Counterfactual Explanations Under Model Multiplicity and Their Use in Computational Argumentation

Gianvincenzo Alfano[1] , **Adam Gould**[2] , **Francesco Leofante**[2] ,
Antonio Rago[2,3] and **Francesca Toni**[2]
[1]DIMES Department, University of Calabria, Rende, Italy
[2]Department of Computing, Imperial College London, United Kingdom
[3]Department of Informatics, King's College London, United Kingdom
g.alfano@dimes.unical.it, {adam.gould19, f.leofante, a.rago, f.toni}@imperial.ac.uk

Abstract

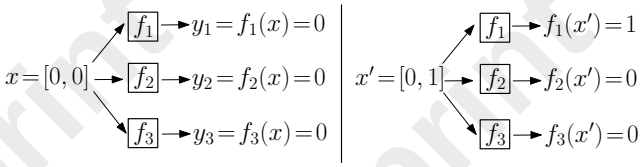
Counterfactual explanations (CXs) are widely recognised as an essential technique for providing recourse recommendations for AI models. However, it is not obvious how to determine CXs in *model multiplicity* scenarios, where equally performing but different models can be obtained for the same task. In this paper, we propose novel qualitative and quantitative definitions of CXs based on explicit, nested quantification over (groups) of model decisions. We also study properties of these notions and identify decision problems of interest therefor. While our CXs are broadly applicable, in this paper we instantiate them within computational argumentation where model multiplicity naturally emerges, e.g. with incomplete and case-based argumentation frameworks. We then illustrate the suitability of our CXs for model multiplicity in legal and healthcare contexts, before analysing the complexity of the associated decision problems.

1 Introduction

Counterfactual reasoning is often leveraged in eXplainable AI (XAI) to shed light on the decision-making process of AI models (e.g. decision trees [Tolomei *et al.*, 2017] and neural networks [Wachter *et al.*, 2017]). In particular, counterfactual explanations (CXs) help users understand the outputs of a model by revealing how minimal changes in its inputs would alter the model's decisions. CXs are often advocated as being useful because they can provide actionable insights into AI models and empower recourse recommendations to users that are negatively affected by the models' decisions. For example, to understand a loan rejection for an applicant asking for £$15k$ to be repaid over 5 years, a CX may suggest a longer repayment period, thus showing the user what needs to be changed for the loan to be approved.

While CXs are widely advocated in XAI for explaining individual models [Guidotti, 2024a], recent work has highlighted the challenges arising when determining CXs in *model multiplicity* scenarios, where equally performing models that slightly differ in their internals can be obtained for the same task [Black *et al.*, 2022]. Indeed, under model multiplicity, CXs computed for individual models may fail to provide valid recourse with the other models [Leofante *et al.*, 2023; Jiang *et al.*, 2024], thus raising questions about their justifiability as explanations. The following example illustrates the problem.

Example 1. Consider a simple loan application scenario modelled by binary features `LoanTermOver10Years`, i.e., whether the loan term exceeds 10 years (1) or not (0), and `LoanAmountBelow10k`, i.e., whether the amount to be borrowed is below £$10k$ (1) or not (0). Assume an individual represented by input $x = [0, 0]$ applies for a loan and the bank uses three equally performing classifiers in $\mathcal{F} = \{f_1, f_2, f_3\}$ to predict whether the loan should be granted. Let $y_1 = f_1(x) = 0$, $y_2 = f_2(x) = 0$ and $y_3 = f_3(x) = 0$ be the classifier's outputs for x as shown below, left:



Then, the loan may be rejected. Suppose the applicant requests an explanation for the rejection and a CX $x' = [0, 1]$ is produced as shown above, right, intuitively encoding 'if-only the applicant were to request less than £$10k$ then their loan would be accepted'. The choice of x' is based on it being a valid CX for f_1, but, since x' is still rejected by f_2 and f_3, it is unclear whether x' is the best CX under model multiplicity since it only changes one model's prediction. □

In this paper, we take the view that computing CXs under model multiplicity requires fine-grained reasoning about the outputs of multiple models simultaneously, and propose a novel approach for this. Specifically, we focus on *Multiplicity Decision Frameworks* (MDFs) where individual model decisions are first-class citizens and propose novel definitions of CXs based on explicit, nested quantification over (groups of) models in the MDFs, as illustrated next.

Example 2. Continuing from Example 1, we observe that x' encodes the following explanation, which we call an $\exists\forall$-CX for x: *there **exists** at least one (non-empty) group $\mathcal{G} = \{f_1\}$ of models among those in $\mathcal{F}$, such that (i) **all** models in $\mathcal{G}$ agree on the outcome for x, and (ii) x' is a valid CX for x and (all*

models in) $\mathcal{G}$. Other CXs may exist, e.g. $x'' = [1, 0]$ for which $f_1(x'') = f_2(x'') = f_3(x'') = 1$, which we call a $\forall\forall$-CX for x as it captures the following intuition: *(i)* ***all*** *models in* $\mathcal{F}$ *agree on the outcome for* x*, and (ii)* x'' *is a CX for* x *in* ***all*** *models in* $\mathcal{F}$. Such x'' is intuitively stronger, and potentially more informative, than x' as it better captures the behaviour of the three models. □

In addition to *qualitative* notions of CXs as in the earlier example, we define *quantitative* notions of CXs intuitively capturing the number of subsets $\mathcal{G}$ of $\mathcal{F}$ in which all functions $f_i \in \mathcal{G}$: (1) agree on $f_i(x)$, and (2) differ in the output for x and x', i.e., $f_i(x) \neq f_i(x')$.

We then study the computational properties of our novel notions and the benefits they provide under model multiplicity, e.g. depending on the quantity of models that change their output, and pointing to trade-offs between the CXs' benefits and their cost to action.

While our notions of CXs are applicable to MDFs with a wide range of AI models, in this paper we explicitly study instantiations thereof in settings where decision models are represented by argumentation frameworks (AFs) as in Computational Argumentation (CA) (see [Atkinson *et al.*, 2017] for an overview). Within CA, the multiplicity that we target may result from the emergence of different AFs with the same arguments (e.g. as in argumentation for case-based reasoning when choosing different default arguments [Cyras *et al.*, 2016; Gould *et al.*, 2024]). Further, this multiplicity naturally arises in AFs that adopt a possible-world semantics, e.g. incomplete AFs [Baumeister *et al.*, 2018; Baumeister *et al.*, 2021].

Contributions. In summary, our main contributions are:

- a general framework for counterfactual reasoning under multiplicity, formalising novel qualitative and quantitative notions of CXs based on explicit, nested quantification;
- a categorization of qualitative definitions of CXs, based on how well they capture multiplicity, and the definition of natural decision problems related to their computation;
- instantiations of our general framework in CA giving novel notions of CXs for incomplete AFs and AFs for case-based reasoning, and the study of the complexity of the decision problems for these instantiations;
- an illustrative exploration of the applicability of our CA instantiation within legal and healthcare contexts.

2 A General Theoretical Framework

In this section we define qualitative and quantitative notions of CXs for generic *Multiplicity Decision Frameworks (MDFs)*, dealing with model multiplicity, and making use of the following notation. Given two functions $f_1 : D_1 \to C_1$ and $f_2 : D_2 \to C_2$, we say that f_1 and f_2 are *similar* if they share the same domain and codomain (i.e., $D_1 = D_2$ and $C_1 = C_2$). Given similar functions $f_1, \ldots, f_n$, we simply denote D and C their domain and codomain, respectively. Given a set $\mathcal{F} = \{f_1, \ldots, f_n\}$ of functions, we denote with $pow(\mathcal{F})$ the powerset of $\mathcal{F}$.

Definition 1 (MDF)**.** A *Multiplicity Decision Framework (MDF)* consists of a set $\mathcal{F} = \{f_1, \ldots f_n\}$ of similar functions that, for any input instance $x \in D$, returns the output $\mathcal{F}(x) = \{y_i = f_i(x) \mid i \in \{1, \ldots, n\}\} \in C^n$.

Note that the output of an MDF may be passed to a method producing a single output for each input as is done in several machine learning contexts, e.g. using majority voting [Black *et al.*, 2022]. In these contexts, if a CX is computed *post-hoc* for the single output, then issues concerning the CX's justifiability may emerge, as we have seen in Example 1. In this paper we take the view that CXs for MDFs should be more granular and allow to reason *ex-ante* about (sets of) decisions corresponding to different elements in the MDF's output explicitly. In the remainder of this section, we define new qualitative and quantitative notions of CXs, where individual model predictions are first-class citizens.

2.1 A Qualitative Approach to CXs

We consider four notions of CXs, based on choosing (i) subsets of the set of functions in an MDF such that the functions all agree on the output for a given input, and (ii) CXs that are valid for all or some of the functions in the chosen subsets while being 'close enough' to the input, according to a distance metric $d : D \times D \to \mathbb{N}$, as standard in the XAI literature when defining CXs [Guidotti, 2024a].

Definition 2 ($\nu\mu$-CX)**.** Let $\mathcal{F} = \{f_1, \ldots f_n\}$ be an MDF, $x \in D$ an input and $\nu,\mu \in \{\forall, \exists\}$ be two quantifiers. Given a threshold $\delta \in \mathbb{N}$, we say that $x' \in D$ is a $\nu\mu$-CX for x, $\mathcal{F}$ and δ iff $d(x, x') \leq \delta$ and

- for $\nu = \forall$ and $\mu \in \{\forall, \exists\}$:
  $\nu\, f_i, f_j \in \mathcal{F}.\ f_i(x) = f_j(x)$ and
  $\mu f_k \in \mathcal{F}.\ f_k(x') \neq f_k(x)$;
- for $\nu = \exists$ and $\mu \in \{\forall, \exists\}$:
  $\nu\, \mathcal{G} \in pow(\mathcal{F}) \setminus \emptyset.\ f_i(x) = f_j(x)$ for all $f_i, f_j \in \mathcal{G}$ and
  $\mu f_k \in \mathcal{G}.\ f_k(x') \neq f_k(x)$;

Intuitively, for $\forall\forall$-CXs and $\forall\exists$-CXs, all functions agree on the output for the given input, and the CXs offer valid recourse for all/some functions in $\mathcal{F}$, resp.; instead, for $\exists\forall$-CXs and $\exists\exists$-CXs, only some functions in $\mathcal{F}$ need to agree on the output for the given input, and the CXs offer valid recourse for these functions.

Note that Definition 2 captures proximity by means of a threshold δ. This is to ease the discussion on computational properties later in the paper, as is often the case in the XAI literature, e.g. [Leofante *et al.*, 2023; Mohammadi *et al.*, 2021]. However, we will also consider optimal counterfactuals, formalised as follows. In the remainder, the set of $\nu\mu$-CXs for x, $\mathcal{F}$ and δ is denoted as $\nu\mu(x, \mathcal{F}, \delta)$. Moreover, $x' \in \nu\mu(x, \mathcal{F}, \delta)$ is said to be *optimal* iff there is no $x'' \in \nu\mu(x, \mathcal{F}, \delta)$ such that $d(x, x'') < d(x, x')$. The set of optimal $\nu\mu$-CXs for x, $\mathcal{F}$ and δ is denoted as $\tilde{\nu\mu}(x, \mathcal{F}, \delta)$.

Informally, Definition 2 allows to specify CXs for subsets of $\mathcal{F}$ that can be chosen dynamically by using the quantifiers ν and μ, as illustrated in the following example.

Example 3. Consider the MDF introduced in Example 1. Formally, we have $\mathcal{F} = \{f_1, f_2, f_3\}$ with $f_i : \{0, 1\}^2 \to \{0, 1\}$ for any $i \in \{1, 2, 3\}$ and thus $D = \{0, 1\}^2$, $C = \{0, 1\}^3$. Consider the distance metric $d([a_1, a_2], [b_1, b_2]) = |a_1 - b_1| + |a_2 -$

$b_2|$, and threshold $\delta = 2$. Then, given the input $x = [0, 0]$, the (optimal) $\nu\mu$-CX of x w.r.t. $\mathcal{F}$ and δ are described in the last column of the following table.

instance	$f_1(\cdot)$	$f_2(\cdot)$	$f_3(\cdot)$
$x = [0,0]$	0	0	0
$x' = [0,1]$	1	0	0
$x'' = [1,0]$	1	1	1
$x''' = [1,1]$	1	1	1

$d(x,\cdot)$	$\nu\mu(x,\mathcal{F},\delta)$
1	$\forall\tilde{\exists}$ $\tilde{\exists}\forall$ $\tilde{\exists}\tilde{\exists}$
1	$\forall\tilde{\forall}$ $\forall\tilde{\exists}$ $\tilde{\exists}\forall$ $\tilde{\exists}\tilde{\exists}$
2	$\forall\forall$ $\forall\exists$ $\exists\forall$ $\exists\exists$

It holds that functions f_is (with $i \in \{1, 2, 3\}$) classify the input instance $x = [0, 0]$ (resp., x'' and x''') with $f_i(x) = 0$ (resp., $f_i(x'') = f_i(x''') = 1$). Thus, x'' and x''' are $\forall\forall$-CXs for x, $\mathcal{F}$ and δ. However, among x'' and x''', only the former is optimal, due to having a lower distance from x (as $d(x, x'') = 1 < d(x, x''') = 2$). □

We will focus on two decision problems regarding $\nu\mu$-CXs, i.e., *existence* and *verification*, as follows.

Definition 3 (Existence)**.** Let $\mathcal{F} = \{f_1, \dots f_n\}$ be an MDF, $x \in D$ be an input, δ a distance threshold, and ν, $\mu \in \{\forall, \exists\}$ be two quantifiers. $\mathsf{E}^{\nu\mu}$ (resp. $\tilde{\mathsf{E}}^{\nu\mu}$) is the problem of deciding whether there exists a $\nu\mu$- (resp., $\tilde{\nu\mu}$-) CX for x, $\mathcal{F}$ and δ.

The following proposition states the relations between the existence problems, also synthesized in Figure 1.

Proposition 1. The following relations hold:

(a) $\mathsf{E}^{\forall\forall}$ implies $\mathsf{E}^{\forall\exists}$ (resp., $\tilde{\mathsf{E}}^{\forall\forall}$ implies $\tilde{\mathsf{E}}^{\forall\exists}$);
(b) $\mathsf{E}^{\forall\exists}$ implies $\mathsf{E}^{\exists\forall}$ (resp., $\tilde{\mathsf{E}}^{\forall\exists}$ implies $\tilde{\mathsf{E}}^{\exists\forall}$);
(c) $\mathsf{E}^{\exists\forall}$ and $\mathsf{E}^{\exists\exists}$ (resp., $\tilde{\mathsf{E}}^{\exists\forall}$ and $\tilde{\mathsf{E}}^{\exists\exists}$) are equivalent;
(d) $\tilde{\mathsf{E}}^{\nu\mu}$ and $\mathsf{E}^{\nu\mu}$ are equivalent, for any $\nu, \mu \in \{\exists, \forall\}$;
(e) The inverse of relations (a) and (b) does not hold.

Note that (d) is trivial, as the existence of a $\nu\mu$-CX for x, $\mathcal{F}$ and δ implies the existence of an optimal $\nu\mu$-CX for x, $\mathcal{F}$ and δ, and vice versa, for any fixed ν,$\mu \in \{\forall, \exists\}$.

Definition 4 (Verification)**.** Let $\mathcal{F} = \{f_1, \dots f_n\}$ be an MDF, $x, x' \in D$ be two inputs, δ a distance threshold, and $\nu, \mu \in \{\forall, \exists\}$ be two quantifiers. $\mathsf{V}^{\nu\mu}$ (resp., $\tilde{\mathsf{V}}^{\nu\mu}$) is the problem of checking whether x' is a $\nu\mu$- (resp., $\tilde{\nu\mu}$-) CX for x, $\mathcal{F}$ and δ.

Differently from the case of the existence problem, the optimality constraint makes the problems $\mathsf{V}^{\nu\mu}$ and $\tilde{\mathsf{V}}^{\nu\mu}$ non-equivalent. That is, $\tilde{\mathsf{V}}^{\nu\mu}$ implies $\mathsf{V}^{\nu\mu}$, while the vice-versa may not hold as there may exist some $\nu\mu$-CX for x, $\mathcal{F}$ and δ that is not optimal, as illustrated next.

Example 4. Continuing from Example 3, for any pair of quantifiers $\nu, \mu \in \{\exists, \forall\}$ it holds that $(x, x'', \mathcal{F}, \delta = 2)$ and $(x, x''', \mathcal{F}, \delta)$ are true instances of $\mathsf{V}^{\nu\mu}$, as both x'' and x''' belongs to $\nu\mu(x, \mathcal{F}, \delta)$. However, as $d(x, x'') < d(x, x''')$, we have that $(x, x'', \mathcal{F}, \delta)$ is the only true instance of $\tilde{\mathsf{V}}^{\nu\mu}$. □

The relationships between verification problems are outlined below and summarised in Figure 1.

Proposition 2. The following relations hold:

(a) $\mathsf{V}^{\forall\forall}$ implies $\mathsf{V}^{\forall\exists}$;
(b) $\mathsf{V}^{\forall\exists}$ implies $\mathsf{V}^{\exists\forall}$;
(c) $\mathsf{V}^{\exists\forall}$ and $\mathsf{V}^{\exists\exists}$ (resp., $\tilde{\mathsf{V}}^{\exists\forall}$ and $\tilde{\mathsf{V}}^{\exists\exists}$) are equivalent;
(d) $\tilde{\mathsf{V}}^{\nu\mu}$ implies $\mathsf{V}^{\nu\mu}$, for any $\nu, \mu \in \{\exists, \forall\}$;
(e) The inverse of relations (a) and (b) does not hold.

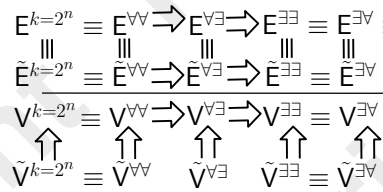


Figure 1: Relations for existence (top), and verification problems (bottom) from Propositions 1-3.

2.2 A Quantitative Approach to CXs

Our qualitative notions of CXs offer limited insight into the numerical strength of an explanation in terms of multiplicity. Moreover, when the functions in $\mathcal{F}$ do not all agree on the outcome for x, the qualitative notions cannot distinguish the type of CX. For illustration, assuming $f_3(x) = 1$ in the previous example, we would no longer be able to distinguish between x' and x'', as they collapse to be both $\exists\forall$-CXs for x.

Here, we introduce a novel, entirely quantitative definition of CXs, making use of the following notation. We denote with $\mathcal{F}_{(x,x')} = |\{\mathcal{G} \in pow(\mathcal{F}).\ f_i(x) = f_j(x)$ and $f_i(x') \neq f_i(x)$ for all $f_i, f_j \in \mathcal{G}\}|$ the number of subsets $\mathcal{G}$ of $\mathcal{F}$ in which all functions $f_i \in \mathcal{G}$: (1) agree on $f_i(x)$, and (2) differ in the output for x and x', i.e., $f_i(x) \neq f_i(x')$. We illustrate this notation next.

Example 5. Consider the MDF $\mathcal{F} = \{f_1, f_2, f_3\}$ and inputs $x = [0, 0]$ and $x' = [0, 1]$ of Example 3 having distance $d(x, x') = 1 \leq \delta = 2$. We have that $\mathcal{F}_{(x,x')} = |\{\varnothing, \{f_1\}\}| = 2$. When considering $x' = [1, 0]$ (or $x' = [1, 1]$) we have that $\mathcal{F}_{(x,x')} = |pow(\mathcal{F})| = 2^3 = 8$. □

Definition 5 (k-CX)**.** Let $\mathcal{F} = \{f_1, \dots f_n\}$ be an MDF, $x \in D$ an input, $\delta \in \mathbb{N}$ a threshold, and $k \in \mathbb{N}^+$ a positive integer. We say that $x' \in D$ is a k-CX for x, $\mathcal{F}$ and δ iff $d(x, x') \leq \delta$ and $\mathcal{F}_{(x,x')} \geq k$.

The set of k-CXs for x, $\mathcal{F}$ and δ is denoted as $k(x, \mathcal{F}, \delta)$. Moreover, $x' \in k(x, \mathcal{F}, \delta)$ is said to be *optimal* iff there is no $x'' \in k(x, \mathcal{F}, \delta)$ such that $d(x, x'') < d(x, x')$. The set of optimal k-CX for x, $\mathcal{F}$ and δ is denoted as $\tilde{k}(x, \mathcal{F}, \delta)$.

Example 6. Continuing from the previous example, we have that x' and x'' = $[1, 1]$ are the only 8-CXs for x, $\mathcal{F}$ and δ. Moreover, only x' is optimal, i.e., $\tilde{8}(x, \mathcal{F}, \delta) = \{x'\}$. □

The natural quantified versions of the existence and verification problems follow.

Definition 6 (Quantified Existence)**.** Let $\mathcal{F} = \{f_1, \dots f_n\}$ be an MDF, $k \in \mathbb{N}$, $x \in D$ be an input, and δ a threshold. E^k (resp., $\tilde{\mathsf{E}}^k$) is the problem of deciding whether there exists a k- (resp., $\tilde{k}$-) CX for x, $\mathcal{F}$ and δ.

Definition 7 (Quantified Verification)**.** Let $\mathcal{F} = \{f_1, \dots f_n\}$ be an MDF, $k \in \mathbb{N}$, $x, x' \in D$ be two inputs, and δ a threshold. V^k (resp., $\tilde{\mathsf{V}}^k$) is the problem of checking whether x' is a k- (resp., $\tilde{k}$-) CX for x, $\mathcal{F}$, and δ.

The next proposition states the relations between the existence and verification problems, also synthesized in Figure 1.

Proposition 3. Let $\mathcal{F} = \{f_1, \dots f_n\}$ be an MDF, x an input, and δ a threshold. The following relations hold:

(a) V^{2^n} and $\mathsf{V}^{\forall\forall}$ (resp., $\tilde{\mathsf{V}}^{2^n}$ and $\tilde{\mathsf{V}}^{\forall\forall}$) are equivalent;
(b) E^{2^n} and $\mathsf{E}^{\forall\forall}$ (resp., $\tilde{\mathsf{E}}^{2^n}$ and $\tilde{\mathsf{E}}^{\forall\forall}$) are equivalent;
(c) $\mathsf{E}^{\exists\exists}$ and $\mathsf{E}^{\exists\forall}$ (resp., $\tilde{\mathsf{E}}^{\exists\exists}$ and $\tilde{\mathsf{E}}^{\exists\forall}$) are equivalent to $\mathsf{E}^{k=1}$ (resp., $\tilde{\mathsf{E}}^{k=1}$);
(d) $\mathsf{V}^{\exists\exists}$ and $\mathsf{V}^{\exists\forall}$ (resp., $\tilde{\mathsf{V}}^{\exists\exists}$ and $\tilde{\mathsf{V}}^{\exists\forall}$) are equivalent to $\mathsf{V}^{k=1}$ (resp., $\tilde{\mathsf{V}}^{k=1}$).

3 An Argumentative Setting

We now instantiate our framework within the context of Computational Argumentation (CA), choosing specifically Abstract Argumentation (AA) as the underlying form of CA [Bench-Capon and Dunne, 2007]. We first give some core background in AA (Section 3.1), specifically on the AA Framework (AF) [Dung, 1995], followed by the corresponding instantiations of the notions introduced in Section 2 (Section 3.2). We then illustrate the emergence of model multiplicity in AA with two case studies: with *incomplete AF* [Baumeister *et al.*, 2018; Fazzinga *et al.*, 2020] (Section 3.3) and with *AA-CBR* [Cyras *et al.*, 2016; Gould *et al.*, 2024], when an AF is used to perform case-based reasoning (Section 3.4). Finally, we study the computational complexity of the existence and verification problems we have defined in the argumentative setting we consider (Section 3.5).

3.1 Background on AFs

Let $\mathcal{A}$ a set of arguments, which we call *universal*. An Abstract Argumentation Framework (AF) [Dung, 1995] is a pair $\langle A, R\rangle$, where $A \subseteq \mathcal{A}$ is a set of *arguments* and $R \subseteq A \times A$ is a set of *attacks*: if $(a, b) \in R$ then we say that a attacks b. Given an AF $\Lambda = \langle A, R\rangle$ and a set $S \subseteq A$ of arguments, the sets of *defeated* and *acceptable* arguments w.r.t. S are as follows (where Λ is understood):
• $\mathit{Def}(S) = \{a \in A \mid \exists (b, a) \in R.\ b \in S\}$;
• $\mathit{Acc}(S) = \{a \in A \mid \forall (b, a) \in R.\ b \in \mathit{Def}(S)\}$.
Then, $S \subseteq A$ is said to be (i) *conflict-free* iff $S \cap Def(S) = \varnothing$; and (ii) *admissible* iff it is conflict-free and $S \subseteq Acc(S)$.

Different semantics have been proposed to characterize collectively acceptable sets of arguments, called *extensions* [Dung, 1995]. Specifically, $S \subseteq A$ is an *extension* called:
• *complete* (co) iff it is admissible and $S = Acc(S)$;
• *preferred* (pr) iff it is a $\subseteq$-maximal complete extension;
• *stable* (st) iff it is a preferred extension s.t. $S \cup Def(S) = A$;
• *grounded* (gr) iff it is a $\subseteq$-minimal complete extension.

Example 7. Let $\Lambda = \langle A, R\rangle$ be an AF where $A = \{\mathtt{a}, \mathtt{b}, \mathtt{c}\}$ and $R = \{(\mathtt{a}, \mathtt{b}), (\mathtt{b}, \mathtt{a}), (\mathtt{b}, \mathtt{c}), (\mathtt{c}, \mathtt{c})\}$. AF Λ has three complete extensions: $E_1 = \varnothing, E_2 = \{\mathtt{a}\}, E_3 = \{\mathtt{b}\}$, where E_2 and E_3 are preferred, E_3 is stable, and E_1 is grounded. □

The set of complete (resp. stable and grounded) extensions of an AF Λ will be denoted by $\mathtt{co}(\Lambda)$ (resp. $\mathtt{pr}(\Lambda)$, $\mathtt{st}(\Lambda)$, $\mathtt{gr}(\Lambda)$). With a small abuse of notation, we also use $\mathtt{gr}(\Lambda)$ to denote the grounded extension.

Several decision problems can be associated to an AF Λ under semantics σ, including the *verification* problem, denoted as $\mathsf{V}^{\sigma,\Lambda}(S)$, that checks whether a given set S of arguments belongs to $\sigma(\Lambda)$. For illustration, considering the AF Λ of Example 7, $\mathsf{V}^{\mathtt{st},\Lambda}(S = \{\mathtt{b}\})$ returns true as $S \in \mathtt{st}(\Lambda)$.

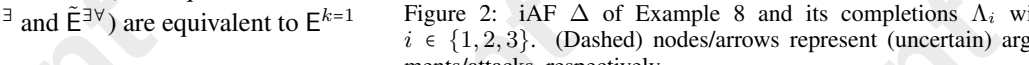


Figure 2: iAF Δ of Example 8 and its completions Λ_i with $i \in \{1, 2, 3\}$. (Dashed) nodes/arrows represent (uncertain) arguments/attacks, respectively.

The complexity of the above problems has been investigated (see [Dvorák and Dunne, 2017] for a survey).

3.2 Argumentative MDFs and CXs

Argumentative Multiplicity Decision Frameworks are MDFs comprising *argumentative functions* able to solve argumentative queries about decision problems in the argumentation frameworks (AFs in this paper) underpinning the functions, as follows. Note that, from now on, for uniformity with Section 2, we use x to denote a set of arguments in $\mathcal{A}$.

Definition 8 (AA-MDF)**.** An *AA Multiplicity Decision Framework (AA-MDF)* is an MDF with $\mathcal{F} = \{f_1, \ldots f_n\}$ consisting of argumentative functions $f_i(x) = \mathsf{P}_i^{\sigma_i,\Lambda_i}(x)$ whose input is a set of arguments $x \subseteq \mathcal{A}$, and whose output is a boolean answer to the question $\mathsf{P}_i^{\sigma_i,\Lambda_i}(x)$, where: P is a decision problem; σ_i is a semantics; and Λ_i is an AF.

Thus, two functions in an AA-MDF are similar in that they share the same domain (i.e., the universal set of arguments $\mathcal{A}$) and the same codomain (i.e., a boolean value corresponding to the output of an argumentative decision problem).

Qualitative and quantitative CXs for AA-MDFs are then as in Section 2, but using the following concrete distance metric whereby the distance between two sets of arguments is the number of changes needed to make them equal.

Definition 9 (Symmetric Difference Distance Metric)**.** Let $\Lambda = \langle A, R\rangle$ be an AF. Given two sets $x \subseteq A$ and $x' \subseteq A$ of arguments, the symmetric difference distance metric $d(x, x')$ between x and x' is given by $d(x, x') = |x \setminus x'| + |x' \setminus x|$.

Considering, e.g., the sets $x = \{\mathtt{a}, \mathtt{b}\}, x' = \{\mathtt{b}\}$ and $x'' = \{\mathtt{a}\}$, it holds that $d(x, x') = d(x, x'') = 1$. We note that a similar distance metric is commonly adopted in several works within logic-based explainable AI to compare vectors of (binary) features [Barceló *et al.*, 2020; Alfano *et al.*, 2025; Guidotti, 2024b].

3.3 AA-MDFs and CXs with Incomplete AF

For our first case study, consider *incomplete AFs* (iAFs), i.e., tuples $\langle A, B, R, T\rangle$, where A and B are disjoint sets of arguments, and R and T are disjoint sets of attacks between arguments in $A \cup B$. Arguments in A and attacks in R are said to be *certain*, while arguments in B and attacks in T are said to be *uncertain* [Baumeister *et al.*, 2018; Fazzinga *et al.*, 2020]. Certain arguments/attacks in A/R are definitely known to exist, while uncertain arguments/attacks in B/T are not known for sure: they may or may not occur.

An iAF compactly represents alternative AF scenarios, called *completions*. A *completion* for an iAF $\Delta = \langle A, B, R, T\rangle$ is an AF $\Lambda = \langle A', R'\rangle$ such that $A \subseteq A' \subseteq A \cup B$ and $R \cap (A' \times A') \subseteq R' \subseteq (R \cup T) \cap (A' \times A')$.

Verification problems for iAF have been investigated in [Baumeister *et al.*, 2018; Fazzinga *et al.*, 2020]. Given an

iAF $\Delta = \langle A, B, R, T\rangle$, a set of arguments $x \subseteq (A \cup B)$, and a semantics $\sigma \in \{\mathtt{gr}, \mathtt{co}, \mathtt{pr}, \mathtt{st}\}$, the *possible/necessary verification* problem under σ (denoted as $\mathsf{PV}^{\sigma,\Delta}/\mathsf{NV}^{\sigma,\Delta}$) consists in deciding whether x is a σ-extension in any/all completions of Δ, respectively.

Example 8. Consider a legal situation where a defendant, Alex, is accused of robbery. Bob, a potential witness, says that they possibly saw Alex. It is unclear whether Bob will testify or whether the jury will consider Bob's eventual testimony. This situation can be modelled by the iAF Δ of Figure 2 (left), where statements 'Alex will not be found guilty' and 'Bob will testify' are represented by arguments $\mathtt{a}$ and $\mathtt{b}$, respectively. The iAF Δ has three completions Λ_1, Λ_2, and Λ_3 (see Figure 2) capturing alternative scenarios. $x = \{\mathtt{a}\}$ is a stable extension for all the three completions (i.e., $\mathsf{NV}^{\mathtt{st},\Delta}(x)$ holds), while $x' = \{\mathtt{b}\}$ is a stable extension in only one of the three completions (i.e., $\mathsf{PV}^{\mathtt{st},\Delta}(x')$ holds). □

However, if PV/NV problems are false, a user may want to know the reasons. For instance, assume the user is interested in knowing whether both Bob will testify and Alex will not be found guilty. As $x = \{\mathtt{a}, \mathtt{b}\}$ is neither a possible nor necessary extension, an interesting question is *why not*. To answer such questions we make use of an AA-MDF where argumentative function $f_i(x)$ solve the problem $\mathsf{V}_i^{\sigma_i,\Lambda_i}(x)$, as outlined next.

Example 9. Consider the iAF Δ of Example 8. To capture such multiplicity , an AA-MDF $\mathcal{F} = \{f_1, f_2, f_3\}$ (see Figure 3) can be instantiated, where functions $f_i(x) = \mathsf{V}^{\mathtt{st},\Lambda_i}(x)$ return true if $x \in \mathtt{st}(\Lambda)$. Thus:

- for $x = \{\mathtt{a}, \mathtt{b}\}$, $f_1(x) = f_2(x) = f_3(x) = 0$, intuitively capturing the fact that $\{\mathtt{a}, \mathtt{b}\}$ is not a stable extension in any of the three AFs;
- for $x' = \{\mathtt{b}\}$, $f_1(x') = 1$ and $f_2(x') = f_3(x') = 0$, intuitively capturing the fact that $\{\mathtt{b}\}$ is a stable extension only for AF Λ_1; and
- for $x'' = \{\mathtt{a}\}$, $f_1(x'') = f_2(x'') = f_3(x'') = 1$, intuitively capturing the fact that $\{\mathtt{a}\}$ is a stable extension in all of the three AFs.

Assume the Symmetric Difference Distance Metric (cf. Definition 9), and a threshold $\delta = 2$. It holds that $x'' = \{\mathtt{a}\}$ belongs to $\nu\tilde{\mu}(x, \mathcal{F}, \delta)$, for any $\nu, \mu \in \{\exists, \forall\}$, intuitively capturing the notion 'if only you remove $\{\mathtt{b}\}$ from $x = \{\mathtt{a}, \mathtt{b}\}$, then you can conclude that x'' is a stable extension in all completions, which tells us that if only Bob does not testify, then Alex will not be found guilty in any possible scenarios.' □

Thus, counterfactual reasoning in AA-MDF encodes explanations in the form of minimal changes to the input set of arguments to obtain a different outcome (i.e., an extension).

Further, we illustrate the usefulness of the resulting instances of our quantitative notions of CXs.

Example 10. Continuing from Example 9, we have that $\mathcal{F}_{(x,x')} = |\{\varnothing, \{f_1\}\}| = 2$, while $\mathcal{F}_{(x,x'')} = |pow(\mathcal{F})| = 2^3 = 8$. Thus, for $\delta = 2$, we have that x' and x'' belong to $\tilde{2}(x, \mathcal{F}, \delta)$, while only x'' belongs to $\tilde{8}(x, \mathcal{F}, \delta)$. Intuitively, x'' captures an explanation of the form: 'if only you remove $\{\mathtt{b}\}$ from x then you can conclude that x'' is a stable extension independently of the chosen set of scenarios'. □

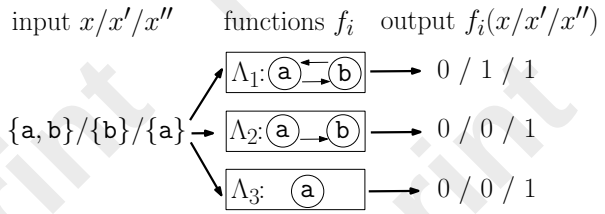


Figure 3: AA-MDF $\mathcal{F} = \{f_1, f_2, f_3\}$ of Example 9. Functions $f_i(x) = \mathsf{V}^{\mathtt{st},\Lambda_i}(x)$ return true if the set x is a stable extension of the completion Λ_i of iAF Δ of Example 8.

Finally, qualitative and quantitative problems in AA-MDF naturally instantiate their MDF counterparts (cf. Definitions 3-4 and 6-7), where x and x' denote sets of arguments.

Example 11. Considering the AA-MDF of Example 9, it holds that for any $\nu, \mu \in \{\exists, \forall\}$: $(x, \mathcal{F}, \delta)$ is a true instance of $\tilde{\mathsf{E}}^{\nu\mu}$ and $\mathsf{E}^{\nu\mu}$; and $(x, x'', \mathcal{F}, \delta)$ is a true instance of $\tilde{\mathsf{V}}^{\nu\mu}$ and $\mathsf{V}^{\nu\mu}$. Moreover, $(x, x', \mathcal{F}, \delta)$ is a true (resp. false) instance of $\tilde{\mathsf{V}}^{\forall\exists}$ (resp. $\tilde{\mathsf{V}}^{\forall\forall}$). Also, quantitatively, for k-CXs, we have that for any $k \in \{1, \ldots, 8\}$: $(x, \mathcal{F}, \delta)$ is a true instance of $\tilde{\mathsf{E}}^k$ and E^k; and $(x, x'', \mathcal{F}, \delta)$ is a true instance of $\tilde{\mathsf{V}}^k$ and V^k. Finally, $(x, x', \mathcal{F}, \delta)$ is a true instance of $\tilde{\mathsf{V}}^k$ and V^k, for $k \in \{1, 2\}$. □

3.4 AA-MDFs and CXs with AA-CBR

We now show how our AA-MDFs can be applied with *Preference-Based Abstract Argumentation for Case-Based Reasoning (AA-CBR-$\mathcal{P}$)* [Gould *et al.*, 2024], which is a binary classification model using an AF as a reasoner constructed from labelled data points, known as *cases*, and user preferences. In AA-CBR-$\mathcal{P}$, cases are structured with a *characterisation* (e.g. a set of features) and a labelled outcome. A *casebase* is a set of past cases, and can be ordered by how *exceptional* cases are as determined by a provided *sequence of preorders* defined over the characterisations. An AF can be obtained from the casebase whereby arguments are cases and attacks occur between cases of opposing outcomes from more exceptional to less exceptional cases; an additional constraint ensures minimal difference between the cases to avoid unnecessary attacks. When defining attacks, we apply the preorders in the sequence lexicographically. Thus, preferences are applied based on how the preorder sequence is specified. Different preferences naturally introduce multiplicity of AFs. An unlabelled *new case*, N, can be classified by adding it to the AF using a notion of *irrelevance*. We then compute the grounded semantics. If a *default case* is accepted, we assign its outcome to the new case; otherwise, we assign the opposing outcome. For example, we may characterize cases by sets of high-priority and low-priority features, with case A considered more exceptional than case B if either A has a superset of high-priority features or they have the same set of high-priority features and A has a superset of low priority features. Different stakeholders may consider which features are high or low priority differently.

Example 12. Consider assessing patients in good (+) or poor health (−) within a simple medical setting. Patients may dis-

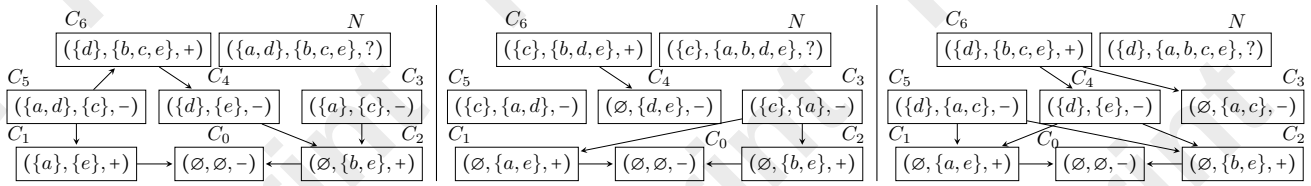


Figure 4: AFs of Clinitian 1 (left), 2 (middle), and 3 (right) generated from AA-CBR-$\mathcal{P}$ according to Example 12. The high-priority features in the AF of Clinitian 1 (resp., 2, and 3) are $\{a, d\}$ (resp., $\{c\}$, and $\{d\}$).

play weight loss (feature a), maintain a healthy diet (feature b), suffer from a chronic condition (feature c), experience appetite loss (feature d), and engage in regular exercise (feature e). When conducting this assessment, three clinicians consider the same set of past patients as their casebase, D_{cb}, to assess a new patient, N, who presents with all features. However, differing clinicians will prioritise each feature in their assessment differently, with clinician 1 considering a and d as high-priority features and b, c, and e as low-priority features; clinician 2 considering feature c as a high-priority feature and a, b, d and e as low-priority features; and clinician 3 considering d as a high-priority feature and a, b, c and e as low-priority features. Figure 4 showcases the AA-CBR-$\mathcal{P}$ models for each clinician. All three models include the default case, C_0, in the grounded extension and, therefore, predict that the new case N is in poor health (−). □

The clinicians may be convinced that the patient is healthy despite what the models suggest. We can leverage AA-CBR-$\mathcal{P}$ as an interpretable argumentative model to generate CXs that can modify the models to accommodate this. This form of counterfactual can be viewed in the light of contestability , which is necessary to identify and repair incorrect predictions or faulty reasoning in algorithmic decision systems [Leofante *et al.*, 2024].

To do so, we can leverage an AA-MDF with the problem $\mathsf{P}_i^{\mathtt{gr},\Lambda_i}(x) = [C_0 \in \mathtt{gr}(AF_{\mathcal{P}_i}(x \cap D_{cb}, N))]$, where $AF_{\mathcal{P}_i}(.)$ is a function constructing the AF with AA-CBR-$\mathcal{P}$ using clinician i's preferences, x is the input set of arguments, D_{cb} is the casebase and N is a new case. We can identify which subsets of the casebase can be used to change the outcome using the symmetric distance metric (Definition 9), $\delta = 5$ and letting $x = D_{cb}$. Intuitively, this asks the counterfactual question of which past patients need to be disregarded from each clinician's casebase for the new patient to be predicted as healthy.

The following table showcases (a subset of) possible CXs.

instance	$f_1(\cdot)$	$f_2(\cdot)$	$f_3(\cdot)$	$d(x,\cdot)$	$\nu\mu(x,\mathcal{F},\delta)$
D_{cb}	-	-	-		
$\{C_1, C_2, C_4, C_6\}$	+	+	+	2	$\tilde{\forall}\forall\ \forall\exists\ \exists\forall\ \exists\exists$
$\{C_1, C_3, C_4, C_6\}$	+	+	+	2	$\tilde{\forall}\forall\ \forall\exists\ \exists\forall\ \exists\exists$
$\{C_1, C_2, C_6\}$	+	+	+	3	$\forall\forall\ \forall\exists\ \exists\forall\ \exists\exists$
$\{C_1, C_2, C_3, C_4, C_6\}$	+	-	+	1	$\tilde{\forall}\exists\ \tilde{\exists}\forall\ \tilde{\exists}\exists$
$\{C_1, C_3, C_4, C_5, C_6\}$	-	+	+	1	$\tilde{\forall}\exists\ \tilde{\exists}\forall\ \tilde{\exists}\exists$
$\{C_1, C_2, C_3, C_4\}$	+	-	-	2	$\forall\exists\ \forall\forall\ \exists\exists$

We see two possible $\tilde{\forall}\forall$-CXs, which require removing two arguments from the casebase to convince all three clinicians to change their outcome for the new case. The first $\tilde{\forall}\forall$-CX $\{C_1,C_2,C_4,C_6\}$ states that if arguments C_3 and C_5 are removed from the casebase, then all three clinicians will predict that the patient is in good health. Similarly, the second $\tilde{\forall}\forall$-CX $\{C_1,C_3,C_4,C_6\}$ requires removing arguments C_2, C_5.

However, disregarding two arguments from the casebase may be extreme, and comes at a greater cost than if we remove solely C_5 or solely C_2. Both scenarios convince two out of three of the clinicians to change their prediction but with only a single argument removed. These are optimal $\tilde{\forall}\exists$-CXs. Our novel CX notions can, therefore, empower a user with a choice between a 'stronger' CX that convinces more clinicians to change their minds or a 'weaker' CX in which fewer clinicians change their minds, but with a smaller and more palatable cost to action.

3.5 Computational Complexity

The parametric nature of AA-MDF is influenced by three key aspects: 1) the class of AFs used, 2) the specific argumentation problem considered, and 3) the semantics applied. These aspects result in numerous possible combinations, making it infeasible to fully address the complexity of all variants. Thus, we focus on a single source of multiplicity, reserving the exploration of other dimensions for future research. Particularly, we focus on multiplicity that comes from choosing multiple AFs, as in the case of Sections 3.3 and 3.4. Moreover, to limit multiplicity in the number of extensions, we adopt the grounded semantics, also motivated from the AA-CBR instance where the grounded semantics is used for data classification. As argumentative queries/functions, we choose $\mathsf{U}^{\mathtt{gr},\Lambda}$, that is a small variant of the verification problem (i.e., $\mathsf{V}^{\mathtt{gr},\Lambda}$) returning true iff the input set of arguments x is contained in the grounded extension of Λ (this last choice is also motivated from the AA-CBR instance). Note that, this different choice of problem does not impact on the complexity results, as problems U and V can both be solved in PTIME [Dvorák and Dunne, 2017].

Theorem 1. The problems $\mathsf{E}^{\nu\mu}$ and $\tilde{\mathsf{E}}^{\nu\mu}$ are in PTIME for any $\nu, \mu \in \{\exists, \forall\}$.

Theorem 2. The problems $\mathsf{V}^{\nu\mu}$ and $\tilde{\mathsf{V}}^{\nu\mu}$ are in PTIME for any $\nu, \mu \in \{\exists, \forall\}$.

From the previous theorems we can conclude that, independently of the quantification, computing a CX in the chosen setting of interest can be done in PTIME. However, this can not generally hold when moving the attention to the quantified versions, as stated in the following theorems.

Theorem 3. The problems E^k and $\tilde{\mathsf{E}}^k$ are in PTIME for $k \in \{1, 2^n\}$; and in NP for $k \notin \{1, 2^n\}$.

	Problems			
$\nu\mu/k$	$\mathsf{E}^{\nu\mu/k}$	$\tilde{\mathsf{E}}^{\nu\mu/k}$	$\mathsf{V}^{\nu\mu/k}$	$\tilde{\mathsf{V}}^{\nu\mu/k}$
$\exists\exists, \exists\forall, \forall\exists, \forall\forall$	PTIME	PTIME	PTIME	PTIME
$k \in \{1, 2^n\}$	PTIME	PTIME	PTIME	PTIME
$k \notin \{1, 2^n\}$	NP	NP	PTIME	coNP

Table 1: Complexity of the (optimal) $\nu\mu$- and k- existence and verification problems in the considered AA-MDF setting.

Theorem 4. The problems V^k, $\tilde{\mathsf{V}}^1$, and $\tilde{\mathsf{V}}^{2^n}$ are in PTIME; while $\tilde{\mathsf{V}}^k$ is in coNP for any $k \notin \{1, 2^n\}$

Thus, these results demonstrate the trade-off between the granularity of the CX and its complexity; specifically, the finer the granularity, the greater the complexity.

4 Related Work

Model multiplicity has been the subject of recent research efforts within the trustworthy machine learning community [Black *et al.*, 2022]. Researchers have shown that among equally accurate models, there could be different fairness characteristics [Wick *et al.*, 2019; Coston *et al.*, 2021], interpretability levels [Rudin, 2019; Semenova *et al.*, 2022] and even inconsistent explanations [Fisher *et al.*, 2019; Mehrer *et al.*, 2020; Marx *et al.*, 2023]. More central to this work, a Mixed-Integer Linear Programming encoding method has been proposed to compute CXs that are provably valid for all feed-forward neural networks in an ensemble [Leofante *et al.*, 2023]. This approach corresponds to a specific instantiation of our framework, namely $\forall\forall$ CXs. Different from us, their work focuses on deep learning models, and their encoding does not apply to our setting in this paper.

Meanwhile, argumentation has been advocated by prominent works in the explainable AI literature [Miller, 2019] as a useful mechanism for explanation, given that a majority of statements made in explanations have been shown to actually be argumentative claim-backings [Antaki and Leudar, 1992]. This has given rise to a whole sub-field of works known as argumentative explainable AI (see [Cyras *et al.*, 2021; Vassiliades *et al.*, 2021; Guo *et al.*, 2023] for overviews), where argumentation is used for explaining AI models. This range of methods includes those explaining argumentative reasoning mechanisms themselves, e.g. [Ulbricht and Wallner, 2021; Brewka and Ulbricht, 2019; Saribatur *et al.*, 2020; Fan and Toni, 2014; Amgoud, 2024; Borg and Bex, 2024; Kampik *et al.*, 2024; Yin *et al.*, 2024]. However, these methods generally assume access to the underlying AF and thus do not qualify as post-hoc explanations. Moreover, while some methods also touch upon counterfactual reasoning [Fan and Toni, 2014; Amgoud, 2024; Borg and Bex, 2024; Kampik *et al.*, 2024; Yin *et al.*, 2024; Alfano *et al.*, 2024], none explicitly consider the challenge posed by model multiplicity, which we address in this paper for the first time.

Within the realm of model multiplicity, an argumentation-based solution to the problem of aggregation to address model multiplicity, while factoring in CXs has been proposed [Jiang *et al.*, 2024]. Their solution is in the spirit of an $\exists\forall$ instance of our approach whereby $\mathcal{F}$ is a set of machine learning models for binary classification. Moreover, their aggregation strategy relies on the availability of CXs, while our goal here is to define CXs for the whole set of functions.

5 Conclusions and Future Work

We introduced a novel method for addressing the challenges of counterfactual reasoning in *Multiplicity Decision Frameworks (MDFs)*. We proposed qualitative and quantitative definitions of CXs for MDFs, categorized by the degree of satisfaction, which indicates the strength of the explanation. Our framework is defined generally, supporting any instance in which model multiplicity arises and providing a path to resolving individual CXs that would otherwise conflict.

Further, we have instantiated our novel framework within abstract argumentation, obtaining the *Argumentative Multiplicity Decision Frameworks (AA-MDFs)*, that are MDFs comprising argumentative functions able to solve argumentative queries about decision problems. While our notion of AA-MDF is designed to capture a wide range of sources of multiplicity in abstract argumentation, it could be naturally extended to address diverse scenarios in which other forms of argumentation are needed. An example amounts to structured argumentation frameworks [Besnard *et al.*, 2014], where the structural composition of arguments may introduce variations. Variations can also stem from differences in preferences used during the construction of argumentation frameworks, such as in preference-based abstract argumentation for case-based reasoning [Gould *et al.*, 2024] or ASPIC$^+$ [Modgil and Prakken, 2014], amongst others. Furthermore, it accounts for distinct choices of argumentation semantics applied within the same framework, which can yield varying conclusions. Additionally, our framework could accommodate multiplicity arising from different choices of probabilities associated to argumentative structures. As future work it would be interesting to instantiate our AA-MDFs in the case of probabilistic AFs since here, as in iAFs, multiplicity arises from the different AFs referred to as possible worlds [Dung and Thang, 2010; Li *et al.*, 2011].

We also provided novel notions of CXs for incomplete AFs and AFs for case-based reasoning, as well as analyzed the complexity of decision problems related to counterfactual existence and verification. Further, we have exemplified the potential of our approach in three different real-world, high-stakes scenarios (i.e., financial, legal, and healthcare) where the validity of recourse recommendations may be critical.

An interesting direction for future research is the investigation of properties of our CXs, as done for (non-post-hoc) explanation strategies in the CA literature [Ulbricht and Wallner, 2021; Jiang *et al.*, 2024; Borg and Bex, 2024]. Moreover, we plan to undertake further analyses of MDFs in argumentation, both within the settings we have described and others, e.g. when multiplicity amounts to alternative dialectical strengths for the same arguments in Quantitative Bipolar AFs [Baroni *et al.*, 2018], due to different choices of initial intrinsic strengths for the arguments or different gradual semantics [Baroni *et al.*, 2019].

As a further line for future work, we plan to instantiate MDFs to other AI domains with pronounced multiplicity, such as ensemble models typically used in machine learning.

Acknowledgments

Alfano was supported by the PNRR MUR project PE0000013-FAIR and project Tech4You ECS0000009. Gould was supported by UK Research and Innovation [UKRI Centre for Doctoral Training in AI for Healthcare grant number EP/S023283/1]. Leofante was supported by Imperial College London through the Imperial College Research Fellowship scheme. Rago and Toni were partially funded by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 101020934) and by J.P. Morgan and the Royal Academy of Engineering under the Research Chairs and Senior Research Fellowships scheme (grant agreement no. RCSRF2021\11\45).

References

[Alfano *et al.*, 2024] Gianvincenzo Alfano, Sergio Greco, Francesco Parisi, and Irina Trubitsyna. Counterfactual and semifactual explanations in abstract argumentation: Formal foundations, complexity and computation. In *KR*, 2024.

[Alfano *et al.*, 2025] Gianvincenzo Alfano, Sergio Greco, Domenico Mandaglio, Francesco Parisi, Reza Shahbazian, and Irina Trubitsyna. Even-if explanations: Formal foundations, priorities and complexity. In *AAAI*, pages 15347–15355, 2025.

[Amgoud, 2024] Leila Amgoud. Post-hoc explanation of extension semantics. In *ECAI*, pages 3276–3283, 2024.

[Antaki and Leudar, 1992] Charles Antaki and Ivan Leudar. Explaining in conversation: Towards an argument model. *Europ. J. of Social Psychology*, 22:181—-194, 1992.

[Atkinson *et al.*, 2017] Katie Atkinson, Pietro Baroni, Massimiliano Giacomin, Anthony Hunter, Henry Prakken, Chris Reed, Guillermo Ricardo Simari, Matthias Thimm, and Serena Villata. Towards artificial argumentation. *AI Magazine*, 38(3):25–36, 2017.

[Barceló *et al.*, 2020] Pablo Barceló, Mikaël Monet, Jorge Pérez, and Bernardo Subercaseaux. Model interpretability through the lens of computational complexity. In *NeurIPS*, 2020.

[Baroni *et al.*, 2018] Pietro Baroni, Antonio Rago, and Francesca Toni. How many properties do we need for gradual argumentation? In *AAAI*, pages 1736–1743, 2018.

[Baroni *et al.*, 2019] Pietro Baroni, Antonio Rago, and Francesca Toni. From fine-grained properties to broad principles for gradual argumentation: A principled spectrum. *Int. J. Approx. Reason.*, 105:252–286, 2019.

[Baumeister *et al.*, 2018] Dorothea Baumeister, Daniel Neugebauer, Jörg Rothe, and Hilmar Schadrack. Verification in incomplete argumentation frameworks. *Artif. Intell.*, 264:1–26, 2018.

[Baumeister *et al.*, 2021] Dorothea Baumeister, Matti Järvisalo, Daniel Neugebauer, Andreas Niskanen, and Jörg Rothe. Acceptance in incomplete argumentation frameworks. *Artif. Intell.*, 295:103470, 2021.

[Bench-Capon and Dunne, 2007] T.J.M. Bench-Capon and P. E. Dunne. Argumentation in artificial intelligence. *Artif. Intell.*, 171:619 – 641, 2007.

[Besnard *et al.*, 2014] Philippe Besnard, Alejandro Javier García, Anthony Hunter, Sanjay Modgil, Henry Prakken, Guillermo Ricardo Simari, and Francesca Toni. Introduction to structured argumentation. *Argument Comput.*, 5(1):1–4, 2014.

[Black *et al.*, 2022] Emily Black, Manish Raghavan, and Solon Barocas. Model multiplicity: Opportunities, concerns, and solutions. In *FAccT*, pages 850–863, 2022.

[Borg and Bex, 2024] Annemarie Borg and Floris Bex. Minimality, necessity and sufficiency for argumentation and explanation. *Int. J. Approx. Reason.*, 168:109143, 2024.

[Brewka and Ulbricht, 2019] Gerhard Brewka and Markus Ulbricht. Strong explanations for nonmonotonic reasoning. In *LNCS*, volume 11560, pages 135–146, 2019.

[Coston *et al.*, 2021] Amanda Coston, Ashesh Rambachan, and Alexandra Chouldechova. Characterizing fairness over the set of good models under selective labels. In *ICML*, pages 2144–2155, 2021.

[Cyras *et al.*, 2016] Kristijonas Cyras, Ken Satoh, and Francesca Toni. Abstract argumentation for case-based reasoning. In *KR*, pages 549–552, 2016.

[Cyras *et al.*, 2021] Kristijonas Cyras, Antonio Rago, Emanuele Albini, Pietro Baroni, and Francesca Toni. Argumentative XAI: A survey. In *IJCAI*, pages 4392–4399, 2021.

[Dung and Thang, 2010] P. M. Dung and P. M. Thang. Towards (probabilistic) argumentation for jury-based dispute resolution. In *COMMA*, pages 171–182, 2010.

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

[Dvorák and Dunne, 2017] Wolfgang Dvorák and Paul E. Dunne. Computational problems in formal argumentation and their complexity. *FLAP*, 4(8), 2017.

[Fan and Toni, 2014] Xiuyi Fan and Francesca Toni. On computing explanations in abstract argumentation. In *ECAI*, pages 1005–1006, 2014.

[Fazzinga *et al.*, 2020] B. Fazzinga, S. Flesca, and F. Furfaro. Revisiting the notion of extension over incomplete abstract argumentation frameworks. In *IJCAI*, pages 1712–1718, 2020.

[Fisher *et al.*, 2019] Aaron Fisher, Cynthia Rudin, and Francesca Dominici. All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *J. Mach. Learn. Res.*, 20(177):1–81, 2019.

[Gould *et al.*, 2024] Adam Gould, Guilherme Paulino-Passos, Seema Dadhania, Matthew Williams, and Francesca Toni. Preference-Based Abstract Argumentation for Case-Based Reasoning. In *KR*, pages 394–404, 8 2024.

[Guidotti, 2024a] Riccardo Guidotti. Counterfactual explanations and how to find them: literature review and benchmarking. *Data Min. Knowl. Discov.*, 38(5):2770–2824, 2024.

[Guidotti, 2024b] Riccardo Guidotti. Counterfactual explanations and how to find them: literature review and benchmarking. *Data Min. Knowl. Discov.*, 38(5):2770–2824, 2024.

[Guo *et al.*, 2023] Yihang Guo, Tianyuan Yu, Liang Bai, Jun Tang, Yirun Ruan, and Yun Zhou. Argumentative explanation for deep learning: A survey. In *ICUS*, pages 1738–1743, 2023.

[Jiang *et al.*, 2024] Junqi Jiang, Francesco Leofante, Antonio Rago, and Francesca Toni. Recourse under model multiplicity via argumentative ensembling. In *AAMAS*, pages 954–963, 2024.

[Kampik *et al.*, 2024] Timotheus Kampik, Kristijonas Cyras, and José Ruiz Alarcón. Change in quantitative bipolar argumentation: Sufficient, necessary, and counterfactual explanations. *Int. J. Approx. Reason.*, 164:109066, 2024.

[Leofante *et al.*, 2023] Francesco Leofante, Elena Botoeva, and Vineet Rajani. Counterfactual explanations and model multiplicity: a relational verification view. In *KR*, pages 763–768, 2023.

[Leofante *et al.*, 2024] Francesco Leofante, Hamed Ayoobi, Adam Dejl, Gabriel Freedman, Deniz Gorur, Junqi Jiang, Guilherme Paulino-Passos, Antonio Rago, Anna Rapberger, Fabrizio Russo, Xiang Yin, Dekai Zhang, and Francesca Toni. Contestable AI Needs Computational Argumentation. In *KR*, pages 888–896, 2024.

[Li *et al.*, 2011] H. Li, N. Oren, and T. J. Norman. Probabilistic argumentation frameworks. In *TAFA*, pages 1–16, 2011.

[Marx *et al.*, 2023] Charles Marx, Youngsuk Park, Hilaf Hasson, Yuyang Wang, Stefano Ermon, and Luke Huan. But are you sure? an uncertainty-aware perspective on explainable AI. In *AISTATS*, pages 7375–7391, 2023.

[Mehrer *et al.*, 2020] Johannes Mehrer, Courtney J Spoerer, Nikolaus Kriegeskorte, and Tim C Kietzmann. Individual differences among deep neural network models. *Nature communications*, 11(1):5725, 2020.

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

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

[Mohammadi *et al.*, 2021] Kiarash Mohammadi, Amir-Hossein Karimi, Gilles Barthe, and Isabel Valera. Scaling guarantees for nearest counterfactual explanations. In *AIES*, pages 177–187, 2021.

[Rudin, 2019] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.*, 1(5):206–215, 2019.

[Saribatur *et al.*, 2020] Zeynep Gozen Saribatur, Johannes Peter Wallner, and Stefan Woltran. Explaining non-acceptability in abstract argumentation. In *Proc. of ECAI*, volume 325, pages 881–888, 2020.

[Semenova *et al.*, 2022] Lesia Semenova, Cynthia Rudin, and Ronald Parr. On the existence of simpler machine learning models. In *FAccT*, pages 1827–1858, 2022.

[Tolomei *et al.*, 2017] Gabriele Tolomei, Fabrizio Silvestri, Andrew Haines, and Mounia Lalmas. Interpretable predictions of tree-based ensembles via actionable feature tweaking. In *SIGKDD*, pages 465–474, 2017.

[Ulbricht and Wallner, 2021] Markus Ulbricht and Johannes Peter Wallner. Strong explanations in abstract argumentation. In *AAAI*, pages 6496–6504, 2021.

[Vassiliades *et al.*, 2021] Alexandros Vassiliades, Nick Bassiliades, and Theodore Patkos. Argumentation and explainable artificial intelligence: a survey. *Knowl. Eng. Rev.*, 36:e5, 2021.

[Wachter *et al.*, 2017] Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the gdpr. *Harv. JL & Tech.*, 31:841, 2017.

[Wick *et al.*, 2019] Michael L. Wick, Swetasudha Panda, and Jean-Baptiste Tristan. Unlocking fairness: a trade-off revisited. In *NeurIPS*, pages 8780–8789, 2019.

[Yin *et al.*, 2024] Xiang Yin, Nico Potyka, and Francesca Toni. Explaining arguments' strength: Unveiling the role of attacks and supports. In *IJCAI*, pages 3622–3630, 2024.