Computational Models of Argument
F. Toni et al. (Eds.)

doi:10.3233/FAIA220139

Explainable Logic-Based Argumentation

Ofer ARIELI [a,1], AnneMarie BORG [b], Matthis HESSE [c], Christian STRASSER [c]
[a] *School of Computer Science, Tel-Aviv Academic College, Israel*
[b] *Dept of Information and Computing Sciences, Utrecht University, the Netherlands*
[c] *Institute for Philosophy II, Ruhr University Bochum, Germany*

Abstract. Explainable artificial intelligence (XAI) has gained increasing interest in recent years in the argumentation community. In this paper we consider this topic in the context of logic-based argumentation, showing that the latter is a particularly promising paradigm for facilitating explainable AI. In particular, we provide two representations of abductive reasoning by sequent-based argumentation frameworks and show that such frameworks successfully cope with related challenges, such as the handling of synonyms, justifications, and logical equivalences.

Keywords. Explainable AI, sequent-based argumentation, abductive logics

1. Introduction

EXplainable Artificial Intelligence (XAI) is an AI research area aimed at providing explanation to inferences and decisions made by intelligent systems [1]. *Argumentative* XAI is a fast growing area that studies XAI by means of computational argumentation (see, e.g., the recent survey papers in [2,3]).

Computational argumentation is based here on *argumentation frameworks* (AFs) [4], which are pairs of set of arguments and attack relation between the arguments, where conclusions are derived by determining subsets of arguments that can collectively be accepted in the framework. In *logic-based argumentation* [5,6] the arguments are instantiated by applying an underlying logic. Studying argumentative XAI from a logic-based perspective has several advantages. Beyond the fact that explanations in this context can be justified in a logical and rational manner, a logic-based setting is especially suitable for modeling *abductive reasoning* [7], which can be viewed as inference to the best explanation. Thus, it allows also for 'backwards reasoning', seeking for explanations for drawing conclusions from a set of observations.

In this work, we show that logic-based argumentation (and in particular sequent-based argumentation [6,8]) provides robust mechanisms for abductive reasoning in argumentative settings. In particular, we consider two ways in which abductive reasoning can be modeled by sequent-based argumentation. The first one is based on the derived argumentative conclusions, where explanations can be determined in terms of entailment relations. In the other approach, abductive reasoning is represented *within* the frameworks, where explanations are incorporated in the arguments and in the attack relations. The two approaches are then related and are used for providing information on how explanations are justified relative to the assumptions.

[1] Supported by the Israel Science Foundation (grant 550/19).

2. Preliminaries; Sequent-Based Argumentation

In this paper, we denote by $\mathfrak{L}$ a propositional language. Atomic formulas in $\mathfrak{L}$ are denoted by p,q,r, formulas are denoted by $\phi,\psi,\delta,\gamma,\varepsilon$, sets of formulas are denoted by X, S, E, and finite sets of formulas are denoted by Γ,Δ,Π,Θ, all of which can be primed or indexed. The set of atomic formulas appearing in the formulas of S is denoted $\mathsf{Atoms}(\mathsf{S})$. The set of the (well-formed) formulas of $\mathfrak{L}$ is denoted $\mathsf{WFF}(\mathfrak{L})$, the power set of $\mathsf{WFF}(\mathfrak{L})$ is denoted $\wp(\mathsf{WFF}(\mathfrak{L}))$. Sequent-based argumentation is then described as follows:

• **The base logic** is an arbitrary propositional logic, namely a pair $\mathsf{L} = \langle \mathfrak{L}, \vdash \rangle$, consisting of a language $\mathfrak{L}$ and a consequence relation $\vdash$ on $\wp(\mathsf{WFF}(\mathfrak{L})) \times \mathsf{WFF}(\mathfrak{L})$. $\vdash$ is assumed to satisfy: *reflexivity* ($\mathsf{S} \vdash \phi$ if $\phi \in \mathsf{S}$), *monotonicity* (if $\mathsf{S}' \vdash \phi$ and $\mathsf{S}' \subseteq \mathsf{S}$, then $\mathsf{S} \vdash \phi$), and *transitivity* (if $\mathsf{S} \vdash \phi$ and $\mathsf{S}', \phi \vdash \psi$ then $\mathsf{S}, \mathsf{S}' \vdash \psi$).

Let $\mathsf{L} = \langle \mathfrak{L}, \vdash \rangle$ be a logic and let S be a set of $\mathfrak{L}$-formulas. The $\vdash$-*closure of* S is the set $\mathsf{CN}_{\mathsf{L}}(\mathsf{S}) = \{\phi \mid \mathsf{S} \vdash \phi\}$. We say that S is $\vdash$-*consistent*, if there are no formulas $\phi_1,\ldots,\phi_n \in \mathsf{S}$ for which $\vdash \neg(\phi_1 \wedge \cdots \wedge \phi_n)$.

• **The language** $\mathfrak{L}$ contains at least a $\vdash$-negation operator $\neg$, satisfying $p \not\vdash \neg p$ and $\neg p \not\vdash p$ (for atomic p), and a $\vdash$-conjunction operator $\wedge$, for which $\mathsf{S} \vdash \psi \wedge \phi$ iff $\mathsf{S} \vdash \psi$ and $\mathsf{S} \vdash \phi$. We denote by $\bigwedge\Gamma$ the conjunction of all the formulas in Γ. We shall sometimes assume the availability of a deductive implication $\rightarrow$, satisfying $\mathsf{S}, \psi \vdash \phi$ iff $\mathsf{S} \vdash \psi \rightarrow \phi$.

• **Arguments** based on a logic $\mathsf{L} = \langle \mathfrak{L}, \vdash \rangle$ are single-conclusioned L-*sequents* [9], namely: expressions of the form $\Gamma \Rightarrow \psi$, where $\Rightarrow$ is a symbol that does not appear in $\mathfrak{L}$, and such that $\Gamma \vdash \psi$. Γ is called the argument's support (also denoted $\mathsf{Supp}(\Gamma \Rightarrow \psi)$) and ψ is the argument's conclusion (denoted $\mathsf{Conc}(\Gamma \Rightarrow \psi)$). Given a set S of $\mathfrak{L}$-formulas (premises), an S-based argument is an L-argument $\Gamma \Rightarrow \psi$, where $\Gamma \subseteq \mathsf{S}$. We denote by $\mathrm{Arg}_{\mathsf{L}}(\mathsf{S})$ the set of all the L-arguments that are based on S.

We distinguish between two types of non-intersecting premises: a $\vdash$-consistent set X of strict (i.e., non-attacked) premises, and a set S of defeasible premises. Their non-defeasible character will give them a special status when we define argumentative attacks below. We write $\mathrm{Arg}^{\mathsf{X}}_{\mathsf{L}}(\mathsf{S})$ for the set $\mathrm{Arg}_{\mathsf{L}}(\mathsf{X} \cup \mathsf{S})$. In particular, $\mathrm{Arg}^{\emptyset}_{\mathsf{L}}(\mathsf{S}) = \mathrm{Arg}_{\mathsf{L}}(\mathsf{S})$.

• **Attack rules** are sequent-based inference rules for representing attacks between sequents. Such rules consist of an attacking argument (the first condition of the rule), an attacked argument (the last condition of the rule), conditions for the attack (the other conditions of the rule) and a conclusion (the eliminated attacked sequent). The outcome of an application of such a rule is that the attacked sequent is 'eliminated' (or 'invalidated'; see below the exact meaning of this). The elimination of $\Gamma \Rightarrow \phi$ is denoted by $\Gamma \not\Rightarrow \phi$.

Given a set X of strict (non-attacked) formulas, some common attack rules are:

- Defeat: $\dfrac{\Gamma_1 \Rightarrow \psi_1 \quad \psi_1 \Rightarrow \neg\bigwedge\Gamma_2 \quad \Gamma_2, \Gamma_2' \Rightarrow \psi_2}{\Gamma_2, \Gamma_2' \not\Rightarrow \psi_2} \quad (\Gamma_2 \neq \emptyset,\ \Gamma_2 \cap \mathsf{X} = \emptyset)$

- Direct Defeat: $\dfrac{\Gamma_1 \Rightarrow \psi_1 \quad \psi_1 \Rightarrow \neg\gamma \quad \Gamma_2, \gamma \Rightarrow \psi_2}{\Gamma_2, \gamma \not\Rightarrow \psi_2} \quad (\gamma \notin \mathsf{X})$

- Undercut: $\dfrac{\Gamma_1 \Rightarrow \psi_1 \quad \psi_1 \Rightarrow \neg\bigwedge\Gamma_2 \quad \neg\bigwedge\Gamma_2 \Rightarrow \psi_1 \quad \Gamma_2, \Gamma_2' \Rightarrow \psi_2}{\Gamma_2, \Gamma_2' \not\Rightarrow \psi_2} \quad (\Gamma_2 \neq \emptyset,\ \Gamma_2 \cap \mathsf{X} = \emptyset)$

- Direct Undercut: $\dfrac{\Gamma_1 \Rightarrow \psi_1 \quad \psi_1 \Rightarrow \neg\gamma \quad \neg\gamma \Rightarrow \psi_1 \quad \Gamma_2, \gamma \Rightarrow \psi_2}{\Gamma_2, \gamma \not\Rightarrow \psi_2}$ $(\gamma \notin \mathsf{X})$
- Consistency Undercut: $\dfrac{\Gamma_1 \Rightarrow \neg \bigwedge \Gamma_2 \quad \Gamma_2, \Gamma_2' \Rightarrow \psi}{\Gamma_2, \Gamma_2' \not\Rightarrow \psi}$ $(\Gamma_2 \neq \emptyset, \Gamma_2 \cap \mathsf{X} = \emptyset, \Gamma_1 \subseteq \mathsf{X})$

For instance, in the particular case where $\Gamma_1 = \emptyset$, consistency undercut indicates that an argument with an inconsistent support is eliminated.

• **A (sequent-based) argumentation framework** (AF), based on the logic L and the attack rules in AR, for a set of defeasible premises S and a $\vdash$-consistent set of strict premises X, is a pair $\mathbb{AF}^{\mathsf{X}}_{\mathsf{L,AR}}(\mathsf{S}) = \langle \mathrm{Arg}^{\mathsf{X}}_{\mathsf{L}}(\mathsf{S}), \mathsf{A} \rangle$ where $\mathsf{A} \subseteq \mathrm{Arg}^{\mathsf{X}}_{\mathsf{L}}(\mathsf{S}) \times \mathrm{Arg}^{\mathsf{X}}_{\mathsf{L}}(\mathsf{S})$ and $(a_1, a_2) \in \mathsf{A}$ iff there is a rule $\mathsf{R}_{\mathsf{X}} \in \mathsf{AR}$, such that a_1 R_{X}-attacks a_2. In what follows we shall use AR and A interchangeably, denoting both of them by A.

• **Semantics** of sequent-based frameworks are defined as usual by Dung-style extensions [4]: Let $\mathbb{AF} = \mathbb{AF}^{\mathsf{X}}_{\mathsf{L,A}}(\mathsf{S}) = \langle \mathrm{Arg}^{\mathsf{X}}_{\mathsf{L}}(\mathsf{S}), \mathsf{A} \rangle$ be an argumentation framework and let $\mathbb{E} \subseteq \mathrm{Arg}^{\mathsf{X}}_{\mathsf{L}}(\mathsf{S})$ be a set of arguments. It is said that: $\mathbb{E}$ *attacks* a if there is an $a' \in \mathbb{E}$ such that $(a', a) \in \mathsf{A}$, $\mathbb{E}$ *defends* a if $\mathbb{E}$ attacks every attacker of a, and $\mathbb{E}$ is *conflict-free* (cf) if for no $a_1, a_2 \in \mathbb{E}$ it holds that $(a_1, a_2) \in \mathsf{A}$. We say that $\mathbb{E}$ is *admissible* if it is conflict-free and defends all of its elements. A *complete (*cmp*) extension* of $\mathbb{AF}$ is an admissible set that contains all the arguments that it defends. By this, various argumentative semantics may be defined. For instance, the *grounded (*grd*) extension* of $\mathbb{AF}$ is the $\subseteq$-minimal complete extension of $\mathrm{Arg}^{\mathsf{X}}_{\mathsf{L}}(\mathsf{S})$, a *preferred (*$\mathsf{prf}$*) extension* of $\mathbb{AF}$ is a $\subseteq$-maximal complete extension of $\mathrm{Arg}^{\mathsf{X}}_{\mathsf{L}}(\mathsf{S})$, and a *stable (*$\mathsf{stb}$*) extension* of $\mathbb{AF}$ is a conflict-free set in $\mathrm{Arg}^{\mathsf{X}}_{\mathsf{L}}(\mathsf{S})$ that attacks every argument not in it.[2] We denote by $\mathsf{Ext}_{\mathsf{sem}}(\mathbb{AF})$ the set of all the extensions of $\mathbb{AF}$ of type sem.

• **Entailments** induced from an argumentation framework $\mathbb{AF} = \mathbb{AF}^{\mathsf{X}}_{\mathsf{L,A}}(\mathsf{S}) = \langle \mathrm{Arg}^{\mathsf{X}}_{\mathsf{L}}(\mathsf{S}), \mathsf{A} \rangle$ are based on the extensions derived from $\mathbb{AF}$ under a semantics sem:

- *Skeptical entailment:* $\mathsf{S} \mid\!\sim^{\cap,\mathsf{sem}}_{\mathsf{L,A,X}} \phi$ if there is an argument $a \in \bigcap \mathsf{Ext}_{\mathsf{sem}}(\mathbb{AF})$ such that $\mathsf{Conc}(a) = \phi$.
- *Weakly skeptical entailment:* $\mathsf{S} \mid\!\sim^{\Cap,\mathsf{sem}}_{\mathsf{L,A,X}} \phi$ if for every extension $\mathbb{E} \in \mathsf{Ext}_{\mathsf{sem}}(\mathbb{AF})$ there is an argument $a \in \mathbb{E}$ such that $\mathsf{Conc}(a) = \phi$.
- *Credulous entailment:* $\mathsf{S} \mid\!\sim^{\cup,\mathsf{sem}}_{\mathsf{L,A,X}} \phi$ iff there is an argument $a \in \bigcup \mathsf{Ext}_{\mathsf{sem}}(\mathbb{AF})$ such that $\mathsf{Conc}(a) = \phi$.

Example 1. Consider an AF, based on classical logic CL and the following set of defeasible assumptions:

$$\mathsf{S} = \left\{ \begin{array}{l} \mathtt{clear\_skies},\ \mathtt{rainy},\ \mathtt{clear\_skies} \to \neg\mathtt{rainy},\ \mathtt{rainy} \to \neg\mathtt{sprinklers}, \\ \mathtt{rainy} \to \mathtt{wet\_grass},\ \mathtt{sprinklers} \to \mathtt{wet\_grass} \end{array} \right\}$$

Suppose further that there are no strict assumptions ($\mathsf{X} = \emptyset$) and that the only attack rule is undercut (Ucut). Then, for instance, the arguments

$$\begin{array}{l} a_1 :\ \mathtt{clear\_skies}, \mathtt{clear\_skies} \to \neg\mathtt{rainy} \Rightarrow \neg\mathtt{rainy}, \\ a_2 :\ \mathtt{rainy}, \mathtt{clear\_skies} \to \neg\mathtt{rainy} \Rightarrow \neg\mathtt{clear\_skies} \end{array}$$

[2] Further extensions and the relations among them are discussed e.g. in [10].

Ucut-attack each other. In this case there are two stable/preferred extensions $\mathbb{E}_1$ and $\mathbb{E}_2$, where $a_1 \in \mathbb{E}_1$ and $a_2 \in \mathbb{E}_2$. It follows, for instance, that with respect to these semantics, `wet_grass` credulously follows from the framework (since, e.g. `rainy`, `rainy` $\rightarrow$ `wet_grass` $\Rightarrow$ `wet_grass` is in $\mathbb{E}_2$), but it does follow skeptically (since there is no argument in $\mathbb{E}_1$ whose conclusion is `wet_grass`).

3. Abductive Reasoning in Sequent-Based Frameworks

Abductive reasoning is a common method of providing explanations in logic-based contexts. Sequent-based formalisms are particularly adequate for this, as instead of the usual understanding of a sequent $\Gamma, \Delta \Rightarrow \phi$ by 'ϕ is a conclusion of $\Gamma \cup \Delta$', one may intuitively read it as 'Δ is a (prima facia) explanation of ϕ in the presence of Γ'. This kind of 'backward reasoning' is also our starting point for showing the usefulness of sequent-based frameworks for abductive reasoning. We then proceed in two directions, external and internal ones, for defining abductive reasoning in sequent-based argumentation.

3.1. Explanations: External View

We start with an 'external' approach, which is based on argumentative entailment relations. Let $\mathsf{L} = \langle \mathfrak{L}, \vdash \rangle$ be a logic and $\mid\!\sim$ a non-monotonic entailment induced by it.[3] Given sets of strict (X) and defeasible (S) assumptions, an *explanation* E of an *explanandum* ϕ with respect to $\mid\!\sim$, is a finite set that satisfies at least the following two properties:

Sufficiency **(w.r.t.** $\mid\!\sim$**):** $\mathsf{X}, \mathsf{S}, \mathsf{E} \mid\!\sim \phi$
Consistency **(w.r.t.** $\vdash$**):** $\mathsf{X} \nvdash \neg \bigwedge \mathsf{E}$

Thus, the set of explanations should be $\vdash$-consistent with the strict assumptions, and together with the strict and defeasible assumptions they are sufficient for $\mid\!\sim$-inferring the explanandum ϕ. We call these two conditions the *basic explanation properties*.

The basic explanation properties per-se may sometimes be too weak, and so they are usually accompanied with further conditions. The following ones are inspired by [11]:

Non-vacuity **(w.r.t.** $\vdash$**):** $\mathsf{E} \nvdash \phi$
Minimality **(w.r.t.** $\mid\!\sim$**):** $\mathsf{S}, \mathsf{E}' \not\mid\!\sim \phi$ for every E' for which $\mathsf{E}, \mathsf{X} \vdash \bigwedge \mathsf{E}'$ and $\mathsf{E}', \mathsf{X} \nvdash \bigwedge \mathsf{E}$.

Non-vacuity prevents self-explanations, and minimality assures the conciseness of the explanations. In order to make sure that the explanation is indeed necessary (i.e., the explanandum cannot be inferred from the assumptions alone), the property of non-idleness ($\mathsf{X}, \mathsf{S} \nvdash \phi$) or strict non-idleness ($\mathsf{X} \nvdash \phi$) may be required. Here it will be convenient to use the following argumentative variations of this property:

Non-idleness **(w.r.t.** sem**):** there is no $a \in \bigcup \mathsf{Ext}_{\mathsf{sem}}(\mathbb{AF}^{\mathsf{X}}_{\mathsf{L,A}}(\mathsf{S}))$ s.t. $\mathsf{Conc}(a) = \phi$.
Strict non-idleness **(w.r.t.** sem**):** there is no $a \in \bigcup \mathsf{Ext}_{\mathsf{sem}}(\mathbb{AF}^{\mathsf{X}}_{\mathsf{L,A}}(\emptyset))$ s.t. $\mathsf{Conc}(a) = \phi$.

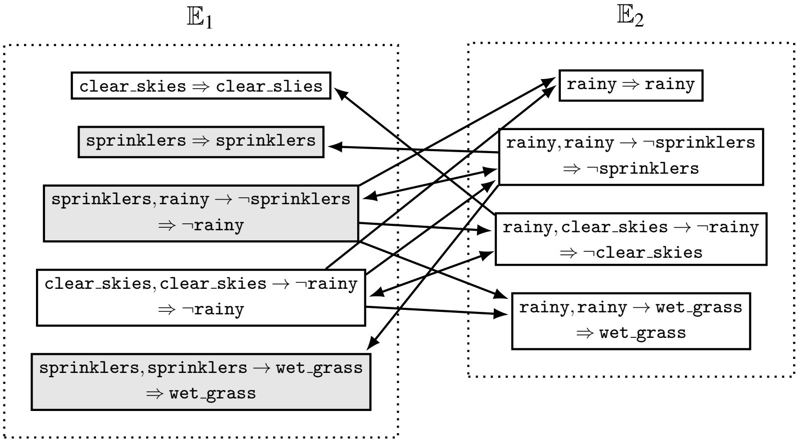
By the above principles, external argumentative explanations are defined as follows:

[3] In our case, $\mid\!\sim$ is the entailment induced from a framework that is based on L.

Definition 1. Given a framework $\mathbb{AF} = \mathbb{AF}^{\mathsf{X}}_{\mathsf{L},\mathsf{A}}(\mathsf{S})$ based on a logic $\mathsf{L} = \langle \mathfrak{L}, \vdash \rangle$, a finite set E of $\mathfrak{L}$-formulas is called:

- *external skeptical* sem-*explanation* of ϕ if it satisfies $\mid\!\sim^{\cap,\mathsf{sem}}_{\mathsf{L},\mathsf{A},\mathsf{X}}$-sufficiency ($\mathsf{X}, \mathsf{S}, \mathsf{E} \mid\!\sim^{\cap,\mathsf{sem}}_{\mathsf{L},\mathsf{A},\mathsf{X}} \phi$), $\vdash$-consistency ($\mathsf{X} \nvdash \neg \bigwedge \mathsf{E}$), and holds in every sem-extension: for every $\mathbb{E} \in \mathsf{Ext}_{\mathsf{sem}}(\mathbb{AF}^{\mathsf{X}}_{\mathsf{L},\mathsf{A}}(\mathsf{S} \cup \mathsf{E}))$ there is $a \in \mathbb{E}$, such that $\mathsf{Conc}(a) = \bigwedge \mathsf{E}$.
- *external weakly-skeptical* sem-*explanation* of ϕ if it satisfies $\mid\!\sim^{\Cap,\mathsf{sem}}_{\mathsf{L},\mathsf{A},\mathsf{X}}$-sufficiency ($\mathsf{X}, \mathsf{S}, \mathsf{E} \mid\!\sim^{\Cap,\mathsf{sem}}_{\mathsf{L},\mathsf{A},\mathsf{X}} \phi$), $\vdash$-consistency ($\mathsf{X} \nvdash \neg \bigwedge \mathsf{E}$), and holds in every sem-extension: for every $\mathbb{E} \in \mathsf{Ext}_{\mathsf{sem}}(\mathbb{AF}^{\mathsf{X}}_{\mathsf{L},\mathsf{A}}(\mathsf{S} \cup \mathsf{E}))$ there is $a \in \mathbb{E}$, such that $\mathsf{Conc}(a) = \bigwedge \mathsf{E}$.
- *external credulous* sem-*explanation* of ϕ if it satisfies $\mid\!\sim^{\cup,\mathsf{sem}}_{\mathsf{L},\mathsf{A},\mathsf{X}}$-sufficiency ($\mathsf{X}, \mathsf{S}, \mathsf{E} \mid\!\sim^{\cup,\mathsf{sem}}_{\mathsf{L},\mathsf{A},\mathsf{X}} \phi$), $\vdash$-consistency ($\mathsf{X} \nvdash \neg \bigwedge \mathsf{E}$), and holds in some sem-extension: there is some $\mathbb{E} \in \mathsf{Ext}_{\mathsf{sem}}(\mathbb{AF}^{\mathsf{X}}_{\mathsf{L},\mathsf{A}}(\mathsf{S} \cup \mathsf{E}))$ and $a \in \mathbb{E}$, such that $\mathsf{Conc}(a) = \bigwedge \mathsf{E}$.

Example 2. Consider again the framework in Example 1. Note that $\mathsf{E} = \{\mathtt{sprinklers}\}$ is a (stable and preferred) credulous explanation for `wet_grass`. Indeed, using the notations of Example 1, the framework that is based on $\mathsf{S} \cup \mathsf{E}$ has two stable/preferred extensions: $\mathbb{E}'_1$ and $\mathbb{E}'_2 = \mathbb{E}_2$ (see Figure 1). In $\mathbb{E}'_1$ the grass is wet since the sprinklers are activated, and in $\mathbb{E}'_2$ the grass is wet since it rains.



$\mathbb{E}_1$
$\mathbb{E}_2$
clear_skies ⇒ clear_slies
sprinklers ⇒ sprinklers
sprinklers, rainy → ¬sprinklers ⇒ ¬rainy
clear_skies, clear_skies → ¬rainy ⇒ ¬rainy
sprinklers, sprinklers → wet_grass ⇒ wet_grass
rainy ⇒ rainy
rainy, rainy → ¬sprinklers ⇒ ¬sprinklers
rainy, clear_skies → ¬rainy ⇒ ¬clear_skies
rainy, rainy → wet_grass ⇒ wet_grass

Figure 1. Part of the AF of Example 2. The arguments with dark background are added by the explanation.

3.2. *Explanations: Internal View*

We now turn to the 'internal' approach, where abductive explanations are handled by ingredients of the framework. We do so by considering another type of sequents, called 'abductive sequents'. These are expressions of the form $\phi \Leftarrow \Gamma, [\varepsilon]$,[4] and it intuitively means that 'ϕ may be inferred from Γ under the assumption that ε holds'. Note that while $\Gamma \subseteq \mathsf{S} \cup \mathsf{X}$, ε may not be an assumption, but rather a hypothetical explanation of the conclusion.

[4] Note the reverse direction of the sequent sign, to emphasize the backward inference in this case.

Abductive sequents may be produced by the following rule that roughly models the usual idea of abductive inference as backwards reasoning:

$$\frac{\varepsilon, \Gamma \Rightarrow \phi}{\phi \Leftarrow \Gamma, [\varepsilon]} \ \text{(Abduction)}$$

In our running example, this rule will allow us to produce abductive sequents such as

$$\texttt{wet\_grass} \Leftarrow [\texttt{sprinklers}], \texttt{sprinklers} \rightarrow \texttt{wet\_grass}$$

that provides an alternative explanation to the wetness of the grass (i.e., `sprinklers`, in addition to `rainy`), or

$$\neg\texttt{rainy} \Leftarrow [\texttt{sprinklers}], \texttt{rainy} \rightarrow \neg\texttt{sprinklers}$$

that provides another possible evidence for refuting the defeasible assumption that it is rainy (i.e., `sprinklers`, in addition to the assumption that the sky is clear).

Since abductive reasoning is a form of non-monotonic reasoning, which in logic-based argumentation is modeled with the attack relations, we need a way to attack abductive sequents. To this end, we consider rules similar to those from Section 2, e.g.:

$$\frac{\Gamma_1 \Rightarrow \phi_1 \quad \phi_1 \Rightarrow \neg\gamma \quad \phi_2 \Leftarrow [\varepsilon], \Gamma_2}{\phi_2 \not\Leftarrow [\varepsilon], \Gamma_2} \ \gamma \in (\Gamma_2 \cup \{\varepsilon\}) \setminus \mathsf{X} \ \text{(Abductive Direct Defeat)}$$

which models an attack on a subset of the assumptions and a hypothetical explanation of an abductive sequent. Note that this attack rule assures, in particular, the consistency of explanations with the strict assumptions, thus it renders the following rule admissible:

$$\frac{\Gamma_1 \Rightarrow \neg\varepsilon \quad \phi \Leftarrow [\varepsilon], \Gamma_2}{\phi \not\Leftarrow [\varepsilon], \Gamma_2} \ \Gamma_1 \subseteq \mathsf{X} \ \text{(Consistency)}$$

Abductive reasoning has to fulfill certain requirements to ensure proper behavior also in the internal view. This time, attack rules may be introduced for obtaining counterparts of the properties discussed in Section 3.1 for the external view. Note that, since abductive sequents are now derived according to the underlying sequent calculus and the abduction rule introduced above, the sufficiency property is automatically satisfied. Attack rules for the other properties are given next.

Non-vacuity Rules for preventing self-explanations:

$$\frac{\vdash \varepsilon \rightarrow \phi \quad \phi \Leftarrow [\varepsilon]}{\phi \not\Leftarrow [\varepsilon]} \ \text{(Non Vacuity)}$$

Thus, in our running example, $\texttt{wet\_grass} \Leftarrow [\texttt{wet\_grass}]$ is excluded.

Minimality Rules for assuring that explanations will be as general as possible.

$$\frac{\phi \Leftarrow [\varepsilon_1], \Gamma_1 \quad \vdash \varepsilon_2 \rightarrow \varepsilon_1 \quad \nvdash \varepsilon_1 \rightarrow \varepsilon_2 \quad \phi \Leftarrow [\varepsilon_2], \Gamma_2}{\phi \not\Leftarrow [\varepsilon_2], \Gamma_2} \ \text{(Minimality)}$$

This rule assures that in our example $\mathtt{sprinklers} \wedge \mathtt{irrelevant\_fact}$ should not explain `wet_grass`, since `sprinklers` is a more general and so more relevant explanation.

Non-Idleness The [strict] assumptions should not already explain the explanandum.

$$\frac{\Gamma_1 \Rightarrow \phi \qquad \phi \Leftarrow [\varepsilon], \Gamma_2}{\phi \not\Leftarrow [\varepsilon], \Gamma_2} \text{ (Defeasible Non Idleness)}$$

$$\frac{\Gamma_1 \Rightarrow \phi \qquad \phi \Leftarrow [\varepsilon], \Gamma_2}{\phi \not\Leftarrow [\varepsilon], \Gamma_2} \;\; \Gamma_1 \subseteq \mathsf{X} \text{ (Strict Non Idleness)}$$

Note that defeasible non-idleness excludes the explanation `sprinklers` for `wet_grass`, since the latter is already inferred from the defeasible assumptions (assuming that it is rainy), while strict non-idleness will allow this alternative explanation (since `wet_grass` cannot be inferred from the strict assumptions).

The next step is to adapt sequent-based argumentation frameworks to an abductive setting, using abductive sequents, the new inference rule, and additional attack rules. Given a sequent-based framework $\mathbb{AF}^{\mathsf{X}}_{\mathsf{L,A}}(\mathsf{S})$, an *abductive sequent-based framework* $\mathbb{AAF}^{\mathsf{X}}_{\mathsf{L,A^\star}}(\mathsf{S})$ is constructed by adding to the arguments in $\mathrm{Arg}^{\mathsf{X}}_{\mathsf{L}}(\mathsf{S})$ also abductive arguments, produced by Abduction, and where $\mathsf{A}^\star$ is obtained by adding to the attack rules in A also (some of) the rules for maintaining explanations that are described above. Explanations according to the internal view are then defined as follows:

Definition 2. Given an abductive sequent-based framework $\mathbb{AAF}^{\mathsf{X}}_{\mathsf{L,A^\star}}(\mathsf{S})$ as described above, a finite set E of $\mathfrak{L}$-formulas is called:

- *internal skeptical* sem-*explanation* of ϕ, if there is $\Gamma \subseteq \mathsf{S}$ such that the abductive argument $\phi \Leftarrow [\bigwedge \mathsf{E}], \Gamma$ is in every sem-extension of $\mathbb{AAF}^{\mathsf{X}}_{\mathsf{L,A^\star}}(\mathsf{S})$.
- *internal weakly-skeptical* sem-*explanation* of ϕ, if in every sem-extension of $\mathbb{AAF}^{\mathsf{X}}_{\mathsf{L,A^\star}}(\mathsf{S})$ there is an abductive argument $\phi \Leftarrow [\bigwedge \mathsf{E}], \Gamma$ for some $\Gamma \subseteq \mathsf{S}$
- *internal credulous* sem-*explanation* of ϕ, if there is $\Gamma \subseteq \mathsf{S}$ such that the abductive argument $\phi \Leftarrow [\bigwedge \mathsf{E}], \Gamma$ is in some sem-extension of $\mathbb{AAF}^{\mathsf{X}}_{\mathsf{L,A^\star}}(\mathsf{S})$.

Example 3. As noted above, $\mathtt{wet\_grass} \Leftarrow [\mathtt{sprinklers}], \mathtt{sprinklers} \rightarrow \mathtt{wet\_grass}$ is producible by the Abduction rule from the sequent-based framework in Example 1, and belongs to a stable/preferred extension of the corresponding abductive sequent-based framework. Therefore, `sprinklers` credulously stb/prf-explains `wet_grass` also according to Definition 2.

3.3. Explanations: Relating the Two Views

Next, we relate the two approaches for producing argumentative explanations by abductive reasoning in sequent-based frameworks. In what follows we restrict ourselves to singleton explanations in the assumptions language.[5] We consider $\{\text{ConUcut}\} \subset \mathsf{A} \subseteq \{\text{ConUcut}, \text{DirectDefeat}, \text{DirectUndercut}\}$. The main results are the following:

[5] Thus, using the notations of the previous sections, $\mathsf{E} = \{\varepsilon\}$, where $\mathsf{Atoms}(\varepsilon) \subseteq \mathsf{Atoms}(\mathsf{S} \cup \mathsf{X})$.

Theorem 1. *Let* $\mathbb{AF} = \mathbb{AF}^{\mathsf{X}}_{\mathsf{L,A}}(\mathsf{S})$ *where* $\mathsf{L} = \mathsf{CL}$*,* A *is as specified above, and* $\mathbb{AAF} = \mathbb{AAF}^{\mathsf{X}}_{\mathsf{L,A^\star}}(\mathsf{S})$ *where* $\mathsf{A}^\star = \mathsf{A} \cup \{\text{Abductive Direct Defeat}\}$*. For* $\mathsf{sem} \in \{\mathsf{stb},\mathsf{prf}\}$*,* E *is an external weakly skeptical (resp. skeptical)* sem*-explanation of* ϕ *w.r.t.* $\mathbb{AF}$ *iff* E *is an internal weakly skeptical (resp. skeptical)* sem*-explanation of* ϕ *w.r.t.* $\mathbb{AAF}$*. Moreover,* E *satisfies non-vacuity and/or strict non-idleness iff the non-vacuity and/or the strict non-idleness attack rule is added to* $\mathsf{A}^\star$*.*

Theorem 2. *Let* $\mathbb{AF} = \mathbb{AF}^{\mathsf{X}}_{\mathsf{L,A}}(\mathsf{S})$ *where* $\mathsf{L} = \mathsf{CL}$*,* A *is as specified above, and* $\mathbb{AAF} = \mathbb{AAF}^{\mathsf{X}}_{\mathsf{L,A^\star}}(\mathsf{S})$ *where* $\mathsf{A}^\star = \mathsf{A} \cup \{\text{Abductive Direct Defeat}\}$*. Then* E *is an external weakly skeptical (resp. skeptical)* grd*-explanation of* ϕ *w.r.t.* $\mathbb{AF}$ *iff* E *is an internal weakly skeptical (resp. skeptical)* grd*-explanation of* ϕ *w.r.t.* $\mathbb{AAF}$*. Moreover,* E *satisfies non-vacuity and/or strict non-idleness iff the non-vacuity and/or the strict non-idleness attack rule is added to* $\mathsf{A}^\star$*.*

The proofs of Theorems 1 and 2 are based on the correspondence to reasoning with maximally consistent sets of assumptions, shown in [12]. Next, we sketch the proof of Theorem 1 for $\mathsf{A}^\star = \mathsf{A} \cup \{\text{Abductive Direct Defeat}\}$ and the weakly skeptical version (the proof for the skeptical version and the proof of Theorem 2 are similar). In the proof, $\mathsf{MCS}^{\mathsf{X}}_{\mathsf{L}}(\mathsf{S})$ is the set of the maximally $\vdash$-consistent subsets of S, which are also $\vdash$-consistent with X.

Proof outline of Theorem 1. [$\Rightarrow$] Suppose that $\mathsf{E} = \{\varepsilon\}$ is an external weakly skeptical sem-explanation of ϕ w.r.t. $\mathbb{AF}^{\mathsf{X}}_{\mathsf{L,A}}(\mathsf{S})$, where $\mathsf{sem} \in \{\mathsf{stb},\mathsf{prf}\}$. In particular, $\mathsf{S},\varepsilon \mathrel{|\!\sim}^{\Cap,\mathsf{sem}}_{\mathsf{L,A,X}} \phi$, and for every $\mathbb{E} \in \mathsf{Ext}_{\mathsf{sem}}(\mathbb{AF}^{\mathsf{X}}_{\mathsf{L,A}}(\mathsf{S} \cup \{\varepsilon\})$ there is $a \in \mathbb{E}$, such that $\mathsf{Conc}(a) = \varepsilon$. By [12, Theorem 1], (†) for all $\Delta \in \mathsf{MCS}^{\mathsf{X}}_{\mathsf{L}}(\mathsf{S} \cup \{\varepsilon\})$ we have that $\mathsf{X},\Delta \vdash \phi$ and $\mathsf{X},\Delta \vdash \varepsilon$.

Let now $\mathbb{E} \in \mathsf{Ext}_{\mathsf{sem}}(\mathbb{AAF}^{\mathsf{X}}_{\mathsf{L,A^\star}}(\mathsf{S}))$. Then $\mathbb{E} \cap \mathsf{Arg}^{\mathsf{X}}_{\mathsf{L}}(\mathsf{S}) \in \mathsf{Ext}_{\mathsf{sem}}(\mathbb{AF}^{\mathsf{X}}_{\mathsf{L,A}}(\mathsf{S}))$, and so, by [12, Theorem 1] again, $\mathbb{E} \cap \mathsf{Arg}^{\mathsf{X}}_{\mathsf{L}}(\mathsf{S}) = \mathsf{Arg}^{\mathsf{X}}_{\mathsf{L}}(\Delta)$ for some $\Delta \in \mathsf{MCS}^{\mathsf{X}}_{\mathsf{L}}(\mathsf{S})$. By (†), for all $\Omega \in \mathsf{MCS}^{\mathsf{X}}_{\mathsf{L}}(\mathsf{S})$, $\Omega,\mathsf{X} \nvdash \neg\varepsilon$. So, $\mathsf{X},\Delta \nvdash \neg\varepsilon$. Thus $\Delta \cup \{\varepsilon\} \in \mathsf{MCS}^{\mathsf{X}}_{\mathsf{L}}(\mathsf{S} \cup \{\varepsilon\})$. By (†), there is some finite $\Gamma \subseteq \Delta \setminus \{\varepsilon\}$, for which $\mathsf{X},\Gamma,\varepsilon \vdash \phi$. It follows that $\phi \Leftarrow [\varepsilon],\Gamma$ is an abductive argument in $\mathbb{AAF}^{\mathsf{X}}_{\mathsf{L,A^\star}}(\mathsf{S})$.

Note that $\mathsf{X},\Delta \nvdash \neg\gamma$ for all $\gamma \in (\Gamma \cup \{\varepsilon\}) \setminus \mathsf{X}$, otherwise $\mathsf{X},\Delta \vdash \neg\varepsilon$, in a contradiction to (†) and the consistency of $\Gamma \subseteq \Delta$. Thus $\phi \Leftarrow [\varepsilon],\Gamma$ is not abductively attacked by any element of $\mathbb{E}$, and so $\phi \Leftarrow [\varepsilon],\Gamma \in \mathbb{E}$. It follows that ε is an internal weakly skeptical sem-explanation of ϕ w.r.t. $\mathbb{AAF}^{\mathsf{X}}_{\mathsf{L,A^\star}}(\mathsf{S})$.

[$\Leftarrow$] Suppose that $\mathsf{E} = \{\varepsilon\}$ is an internal weakly skeptical sem-explanation of ϕ w.r.t. $\mathbb{AAF}^{\mathsf{X}}_{\mathsf{L,A^\star}}(\mathsf{S})$. Let $\mathbb{E} \in \mathsf{Ext}_{\mathsf{sem}}(\mathsf{Arg}^{\mathsf{X}}_{\mathsf{L}}(\mathsf{S} \cup \{\varepsilon\}))$. By [12, Theorem 1], $\mathbb{E} = \mathsf{Arg}^{\mathsf{X}}_{\mathsf{L}}(\Delta)$ for some $\Delta \in \mathsf{MCS}^{\mathsf{X}}_{\mathsf{L}}(\mathsf{S} \cup \{\varepsilon\})$. Then $\mathsf{X},\Delta \nvdash \neg\varepsilon$, and so $\Delta' = \Delta \cap \mathsf{S} \in \mathsf{MCS}^{\mathsf{X}}_{\mathsf{L}}(\mathsf{S})$. Let $\mathbb{E}'$ be the set of all the $(\mathsf{X} \cup \Delta)$-based sequents and $(\mathsf{X} \cup \Delta)$-based abducitive sequents. It can be shown that $\mathbb{E}' \in \mathsf{Ext}_{\mathsf{sem}}(\mathbb{AAF}^{\mathsf{X}}_{\mathsf{L,A^\star}}(\mathsf{S}))$. Thus, there is an $\phi \Leftarrow [\varepsilon],\Gamma \in \mathbb{E}$, and $\Gamma,\varepsilon \vdash \phi$ (for $\Gamma \subseteq \Delta \cup \mathsf{X}$). Thus, $\mathsf{X},\Delta \vdash \phi$ and $\mathsf{X},\Delta \vdash \varepsilon$. It follows that ε is an external weakly skeptical sem-explanation of ϕ w.r.t. $\mathbb{AF}^{\mathsf{X}}_{\mathsf{L,A}}(\mathsf{S})$. □

We note that not in all cases the external and internal explanations coincide, even when $\mathsf{L} = \mathsf{CL}$ and $\mathsf{A} = \{\text{Direct Defeat}, \text{ConUcut}\}$. The next example illustrates this:

Example 4. Let $\mathsf{L} = \mathsf{CL}$, $\mathsf{A} = \{\text{Direct Defeat}, \text{ConUcut}\}$, $\mathsf{S} = \{p, \neg p \wedge q\}$ and $\mathsf{X} = \{q \wedge r \to s\}$. Then:

1. $q \wedge r$ is an external weakly-skeptical stb-explanation of s, since the corresponding sequent-based framework has two stable extensions: $\mathrm{Arg}_{\mathsf{L}}^{\mathsf{X}}(\{p, q \wedge r\})$ and $\mathrm{Arg}_{\mathsf{L}}^{\mathsf{X}}(\{\neg p \wedge q, q \wedge r\})$, both of which contain arguments for $q \wedge r$ and for $q \wedge r \to s$. Note that this explanation satisfies Non-vacuity (s does not follow from $q \wedge r$).
2. $q \wedge r$ is an internal weakly-skeptical stb-explanation of s, since the corresponding abductive sequent framework also has two stable extensions, both with the abducible sequent $s \Leftarrow [q \wedge r], q \wedge r \to s$.This holds also when the non-vacuity and/or strict non-idleness attack rules are part of the framework.

This is in accordance with Theorem 1. Suppose now that minimality is imposed. Then:

1. $q \wedge r$ remains an external weakly-skeptical stb-explanation of s, since it satisfies the minimality condition.
2. $q \wedge r$ is *no longer* an internal weakly-skeptical stb-explanation of s, since one extension also contains a minimality attacker of $s \Leftarrow [q \wedge r], q \wedge r \to s$, namely: $s \Leftarrow [r], \neg p \wedge q, q \wedge r \to s$.

4. Some Further Considerations

In this section we briefly comment on some other aspects of argumentation explanation.

4.1. Handling of Synonyms and Antonyms

Synonyms and antonyms may be handled by the strict assumptions, as they should not be revised. This may be done either to clarify the meaning of some terminology used by defeasible formulas, or for extending the vocabulary describing the domain of discourse. For instance, suppose that in our running example we add the strict assumption $\mathsf{X} = \{\mathtt{blue\_skies} \leftrightarrow \mathtt{clear\_skies}\}$. Then, since

$$\mathtt{blue\_skies}, \mathtt{blue\_skies} \leftrightarrow \mathtt{clear\_skies}, \mathtt{clear\_skies} \to \neg\mathtt{rainy} \vdash \neg\mathtt{rainy}$$

we derive, by the Abduction rule, the abductive sequent

$$\neg\mathtt{rainy} \Leftarrow [\mathtt{blue\_skies}], \mathtt{blue\_skies} \leftrightarrow \mathtt{clear\_skies}, \mathtt{clear\_skies} \to \neg\mathtt{rainy}$$

Thus, under stable or preferred semantics, $\mathtt{blue\_skies}$ explains $\neg\mathtt{rainy}$. Similarly, $\mathtt{blue\_skies}$ explains $\neg\mathtt{wet\_grass}$, etc.

4.2. Keeping Track of Explanations; Explanations Justifications

In the context of defeasible reasoning explanatory arguments are threatened by defeaters. While abductive sequents $\phi \Leftarrow [\varepsilon], \Gamma$ state that in the context Γ the explanandum ϕ is deducible from the explanation ε, it contains no information of how this explanation is justified against the background of possible defeaters. In the terminology of argumentation theory, abductive sequents cover the illative tier (support) but not the dialectic tier (defeating defeaters) of argumentation [13,14]. In order to keep track of the latter, we incorporate some ideas in the spirit of [15], adapted to logic-based argumentation in general and abductive argumentation frameworks in particular.

Let $\mathbb{AAF}^{\mathsf{X}}_{\mathsf{L},\mathsf{A}^\star}(\mathsf{S})$ be an abductive sequent-based framework with a set $\mathsf{Arg}^{\mathsf{X}}_{\mathsf{L}}(\mathsf{S})$ of ordinary and abductive arguments, and a set A of attack rules on $\mathsf{Arg}^{\mathsf{X}}_{\mathsf{L}}(\mathsf{S}) \times \mathsf{Arg}^{\mathsf{X}}_{\mathsf{L}}(\mathsf{S})$. For a semantics sem and operator $\Box \in \{\cup, \cap\}$, we consider the following sets:

- $\mathsf{AbdArg}(\phi, [\varepsilon]) = \{a \in \mathsf{Arg}^{\mathsf{X}}_{\mathsf{L}}(\mathsf{S}) \mid a \text{ is of the form } \phi \Leftarrow \Gamma, [\varepsilon] \text{ for some } \Gamma \subseteq \mathsf{S}\}$
- $\mathsf{AbdArg}^{\Box}_{\mathsf{sem}}(\phi, [\varepsilon]) = \{a \in \mathsf{AbdArg}(\phi, [\varepsilon]) \mid a \in \Box\mathsf{Ext}_{\mathsf{sem}}(\mathbb{AAF}^{\mathsf{X}}_{\mathsf{L},\mathsf{A}^\star}(\mathsf{S}))\}$

Thus, $\mathsf{AbdArg}^{\Box}_{\mathsf{sem}}(\phi, [\varepsilon])$ consists of all the abductive arguments in which ε explains ϕ (namely, the elements of $\mathsf{AbdArg}(\phi, [\varepsilon])$), and that belong to the intersection (if $\Box = \cap$) or the union (if $\Box = \cup$) of all the sem-extensions of $\mathbb{AAF}^{\mathsf{X}}_{\mathsf{L},\mathsf{A}^\star}(\mathsf{S})$.

To justify the explanation of ϕ by ε with respect to sem and $\Box$, we therefore need to compute the supports of the arguments that defend the elements in $\mathsf{AbdArg}^{\Box}_{\mathsf{sem}}(\phi, [\varepsilon])$ (divided by sem-extensions)

- $\mathsf{Def}_{\mathbb{E}}(a) = \{\mathsf{Supp}(b) \mid b \in \mathbb{E}, b \text{ defends } a \text{ in } \mathbb{AAF}^{\mathsf{X}}_{\mathsf{L},\mathsf{A}^\star}(\mathsf{S})\}$
- $\mathsf{Justify}^{\Box}_{\mathsf{sem}}(\phi, [\varepsilon]) = \{\mathsf{Def}_{\mathbb{E}}(a) \mid a \in \mathsf{AbdArg}^{\Box}_{\mathsf{sem}}(\phi, [\varepsilon]), \mathbb{E} \in \mathsf{Ext}_{\mathsf{sem}}(\mathbb{AAF}^{\mathsf{X}}_{\mathsf{L},\mathsf{A}^\star}(\mathsf{S}))\}$

Example 5. Suppose that we want to justify the comment in Example 3 that `sprinklers` credulously stb-explains `wet_grass`. For this, note that:

1. The abductive sequent $a = \texttt{wet\_grass} \Leftarrow [\texttt{sprinklers}], \texttt{sprinklers} \to \texttt{wet\_grass}$ is in $\mathsf{AbdArg}(\texttt{wet\_grass}, [\texttt{sprinklers}])$ and $\mathsf{AbdArg}^{\cup}_{\mathsf{stb}}(\texttt{wet\_grass}, [\texttt{sprinklers}])$.

2. By Abductive Defeat, the abductive sequent a in Item 1 is attacked by the sequent $b = \texttt{rainy}, \texttt{rainy} \to \neg\texttt{sprinklers} \Rightarrow \neg\texttt{sprinklers}$, which in turn is counter-attacked (using Defeat) by $c = \texttt{clear\_skies}, \texttt{clear\_skies} \to \neg\texttt{rainy} \Rightarrow \neg\texttt{rainy}$. It follows that c defends a.

3. By the introduced notation, $\mathsf{Supp}(c) = \{\texttt{clear\_skies}, \texttt{clear\_skies} \to \neg\texttt{rainy}\}$ is in $\mathsf{Def}_{\mathbb{E}}(a)$, where $\mathbb{E}$ is one of the two stable extensions of the abductive argumentation framework under consideration. Thus, for these a and $\mathbb{E}$, we have:

$$(\star) \quad \begin{array}{l} \mathsf{Def}_{\mathbb{E}}(a) \in \mathsf{Justify}^{\cup}_{\mathsf{stb}}(\texttt{wet\_grass}, [\texttt{sprinklers}]), \\ \{\texttt{clear\_skies}, \texttt{clear\_skies} \to \neg\texttt{rainy}\} \in \mathsf{Def}_{\mathbb{E}}(a). \end{array}$$

An intuitive description of $(\star)$ is the following: `sprinklers` is an explanation for `wet_grass`. The set $\{\texttt{clear\_skies}, \texttt{clear\_skies} \to \neg\texttt{rainy}\}$ is a justification for this explanation. Indeed, it is assumed that the sky is clear, and in that case there is no rain. Therefore, the wetness of the grass can be explained by the operation of the sprinklers.

4.3. Explanations Reduction; Avoiding Logically Equivalent Explanations

By its definition, if ε explains ϕ (either internally or externally), then – unless the range of the explanations is restricted – every formula that is logically equivalent to ε according to the base logic L also explains ϕ. This 'explosion' in the number of explanations may be avoided in several ways, e.g., by introducing appropriate attack rules that exclude logically equivalent alternatives of a derived explanation, or by switching to equivalence classes of logically equivalent formulas (see, e.g., [16]). Briefly, the idea is the following:

1. equivalence in L is defined as usual by: $\psi \equiv \phi$ iff $\psi \vdash \phi$ and $\phi \vdash \psi$.
2. classes of arguments are defined by: $[\![\Gamma \Rightarrow \psi]\!] = \{\Delta \Rightarrow \phi \mid \Delta \in [\![\Gamma]\!],\ \phi \in [\![\psi]\!]\}$, where: $[\![\psi]\!] = \{\phi \mid \phi \equiv \psi\}$ and $[\![\psi_1, \dots, \psi_n]\!] = \{\{\phi_1, \dots, \phi_n\} \mid \forall 1 \leq i \leq n\ \phi_i \in [\![\psi_i]\!]\}$.

Now, given a framework $\mathbb{AF}^{\mathsf{X}}_{\mathsf{L},\mathsf{A}}(\mathsf{S}) = \langle \mathrm{Arg}^{\mathsf{X}}_{\mathsf{L}}(\mathsf{S}), \mathsf{A} \rangle$ we switch to a framework whose arguments are classes $[\![a]\!]$ for $a \in \mathrm{Arg}^{\mathsf{X}}_{\mathsf{L}}(\mathsf{S})$, and where $[\![a]\!]$ attacks $[\![b]\!]$ if there are some $a' \in [\![a]\!] \cap \mathrm{Arg}^{\mathsf{X}}_{\mathsf{L}}(\mathsf{S})$ and $b' \in [\![b]\!] \cap \mathrm{Arg}^{\mathsf{X}}_{\mathsf{L}}(\mathsf{S})$ such that $(a', b') \in \mathsf{A}$. As usual, one has to show *independence of the choice of representatives*. This is rather routine.

5. Discussion and Conclusion

Abduction has been widely applied in different deductive systems, such as adaptive logics (see, e.g., [17,18]), and AI-based disciplines, perhaps the most prominent one is logic programing (see [19,20] for surveys). Argumentation-based approaches include frameworks for agent-based dialogues [21,22] and assumption-based argumentation frameworks [23]. In [24,25] abduction is studied as the problem of adding arguments to a given argumentation framework so that a given argument is rendered acceptable.

Our approach offers several novelties. In terms of knowledge representation we transparently represent abductive inferences by an explicit inference rule that produces abductive arguments. The latter are a new type of hypothetical arguments that are subjected to potential defeat. A variety attack rules address the quality of the offered explanation and thereby model critical questions [26] and meta-argumentative reasoning [27]. This is both natural and philosophically motivated, as argued in [28], where also a gap in argumentative accounts of abduction is identified. Instead of imposing desiderata on abductive inferences from the outside we incorporate them in the argumentative reasoning process. Our framework offers a high degree of modularity, and in comparison to approaches in logic programming we allow for fully propositional base logics. Desiderata on abductive arguments can be disambiguated in various ways by simply changing the attack rules, all in the same base framework. This allows for a thorough logical analysis and disambiguation of these properties as demonstrated in Theorems 1, 2 and Example 4.

The presented work is mainly focused on representation considerations. In future work we plan to take advantage of the uniformity of the sequent-based methods for explanation, and carry them on to more expressive logics (involving, e.g., preference relations among arguments) and to other types of explanations. We also plan to further develop meta-theoretical results concerning our setting and incorporate other approaches to the dialectic tier of explanation, such as *related admissibility* [14] or *strong explanation* [29].

References

[1] Adadi A, Berrada M. Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). IEEE Access. 2018;6:52138–52160.

[2] Čyras K, Rago A, Albini E, Baroni P, Toni F. Argumentative XAI: A Survey. In: Proc. of the 30th International Joint Conference on Artificial Intelligence, (IJCAI'21). ijcai.org; 2021. p. 4392–4399.

[3] Vassiliades A, Bassiliades N, Patkos T. Argumentation and explainable artificial intelligence: a survey. The Knowledge Engineering Review. 2021;36:e5.

[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):321–357.

[5] Besnard P, Hunter A. A review of argumentation based on deductive arguments. In: Baroni P, Gabbay D, Giacomin M, van der Torre L, editors. Handbook of Formal Argumentation. vol. 1. College Publications; 2018. p. 437–484.

[6] Arieli O, Straßer C. Sequent-based logical argumentation. Argument & Computation. 2015;6(1):73–99.

[7] Lipton P. Inference to the Best Explanation. Routledge; 2004. Second edition.

[8] Borg A. Assumptive sequent-based argumentation. Journal of Applied Logics – IfCoLog Journal of Logics and Their Applications. 2020;7(3):227–294.

[9] Gentzen G. Untersuchungen über das logische Schließen I, II. Mathematische Zeitschrift. 1934;39:176–210, 405–431.

[10] Baroni P, Caminada M, Giacomin M. Abstract argumentation frameworks and their semantics. In: Baroni P, Gabbay D, Giacomin M, van der Torre L, editors. Handbook of Formal Argumentation. vol. 1. College Publications; 2018. p. 159–236.

[11] Meheus J, Verhoeven L, Van Dyck M, Provijn D. Ampliative Adaptive Logics and the Foundation of Logic-Based Approaches to Abduction. Logical and Computational Aspects of Model-Based Reasoning. 2002;25:39–71.

[12] Arieli O, Borg A, Straßer C. Characterizations and classifications of argumentative entailments. In: Proc. 18th Conference on Knowledge Representation and Reasoning (KR'21); 2021. p. 52–62.

[13] Johnson RH. Manifest rationality: A pragmatic theory of argument. Routledge; 2000.

[14] Fan X, Toni F. On computing explanations in argumentation. In: Proc. 29th AAAI Conference on Artificial Intelligence (AAAI'15); 2015. p. 1496–1502.

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

[16] Amgoud L, Besnard P, Vesic S. Equivalence in logic-based argumentation. Journal of Applied Non Classical Logics. 2014;24(3):181–208.

[17] Lycke H. The Adaptive Logics Approach to Abduction. In: Logic, Philosophy and History of Science in Belgium; 2008. p. 35–41.

[18] Meheus J, Batens D. A Formal Logic for Abductive Reasoning. Logic Journal of the IGPL. 2006;14(2):221–236.

[19] Denecker M, Kakas A. Abduction in Logic Programming. In: Computational Logic: Logic Programming and Beyond. vol. 2407 of Lecture Notes in Computer Science. Springer; 2002. p. 402–436.

[20] Kakas A, Michael L. Abduction and argumentation for explainable machine learning: A Position Survey. arXiv preprint arXiv:201012896. 2020;.

[21] Bex F, Budzynska K, Walton D. Argumentation and explanation in the context of dialogue. Explanation-aware Computing ExaCt 2012. 2012;9:6.

[22] Arioua A, Croitoru M. Formalizing Explanatory Dialogues. In: Proc. 9th Conf. on Scalable Uncertainty Management (SUM'15). vol. 9310 of Lecture Notes in Computer Science. Springer; 2015. p. 282–297.

[23] Wakaki T. Extended abduction in assumption-based argumentation. In: Proc. IEA/AIE. vol. 11606 of Lecture Notes in Computer Science. Springer; 2019. p. 593–607.

[24] Sakama C. Abduction in Argumentation Frameworks. Journal of Applied Non-Classical Logics. 2018;28(2-3):218–239.

[25] Booth R, Gabbay DM, Kaci S, Rienstra T, Van Der Torre LW. Abduction and Dialogical Proof in Argumentation and Logic Programming. In: Proc. 21st European Conf. on Artificial Intelligence (ECAI'14). IOS Press; 2014. p. 117–122.

[26] Walton D, Reed C, Macagno F. Argumentation Schemes. Cambridge University Press; 2008.

[27] Boella G, Gabbay D, van der Torre L, Villata S. Meta-Argumentation Modelling I: Methodology and Techniques. Studia Logica. 2009;93(2–3):297–355.

[28] Olmos P. Abduction and comparative weighing of explanatory hypotheses: an argumentative approach. Logic Journal of the IGPL. 2019;.

[29] Ulbricht M, Wallner JP. Strong explanations in abstract argumentation. In: Proc. 23rd AAAI Conference on Artificial Intelligence. AAAI Press; 2021. p. 6496–6504.