Beyond But-for Test: Counterfactual Explanation in Abstract Argumentation via Actual Causality (Extended Version)

Siyi LIU [a], Muyun SHAO [a] and Beishui LIAO [a,1]
[a] *Zhejiang University, Hangzhou, China*

Abstract. Counterfactual explanation in abstract argumentation calls for an answer to the *what-if* query: would the topic argument still be accepted if the status of certain other arguments were changed? Existing approaches are limited to the but-for test and fail to accommodate more refined counterfactual conditions. To overcome these limitations, we introduce an intervention-based counterfactual reasoning framework in abstract argumentation. Our approach encodes the acceptance conditions of arguments as equations, then defines an intervention operator that supports (1) changing sets of arguments *simultaneously*, and (2) fixing *witness* arguments to their actual labels. Guided by the refined counterfactual condition introduced in the Halpern-Pearl definition, our method goes beyond the but-for test, thereby correctly identifying causes in argumentation structures such as Preemption and Overdetermination. Through comparison, we show that our method surpasses prior methods in both expressiveness and reliability.

Keywords. Argumentation, Counterfactual, Actual Causality, Explainable AI

1. Introduction

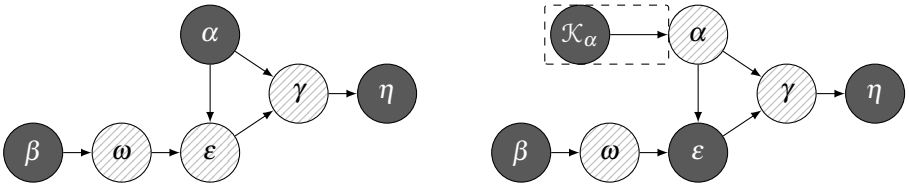
In the field of eXplainable AI (XAI), *abstract argumentation* (AA) has been shown to have advantages in providing structured and verifiable explanations [1,2,3]. AA formalises conflicts within an argumentation framework (AF) and selects acceptable arguments via semantics [4], which naturally supports answering the post-hoc explanatory question [5]: *why is an argument accepted under specific semantics?* Existing argument-based explanation methods answer this question by identifying sets of arguments that satisfy formal properties, such as Relevance [6,7], Monotonicity [8], Sufficiency and Necessity [9,10,11].

In everyday reasoning, however, individuals rarely rely on static lists of reasons; they instinctively engage in counterfactual thinking, asking "*what-if*" questions to probe causes and dependencies [12,13]. In AA terms, this becomes a direct inquiry:

> *Would the topic argument remain accepted if we were to change the acceptance status of one or more arguments suspected to be its causes?*

[1]Corresponding Author: Beishui Liao, baiseliao@zju.edu.cn

This calls for a counterfactual reasoning account in AA. Early work addresses this via the most primitive form of counterfactual reasoning, the *but-for test*, which conducts a counterfactual analysis by asking whether the effect would have occurred without the cause. In AA, this analysis is implemented by modifying the AF [14,15,16]. Sakama [14], for instance, defines two primitive updates: adding a new initial argument that attacks an accepted argument to render it rejected; and deleting all incoming attacks on a rejected argument to render it accepted. In essence, these operations ask whether changing a *single* argument's status would change the acceptance of the topic argument. In the resulting counterfactual, only the *suspected cause* is altered, with no mechanism to hold other arguments fixed as *witnesses*. Consider Preemption (Fig. 1a), which illustrates the difficulty that arises when witnesses cannot be held fixed.



(a) $\mathcal{F}_1$: Preemption (b) Modified $\mathcal{F}_1$ with α rejected by $\mathcal{K}_\alpha$

Figure 1. Counterexample to the but-for test in AA. The status of each argument is visually encoded as: **in**=dark gray (accepted), **out**=cross-hatched (rejected), **und**=light gray (undecided). The dashed box encloses the newly introduced argument and the attack relation. These visual conventions are used throughout the paper.

Example 1. *Consider the AF $\mathcal{F}_1$ in Fig. 1a, where η is the accepted argument to be explained. α and β are initial and thus accepted. From each, an even-length path reaches η. However, only the "protecting" path from α is active, since α also attacks ε, keeping it rejected and thereby blocking the defensive path from β to η.*

Intuitively, α contributes to η's acceptance, whereas β does not, leading to the conclusion that α should be qualified as a cause of η's acceptance. However, rejecting α (by adding an attacker) does not alter η's acceptance: with α out, it no longer keeps ε rejected, reactivating the defensive path from β via ω–ε–γ to η (Fig. 1b). The problem is that the counterfactual scenario does not keep ε fixed to its actual rejected status, allowing the alternative pathway to distort the counterfactual evaluation.

To address the limitations of the but-for test highlighted above, we build on the expressive framework of structural equation models (SEMs) [17] and the Halpern–Pearl (HP) definition of actual causality [18,19,20]. We encode argument acceptance via equations and define an intervention operator that supports manipulating *sets* of arguments while keeping *witness* arguments fixed. Building on this, our approach yields a refined counterfactual condition that goes beyond the but-for test and properly handles cases where the but-for test fails, namely Preemption and Overdetermination. A detailed analysis of these cases is provided in Section 4.

The structure of this paper is as follows. Section 2 recalls the necessary background. Section 3 introduces the acceptability model and the intervention operator. Section 4 defines the notion of actual cause via the refined counterfactual condition AC2^m and forms the basis for the counterfactual explanation method advocated in this paper. Section 5 presents the graphical representation for visualization. Section 6 compares prior meth-

ods. Section 7 concludes and outlines future work. All proofs of theorems and propositions are provided in the Appendix.

2. Preliminaries

An *argumentation framework* (AF) is a directed graph $\mathcal{F} = (\mathcal{A}, \mathcal{R})$, where $\mathcal{A}$ is a finite set of *arguments* and $\mathcal{R} \subseteq \mathcal{A} \times \mathcal{A}$ represents the *attack relation* [4]. The universe of AF is represented by the set $\mathcal{UF}$, while the set of arguments that appear within $\mathcal{UF}$ is denoted as $\mathcal{UA}$. For an AF $\mathcal{F} = (\mathcal{A}, \mathcal{R})$ and $\alpha, \beta \in \mathcal{A}$, define $\alpha^+ = \{\beta \in \mathcal{A} \mid (\alpha, \beta) \in \mathcal{R}\}$ and $\alpha^- = \{\beta \in \mathcal{A} \mid (\beta, \alpha) \in \mathcal{R}\}$.

Given an AF $\mathcal{F} = (\mathcal{A}, \mathcal{R})$, a set $E \subseteq \mathcal{A}$ is *conflict-free*, denoted as $E \in \mathcal{E}_{cf}(\mathcal{F})$ iff for all $\alpha, \beta \in E$, $(\alpha, \beta) \notin \mathcal{R}$. *E defends* an argument $\alpha \in \mathcal{A}$, iff for all $\beta \in \alpha^-$, there exists $\gamma \in E$, s.t. $(\gamma, \beta) \in \mathcal{R}$. The characteristic function of an AF $\mathcal{F} = (\mathcal{A}, \mathcal{R})$ is a function Γ s.t. for every $E \subseteq A$, we have $\Gamma_{\mathcal{F}}(E) = \{\alpha \in \mathcal{A} \mid E \text{ defends } \alpha\}$.

To characterize how an argument *(in)directly defends* another, we introduce the notion of directed and defense paths.

Definition 1. *Let $\mathcal{F} = (\mathcal{A}, \mathcal{R})$ be an AF, and $\alpha, \beta \in \mathcal{A}$. A sequence $P = (\alpha_0, \cdots, \alpha_n)$ is a directed path from α to β iff $\alpha_0 = \alpha$, $\alpha_n = \beta$, and $(\alpha_{i-1}, \alpha_i) \in \mathcal{R}$ for all $i \in \{1, \cdots, n\}$; its length is n (number of edges). If such a path exists, α is also said to be relevant to β (by convention, each α is relevant to itself via a path of length 0). For a set $S \subseteq \mathcal{A}$, S is relevant for α if every $\beta \in S$ is relevant to α.*

A directed path from α to β is a defense path if its length is even. Furthermore, if there exists a directed path $P = (\alpha, \gamma, \beta)$ of length two, s.t. γ is distinct from α and β, we say that α directly defends β.

The acceptability of an argument is prescribed by specific *argumentation semantics* [21]. Formally, a semantics σ assigns to each AF $\mathcal{F} = (\mathcal{A}, \mathcal{R})$ a subset of $2^{\mathcal{A}}$, denoted as $\mathcal{E}_\sigma(\mathcal{F})$. Classical semantics includes *admissible*, *complete*, *grounded*, *preferred*, and *stable* semantics (abbr. ad, co, gr, pr, st).

An extension under the semantics σ is called a σ-extension. While we focus on *complete* (*co*) semantics as our baseline, we outline in Section. 3 how the formal framework for counterfactual explanation can be adapted to other semantics. Let $\mathcal{F} = (\mathcal{A}, \mathcal{R})$ be an AF, and $E \in \mathcal{E}_{cf}(\mathcal{F})$, $E \in \mathcal{E}_{co}(\mathcal{F})$ iff E defends all its elements ($E \in \mathcal{E}_{ad}(\mathcal{F})$), and $E = \Gamma_{\mathcal{F}}(E)$.

Argumentation semantics can be defined through two parallel methods: the above *extension-based* method, and the *labelling-based* method, which formulates semantics in terms of σ-labelling [22].

Definition 2. *Given an AF $\mathcal{F} = (\mathcal{A}, \mathcal{R})$, a labelling for $\mathcal{F}$ is a total function $\mathcal{L} : \mathcal{A} \mapsto \mathbb{V}$, that assigns each argument a label from the set $\mathbb{V} = \{\mathbf{in}, \mathbf{out}, \mathbf{und}\}$, denoted as $\mathcal{L}_\sigma(\mathcal{F})$ for $\sigma \in \{ad, co, gr, pr, st\}$.*

Let $\mathbf{in}(\mathcal{L})$, $\mathbf{out}(\mathcal{L})$, and $\mathbf{und}(\mathcal{L})$ be the sets of arguments labelled **in**, **out**, and **und** respectively. We use the triple $\langle \mathbf{in}(\mathcal{L}), \mathbf{out}(\mathcal{L}), \mathbf{und}(\mathcal{L}) \rangle$ to represent the labelling $\mathcal{L}$.

A labelling $\mathcal{L}$ of $\mathcal{F} = (\mathcal{A}, \mathcal{R})$ is said to be *admissible* ($\mathcal{L} \in \mathcal{L}_{ad}(\mathcal{F})$) iff $\forall \alpha \in \mathbf{in}(\mathcal{L}) \cup \mathbf{out}(\mathcal{L})$ it holds that: (i) $\mathcal{L}(\alpha) = \mathbf{out}$ iff $\exists (\beta, \alpha) \in \mathcal{R}$ s.t. $\mathcal{L}(\beta) = \mathbf{in}$; and (ii) $\mathcal{L}(\alpha) = \mathbf{in}$ iff $\forall (\beta, \alpha) \in \mathcal{R}$, $\mathcal{L}(\beta) = \mathbf{out}$. Moreover, $\mathcal{L}$ is a *complete labelling* ($\mathcal{L} \in \mathcal{L}_{co}(\mathcal{F})$) iff conditions (i) and (ii) hold for all arguments in $\mathcal{A}$.

3. Intervention-based Counterfactual Reasoning in AA

3.1. Acceptability Model and Intervention

We start by defining a language of the logical system for reasoning about *argument-label pairs*. The *atomic propositions* are of the form $\mathbf{v}(\alpha)$, asserting that "argument α has the label $\mathbf{v}$". The set of logical connectives $\{\neg, \wedge\}$ forms an adequate set, where $\neg\mathbf{v}(\alpha)$ is interpreted as "the label of α is not $\mathbf{v}$". The language further includes *counterfactual formulas* $[\psi]\varphi$, where $[\psi]$ is an *intervention operator*, read as "after intervening to make ψ true, φ holds". Here, ψ, called a *hypothesis*, is a consistent conjunction of atomic propositions, ensured by the constraint that no argument receives two different labels.

Definition 3 (Syntax)**.** *Given a universal set of arguments $\mathcal{UA}$ and the range of label $\mathbb{V}$, with α ranging over $\mathcal{UA}$ and $\mathbf{v}$ ranging over $\mathbb{V}$, the syntax of the language $L_{\mathcal{UA}}$ is defined recursively as follows:*

$$\varphi ::= \mathbf{v}(\alpha) \mid \neg\varphi \mid \varphi \wedge \varphi \mid [\psi]\varphi,$$

where in $[\psi]\varphi$, the hypothesis ψ is defined by the grammar: $\psi ::= \top \mid \mathbf{v}(\alpha) \mid \psi \wedge \psi$, with the constraint that for any distinct atomic subformula $\mathbf{v_1}(\alpha_1), \mathbf{v_2}(\alpha_2) \in \mathsf{Sub}(\psi)$, $\alpha_1 \neq \alpha_2$, where $\mathsf{Sub}(\psi)$ denotes the set of all atomic subformulas of ψ.

A hypothesis ψ is called an *atomic hypothesis* if $\psi = \mathbf{v}(\alpha)$ for $\alpha \in \mathcal{UA}$ and $\mathbf{v} \in \mathbb{V}$; otherwise, it is a *compound hypothesis*. And for all $\mathbf{v_i}(\alpha_i) \in \mathsf{Sub}(\psi)$, denote the set of α_i by $\mathcal{A}_\psi$. We write $L^-_{\mathcal{UA}}$ for the fragment of $L_{\mathcal{UA}}$ without counterfactual formulas.

Given an AF $\mathcal{F}$, there exists a *unique acceptability model*, denoted as $\mathcal{M}_\mathcal{F}$, which is a set of *acceptability equations*. For any argument α in $\mathcal{F}$, its equation f_α is a function from the labels of α's attackers to the label of α.

Definition 4 (Acceptability Model)**.** *Given an AF $\mathcal{F} = (\mathcal{A}, \mathcal{R})$, its acceptability model $\mathcal{M}_\mathcal{F} = \{f_\alpha\}_{\alpha\in\mathcal{A}}$ is the unique set of functions, where each f_α is called an acceptability equation for α, defined as follows.*

For each $\alpha \in \mathcal{A}$. The function f_α takes as input an assignment of labels to each attacker $\beta \in \alpha^-$, denoted by $(\mathbf{v}_\beta)_{\beta\in\alpha^-}$ with $\mathbf{v}_\beta \in \mathbb{V}$, and returns a label in $\mathbb{V}$ as follows:

$$f_\alpha\big((\mathbf{v}_\beta)_{\beta\in\alpha^-}\big) = \begin{cases} \textbf{in} & \text{if } \forall\beta \in \alpha^- : \mathbf{v}_\beta = \textbf{out}, \\ \textbf{out} & \text{if } \exists\beta \in \alpha^- : \mathbf{v}_\beta = \textbf{in}, \\ \textbf{und} & \text{otherwise}. \end{cases}$$

Our framework adopts the *intervention-based semantics* of SEMs [17] to define counterfactual reasoning. An intervention modifies the model $\mathcal{M}_\mathcal{F}$ according to a hypothesis ψ, producing an *intervened model* $\mathcal{M}^\psi_\mathcal{F}$ where the acceptability equation of each argument appearing in ψ is replaced by a constant function returning its stipulated label, while all other equations remain unchanged.

Definition 5 (Intervention)**.** *Given an AF $\mathcal{F} = (\mathcal{A}, \mathcal{R})$ and its acceptability model $\mathcal{M}_\mathcal{F} = \{f_\alpha\}_{\alpha\in\mathcal{A}}$, let $\psi = \bigwedge_{i=1}^k \mathbf{v_i}(\alpha_i)$ (with $\alpha_i \in \mathcal{A}$ and $\mathbf{v_i} \in \mathbb{V}$) be a hypothesis. The intervened*

model $\mathcal{M}_{\mathcal{F}}^{\psi} = \{f_{\alpha}^{\psi}\}_{\alpha \in \mathcal{A}}$ *is defined by: for each* $\alpha \in \mathcal{A}$*, a function* f_{α}^{ψ} *on the same domain as* f_{α}*:*

$$f_{\alpha}^{\psi} = \begin{cases} \mathbf{v_i}, & \exists i \in \{1, \cdots, k\} : \alpha = \alpha_i, \\ f_{\alpha} & \textit{otherwise}. \end{cases}$$

Recalled from Def. 2, a labelling $\mathcal{L}$ is a total function that assigns a label to each argument in an AF $\mathcal{F}$. To unify non-intervened and intervened models, we have introduced $\top$ and let $\mathcal{M}_{\mathcal{F}} = \mathcal{M}_{\mathcal{F}}^{\top}$. A *solution* of an (intervened) model $\mathcal{M}_{\mathcal{F}}$ is a labelling that satisfies all its equations. Unlike classic SEMs, which are typically acyclic and yield a unique solution once the exogenous variables are fixed, an AF may contain cycles and thus admit multiple solutions, which correspond to the multiple *co*-labellings.

Definition 6 (Solution)**.** *Given an AF* $\mathcal{F} = (\mathcal{A}, \mathcal{R})$*, and its acceptability model* $\mathcal{M}_{\mathcal{F}}$*. A solution of the (intervened) model* $\mathcal{M}_{\mathcal{F}}^{\psi}$ *is a labelling* $\mathcal{L} : \mathcal{A} \mapsto \mathbb{V}$*, s.t. for all* $\alpha \in \mathcal{A}$*:*

$$\mathcal{L}(\alpha) = f_{\alpha}^{\psi}\big((\mathcal{L}(\beta))_{\beta \in \alpha^{-}}\big), \textit{ where } f_{\alpha}^{\psi} \in \mathcal{M}_{\mathcal{F}}^{\psi}.$$

Example 2. *Consider the AF* $\mathcal{F}_2$ *in Fig. 2. For* α *whose only attacker is* $\gamma \in \alpha^{-}$*, the acceptability equation* f_{α} *can be expressed as:*

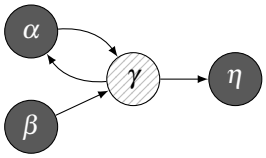


Figure 2.: $\mathcal{F}_2$ with Even-Cycle

$$f_{\alpha} = \big\{ \big((\mathbf{in}), \mathbf{out}\big),\ \big((\mathbf{out}), \mathbf{in}\big),\ \big((\mathbf{und}), \mathbf{und}\big) \big\},$$

where each pair $\big((\mathbf{v}_{\gamma}), \mathbf{v}_{\alpha}\big)$ *denotes that input label* $\mathbf{v}_{\gamma}$ *yields output label* $\mathbf{v}_{\alpha}$*.*

Intervening with the hypothesis $\psi = \mathbf{out}(\alpha)$*, the equation* f_{α} *is replaced by the constant function* f_{α}^{ψ} *that always returns* **out***. The unique solution of the intervened model* $\mathcal{M}_{\mathcal{F}_2}^{\psi}$ *is the labelling* $\mathcal{L} = \langle \{\beta, \eta\}, \{\alpha, \gamma\}, \emptyset \rangle$*.*

The set of all labellings that are solutions of $\mathcal{M}_{\mathcal{F}}^{\psi}$ is denoted as $\mathsf{Sol}(\mathcal{M}_{\mathcal{F}}^{\psi})$. Without intervention, the solutions of the model $\mathcal{M}_{\mathcal{F}}$ coincide with the *co*-labellings of the AF.

Theorem 1. *For any AF* $\mathcal{F} = (\mathcal{A}, \mathcal{R})$*, and its acceptability model* $\mathcal{M}_{\mathcal{F}}$*, it holds that* $\mathsf{Sol}(\mathcal{M}_{\mathcal{F}}) = \mathcal{L}_{co}(\mathcal{F})$*.*

The solution of a model $\mathcal{M}_{\mathcal{F}}^{\psi}$ (Def. 6) is tied to *complete* semantics but can extend to others (e.g., *gr*, *pr*, *st*) by constraining the solution set $\mathsf{Sol}(\mathcal{M}_{\mathcal{F}}^{\psi})$. For *grounded* semantics, the solution is restricted to the one where $\mathbf{in}(\mathcal{L})$ is minimal w.r.t. set-inclusion. In Ex. 2, the intervened model $\mathcal{M}_{\mathcal{F}_2}^{\psi}$ admits a unique solution $\mathcal{L}$, which strictly satisfies set-inclusion minimality and thus represents the exact grounded solution.

3.2. Semantics: Base and Model-level Satisfaction

Next, we define the semantics of this logical system. We introduce two levels of satisfaction: *base satisfaction* and *model-level satisfaction*. The base satisfaction relation $(\mathcal{M}, \mathcal{L}) \Vdash \varphi$ evaluates a formula $\varphi \in L_{\mathcal{U}\mathcal{A}}^{-}$ which is not a counterfactual formula, w.r.t. a specific solution $\mathcal{L}$ of a (possibly intervened) model $\mathcal{M}$.

Definition 7 (Base Satisfaction)**.** *Given an AF $\mathcal{F} = (\mathcal{A}, \mathcal{R})$ and its acceptability model $\mathcal{M}_\mathcal{F}$. Consider any (intervened) model $\mathcal{M} = \mathcal{M}_\mathcal{F}^\psi$ and a solution $\mathcal{L}$ of $\mathcal{M}$, the base satisfaction relation $(\mathcal{M}, \mathcal{L}) \Vdash \varphi$ for formula $\varphi \in L^-_{\mathcal{UA}}$ is defined recursively:*

$$
\begin{aligned}
(\mathcal{M}, \mathcal{L}) \Vdash \mathbf{v}(\alpha) &\iff \mathcal{L}(\alpha) = \mathbf{v}, \\
(\mathcal{M}, \mathcal{L}) \Vdash \neg\varphi &\iff (\mathcal{M}, \mathcal{L}) \nVdash \varphi, \\
(\mathcal{M}, \mathcal{L}) \Vdash \varphi \wedge \psi &\iff (\mathcal{M}, \mathcal{L}) \Vdash \varphi \text{ and } (\mathcal{M}, \mathcal{L}) \Vdash \psi.
\end{aligned}
$$

Model-level satisfaction lifts the base evaluation, considering either *all* ($\models_\forall$) or *some* ($\models_\exists$) of its solutions, and defines the meaning of counterfactual formulas.

Definition 8 (Model-level Satisfaction)**.** *Let $\varphi \in L_{\mathcal{UA}}$ be a formula, $\mathcal{F} = (\mathcal{A}, \mathcal{R})$ an AF, and $\mathcal{M} = \mathcal{M}_\mathcal{F}^\psi$ any (intervened) acceptability model.* ***Strong satisfaction*** *($\models_\forall$) and* ***weak satisfaction*** *($\models_\exists$) are defined recursively over the structure of φ:*

$$
\begin{aligned}
\text{For } \varphi \in L^-_{\mathcal{UA}}: \quad \mathcal{M} \models_\forall \varphi &\iff \forall \mathcal{L} \in \mathsf{Sol}(\mathcal{M}) : (\mathcal{M}, \mathcal{L}) \Vdash \varphi, \\
\mathcal{M} \models_\exists \varphi &\iff \exists \mathcal{L} \in \mathsf{Sol}(\mathcal{M}) : (\mathcal{M}, \mathcal{L}) \Vdash \varphi. \\
\text{For } \varphi = [\psi]\chi \text{ with } \psi, \chi \in L^-_{\mathcal{UA}}: \quad \mathcal{M} \models_\forall [\psi]\chi &\iff \mathcal{M}^\psi \models_\forall \chi, \\
\mathcal{M} \models_\exists [\psi]\chi &\iff \mathcal{M}^\psi \models_\exists \chi.
\end{aligned}
$$

Example 3 (Cont. of Ex. 2)**.** *Recall $\mathcal{F}_2$ in Fig. 2 and its unique co-labelling $\mathcal{L} \in \mathcal{L}_{co}(\mathcal{F}_2)$ depicted therein. Base satisfaction gives $(\mathcal{M}_{\mathcal{F}_2}, \mathcal{L}) \Vdash \mathbf{in}(\alpha)$. Consider two interventions: $\psi_1 = \mathbf{out}(\alpha)$ and $\psi_2 = \mathbf{out}(\beta)$.*

- *Under $\psi_1 = \mathbf{out}(\alpha)$, $\mathcal{M}_{\mathcal{F}_2}^{\psi_1}$ has the unique solution $\mathcal{L}_1 = \langle \{\beta, \eta\}, \{\alpha, \gamma\}, \emptyset \rangle$, where η labelled* **in***. Thus, $\mathcal{M}_{\mathcal{F}_2} \models_\forall [\mathbf{out}(\alpha)]\mathbf{in}(\eta) \iff \mathcal{M}_{\mathcal{F}_2}^{\mathbf{out}(\alpha)} \models_\forall \mathbf{in}(\eta)$.*
- *Under $\psi_2 = \mathbf{out}(\beta)$, one of the solutions of $\mathcal{M}_{\mathcal{F}_2}^{\psi_2}$ is $\mathcal{L}_2 = \langle \emptyset, \{\beta\}, \{\alpha, \gamma, \eta\} \rangle$, where η labelled* **und***. Thus, $\mathcal{M}_{\mathcal{F}_2} \models_\exists [\mathbf{out}(\beta)]\mathbf{out}(\eta) \iff \mathcal{M}_{\mathcal{F}_2}^{\mathbf{out}(\beta)} \models_\exists \mathbf{out}(\eta)$.*

4. Counterfactual Explanation based on Actual Causality

In general, an *explanation strategy* maps an AF $\mathcal{F}$ and a topic argument α in $\mathcal{F}$ to a set of explanations of a specific type. Here, we focus on the *formula-based* strategy. A formula-based explanation method is a function F-Expl : $\mathcal{UF} \times \mathcal{UA} \to 2^{L^-_{\mathcal{UA}}}$ associating with each AF $\mathcal{F} = (\mathcal{A}, \mathcal{R})$ and $\alpha \in \mathcal{A}$ a subset of $L^-_{\mathcal{UA}}$, denoted as F-Expl$(\mathcal{F}, \alpha)$. Each $\psi \in$ F-Expl$(\mathcal{F}, \alpha)$ is called a *formula-based explanation*.

4.1. From But-for Cause to Actual Causes

To formalise "changing the cause(s)", we introduce the set of *modified hypotheses* for the hypothesis ψ: those obtained by changing the label of at least one argument in ψ to a different label while keeping the other arguments fixed.

Definition 9 (Modified Hypotheses)**.** *Let $\psi = \bigwedge_{i=1}^{k} \mathbf{v_i}(\alpha_i)$ be a hypothesis. The set of modified hypotheses of ψ, denoted by $\mathsf{Mod}(\psi)$, comprises all $\psi' = \bigwedge_{i=1}^{k} \mathbf{v'_i}(\alpha_i)$ s.t. there exists at least one index $i \in \{1,\ldots,k\}$ with $\mathbf{v'_i} \neq \mathbf{v_i}$.*

We consider two counterfactual conditions: the but-for test BFT (underlying graph-modification, Ex. 1) and the modified condition AC2^m. A hypothesis passing BFT is called a but-for cause, and one passing AC2^m an actual cause. Both admit strong ($\forall$) and weak ($\exists$) versions, as an AF may admit multiple *co*-labellings due to cycles.

The *but-for* test (BFT) applies only to an atomic hypothesis of the form $\mathbf{v}(\alpha)$ and involves no witness. It asks whether the effect would have occurred without the cause and, in the AA setting, modifies the suspected argument's label to either **in** or **out**. We define the restricted modification set $\mathsf{Mod}_{\mathbf{in/out}}(\psi) \subseteq \mathsf{Mod}(\psi)$, which excludes modifications to **und**, following the graph-modification approach of Sakama [14].

Definition 10 (But-for Cause)**.** *Let $\mathcal{F} = (\mathcal{A}, \mathcal{R})$ be an AF, $\mathcal{M}_\mathcal{F}$ its acceptability model, $\alpha \in \mathcal{A}$ the topic, and $\mathcal{L}$ a solution of $\mathcal{M}_\mathcal{F}$. An atomic hypothesis $\psi \in L^{-}_{\mathcal{U}\mathcal{A}}$ is a but-for cause for $\mathbf{v}(\alpha)$ with $(\mathcal{M}, \mathcal{L}) \Vdash \psi \wedge \mathbf{v}(\alpha)$, iff it passes the but-for test* BFT$_q$*, formally:*

$$\exists \psi' \in \mathsf{Mod}_{\mathbf{in/out}}(\psi) : \mathcal{M}_\mathcal{F} \models_q [\psi'] \neg \mathbf{v}(\alpha) \text{ for } q \in \{\forall, \exists\}.$$

Our definition of an actual cause in AA preserves the three HP-conditions of actual causality [19, Def. 2.1]: Actuality requires the cause and effect to hold in the original labelling; Counterfactual Condition uses the modified version AC2^m adopted in this paper; [2] Minimality ensures that the cause contains no redundant parts.

Definition 11 (Actual Cause)**.** *Let $\mathcal{F} = (\mathcal{A}, \mathcal{R})$ be an AF and $\mathcal{M}_\mathcal{F}$ its acceptability model. Consider $\alpha \in \mathcal{A}$ the topic, and $\mathcal{L}$ a solution of $\mathcal{M}_\mathcal{F}$. A hypothesis $\psi \in L^{-}_{\mathcal{U}\mathcal{A}}$ is an actual cause for $\mathbf{v}(\alpha)$ iff it satisfies the following three conditions:*

Actuality (AC1**):** $(\mathcal{M}_\mathcal{F}, \mathcal{L}) \Vdash \psi \wedge \mathbf{v}(\alpha)$.
Modified Counterfactual Condition (AC2^m**):** *In Def. 12 below.*
Minimality (AC3**):** *Let $\psi = \bigwedge_{i=1}^{k} \mathbf{v_i}(\alpha_i)$. Then for every proper sub-conjunction $\psi_J = \bigwedge_{i \in J} \mathbf{v_i}(\alpha_i)$ ($J \subsetneq \{1,\ldots,k\}$), ψ_J fails to satisfy both* AC1 *and* AC2*.*

From Def. 10, a but-for cause trivially satisfies AC1 and AC3. The distinction between a but-for cause and an actual cause lies solely in the counterfactual condition.

Proposition 1. *Let $\mathcal{M}_\mathcal{F}$ be an acceptability model with a topic argument α. If hypothesis ψ is a but-for cause for $\mathbf{v}(\alpha)$, then ψ satisfies Actuality (*AC1*) and Minimality (*AC3*).*

AC2^m improves upon BFT by allowing *compound hypothesis* and keeping *witnesses* fixed to their actual labels. It validates a counterfactual if there *exist* fixed witnesses that enable the intervention to alter the topic's label. Analyzing the specific topological properties of such witnesses is left for future work.

Definition 12 (Modified Counterfactual Condition (AC2^m))**.** *Let ψ and χ be two hypotheses with $\mathcal{A}_\psi \cap \mathcal{A}_\chi = \emptyset$, $\alpha \notin \mathcal{A}_\chi$, and $(\mathcal{M}_\mathcal{F}, \mathcal{L}) \Vdash \chi$. We say ψ satisfies* AC2^m_q *iff:*

$$\exists \psi' \in \mathsf{Mod}(\psi) : \mathcal{M}_\mathcal{F} \models_q [\psi' \wedge \chi] \neg \mathbf{v}(\alpha) \text{ for } q \in \{\forall, \exists\}.$$

[2] In the original HP-definition, this condition is denoted by AC2(a^m) [19].

Example 4 (But-for vs. Actual Cause)**.** *Recall the AF $\mathcal{F}_1$ in Fig. 1a and its co-labelling $\mathcal{L}$ depicted therein, where the topic argument η is labelled* **in***. Let the suspected cause of* $\mathbf{in}(\eta)$ *be the hypothesis* $\psi = \mathbf{in}(\alpha)$.

$\mathbf{in}(\alpha)$ *fails the but-for test* $\mathrm{BFT}_{\exists}$*: after intervening with* $\mathbf{out}(\alpha)$, $\mathcal{M}_{\mathcal{F}_1}^{\mathbf{out}(\alpha)}$ *has the unique solution* $\mathcal{L}_1 = \langle\{\beta,\varepsilon,\eta\},\{\alpha,\omega,\gamma\},\emptyset\rangle$*, in which* η *remains* **in***. Hence,* $\neg\mathbf{in}(\eta)$ *does not hold, and* $\mathbf{in}(\alpha)$ *does not qualify as a but-for cause.*

Yet $\mathbf{in}(\alpha)$ *passes* $\mathrm{AC2}_{\exists}^{m}$ *with witness* $\chi = \mathbf{out}(\varepsilon)$*. Specifically, there exists a witness* $\chi = \mathbf{out}(\varepsilon)$ *fixed to its actual label in* $\mathcal{L}$*, s.t. once we intervene with* $\psi_1 = \mathbf{out}(\alpha) \in \mathsf{Mod}(\psi)$ *while holding* χ *fixed, the compound hypothesis* $\psi_2 = \mathbf{out}(\alpha) \wedge \mathbf{out}(\varepsilon)$ *is formed. The intervened model* $\mathcal{M}_{\mathcal{F}_1}^{\psi_2}$ *then yields the unique solution* $\mathcal{L}_2 = \langle\{\beta,\gamma\},\{\alpha,\omega,\varepsilon,\eta\},\emptyset\rangle$*, in which* η *is labelled* **out***. Thus* $\mathbf{in}(\alpha)$ *is identified as an actual cause of* $\mathbf{in}(\eta)$*, overcoming the counter-intuition of the but-for test illustrated in Ex. 1.*

Theorem 2 establishes that for both BFT and $\mathrm{AC2}^m$ the strong variant implies the weak variant, while Theorem 3 shows that every but-for cause is also an actual cause. Together, they characterize the logical relationships among the counterfactual conditions.

Theorem 2. *Let $\mathcal{M}_{\mathcal{F}}$ be an acceptability model with a topic argument α. For any atomic hypothesis $\psi \in L^{-}_{\mathcal{UA}}$, if ψ is a but-for cause for $\mathbf{v}(\alpha)$ under $\mathrm{BFT}_{\forall}$, then it is also a but-for cause for $\mathbf{v}(\alpha)$ under $\mathrm{BFT}_{\exists}$. For any hypothesis $\psi \in L^{-}_{\mathcal{UA}}$, if ψ is an actual cause for $\mathbf{v}(\alpha)$ under $\mathrm{AC2}_{\forall}^{m}$, then it is also an actual cause for $\mathbf{v}(\alpha)$ under $\mathrm{AC2}_{\exists}^{m}$.*

Theorem 3. *Let $\mathcal{M}_{\mathcal{F}}$ be an acceptability model with a topic argument α. For any atomic hypothesis $\psi \in L^{-}_{\mathcal{UA}}$ and $q \in \{\forall,\exists\}$, if ψ is a but-for cause for $\mathbf{v}(\alpha)$ under BFT_q, then it is also an actual cause for $\mathbf{v}(\alpha)$ under $\mathrm{AC2}_q^m$.*

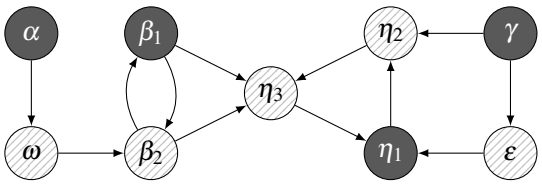


Figure 3. For Ex. 5 ($\mathcal{F}_3$)

Example 5. *Consider the AF $\mathcal{F}_3$ and its co-labelling $\mathcal{L}$ in Fig. 3. Although α, β_1, and γ each have an even-length path to η_1, not all are actual causes: α's defense is "blocked" by β_1, whereas β_1 and γ are both necessary for η_1's acceptance.*

First, test $\mathbf{in}(\gamma)$*. Intervening with* $\mathbf{out}(\gamma)$ *yields a intervened model* $\mathcal{M}_{\mathcal{F}_3}^{\mathbf{out}(\gamma)}$*, one of whose solution is* $\mathcal{L}' = \langle\{\varepsilon,\eta_2,\beta_1,\alpha\},\{\gamma,\eta_1,\eta_3,\beta_2,\omega\},\emptyset\rangle$*, where η_1's label becomes* **out***. Hence,* $\mathbf{in}(\gamma)$ *is a but-for cause (and thus also satisfies* $\mathrm{AC2}^m$*).*

Next, test $\mathbf{in}(\beta_1)$*. Under a but-for intervention* $\mathbf{out}(\beta_1)$ *alone, the unique labelling is* $\mathcal{L}'' = \langle\{\alpha,\beta_2,\eta_1,\gamma\},\{\omega,\beta_1,\eta_3,\eta_2,\varepsilon\},\emptyset\rangle$*, and η_1's label remains. The blocked path from α is now active, preserving η_1's acceptance.*

To prevent this, $\mathrm{AC2}^m$ *also fixes β_2's label to its original label* **out***. In* $\mathcal{M}_{\mathcal{F}_3}^{\mathbf{out}(\beta_1)\wedge\mathbf{out}(\beta_2)}$*, one solution is* $\mathcal{L}''' = \langle\{\alpha,\gamma\},\{\omega,\beta_1,\beta_2,\varepsilon\},\{\eta_1,\eta_2,\eta_3\}\rangle$*, where η_1 is now labelled* **und***. Thus,* $\mathbf{in}(\beta_1)$ *qualifies as an actual cause under* $\mathrm{AC2}^m$*, while the stricter but-for condition* BFT *fails.*

4.2. Counterfactual Explanation

We now instantiate the formula-based explanation strategy F-Expl with *counterfactual explanation* (Counter) and *but-for explanation* (ButFor). The latter serves as a baseline to further justify the former as the appropriate instantiation.

While the notion of actual cause (Def. 11) is independent of specific labels and accommodates any status of the topic argument (**in**, **out**, or **und**), we restrict our focus here to explaining its *acceptance* (**in**). This restriction is motivated by the fact that standard formal properties (e.g., Relevance, Existence) and the prevailing explanation methods compared in Section 6 are predominantly formulated for explaining *accepted* arguments.

We adopt AC2^m for actual causes, and call the resulting method *counterfactual explanation*. Let $\mathfrak{Act}(\mathcal{M}_\mathcal{F}, \mathcal{L}, \mathbf{in}(\alpha))$ be the set of hypotheses $\psi \in L^-_{\mathcal{U}\mathcal{A}}$ that satisfy AC2$^m_\exists$ (the strong version AC2$^m_\forall$ is left for future work) for $\mathbf{in}(\alpha)$ under the solution $\mathcal{L}$ of $\mathcal{M}_\mathcal{F}$.

Definition 13 (Counterfactual Explanation)**.** *Let $\mathcal{F} = (\mathcal{A}, \mathcal{R})$ be an AF and $\mathcal{M}_\mathcal{F}$ its acceptability model. Consider $\alpha \in \mathcal{A}$ the topic, $\mathcal{L}$ a solution of $\mathcal{M}_\mathcal{F}$ with $\mathcal{L}(\alpha) = \mathbf{in}$. Then,*

$$\text{Counter}(\mathcal{F}, \alpha) = \bigcup_{\mathcal{L} \in \mathsf{Sol}(\mathcal{M}_\mathcal{F})} \mathfrak{Act}(\mathcal{M}_\mathcal{F}, \mathcal{L}, \mathbf{in}(\alpha)).$$

A basic requirement for any explanation is that the arguments it involves be relevant (recall Def. 1) to the topic. The following proposition confirms that our counterfactual explanation method satisfies the *Relevance* property.

Proposition 2. *Given an AF $\mathcal{F} = (\mathcal{A}, \mathcal{R})$ and the topic $\alpha \in \mathcal{A}$. If the hypothesis $\psi \in$ Counter$(\mathcal{F}, \alpha)$, then for all $\alpha_i \in \mathcal{A}_\psi$, it holds that α_i is relevant to α.*

Monotonicity is a well-known property in argument-based explanation strategy [8]: if a set of arguments qualifies as an explanation, any superset should also be one. The counterfactual explanation method does not satisfy *Monotonicity*, as Minimality requires that no proper sub-conjunction of the cause satisfies Actuality (AC1) and Modified Counterfactual Condition (AC2^m).

Proposition 3. *Given an AF $\mathcal{F} = (\mathcal{A}, \mathcal{R})$ and the topic $\alpha \in \mathcal{A}$, for any hypotheses $\psi, \psi' \in L^-_{\mathcal{U}\mathcal{A}}$ with $\mathcal{A}_\psi \subsetneq \mathcal{A}_{\psi'}$, if $\psi \in$ Counter$(\mathcal{F}, \alpha)$ then $\psi' \notin$ Counter$(\mathcal{F}, \alpha)$.*

For comparison, we also define the *but-for explanation*. Let $\mathfrak{BF}(\mathcal{M}_\mathcal{F}, \mathcal{L}, \mathbf{in}(\alpha))$ denote the set of all atomic hypotheses $\psi \in L^-_{\mathcal{U}\mathcal{A}}$ that satisfy BFT$_\exists$ for $\mathbf{in}(\alpha)$ under the solution $\mathcal{L}$ of $\mathcal{M}_\mathcal{F}$. Consistent with the counterfactual explanation method, we restrict our focus to the weak version BFT$_\exists$. Then, ButFor$(\mathcal{F}, \alpha) = \bigcup_{\mathcal{L} \in \mathsf{Sol}(\mathcal{M}_\mathcal{F})} \mathfrak{BF}(\mathcal{M}_\mathcal{F}, \mathcal{L}, \mathbf{in}(\alpha))$.

The following proposition compares the two explanation methods w.r.t *Non-trivial Explanation*: the existence of an explanation that involves accepted arguments beyond the topic itself when the topic is a sink node. It shows that counterfactual explanation always provides such an explanation, whereas the but-for explanation may fail to do so.

Proposition 4. *Let $\mathcal{F} = (\mathcal{A}, \mathcal{R})$ be an AF and $\alpha \in \mathcal{A}$ the topic s.t. $\alpha^- \neq \emptyset$, $\alpha^+ = \emptyset$ and $\mathcal{L}(\alpha) = \mathbf{in}$ for some $\mathcal{L} \in \mathcal{L}_{co}(\mathcal{F})$.*

1. *For any $\mathcal{F}$ and α satisfying the above conditions, there exists $\psi \in$ Counter$(\mathcal{F}, \alpha)$, s.t. (i) $\mathcal{A}_\psi \setminus \{\alpha\} \neq \emptyset$, and (ii) for all $\beta \in \mathcal{A}_\psi$, $\mathcal{L}(\beta) = \mathbf{in}$.*

2. *There exists $\mathcal{F}$ and α satisfying the above conditions, s.t. for all $\psi = \mathbf{in}(\beta)$, if $\psi \in \mathrm{ButFor}(\mathcal{F}, \alpha)$, then $\mathcal{A}_\psi \setminus \{\alpha\} = \emptyset$.*

Example 6. *Consider the AF $\mathcal{F}_4$ in Fig. 4 and its unique co-labelling $\mathcal{L}$ depicted therein. Let the topic be η labelled* **in***. Although η is attacked by γ, it is accepted because both α and β defend it, each sufficient to protect η from γ's attack.*

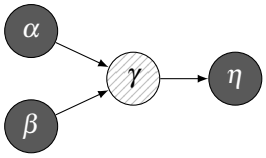


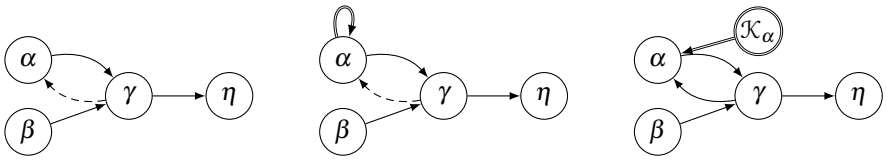
Figure 4.: $\mathcal{F}_4$: Overdetermination

Neither $\mathbf{in}(\alpha)$ nor $\mathbf{in}(\beta)$ passes the but-for test $\mathrm{BFT}_\exists$*. By contrast, the compound hypothesis $\psi = \mathbf{in}(\alpha) \wedge \mathbf{in}(\beta)$ satisfies* $\mathrm{AC2}^m_\exists$ *and therefore belongs to* $\mathrm{Counter}(\mathcal{F}_4, \eta)$*, providing a non-trivial explanation with $\mathcal{A}_\psi \setminus \{\eta\} \neq \emptyset$.*

5. Graph Mutilations of Intervention

Following Pearl's graph mutilations for interventions in causal networks [17, p. 23], we define a corresponding graph operation on AFs for each atomic hypothesis $\mathbf{v}(\alpha)$. These operations are illustrated in Fig. 5 and formalised in the following Def. 14.

Notably, this section provides a graphical counterpart of the intervention (Def. 5). The core contribution of the intervention and $\mathrm{AC2}^m$ lies in distinguishing causes from witnesses, a capability absent from earlier graph-modification approaches and essential for moving beyond the but-for test, as demonstrated in Section 4.



(a) $\mathcal{F}_{2,\mathbf{in}(\alpha)}$ (b) $\mathcal{F}_{2,\mathbf{und}(\alpha)}$ (c) $\mathcal{F}_{2,\mathbf{out}(\alpha)}$

Figure 5. Atomic Mutilation of $\mathcal{F}_2$ in Fig. 2. Symbols: dashed arrows = removed attacks, double arrows = newly introduced attacks, and double circles = newly introduced arguments.

Definition 14 (Atomic Mutilation)**.** *Given an AF $\mathcal{F} = (\mathcal{A}, \mathcal{R})$ and an atomic hypothesis $\mathbf{v}(\alpha)$ of $\mathcal{F}$, three types of atomic mutilated $\mathcal{F}$ are defined as follows:*

- $\mathcal{F}_{\mathbf{out}(\alpha)} = \big(\mathcal{A} \cup \{\mathcal{K}_\alpha\},\ \mathcal{R} \cup \{(\mathcal{K}_\alpha, \alpha)\}\big)$,
- $\mathcal{F}_{\mathbf{in}(\alpha)} = \big(\mathcal{A},\ \mathcal{R} \setminus \{(\beta, \alpha) \mid \beta \in \alpha^-\}\big)$,
- $\mathcal{F}_{\mathbf{und}(\alpha)} = \big(\mathcal{A},\ \mathcal{R} \cup \{(\alpha, \alpha)\} \setminus \{(\beta, \alpha) \mid \beta \in \alpha^-\}\big)$.

These atomic mutilations are not unique: $\mathcal{F}_{\mathbf{out}(\alpha)}$ and $\mathcal{F}_{\mathbf{in}(\alpha)}$ match Sakama [14], while $\mathcal{F}_{\mathbf{und}(\alpha)}$ follows Rienstra [15]. Unlike their metaphysical interpretations, our intervention is purely instrumental for probing dependencies, treating the graph representation merely as a technical convenience. Any admissible formulation must respect two properties:

Node Independence : In $\mathcal{M}_{\mathcal{F}}^{\mathbf{v}(\alpha)}$, the label of α no longer depends on its attackers. This is realized in the graph mutilation as follows: for $\mathcal{F}_{\mathbf{out}(\alpha)}$, adding $\mathcal{K}_\alpha$ overrides all incoming attacks; for $\mathcal{F}_{\mathbf{in}(\alpha)}$, incoming attacks are removed; for $\mathcal{F}_{\mathbf{und}(\alpha)}$, incoming attacks are removed and a self-attack is added.

Status Uniqueness : The intervention fixes the label of α to $\mathbf{v}$ in $\mathcal{M}_{\mathcal{F}}^{\mathbf{v}(\alpha)}$. The same holds in the mutilated AF.

The two properties discussed above jointly guarantee that the application of intervention hypotheses is order-invariant.

Proposition 5. *Let $\mathcal{M}_{\mathcal{F}}$ be an acceptability model and $\psi = \psi_1 \wedge \psi_2$ a compound hypothesis. Then, $\mathcal{M}_{\mathcal{F}}^{\psi} = \left(\mathcal{M}_{\mathcal{F}}^{\psi_1}\right)^{\psi_2} = \left(\mathcal{M}_{\mathcal{F}}^{\psi_2}\right)^{\psi_1}$.*

This result also holds for framework mutilation. Based on this, we define mutilation of an AF by a compound hypothesis through the recursion of atomic mutilations.

Definition 15 (Compound Mutilation)**.** *Let $\mathcal{F}$ be an AF and $\psi \in L^{-}_{\mathcal{UA}}$ a hypothesis. The compound mutilated AF $\mathcal{F}_\psi$ is defined recursively w.r.t the possible forms of ψ:*

- *If $\psi = \mathbf{v}(\alpha)$, then $\mathcal{F}_\psi = \mathcal{F}_{\mathbf{v}(\alpha)}$,*
- *If $\psi = \psi_1 \wedge \psi_2$, then $\mathcal{F}_\psi = (\mathcal{F}_{\psi_1})_{\psi_2} = (\mathcal{F}_{\psi_2})_{\psi_1}$.*

We end this section by establishing a mapping from solutions of the intervened models to labellings of the mutilated AFs.

Theorem 4. *Let $\mathcal{F} = (\mathcal{A}, \mathcal{R})$ be an AF, $\mathcal{M}_{\mathcal{F}}$ its acceptability model, and $\psi \in L^{-}_{\mathcal{UA}}$ a hypothesis. For every $\mathcal{L} \in \mathsf{Sol}(\mathcal{M}_{\mathcal{F}}^{\psi})$, there exists a labelling $\mathcal{L}'$ of the mutilated AF $\mathcal{F}_\psi$ s.t., for any argument $\alpha \in \mathcal{A}$, it holds that $\mathcal{L}(\alpha) = \mathcal{L}'(\alpha)$.*

6. Discussion

Most explanation methods in the literature are *argument-based*, defined as a function A-Expl : $\mathcal{UF} \times \mathcal{UA} \to 2^{2^{\mathcal{A}}}$ that returns sets of arguments as explanations. We recall four prominent instantiations Root_σ, Suf, Nec, and $\text{Close-Count}_\sigma$ below.

Root Reason. Liao and Van Der Torre [23, Def. 14] introduce *root reason*. Given an AF $\mathcal{F} = (\mathcal{A}, \mathcal{R})$, an extension $E \in \mathcal{E}_\sigma(\mathcal{F})$, and a topic $\alpha \in E$, let $\overline{E} = \{(\gamma, \delta) \mid \gamma, \delta \in E, \gamma \text{ directly defends } \delta\}$ with transitive closure $\overline{E}^{+}$. An argument $\beta \in E$ is *initial* if it has no attackers in $\mathcal{F}$, and *self-explanatory* if $(\beta, \beta) \in \overline{E}^{+}$ (denoted E_I and E_S). Then,

$$\text{Root}_\sigma(\mathcal{F}, \alpha) = \left\{ \{\beta \in E \mid (\beta, \alpha) \in \overline{E}^{+},\ \beta \in E_I \cup E_S\} \;\middle|\; E \in \mathcal{E}_\sigma(\mathcal{F}),\ \alpha \in E \right\}.$$

Sufficient and Necessary Explanation. Borg and Bex [9, Def. 28] formalise *sufficient and necessary* explanations in AA using the notion of relevance (Def. 1). A set $S \subseteq \mathcal{A}$ is *sufficient* for α iff S is relevant for α, conflict-free, and defends $S \cup \{\alpha\}$. An argument $\beta \in \mathcal{A}$ is *necessary* for α iff β is relevant for α and, for every admissible set $E \in \mathcal{E}_{ad}(\mathcal{F})$, $\beta \notin E$ implies $\alpha \notin E$. Then,

$$\mathrm{Suf}(\mathcal{F},\alpha) = \big\{\, S \cup \{\alpha\} \;\big|\; S \text{ is sufficient for } \alpha \,\big\},$$
$$\mathrm{Nec}(\mathcal{F},\alpha) = \big\{\, \{\alpha\} \cup \{\beta \in \mathcal{A} \mid \beta \text{ is necessary for } \alpha\} \,\big\}.$$

The next proposition relates counterfactual explanation to root reasons and sufficient explanation. When causes are restricted to initial or self-explanatory arguments, counterfactual explanation is more precise than root reasons. When restricted to accepted arguments, it is more selective than the sufficient explanation.

Proposition 6. *Let $\mathcal{F} = (\mathcal{A}, \mathcal{R})$ be an AF, $\alpha \in \mathcal{A}$ the topic, and $\mathcal{L} \in \mathcal{L}_{co}(\mathcal{F})$ a co-labelling with $\mathcal{L}(\alpha) = \mathbf{in}$. Then,*

1. *For all $\psi \in \mathrm{Counter}(\mathcal{F},\alpha)$ with $\mathcal{A}_\psi \subseteq \mathbf{in}(\mathcal{L})_I \cup \mathbf{in}(\mathcal{L})_S$ and $(\alpha_i,\alpha) \in \overline{\mathbf{in}(\mathcal{L})}^{+}$ for all $\alpha_i \in \mathcal{A}_\psi$, there exists $E \in \mathrm{Root}_{co}(\mathcal{F},\alpha)$ s.t. $\mathcal{A}_\psi \subseteq E$. The converse does not hold.*
2. *For all $\psi \in \mathrm{Counter}(\mathcal{F},\alpha)$ with $\mathcal{A}_\psi \subseteq \mathbf{in}(\mathcal{L})$, there exists $E \in \mathrm{Suf}(\mathcal{F},\alpha)$ s.t. $\mathcal{A}_\psi \subseteq E$. The converse does not hold: there exist $E \in \mathrm{Suf}(\mathcal{F},\alpha)$ for which no such ψ satisfies $\mathcal{A}_\psi \subseteq E \setminus \{\alpha\}$.*

Example 7 (Cont. of Ex. 2)**.** *Recall the AF $\mathcal{F}_2$ in Fig. 2 and its unique co-labelling $\mathcal{L}$ depicted therein, with η the topic argument.* $\mathrm{Root}_{co}(\mathcal{F}_2,\eta) = \{\{\alpha,\beta\}\}$*, since $\alpha \in \mathbf{in}(\mathcal{L})_S$ and $\beta \in \mathbf{in}(\mathcal{L})_I$, and both directly defend η.* $\mathrm{Suf}(\mathcal{F}_2,\eta) = \{\{\alpha,\beta,\eta\},\{\alpha,\eta\},\{\beta,\eta\}\}$*, each $S \setminus \{\eta\}$ for $S \in \mathrm{Suf}(\mathcal{F}_2,\eta)$ being conflict-free, relevant, and defending $S \cup \{\eta\}$.*

For counterfactual explanation, $\mathbf{in}(\beta)$ is an actual cause, whereas $\mathbf{in}(\alpha)$ fails $\mathrm{AC2}^m_\exists$*. First, after intervening with $\mathbf{out}(\alpha)$, $\mathcal{M}^{\mathbf{out}(\alpha)}_{\mathcal{F}_2}$ has the unique solution $\mathcal{L}_1 = \langle\{\beta,\eta\},\{\alpha,\gamma\},\emptyset\rangle$, where η remains $\mathbf{in}$. Second, after intervening with $\mathbf{und}(\alpha)$, $\mathcal{M}^{\mathbf{und}(\alpha)}_{\mathcal{F}_2}$ has the unique solution $\mathcal{L}_2 = \langle\{\beta,\eta\},\{\gamma\},\{\alpha\}\rangle$, with η still $\mathbf{in}$. Third, no admissible witness can rescue $\mathbf{in}(\alpha)$: in both solutions, γ retains its original label $\mathbf{out}$, so fixing any other arguments as witnesses does not alter the outcome. In contrast, $\mathcal{M}^{\mathbf{out}(\beta)}_{\mathcal{F}_2}$ admits a solution $\mathcal{L}_3 = \langle\emptyset,\{\beta\},\{\alpha,\gamma,\eta\}\rangle$, where η is $\mathbf{und}$. Hence $\mathbf{in}(\beta)$ qualifies as an actual cause. Furthermore, Minimality excludes any ψ' with $\mathcal{A}_{\psi'} \supsetneq \mathcal{A}_{\mathbf{in}(\beta)}$; hence, $\psi' \notin \mathrm{Counter}(\mathcal{F}_2,\eta)$.*

Thus, for $\psi = \mathbf{in}(\beta) \in \mathrm{Counter}(\mathcal{F}_2,\eta)$, $\mathcal{A}_\psi = \beta$ is a subset of every $E \in \mathrm{Root}_{co}(\mathcal{F}_2,\eta)$ and every $S \in \mathrm{Suf}(\mathcal{F}_2,\eta)$, yet the converse fails.

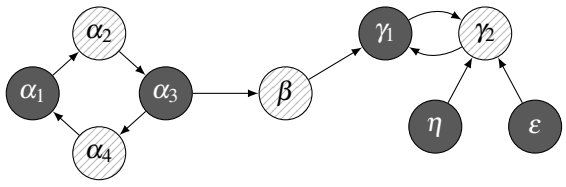


Figure 6. For Ex. 8 ($\mathcal{F}_5$)

Example 8. *Consider the AF $\mathcal{F}_5$ and the co-labelling $\mathcal{L}$ in Fig. 6, with γ_1 as the topic argument. The set $E = \{\alpha_1,\alpha_3,\gamma_1\}$ is a sufficient explanation. And the set of $\mathbf{in}$ labelled arguments $E' = \mathbf{in}(\mathcal{L})$ is a root reason. $\psi_1 = \mathbf{in}(\alpha_1)$, $\psi_2 = \mathbf{in}(\alpha_3)$, and $\psi_3 = \mathbf{in}(\eta) \wedge \mathbf{in}(\varepsilon)$ are all actual causes. Moreover, $\mathcal{A}_{\psi_i}$ is the subset of both E and E' for $i \in \{1,2,3\}$.*

Closest-World-Based Counterfactual Explanation. In contrast to our intervention-based framework, Alfano et al. [24, Def. 1] introduce an alternative counterfactual explanation rooted in the *closest-world semantics* [25]. In this approach, a labelling $\mathcal{L}$ of an AF $\mathcal{F} = (\mathcal{A}, \mathcal{R})$ is seen as the *actual world* and other σ-labellings as *possible worlds*. The similarity between two worlds is measured by the Hamming distance $\delta(\mathcal{L}, \mathcal{L}') = |\{\alpha \in \mathcal{A} \mid \mathcal{L}(\alpha) \neq \mathcal{L}'(\alpha)\}|$. A labelling $\mathcal{L}'$ is a counterfactual of $\mathcal{L}$ w.r.t. α iff:

1. $\mathcal{L}(\alpha) \neq \mathcal{L}'(\alpha)$, and
2. there exists no $\mathcal{L}'' \in \mathcal{E}_\sigma(\mathcal{F})$ s.t. $\mathcal{L}(\alpha) \neq \mathcal{L}''(\alpha)$ and $\delta(\mathcal{L}, \mathcal{L}'') < \delta(\mathcal{L}, \mathcal{L}')$.

The set of counterfactuals of $\mathcal{L}$ in an AF w.r.t. α is denoted as $\mathfrak{CF}^\sigma(\alpha, \mathcal{L})$, and each such counterfactual labelling serves as an explanation for the topic's acceptance. Then,

$$\text{Close-Count}_\sigma(\mathcal{F}, \alpha) = \left\{\mathbf{in}(\mathcal{L}') \mid \mathcal{L} \in \mathcal{E}_\sigma(\mathcal{F}), \alpha \in \mathbf{in}(\mathcal{L}), \text{and } \exists \mathcal{L}' \text{ s.t. } \mathcal{L}' \in \mathfrak{CF}^\sigma(\alpha, \mathcal{L})\right\}.$$

Following the methodology of actual causality [26], human intuitions about causes can be formalised as *structural principles* on a given graph topology. While an exhaustive principle-based analysis is beyond the scope of this discussion, Fig. 1a (Preemption) and Fig. 4 (Overdetermination) instantiate the key structural patterns behind those principles. The next example compares our method with Root$_{co}$, Suf, Nec, and Close-Count$_{co}$ on these two structures.

Example 9 (Root$_{\text{co}}$, Suf, Nec, Close-Count$_{co}$ vs. Counter)**.** *Consider again the AFs $\mathcal{F}_1$ (Preemption, Fig. 1a) and $\mathcal{F}_4$ (Overdetermination, Fig. 4), each with the unique co-labelling depicted therein. Let the topic be η in both AFs.*

Preemption ($\mathcal{F}_1$): Root$_{co}(\mathcal{F}_1, \eta) = \{\{\alpha\}\}$, $\{\alpha, \beta, \eta\} \in \text{Suf}(\mathcal{F}_1, \eta)$, $\text{Nec}(\mathcal{F}_1, \eta) = \{\{\alpha, \eta\}\}$, Close-Count$_{co}(\mathcal{F}_1, \eta) = \emptyset$, $\mathbf{in}(\alpha) \in \text{Counter}(\mathcal{F}_1, \eta)$ *while* $\mathbf{in}(\beta) \notin \text{Counter}(\mathcal{F}_1, \eta)$.

Overdetermination ($\mathcal{F}_4$): Root$_{co}(\mathcal{F}_4, \eta) = \{\{\alpha, \beta\}\}$, $\{\alpha, \eta\}, \{\beta, \eta\} \in \text{Suf}(\mathcal{F}_4, \eta)$, $\text{Nec}(\mathcal{F}_4, \eta) = \{\{\eta\}\}$, Close-Count$_{co}(\mathcal{F}_4, \eta) = \emptyset$, $\mathbf{in}(\alpha) \wedge \mathbf{in}(\beta) \in \text{Counter}(\mathcal{F}_4, \eta)$ *while* $\mathbf{in}(\alpha), \mathbf{in}(\beta) \notin \text{Counter}(\mathcal{F}_4, \eta)$.

Three observations follow.

1. Root$_{co}$ *performs well in Preemption (singling out α) but fails in Overdetermination, as it cannot distinguish between the two symmetric defenders α and β.*
2. Nec *is closer to* Counter *than* Suf *is: in $\mathcal{F}_1$,* Nec *returns $\{\alpha, \eta\}$, whose non-topic element matches the actual cause $\mathbf{in}(\alpha)$; in $\mathcal{F}_4$,* Nec *returns $\{\eta\}$, reflecting the absence of any single necessary argument, while* Counter *captures joint necessity via $\mathbf{in}(\alpha) \wedge \mathbf{in}(\beta)$. This confirms that* Counter *merely tracks necessity, not sufficiency, which is a perspective largely overlooked in prior work.*
3. Close-Count$_{co}$ *fails to satisfy* Existence*, which is a foundational principle for any explanation method. Because both $\mathcal{F}_1$ and $\mathcal{F}_4$ admit only a unique co-labelling, there are no alternative "possible worlds (counterfactual labellings)" where the status of η changes. Consequently,* Close-Count$_{co}$ *yields an empty set ($\emptyset$) for both frameworks, which directly contradicts the intuitive premise that the acceptance of η has identifiable causes.*

7. Conclusion and Future Work

We introduce an *intervention-based* counterfactual reasoning framework in abstract argumentation and develop counterfactual explanations within this setting. By fixing *witness* arguments, our method overcomes the limitations of the but-for approach. We further show that the intervention operation can be visualized through graph mutilation, thereby improving our method's explainability for human-aligned interaction. A comparison with existing argument-based explanation methods shows that our counterfactual explanation is more selective while maintaining connections to these approaches.

Several research directions follow naturally. *First*, our method and existing approaches warrant a thorough principle-based analysis. This includes different paradigms of counterfactual reasoning, such as the *non-intervention-based* counterfactual explanation of Alfano et al. [24] rooted in *closest-world* semantics, and the *post-hoc* explanation proposed by Amgoud [5]. We plan to develop *structural principles* that describe specific graph patterns, thereby examining how these methods perform under those patterns. *Second*, the condition AC2^m is rooted in *acyclic* structural equation models. Although it fits well, we will explore counterfactual conditions that are more compatible with *cyclic* argumentation frameworks. The intervention-based counterfactual reasoning framework introduced in this paper offers an expressive logical language for this endeavor. *Third*, an analysis of the computational complexity and the development of algorithmic implementations for our framework remain crucial directions for practical deployment. *Fourth*, we plan to extend the approach to other semantics, and to extended frameworks (PAF, ADF) [27,28], as well as structured frameworks (ASPIC$^+$, ABA) [29,30,31].

Acknowledgements

The authors are thankful to the anonymous reviewers for their helpful comments and suggestions. This work is supported by the National Natural Science Foundation of China (No. 62576309).

References

[1] Čyras K, Rago A, Albini E, Baroni P, Toni F. Argumentative XAI: A Survey. In: Proceedings of the 13th International Joint Conference on Artificial Intelligence (IJCAI); 2021. p. 4392-9.

[2] Engelmann D, Damasio J, Panisson AR, Mascardi V, Bordini RH. Argumentation as a Method for Explainable AI : A Systematic Literature Review. In: 2022 17th Iberian Conference on Information Systems and Technologies (CISTI); 2022. p. 1-6.

[3] Scheffers R, Bex F, Borg A. Related Explanations in Formal Argumentation, an Empirical Study. In: Proceedings of COMMA; 2024. p. 265-76.

[4] Dung PM. On the Acceptability of Arguments and Its Fundamental Role in Nonmonotonic Reasoning, Logic Programming and n-Person Games. Artificial Intelligence. 1995;77(2):321-57.

[5] Amgoud L. Post-hoc explanation of extension semantics. In: 27th European Conference on Artificial Intelligence (ECAI). vol. 392 of Frontiers in Artificial Intelligence and Applications; 2024. p. 3276-83.

[6] Fan X, Toni F. On Computing Explanations in Argumentation. In: Proceedings of the 29th AAAI Conference on Artificial Intelligence. vol. 29 of 15; 2015. p. 1496-502.

[7] Liao B, Van Der Torre L. Explanation Semantics for Abstract Argumentation. In: Proceedings of COMMA; 2020. p. 271-82.

[8] Ulbricht M, Wallner JP. Strong Explanations in Abstract Argumentation. In: Proceedings of the 35th AAAI Conference on Artificial Intelligence. vol. 35 of 7; 2021. p. 6496-504.

[9] Borg A, Bex F. Minimality, Necessity and Sufficiency for Argumentation and Explanation. International Journal of Approximate Reasoning. 2024;168:109143.

[10] Borg A, Bex F. A Basic Framework for Explanations in Argumentation. IEEE Intelligent Systems. 2021;36(2):25-35.

[11] Borg A, Bex F. Necessary and Sufficient Explanations for Argumentation-Based Conclusions. In: Symbolic and Quantitative Approaches to Reasoning with Uncertainty, ECSQARU 2021. vol. 12897. Cham: Springer; 2021. p. 45-58.

[12] Miller T. Explanation in Artificial Intelligence: Insights from the Social Sciences. Artificial Intelligence. 2019;267:1-38.

[13] Byrne RMJ. Counterfactuals in explainable artificial intelligence (XAI): Evidence from human reasoning. In: Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI); 2019. p. 6276-82.

[14] Sakama C. Counterfactual Reasoning in Argumentation Frameworks. In: Proceedings of COMMA; 2014. p. 385-96.

[15] Rienstra T. Argumentation In Flux (Modelling Change in the Theory of Argumentation) [Ph.D. thesis]. Université Montpellier II-Sciences et Techniques du Languedoc; Université du Luxembourg; 2014.

[16] Sakama C. Abduction in Argumentation Frameworks. Journal of Applied Non-Classical Logics. 2018;28(2-3):218-39.

[17] Pearl J. Causality: Models, Reasoning, and Inference. Cambridge University Press; 2009.

[18] Halpern JY, Pearl J. Causes and Explanations: A Structural-Model Approach. Part II: Explanations. The British Journal for the Philosophy of Science. 2005;56(4):889-911.

[19] Halpern JY. A modification of the Halpern-Pearl definition of causality. In: Proceedings of the 24th International Joint Conference on Artificial Intelligence (IJCAI); 2015. p. 3022-33.

[20] Halpern JY. Actual Causality. MIT Press; 2016.

[21] Baroni P, Caminada M, Giacomin M. Abstract Argumentation Frameworks and Their Semantics. In: Handbook of Formal Argumentation. 1st ed. London, England: College Publications; 2018. p. 159-236.

[22] Caminada M, Gabbay DM. A Logical Account of Formal Argumentation. Studia Logica. 2009;93(2-3):109-45.

[23] Liao B, Van Der Torre L. Attack-Defense Semantics of Argumentation. In: Proceedings of COMMA; 2024. p. 133-44.

[24] Alfano G, Greco S, Parisi F, Trubitsyna I. Counterfactual and Semifactual Explanations in Abstract Argumentation: Formal Foundations, Complexity and Computation. In: Proceedings of the 21st International Conference on Principles of Knowledge Representation and Reasoning (KR); 2024. p. 14-26.

[25] Lewis D. Counterfactuals. Blackwell Publishing; 1973.

[26] Beckers S, Vennekens J. A principled approach to defining actual causation. Synthese. 2018;195(2):835-62.

[27] Alfano G, Calautti M, Greco S, Parisi F, Trubitsyna I. Explainable Acceptance in Probabilistic Abstract Argumentation: Complexity and Approximation. In: Proceedings of the 17th International Conference on Principles of Knowledge Representation and Reasoning (KR); 2020. p. 33-43.

[28] Rienstra T, Heyninck JLA, Kern-Isberner G, Skiba K, Thimm M. Explaining Argument Acceptance in ADFs. In: 1st International Workshop on Argumentation for eXplainable AI (ArgXAI, co-located with COMMA 22). vol. 3209 of CEUR Workshop Proceedings; 2022. .

[29] Bengel L, Blümel L, Rienstra T, Thimm M. Argumentation-based Causal and Counterfactual Reasoning. In: 1st International Workshop on Argumentation for eXplainable AI (ArgXAI, co-located with COMMA 22). vol. 3209 of CEUR Workshop Proceedings; 2022. .

[30] Pisano G, Prakken H, Sartor G, Liepina R. Modelling cause-in-fact in legal cases through defeasible argumentation. In: Proceedings of the 20th International Conference on Artificial Intelligence and Law (ICAIL); 2025. p. 268-77.

[31] Bochman A, Cerutti F, Rienstra T. Causation and Argumentation. Journal of Applied Logics. 2025;12(3):713-86.

Appendix: Proofs

Proof of Theorem 1.

Proof. **(⇒) $\mathsf{Sol}(\mathcal{M}_\mathcal{F}) \subseteq \mathcal{L}_{co}(\mathcal{F})$:** Let $\mathcal{L} \in \mathsf{Sol}(\mathcal{M}_\mathcal{F})$ be a solution of the acceptability model $\mathcal{M}_\mathcal{F}$. Take an arbitrary argument $\alpha \in \mathcal{A}$. Since $\mathcal{L}$ satisfies every equation in $\mathcal{M}_\mathcal{F}$, we have $\mathcal{L}(\alpha) = f_\alpha\big((\mathcal{L}(\beta)_{\beta \in \alpha^-})\big)$, where f_α is the acceptability equation of α.

According to Def. 4, the output of the function f_α is completely determined by the labels of the attackers $\beta \in \alpha^-$. Now, examine the three possible cases for $\mathcal{L}(\alpha)$:

Case 1 If $\mathcal{L}(\alpha) = \textbf{in}$, then by the equation, every attacker $\beta \in \alpha^-$ must satisfy $\mathcal{L}(\beta) = \textbf{out}$. Hence, condition (ii) of a complete labelling holds for α.

Case 2 If $\mathcal{L}(\alpha) = \textbf{out}$, then there exists $\beta \in \alpha^-$ with $\mathcal{L}(\beta) = \textbf{in}$. Thus, condition (i) of a complete labelling holds for α.

Case 3 If $\mathcal{L}(\alpha) = \textbf{und}$, then it is neither the case that all are **out** nor that some attacker is **in**, which indicates that α must satisfy the condition (i) and (ii) of a complete labelling.

Since α was arbitrary, $\mathcal{L}$ satisfies the complete labelling conditions for every argument, i.e., $\mathcal{L} \in \mathcal{L}_{co}(\mathcal{F})$. Given that $\mathcal{L}$ is an arbitrary element of $\mathsf{Sol}(\mathcal{M}_\mathcal{F})$, we conclude that $\mathsf{Sol}(\mathcal{M}_\mathcal{F}) \subseteq \mathcal{L}_{co}(\mathcal{F})$.

(⇐) $\mathcal{L}_{co}(\mathcal{F}) \subseteq \mathsf{Sol}(\mathcal{M}_\mathcal{F})$: Let $\mathcal{L} \in \mathcal{L}_{co}(\mathcal{F})$ be a complete labelling of $\mathcal{F}$. For any $\alpha \in \mathcal{A}$, we show that $\mathcal{L}(\alpha) = f_\alpha\big((\mathcal{L}(\beta)_{\beta \in \alpha^-})\big)$:

Case 1 If $\mathcal{L}(\alpha) = \textbf{in}$, then every attacker $\beta \in \alpha^-$ is labelled **out**. According to Def. 4, f_α returns **in** in this situation.

Case 2 If $\mathcal{L}(\alpha) = \textbf{out}$, then there exists an attacker $\beta \in \alpha^-$ with $\mathcal{L}(\beta) = \textbf{in}$. By Def. 4, f_α returns **out**.

Case 3 If $\mathcal{L}(\alpha) = \textbf{und}$, then it is not true that all attackers are **out** and it is also not true that some attackers is in. Hence, by Def. 4, f_α returns **und**.

Thus, $\mathcal{L}$ satisfies the acceptability equation of α for all $\alpha \in \mathcal{A}$. Consequently, $\mathcal{L}$ is a solution of $\mathcal{M}_\mathcal{F}$, i.e., $\mathcal{L} \in \mathsf{Sol}(\mathcal{M}_\mathcal{F})$. Given that $\mathcal{L}$ is an arbitrary element of $\mathcal{L}_{co}(\mathcal{F})$, we conclude that $\mathcal{L}_{co}(\mathcal{F}) \subseteq \mathsf{Sol}(\mathcal{M}_\mathcal{F})$

Therefore, $\mathsf{Sol}(\mathcal{M}_\mathcal{F}) = \mathcal{L}_{co}(\mathcal{F})$ for every AF $\mathcal{F}$.

□

Proof of Theorem 2

Proof. Suppose that ψ is a but-for cause for $\mathbf{v}(\alpha)$ under $\mathrm{BFT}_\forall$. It is a fundamental result in abstract argumentation that every argumentation framework admits at least one complete extension (specifically, the grounded extension is always a complete extension). Consequently, the set of complete extensions is strictly non-empty. In any non-empty domain, universal quantification logically entails existential quantification. Therefore, by Def. 8, $\mathcal{M} \models_\forall \phi$ implies $\mathcal{M} \models_\exists \phi$ for any arbitrary formula ϕ. Hence, the counterfactual

condition of $\text{BFT}_\forall$ directly entails the corresponding weak version $\text{BFT}_\exists$ for the same modified hypothesis ψ'. By an identical line of reasoning, this implication strictly holds from $\text{AC2}^m_\forall$ to $\text{AC2}^m_\exists$. □

Proof of Theorem 3

Proof. Assume that ψ is an atomic hypothesis and that ψ is a but-for cause for $\mathbf{v}(\alpha)$ under BFT_q for some $q \in \{\forall, \exists\}$. By Def. 10,

$$\exists \psi' \in \mathsf{Mod}_{\mathbf{in}/\mathbf{out}}(\psi) : \mathcal{M}_\mathcal{F} \models_q [\psi']\neg\mathbf{v}(\alpha) \text{ for } q \in \{\forall, \exists\}.$$

We now verify that ψ satisfies AC2^m_q.

Choose χ to be the empty conjunction, i.e., a tautology $\top$. Trivially, $(\mathcal{M}_\mathcal{F}, \mathcal{L}) \Vdash \chi$ holds for any solution $\mathcal{L}$. Since $\psi' \wedge \chi$ is logically equivalent to ψ', the intervened model $\mathcal{M}_\mathcal{F}^{\psi' \wedge \chi}$ is identical to $\mathcal{M}_\mathcal{F}^{\psi'}$. Therefore,

$$\exists \psi' \in \mathsf{Mod}_{\mathbf{in}/\mathbf{out}}(\psi) : \mathcal{M}_\mathcal{F} \models_q [\psi' \wedge \chi]\neg\mathbf{v}(\alpha)$$

holds as well. Note that $\mathsf{Mod}_{\mathbf{in}/\mathbf{out}}(\psi)$ is a subset of $\mathsf{Mod}(\psi)$, it holds that

$$\exists \psi' \in \mathsf{Mod}(\psi) : \mathcal{M}_\mathcal{F} \models_q [\psi' \wedge \chi]\neg\mathbf{v}(\alpha).$$

Hence, ψ meets condition AC2^m_q with the same ψ' and the chosen $\chi = \top$. Conditions AC1 and AC3 are already satisfied because (1) AC1 is given in the condition of BFT_q where $(\mathcal{M}, \mathcal{L}) \Vdash \psi \wedge \mathbf{v}(\alpha)$, and (2) ψ is an atomic hypothesis, which suggests that ψ is minimal w.r.t set inclusion. Consequently, ψ is an actual cause for $\mathbf{v}(\alpha)$ under AC2^m_q. □

Proof of Theorem 4

Proof. Let $\mathcal{L} \in \mathsf{Sol}(\mathcal{M}_\mathcal{F}^\psi)$ be a solution of the intervened acceptability model $\mathcal{M}_\mathcal{F}^\psi$. Take an arbitrary argument $\alpha \in \mathcal{A}$. Since $\mathcal{L}$ satisfies every equation in $\mathcal{M}_\mathcal{F}^\psi$, we have

$$\mathcal{L}(\alpha) = f_\alpha^\psi\big((\mathcal{L}(\beta)_{\beta \in \alpha^-})\big)$$

where f_α^ψ is the intervened acceptability equation of α.

We prove that $\mathcal{L}$ satisfies the complete labelling conditions for every argument $\alpha \in \mathcal{A}$. With additional arguments such as $\mathcal{K}_\alpha$ always labeled **in**, a complete labelling $\mathcal{L}'$ of the modified AF $\mathcal{F}_\psi$ could be construct from $\mathcal{L}$.

We prove this by cases.

Case 1 For intervened arguments, which suggests there exists $\mathbf{v} \in \mathbb{V}$ such that $\mathbf{v}(\alpha) \in \mathsf{Sub}(\psi)$, therefore the equation $f_\alpha^\psi = \mathbf{v}$. Now, we examine the three possible cases in the modified AF $\mathcal{F}_\psi$ to show that in every complete labelling of $\mathcal{F}_\psi$, α is always labeled $\mathbf{v}$:

1.1 If $\mathbf{v} = \textbf{out}$, then there is an additional initial argument $\mathcal{K}_\alpha$ attacks α. Given that $\mathcal{K}_\alpha \notin \mathcal{A}$, $\mathcal{K}_\alpha$ could not be intervened according to Def. 5. So $\mathcal{K}_\alpha$ is initial in $\mathcal{F}_\psi$, and α is attacked by an initial argument $\mathcal{K}_\alpha$, so α is labeled **out** in every complete labelling of $\mathcal{F}_\psi$.

1.2 If $\mathbf{v} = \textbf{in}$, then all attacks from its parent nodes are removed from $\mathcal{F}_\psi$, which suggests α is now an initial argument in $\mathcal{F}_\psi$, this leads to the consequence that α is labeled **in** in every complete labelling of $\mathcal{F}_\psi$.

1.3 If $\mathbf{v} = \textbf{und}$, then all attacks from its parent nodes are removed from $\mathcal{F}_\psi$, and a new self-attack is introduced. In every complete labelling of $\mathcal{F}_\psi$, α could not be labeled **in**, since one of its attacker, α, is labeled **in**; α could not be labeled **out** either, because in that case all attackers of α are labeled **out**, which suggests α should be labeled **in**. So α could only be labeled **und** in every complete labelling of $\mathcal{F}_\psi$.

Case 2 For non-intervened arguments, the intervened function $f_\alpha^\psi = f_\alpha$, which suggests that the label of α is completely determined by the labels of the attackers $\beta \in \alpha^-$. Now, examine the three possible cases for $\mathcal{L}(\alpha)$:

2.1 If $\mathcal{L}(\alpha) = \textbf{in}$, then every attacker $\beta \in \alpha^-$ must satisfy $\mathcal{L}(\beta) = \textbf{out}$. Hence, condition (ii) of a complete labelling holds for α;

2.2 If $\mathcal{L}(\alpha) = \textbf{out}$, then there exists $\beta \in \alpha^-$ with $\mathcal{L}(\beta) = \textbf{in}$. Thus, condition (i) of a complete labelling holds for α;

2.3 If $\mathcal{L}(\alpha) = \textbf{und}$, then it is neither the case that all are **out** nor that some attacker is **in**, which indicates that α must satisfy the condition (i) and (ii) of a complete labelling.

Since α was arbitrary, $\mathcal{L}$ satisfies the complete labelling conditions for every argument $\alpha \in \mathcal{A}$. With the additional arguments such as $\mathcal{K}_\alpha$ labeled **in**, a complete labelling $\mathcal{L}'$ of modified AF $\mathcal{F}_\psi$ could be constructed from $\mathcal{L}$ with $\mathcal{L}(\alpha) = \mathcal{L}'(\alpha)$ holds for $\alpha \in \mathcal{A}$. □

Proof of Proposition 1

Proof. Suppose ψ is a but-for cause for $\mathbf{v}(\alpha)$. By Def. 10, this implies that $(\mathcal{M}_\mathcal{F}, \mathcal{L}) \Vdash \psi \wedge \mathbf{v}(\alpha)$. Consequently, ψ satisfies the Actuality (AC1). Furthermore, the definition of but-for cause requires ψ to be an atomic hypothesis. As an atomic hypothesis, ψ possesses no proper sub-conjunctions. Therefore, ψ satisfies the Minimality (AC3). □

Proof of Proposition 2

Proof. We proceed by case analysis on the structure of the hypothesis ψ.

Case 1 Assume ψ is an atomic hypothesis, let $\psi = \mathbf{v}'(\beta)$. Suppose, for the sake of contradiction, that the set of arguments $\mathcal{A}_\psi$ contains at least one argument that is not relevant to α. Because ψ is atomic, β is the only element that is not relevant to α. Since the label of argument is only affected by its parents' label, if changing the label of β would affect the label of α, then β should be an ancestor of α, which contradicts the fact that β is not relevant to α.

Case 2 Assume ψ is a compound hypothesis. Suppose, for the sake of contradiction, that there exists $\beta \in \mathcal{A}_\psi$ s.t. β is not relevant to α. Consider the proper sub-conjunction $\psi' = \bigwedge_{\gamma \in \mathcal{A}_\psi \setminus \{\beta\}} \mathbf{v}(\gamma)$. If ψ is an actual cause of $\mathbf{v}(\alpha)$, then ψ satisfies AC3, which suggests that the proper sub-conjunction ψ' is not an actual cause of $\mathbf{v}(\alpha)$. Given that ψ' is a proper sub-conjunction of ψ, if ψ satisfies AC1, then ψ' satisfies AC1. Consequently, the reason ψ' fails to be an actual cause is that it violates AC2. Therefore, there exists no modified hypothesis of ψ' that results in $\neg\mathbf{v}(\alpha)$. However, because ψ is an actual cause of $\mathbf{v}(\alpha)$, there exists a modified hypothesis of ψ that causes $\neg\mathbf{v}(\alpha)$. The only difference between these two hypotheses is $\mathbf{v}'(\beta)$. As established, changing the label of β can only affect the label of α if β is an ancestor of α, which contradicts the assumption that β is not relevant to α.

□

Proof of Proposition 3

Proof. We proceed by contradiction. Suppose ψ' is an actual cause of $\mathbf{v}(\alpha)$, then there exists a proper sub-conjunction ψ of ψ' such that ψ is again an actual cause of $\mathbf{v}(\alpha)$. This contradicts the fact that ψ' satisfies AC3. □

Proof of Proposition 4

Proof. This proposition consists of two statements, which we prove respectively.

Proof of 1. This is proved by constructing a hypothesis ψ that satisfies AC1 and AC2$^m_\exists$.

Given that $\alpha^- \neq \emptyset$ and $\mathcal{L}(\alpha) = \text{in}$, for all $\gamma \in \alpha^-$, it holds that $\mathcal{L}(\gamma) = \text{out}$. For each $\gamma \in \alpha^-$, there exists at least one parent $\eta \in \gamma^-$ s.t $\mathcal{L}(\eta) = \text{in}$. Let $\gamma \in \alpha^-$ be a parent of α, let $\mathcal{C} = \{\eta \mid \mathcal{L}(\eta) = \text{in}, \eta \in \gamma^-\}$ be the set and $\psi = \bigwedge_{\eta \in \mathcal{C}} \text{in}(\eta)$ be the conjunction. This construction ensures that ψ satisfies AC1.

Then we prove that ψ satisfies AC2$^m_\exists$. Let $\psi' = \bigwedge_{\eta \in \mathcal{C}} \text{out}(\eta)$. By Def. 9, ψ' is a modified hypothesis of ψ. If the model is intervened by ψ', then $\mathcal{L}(\gamma) \neq \text{out}$ since no parent of γ is labeled in. This leads to the fact that $\mathcal{L}(\alpha) \neq \text{in}$. So ψ satisfies AC2$^\text{m}_\exists$.

The constructed ψ is a finite set since the model $\mathcal{M}$ is a finite model. By checking whether there exists a proper sub-conjunction of ψ that satisfies AC1 and AC2$^m_\exists$. If not, then ψ is the actual cause; if so, then again checking whether there exists a proper sub-conjunction. Within finite steps we could end with a ψ' that is a proper sub-conjunction that satisfies AC3, and the ψ' is the actual cause.

Proof of 2. The Overdetermination presented in Ex. 6 serves as a counterexample.

□

Proof of Proposition 5

Proof. According to the restriction on hypotheses in Def. 3, for any atomic formulas $\mathbf{v_1}(\alpha_1), \mathbf{v_2}(\alpha_2) \in \text{Sub}(\psi)$, we have $\alpha_1 \neq \alpha_2$. Therefore, for any atomic formula $\mathbf{v}(\alpha) \in$

$\mathsf{Sub}(\psi)$, if $\mathbf{v}(\alpha) \in \mathsf{Sub}(\psi_1)$, then there does not exist any $\mathbf{v}' \in \mathbb{V}$ such that $\mathbf{v}'(\alpha) \in \mathsf{Sub}(\psi_2)$. This shows that for any argument α, if a conjunct ψ_1 of ψ intervenes on α_i, then there is no atomic formula related to α_i in the other conjunct ψ_2. (Denoted as Sub-conclusion 1.)

Based on this, we may assume that $\psi = \mathbf{v_1}(\alpha_1) \wedge \mathbf{v_2}(\alpha_2) \wedge \cdots \wedge \mathbf{v_n}(\alpha_n)$ is an intervention consisting of n ($n \geq 2$) atomic formulas. After an arbitrary partition, the two conjuncts are denoted as

$$\psi_1 = \mathbf{v_1}(\alpha_1) \wedge \mathbf{v_2}(\alpha_2) \wedge \cdots \wedge \mathbf{v_j}(\alpha_j),$$
$$\psi_2 = \mathbf{v_{j+1}}(\alpha_{j+1}) \wedge \cdots \wedge \mathbf{v_n}(\alpha_n),$$

where $1 \leq j < n$. For an acceptability model $\mathcal{M}_\mathcal{F} = \{f_\alpha\}_{\alpha \in \mathcal{A}}$, by Def. 5, we have

$$\mathcal{M}_\mathcal{F}^\psi = \{f_\alpha^\psi\}_{\alpha \in \mathcal{A}}, \quad f_\alpha^\psi = \begin{cases} \mathbf{v_i}, & \exists i \in \{1, \cdots, k\} : \alpha = \alpha_i; \\ f_\alpha, & \text{otherwise.} \end{cases}$$

Next, we prove that $\mathcal{M}_\mathcal{F}^\psi = (\mathcal{M}_\mathcal{F}^{\psi_1})^{\psi_2} = (\mathcal{M}_\mathcal{F}^{\psi_2})^{\psi_1}$, i.e., for any $f_\alpha \in \mathcal{M}_\mathcal{F}$,

$$f_\alpha^\psi = (f_\alpha^{\psi_1})^{\psi_2} = (f_\alpha^{\psi_2})^{\psi_1}.$$

We proceed by case analysis on an arbitrary argument $\alpha \in \mathcal{A}$.

Case 1 For an argument α that does *not* appear in ψ, i.e., there is no $\mathbf{v}$ such that $\mathbf{v}(\alpha) \in \mathsf{Sub}(\psi)$. By Def. 5, $f_\alpha^\psi = f_\alpha$. Similarly, since ψ_1 and ψ_2 are conjuncts of ψ, from $\mathbf{v}(\alpha) \notin \mathsf{Sub}(\psi)$ we obtain $\mathbf{v}(\alpha) \notin \mathsf{Sub}(\psi_1)$ and $\mathbf{v}(\alpha) \notin \mathsf{Sub}(\psi_2)$. By Def. 5, $(f_\alpha^{\psi_1})^{\psi_2} = f_\alpha^{\psi_2} = f_\alpha$. Likewise, $(f_\alpha^{\psi_2})^{\psi_1} = f_\alpha$.

Case 2 For an argument α that appears in ψ_1, i.e., there exists $\mathbf{v}$ such that $\mathbf{v}(\alpha) \in \mathsf{Sub}(\psi_1)$. By Def. 5, $f_\alpha^{\psi_1} = \mathbf{v}$. By Sub-conclusion 1, there is no $\mathbf{v}'$ such that $\mathbf{v}'(\alpha) \in \mathsf{Sub}(\psi_2)$. Hence, $(f_\alpha^{\psi_1})^{\psi_2} = f_\alpha^{\psi_1} = \mathbf{v}$. Since ψ_1 is a conjunct of ψ and $\mathbf{v}(\alpha) \in \mathsf{Sub}(\psi_1)$, we have $\mathbf{v}(\alpha) \in \mathsf{Sub}(\psi)$. Then, $f_\alpha^\psi = \mathbf{v}$. Thus, $f_\alpha^\psi = (f_\alpha^{\psi_1})^{\psi_2} = \mathbf{v}$. Similarly, $(f_\alpha^{\psi_2})^{\psi_1} = \mathbf{v}$.

Case 3 For an argument α that appears in ψ_2, i.e., there exists $\mathbf{v}$ such that $\mathbf{v}(\alpha) \in \mathsf{Sub}(\psi_2)$. By Sub-conclusion 1, there is no $\mathbf{v}'$ such that $\mathbf{v}'(\alpha) \in \mathsf{Sub}(\psi_1)$. By Def. 5, $f_\alpha^{\psi_1} = f_\alpha$. Moreover, since $\mathbf{v}(\alpha) \in \mathsf{Sub}(\psi_2)$, $(f_\alpha^{\psi_1})^{\psi_2} = f_\alpha^{\psi_2} = \mathbf{v}$. Because ψ_2 is a conjunct of ψ and $\mathbf{v}(\alpha) \in \mathsf{Sub}(\psi_2)$, we have $\mathbf{v}(\alpha) \in \mathsf{Sub}(\psi)$. Then, $f_\alpha^\psi = \mathbf{v}$. Thus, $f_\alpha^\psi = (f_\alpha^{\psi_1})^{\psi_2} = \mathbf{v}$. Similarly, $(f_\alpha^{\psi_2})^{\psi_1} = \mathbf{v}$.

In summary, for any argument α, we have

$$f_\alpha^\psi = (f_\alpha^{\psi_1})^{\psi_2} = (f_\alpha^{\psi_2})^{\psi_1}.$$

Given that $\mathcal{M}_\mathcal{F}^\psi = \{f_\alpha^\psi\}$, this results in

$$\mathcal{M}_\mathcal{F}^\psi = (\mathcal{M}_\mathcal{F}^{\psi_1})^{\psi_2} = (\mathcal{M}_\mathcal{F}^{\psi_2})^{\psi_1}.$$

□

Proof of Proposition 6

Proof. This proposition consists of two statements, which we prove respectively.

Proof of 1. Assume that $\psi \in \text{Counter}(\mathcal{F}, \alpha)$ with $\mathcal{A}_\psi \subseteq \mathbf{in}(\mathcal{L})_I \cup \mathbf{in}(\mathcal{L})_S$ and $(\alpha_i, \alpha) \in \overline{\mathbf{in}(\mathcal{L})}^+$ for all $\alpha_i \in \mathcal{A}_\psi$. There exists a complete labelling $\mathcal{L} \in \mathcal{L}_{co}(\mathcal{F})$ s.t.:

- $\mathcal{L}(\alpha) = \mathbf{in}$;
- $\mathcal{A}_\psi \subseteq E_0$ with $E_0 = \mathbf{in}(\mathcal{L})$;
- for every $\beta \in \mathcal{A}_\psi$, $\beta \in E_{0_I} \cup E_{0_S}$ and $(\beta, \alpha) \in \overline{E}^+$.

Now define:

$$R_{E_0} := \{\beta \in E_0 \mid (\beta, \alpha) \in \overline{E}_0^+ \text{ and } \beta \in E_{0_I} \cup E_{0_S}\},$$

which is exactly the root reason obtained from the complete extension E_0 according to the definition of Root Reasons. The condition above directly yields $\mathcal{A}_\psi \subseteq R_{E_0}$. Hence, the required root reason exists.

The converse fails (counterexample).

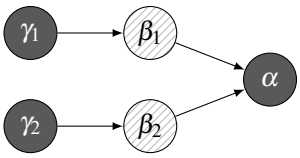


Figure 7.: $\mathcal{F}_6$: V-Shape

Consider the AF $\mathcal{F}_6$ in Fig. 7 with the topic α. The set $E = \{\gamma_1, \gamma_2\}$ is the only root reason for α, while $\psi_1 = \mathbf{in}(\gamma_1)$, $\psi_2 = \mathbf{in}(\gamma_2)$ and $\psi_3 = \mathbf{in}(\alpha)$ are the actual cause for $\mathbf{in}(\alpha)$. It is not the case that $E \subseteq \mathcal{A}_\psi$ for $i \in \{1, 2, 3\}$.

Proof of 2. Assume that $\psi \in \text{Counter}(\mathcal{F}, \alpha)$ with $\mathcal{A}_\psi \subseteq \mathbf{in}(\mathcal{L})$. There exists a complete labelling $\mathcal{L} \in \mathcal{L}_{co}(\mathcal{F})$ s.t., $\mathcal{L}(\alpha) = \mathbf{in}$ and for every $\alpha_i \in \mathcal{A}_\psi$, we have $\mathcal{L}(\alpha_i) = \mathbf{in}$. Let $E_0 \in \mathcal{E}_{co}(\mathcal{F})$ with $\mathcal{A}_\psi \subseteq E_0$ and $\alpha \in E_0$, hence E_0 is admissible and conflict-free.

Define:

$$S = \{\beta \in E_0 \mid \beta \text{ is relevant to } \alpha \text{ in } \mathcal{F}\}.$$

By Proposition 2, each argument in $\mathcal{A}_\psi$ is relevant to α; therefore, $\mathcal{A}_\psi \subseteq S$.

We now show that S is sufficient for α.

- **Relevance:** By construction, we know that every $\beta \in S$ is relevant for α.
- **Conflict-freeness:** $S \subseteq E_0$ and E_0 is conflict-free, so S is conflict-free.
- **Defence:** Take any $\beta \in S \cup \{\alpha\}$.
  - If $\beta \in S$, then $\beta \in E_0$ and E_0 is admissible. For every attacker $\delta \in \beta^-$, there exists $\gamma \in E_0$ with $(\gamma, \delta) \in \mathcal{R}$. Because β is relevant to α, there exists a directed path from β to α. Extending this path via $(\delta, \beta) \in \mathcal{R}$ and $(\gamma, \delta) \in \mathcal{R}$ yields a directed path from γ to α. Thus γ is relevant to α and consequently $\gamma \in S$.
  - If $\beta = \alpha$, the same argument applies.

Therefore, every attacker of β is counter-attacked by an argument in S, i.e., S defends $S \cup \{\alpha\}$.

Thus, S satisfied the three conditions, the set $E = S \cup \{\alpha\}$ belongs to $\mathrm{Suf}(\mathcal{F}, \alpha)$. Since $\mathcal{A}_\psi \subseteq S \subseteq E_0$, the required sufficient explanation exists.

Failure of the converse. Consider the AF in Fig. 4 with the topic η. The set $\{\alpha, \eta\} \in \mathrm{Suf}(\mathcal{F}, \eta)$; similarly, $\{\beta, \eta\} \in \mathrm{Suf}(\mathcal{F}, \eta)$. Take the sufficient explanation $E = \{\alpha, \eta\}$. There exists no $\psi \in \mathrm{Counter}(\mathcal{F}, \alpha)$ with $\mathcal{A}_\psi \subseteq \mathbf{in}(\mathcal{L})$ and $\mathcal{A}_\psi \subseteq E \setminus \{\eta\}$, because for the candidates $\psi_1 = \mathbf{in}(\alpha) \wedge \mathbf{in}(\beta)$ and $\psi_2 = \mathbf{in}(\eta)$, it holds that $\mathcal{A}_{\psi_i} \not\subseteq \{\alpha\}$ for $i \in \{1, 2\}$. Hence the inclusion $\mathcal{A}_\psi \subseteq E \setminus \{\eta\}$ is impossible for this E. This demonstrates that the converse of the statement does not hold in general.

□