, we focused on a single goal argument $g \in \mathrm{A}$. However, a set $S \subseteq \mathrm{A}$ of goal arguments whose status is required to change (resp., to not change) could be considered in Definitions 1 and 5, respectively, or their generalizations introduced earlier. More formally, given an AF $\Lambda = \langle \mathrm{A}, \mathrm{R} \rangle$, distance measure η, semantics $\sigma \in \{\mathtt{co}, \mathtt{st}, \mathtt{pr}, \mathtt{sst}\}$, goal arguments $S \subseteq \mathrm{A}$, and σ-labelling $\mathcal{L}$, we say that $\mathcal{L}'$ is an η-counterfactual (resp., η-semifactual) of $\mathcal{L}$ w.r.t. S if condition (i) and (ii) of Definition 1 (resp., Definition 5) are satisfied for all $g \in S$, where measure η is used (instead of δ). Still, adopting these definitions of counterfactuals and semifactuals does not alter the complexity bounds of Theorems 1–8 as conditions $(i - ii)$ can be checked in PTIME.

7 Related Work

Several researchers explored how to deal with explanations with in formal argumentation (Cyras et al. 2021; Vassiliades, Bassiliades, and Patkos 2021). Important work includes e.g. (Fan and Toni 2015), where a new argumentation semantics is proposed for capturing explanations in AF, and (Craven and Toni 2016) that focuses on ABA frameworks (Craven and Toni 2016; Dung, Kowalski, and Toni 2009; Hung 2016). They treat an explanation as a semantics to answer why an argument is accepted or not. In (Fan and Toni 2015) an explanation is as a set of arguments justifying a given argument by means of a proponent-opponent dispute-tree (Dung, Mancarella, and Toni 2007). An approach based on debate trees as proof procedure for computing grounded, ideal, and preferred semantics, has been proposed in (Thang, Dung, and Hung 2009). The approach in (Alfano et al. 2023a) build explanations that are sequences of arguments by exploiting topological dependencies among arguments. An alternative definition for explaining complete extensions has been proposed in (Baumann and Ulbricht 2021). It exploits the concept of *reduct*, i.e. a subframework obtained by removing true and false arguments w.r.t. a complete extension. The concept of *strong explanation* is proposed in (Ulbricht and Wallner 2021), inspired by the related notions introduced in (Brewka and Ulbricht 2019; Brewka, Thimm, and Ulbricht 2019; Saribatur, Wallner, and Woltran 2020). Finally, in (Cocarascu, Rago, and Toni 2019) quantitative argumentation frameworks have proven effective in generating explanations for review aggregations.

Counterfactual reasoning in AF has been firstly introduced in (Sakama 2014), 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. An interesting problem related to this is *enforcement* (Niskanen, Wallner, and Järvisalo 2016; 2018; Wallner, Niskanen, and Järvisalo 2017), that is how changes in the AF affects the acceptability of arguments, and how to modify an AF to guarantee that some arguments get a given labelling. In particular, extension enforcement concerns how to modify an AF to ensure that a given set of arguments becomes (part of) an extension (Baumann et al. 2021). Moreover, in (Borg and Bex 2024), a framework for determining argument-based explanations in both abstract and structured settings is proposed. The framework is able to answer why-question, such as 'why is argument a credulously accepted under pr'? Finally, an approach to explain the relative change in strength of specified arguments in a Quantitative Bipolar AF updated by changing its arguments, their initial strengths and/or relationships has been recently proposed in (Kampik, Cyras, and Alarcón 2024).

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.

8 Conclusions and Future Work

We have proposed the concept of counterfactual and semifactual explanations in abstract argumentation, and investigated the complexity of counterfactual- and semifactual-based reasoning in AF. 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. It is worth mentioning that, though their formulation is similar and they share the same complexity bounds, the considered counterfactual-based and semifactual-based problems are not dual problems—we do not see how to naturally reduce one to the other; however, a (possibly complex) reduction may exist as our complexity results do not rule this out.

Although counterfactual- and semifactual-based reasoning suffers from high computational complexity (as many other computational problems in argumentation (Alfano, Greco, and Parisi 2019; 2021; Alfano et al. 2023c; 2023b; 2024c)), 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,[3] which aims at nurturing research and development of implementations for computational models of argumentation. In this regard, we have proposed an asprin-based approach that enables leveraging existing tools and techniques to deal with complex problems by reducing to weak-constrained AF and then to ASP with preferences, enabling implementations by using *asprin* (Brewka et al. 2015b).

It can be shown that our complexity results carry over to other frameworks whose complexity is as that of AF, such as Bipolar AF (Cohen et al. 2014; Alfano et al. 2020) and AF with recursive attacks and supports (Cohen et al. 2015; Cayrol et al. 2018; Alfano et al. 2024a), among others (Villata et al. 2012; Gottifredi et al. 2018; Dvorák et al. 2024).

Future work will be devoted to considering more general forms of AF, such as incomplete and probabilistic AF (Li, Oren, and Norman 2011; Hunter 2012; 2013; Baumeister et al. 2021; Alfano et al. 2022) as well as structured argumentation formalisms (Modgil and Prakken 2014; Cyras, Heinrich, and Toni 2021; Garcia, Prakken, and Simari 2020; Alfano et al. 2018; 2021a).

[3] https://argumentationcompetition.org

Acknowledgements

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

References

Albini, E.; Rago, A.; Baroni, P.; and Toni, F. 2020. Relation-based counterfactual explanations for bayesian network classifiers. In *Proc. of International Joint Conference on Artificial Intelligence (IJCAI)*, 451–457.

Alfano, G.; Greco, S.; Parisi, F.; Simari, G. I.; and Simari, G. R. 2018. An incremental approach to structured argumentation over dynamic knowledge bases. In *Proc. of KR*, 78–87.

Alfano, G.; Greco, S.; Parisi, F.; and Trubitsyna, I. 2020. On the semantics of abstract argumentation frameworks: A logic programming approach. *Theory Pract. Log. Program.* 20(5):703–718.

Alfano, G.; Greco, S.; Parisi, F.; Simari, G. I.; and Simari, G. R. 2021a. Incremental computation for structured argumentation over dynamic delp knowledge bases. *Artif. Intell.* 300:103553.

Alfano, G.; Greco, S.; Parisi, F.; and Trubitsyna, I. 2021b. Argumentation frameworks with strong and weak constraints: Semantics and complexity. In *Proc. of AAAI Conference on Artificial Intelligence*, 6175–6184.

Alfano, G.; Greco, S.; Parisi, F.; and Trubitsyna, I. 2022. Incomplete argumentation frameworks: Properties and complexity. In *Proc. of AAAI Conference on Artificial Intelligence*, 5451–5460.

Alfano, G.; Calautti, M.; Greco, S.; Parisi, F.; and Trubitsyna, I. 2023a. Explainable acceptance in probabilistic and incomplete abstract argumentation frameworks. *Artif. Intell.* 323:103967.

Alfano, G.; Greco, S.; Parisi, F.; and Trubitsyna, I. 2023b. On acceptance conditions in abstract argumentation frameworks. *Inf. Sci.* 625:757–779.

Alfano, G.; Greco, S.; Parisi, F.; and Trubitsyna, I. 2023c. Preferences and constraints in abstract argumentation. In *Proc. of International Joint Conference on Artificial Intelligence (IJCAI)*, 3095–3103. ijcai.org.

Alfano, G.; Cohen, A.; Gottifredi, S.; Greco, S.; Parisi, F.; and Simari, G. R. 2024a. Credulous acceptance in high-order argumentation frameworks with necessities: An incremental approach. *Artif. Intell.* 333:104159.

Alfano, G.; Greco, S.; Mandaglio, D.; Parisi, F.; and Trubitsyna, I. 2024b. Abstract argumentation frameworks with strong and weak constraints. *Artif. Intell.* 336:104205.

Alfano, G.; Greco, S.; Parisi, F.; and Trubitsyna, I. 2024c. Complexity of credulous and skeptical acceptance in epistemic argumentation framework. In *Proc. of AAAI Conference on Artificial Intelligence*, 10423–10432.

Alfano, G.; Greco, S.; and Parisi, F. 2019. On scaling the enumeration of the preferred extensions of abstract argumentation frameworks. In *Proceedings of ACM/SIGAPP Symposium on Applied Computing (SAC)*, 1147–1153.

Alfano, G.; Greco, S.; and Parisi, F. 2021. Incremental computation in dynamic argumentation frameworks. *IEEE Intell. Syst.* 36(6):80–86.

Alviano, M.; Calimeri, F.; Dodaro, C.; Fuscà, D.; Leone, N.; Perri, S.; Ricca, F.; Veltri, P.; and Zangari, J. 2017. The ASP system DLV2. In *Proc. of International Conference on Logic Programming and Nonmonotonic Reasoning (LPNMR)*, 215–221.

Amgoud, L.; Dimopoulos, Y.; and Moraitis, P. 2007. A unified and general framework for argumentation-based negotiation. In *Proc. of International Joint Conference on Autonomous Agents and Multiagent Systems*, 158.

Arieli, O. 2015. Conflict-free and conflict-tolerant semantics for constrained argumentation frameworks. *J. Appl. Log.* 13(4):582–604.

Aryal, S., and Keane, M. T. 2023. Even if explanations: Prior work, desiderata & benchmarks for semi-factual xai. In *Proc. of International Joint Conference on Artificial Intelligence (IJCAI)*, 6526–6535.

Barceló, P.; Monet, M.; Pérez, J.; and Subercaseaux, B. 2020. Model interpretability through the lens of computational complexity. In *Proc. of Advances in Neural Information Processing Systems*.

Baroni, P.; Caminada, M.; and Giacomin, M. 2011. An introduction to argumentation semantics. *Knowl. Eng. Rev.* 26(4):365–410.

Baumann, R., and Ulbricht, M. 2021. Choices and their consequences - explaining acceptable sets in abstract argumentation frameworks. In *Proc. of International Conference on Principles of Knowledge Representation and Reasoning (KR)*, 110–119.

Baumann, R.; Doutre, S.; Mailly, J.; and Wallner, J. P. 2021. Enforcement in formal argumentation. *Journal of Applied Logics - IfCoLog Journal* 8(6):1623–1678.

Baumeister, D.; Järvisalo, M.; Neugebauer, D.; Niskanen, A.; and Rothe, J. 2021. Acceptance in incomplete argumentation frameworks. *Artif. Intell.* 103470.

Bex, F., and Walton, D. 2016. Combining explanation and argumentation in dialogue. *Argument & Computation* 7(1):55–68.

Borg, A., and Bex, F. 2024. Minimality, necessity and sufficiency for argumentation and explanation. *International Journal of Approximate Reasoning* 109143.

Brewka, G., and Ulbricht, M. 2019. Strong explanations for nonmonotonic reasoning. In *Description Logic, Theory Combination, and All That*, volume 11560 of *Lecture Notes in Computer Science*, 135–146.

Brewka, G.; Delgrande, J. P.; Romero, J.; and Schaub, T. 2015a. asprin: Customizing answer set preferences without a headache. In *Proc. of AAAI Conference on Artificial Intelligence*, 1467–1474.

Brewka, G.; Delgrande, J. P.; Romero, J.; and Schaub, T. 2015b. Implementing preferences with asprin. In *Proc. of International Conference on Logic Programming and Nonmonotonic Reasoning (LPNMR)*, 158–172.

Brewka, G.; Thimm, M.; and Ulbricht, M. 2019. Strong inconsistency. *Artif. Intell.* 267:78–117.

Caminada, M., and Pigozzi, G. 2011. On judgment aggregation in abstract argumentation. *Auton. Agents Multi Agent Syst.* 22(1):64–102.

Caminada, M. 2006. Semi-stable semantics. In *Proc. of COMMA*, 121–130.

Cayrol, C.; Fandinno, J.; del Cerro, L. F.; and Lagasquie-Schiex, M. 2018. Structure-based semantics of argumentation frameworks with higher-order attacks and supports. In *Proc. of COMMA*, 29–36.

Cocarascu, O.; Rago, A.; and Toni, F. 2019. Extracting dialogical explanations for review aggregations with argumentative dialogical agents. In *Proc. of AAMAS*, 1261–1269.

Cohen, A.; Gottifredi, S.; Garc´ina, A. J.; and Simari, G. R. 2014. A survey of different approaches to support in argumentation systems. *Knowl. Eng. Rev.* 29(5):513–550.

Cohen, A.; Gottifredi, S.; Garcia, A. J.; and Simari, G. R. 2015. An approach to abstract argumentation with recursive attack and support. *J. Appl. Log.* 13(4):509–533.

Coste-Marquis, S.; Devred, C.; and Marquis, P. 2006. Constrained argumentation frameworks. In *Proc. of International Conference on Principles of Knowledge Representation and Reasoning (KR)*, 112–122.

Craven, R., and Toni, F. 2016. Argument graphs and assumption-based argumentation. *Artif. Intell.* 233:1–59.

Cyras, K.; Birch, D.; Guo, Y.; Toni, F.; Dulay, R.; Turvey, S.; Greenberg, D.; and Hapuarachchi, T. 2019. Explanations by arbitrated argumentative dispute. *Expert Systems with Applications* 127:141–156.

Cyras, K.; Rago, A.; Albini, E.; Baroni, P.; and Toni, F. 2021. Argumentative XAI: A survey. In *Proc. of IJCAI*, 4392–4399.

Cyras, K.; Heinrich, Q.; and Toni, F. 2021. Computational complexity of flat and generic assumption-based argumentation, with and without probabilities. *Artif. Intell.* 293:103449.

Dandl, S.; Casalicchio, G.; Bischl, B.; and Bothmann, L. 2023. Interpretable regional descriptors: Hyperbox-based local explanations. In *Proc. of Machine Learning and Knowledge Discovery in Databases*, volume 14171, 479–495.

Dimopoulos, Y.; Mailly, J.; and Moraitis, P. 2019. Argumentation-based negotiation with incomplete opponent profiles. In *Proc. of International Joint Conference on Autonomous Agents and Multiagent Systems*, 1252–1260.

Dung, P. M.; Kowalski, R. A.; and Toni, F. 2009. Assumption-based argumentation. In *Argumentation in Artificial Intelligence*. Springer. 199–218.

Dung, P.; Mancarella, P.; and Toni, F. 2007. Computing ideal sceptical argumentation. *Artif. Intell.* 171(10-15):642–674.

Dung, P. M. 1995. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artif. Intell.* 77:321–358.

Dunne, P. E., and Bench-Capon, T. J. M. 2002. Coherence in finite argument systems. *Artif. Intell.* 141(1/2):187–203.

Dvorák, W., and Dunne, P. E. 2017. Computational problems in formal argumentation and their complexity. *IFCoLog Journal of Logics and their Applications* 4(8).

Dvorák, W.; Gaggl, S. A.; Rapberger, A.; Wallner, J. P.; and Woltran, S. 2020. The ASPARTIX system suite. In *Proc. of COMMA*, volume 326, 461–462.

Dvorák, W.; König, M.; Ulbricht, M.; and Woltran, S. 2024. Principles and their computational consequences for argumentation frameworks with collective attacks. *J. Artif. Intell. Res.* 79:69–136.

Fan, X., and Toni, F. 2015. On computing explanations in argumentation. In *Proc. of AAAI Conference on Artificial Intelligence*, 1496–1502.

Gabbay, D.; Giacomin, M.; Simari, G. R.; and Thimm, M., eds. 2021. *Handbook of Formal Argumentation*, volume 2. College Publications.

Garcia, A. J.; Prakken, H.; and Simari, G. R. 2020. A comparative study of some central notions of ASPIC+ and DeLP. *Theory Pract. Log. Program.* 20(3):358–390.

Gottifredi, S.; Cohen, A.; Garcia, A. J.; and Simari, G. R. 2018. Characterizing acceptability semantics of argumentation frameworks with recursive attack and support relations. *Artif. Intell.* 262:336–368.

Guidotti, R. 2022. Counterfactual explanations and how to find them: literature review and benchmarking. *Data Mining and Knowledge Discovery* 1–55.

Hung, N. D. 2016. Computing probabilistic assumption-based argumentation. In *Proc. of Pacific Rim International Conference on Artificial Intelligence (PRICAI)*, 152–166.

Hunter, A. 2012. Some foundations for probabilistic abstract argumentation. In *Proc. of COMMA*, 117–128.

Hunter, A. 2013. A probabilistic approach to modelling uncertain logical arguments. *International Journal of Approximate Reasoning* 54(1):47–81.

Kahneman, D., and Tversky, A. 1981. *The simulation heuristic*. National Technical Information Service.

Kampik, T.; Cyras, K.; and Alarcón, J. R. 2024. Change in quantitative bipolar argumentation: Sufficient, necessary, and counterfactual explanations. *International Journal of Approximate Reasoning* 164:109066.

Kenny, E. M., and Keane, M. T. 2021. On generating plausible counterfactual and semi-factual explanations for deep learning. In *Proc. of AAAI Conference on Artificial Intelligence*, 11575–11585.

Li, H.; Oren, N.; and Norman, T. J. 2011. Probabilistic argumentation frameworks. In *Proc. of TAFA*, 1–16.

McCloy, R., and Byrne, R. M. 2002. Semifactual "even if" thinking. *Thinking & reasoning* 8:41–67.

Miller, T. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artif. Intell.* 267:1–38.

Modgil, S., and Prakken, H. 2014. The *ASPIC*$^+$ framework for structured argumentation: a tutorial. *Argument Comput.* 5(1):31–62.

Moulin, B.; Irandoust, H.; Bélanger, M.; and Desbordes, G. 2002. Explanation and argumentation capabilities: Towards the creation of more persuasive agents. *Artificial Intelligence Review* 17(3):169–222.

Niskanen, A.; Wallner, J. P.; and Järvisalo, M. 2016. Optimal status enforcement in abstract argumentation. In *Proc. of International Joint Conference on Artificial Intelligence (IJCAI)*, 1216–1222.

Niskanen, A.; Wallner, J. P.; and Järvisalo, M. 2018. Extension enforcement under grounded semantics in abstract argumentation. In *Proc. of International Conference on Principles of Knowledge Representation and Reasoning (KR)*, 178–183.

Papadimitriou, C. H. 1994. *Computational complexity*. Addison-Wesley.

Prakken, H. 2009. Models of persuasion dialogue. In *Argumentation in Artificial Intelligence*. 281–300.

Romashov, P.; Gjoreski, M.; Sokol, K.; Martinez, M. V.; and Langheinrich, M. 2022. Baycon: Model-agnostic bayesian counterfactual generator. In *Proc. of International Joint Conference on Artificial Intelligence (IJCAI)*, 23–29.

Sakama, C., and Inoue, K. 2000. Prioritized logic programming and its application to commonsense reasoning. *Artif. Intell.* 123(1-2).

Sakama, C., and Son, T. C. 2020. Epistemic argumentation framework: Theory and computation. *J. Artif. Intell. Res.* 69:1103–1126.

Sakama, C. 2014. Counterfactual reasoning in argumentation frameworks. In *COMMA*, 385–396.

Saribatur, Z. G.; Wallner, J. P.; and Woltran, S. 2020. Explaining non-acceptability in abstract argumentation. In *Proc. of European Conference on Artificial Intelligence (ECAI)*, 881–888.

Thang, P. M.; Dung, P. M.; and Hung, N. D. 2009. Towards a common framework for dialectical proof procedures in abstract argumentation. *Journal of Logic and Computation* 19(6):1071–1109.

Ulbricht, M., and Wallner, J. P. 2021. Strong explanations in abstract argumentation. In *Proc. of AAAI Conference on Artificial Intelligence*, 6496–6504.

Vassiliades, A.; Bassiliades, N.; and Patkos, T. 2021. Argumentation and explainable artificial intelligence: a survey. *Knowl. Eng. Rev.* 36:e5.

Villata, S.; Boella, G.; Gabbay, D. M.; and van der Torre, L. W. N. 2012. Modelling defeasible and prioritized support in bipolar argumentation. *Ann. Math. Artif. Intell.* 66(1-4).

Wallner, J. P.; Niskanen, A.; and Järvisalo, M. 2017. Complexity results and algorithms for extension enforcement in abstract argumentation. *J. Artif. Intell. Res.* 60:1–40.

Wu, Y.; Zhang, L.; and Wu, X. 2019. Counterfactual fairness: Unidentification, bound and algorithm. In *Proc. of International Joint Conference on Artificial Intelligence (IJCAI)*, 1438–1444.