A Principle-Based Robustness Analysis of Labeling-Based Bipolar Argumentation Semantics

Caren Al Anaissy[1], Chen Chen[2], Srdjan Vesic[3],
Leendert van der Torre[2,4], Liuwen Yu[5]

[1]*Sorbonne Université, France*
[2]*Zhejiang University, China*
[3]*CRIL CNRS Univ. Artois, France*
[4]*University of Luxembourg, Luxembourg*
[5]*Luxembourg Institute of Science and Technology, Luxembourg*

Abstract

Bipolar argumentation frameworks (BAFs) instantiate argumentation as balancing by modeling both attacks and supports between arguments, and are increasingly used in systems such as chatbots and debate platforms. Yet their semantics are less explored than those of Dung's abstract frameworks, especially under dynamic change. We present a principle-based robustness analysis of seven variants of labeling-based complete semantics for BAFs. The variants are grounded in three interpretations of support—deductive, necessary, and evidential. We introduce four robustness principles that assess how these semantics respond to changes, namely the addition or removal of attacks and supports. Our classification identifies, for each semantics, the change patterns that preserve labels and those that force changes, yielding a fine-grained picture of robustness across dynamic scenarios. Looking forward, these findings inform the design of efficient algorithms and enforcement procedures for dynamic BAFs (e.g., in argumentation-based chatbots), and articulate a bridge to reasoning alignment between argumentation as dialogue and argumentation as balancing.

Keywords: Artificial intelligence, knowledge representation and reasoning, bipolar argumentation, robustness principle-based approach

1 Introduction

Dung's abstract argumentation framework [14] initiates the attack–defense paradigm shift in formal argumentation [35], where the acceptability of arguments depends on their attack and defense relations rather than their internal structure. This attack-defense paradigm shift provides a unified foundation for reasoning in AI. *Bipolar argumentation frameworks (BAFs)* extend the attack–defense paradigm shift with *support* relations among arguments, thereby instantiating *argumentation as balancing* [34], which has been discussed in the first volume of Handbook of Formal Argumentation [1, Chap.3]: *where both pros and cons are in which arguments for and against alternative resolutions of the issues (options or positions) are put forward, evaluated, resolved, and balanced*. This makes BAFs particularly relevant in contexts

where decisions must be made in the presence of competing interests or values, such as legal reasoning [32], ethical deliberation, and policy-making.

Among the many proposed BAF semantics [9,31,13,25], those based on *interpretations of support* are the most discussed. Under standard interpretations—*deductive*, *necessary*—supports can be used to introduce *indirect attacks* and thus reduce BAFs to abstract attack graphs, illustrating the *universality of attack* [1, Chap. 3]. For instance, under the deductive interpretation [5], if argument A is acceptable and A supports argument B, then B must also be acceptable. In contrast, under the necessary interpretation [22], if B is acceptable and A supports B, then A must be accepted. Both interpretations treat supports as an intermediary step to introduce indirect attacks, reducing a BAF to an abstract argumentation framework. A third notion, evidential support [23], takes a different approach: support relations link arguments to pieces of evidence without enforcing acceptability, representing a qualitatively distinct form of support.

Despite growing interest, the behavior of BAF semantics is still less well understood than that of abstract argumentation frameworks, in particular in *dynamic* settings. In practical systems, agents interact through dialogue (asking, challenging, justifying), which continually *change* the underlying information: new arguments are added, attacks are discovered or retracted, and supports are introduced or withdrawn. A common architecture is therefore two–layered: *dialogue* governs the interaction between agents, while each agent maintains an *individual* reasoning state represented as a BAF that balances pros and cons for its current stance (e.g., in explainable AI assistants and argumentation-based chatbots [4,7,16,18,26,30]). As the dialogue progresses, the agent's BAF evolves—attacks and supports are added or removed, and the interpretation of support may change—requiring repeated semantic evaluation. This raises a question: *how robust are different BAF semantics to such changes (addition/removal of attacks and supports)?*

In this paper, we specifically focus on the robustness of labeling-based bipolar argumentation semantics with necessary, deductive, and evidential support. Starting with necessary and deductive interpretations, we initially follow the reduction from bipolar argumentation framework to abstract argumentation framework [14] by introducing indirect attacks. Subsequently, we define new labeling-based semantics for bipolar argumentation framework with evidential support. In alignment with the work by Rienstra et al. [28], the principles we investigate are also useful in the design of algorithms, thus bridges from formal argumentation to computational argumentation [17, Chap.13]. For example, Niskanen et al. [20] use robustness principles in the design of an algorithm for computing semantics of incomplete argumentation frameworks, where one can specify that attacks between certain arguments may or may not exist. Robustness principles are also useful in addressing enforcement problems in abstract argumentation [3]. This issue involves determining minimal sets of changes to an argumentation framework in order to enforce some result, such as the acceptance of a given set of arguments. Because robustness principles can be used to determine which changes to the attack and support relations of an argumentation framework do or do not change its evaluation, these principles can be used to guide the search for sets of changes in the enforcement problem. This idea has already been used for extension

enforcement under the grounded semantics [21].

The layout of this paper is as follows. We first present the complete semantics of BAF with necessary and deductive interpretations that is based on a reduction approach. In Section 3, we introduce the labeling-based semantics of BAF with evidential support. In Section 4, we conduct the principle-based analysis to robustness of BAF semantics. Section 5 discusses related work and Section 6 concludes and identifies some future work directions.

2 Necessary and Deductive support

This section gives the concept of indirect attack in bipolar argumentation. Dung's argumentation framework [14] consists of a set of arguments and a relation between arguments, which is called attack.

Definition 2.1 An *argumentation framework* is a pair $AF = \langle \mathcal{A}, \mathcal{R} \rangle$ where $\mathcal{A}$ is a set of arguments and $\mathcal{R} \subseteq \mathcal{A} \times \mathcal{A}$ is a binary attack relation over $\mathcal{A}$. We denote by $\mathcal{AF}$ the set of all argumentation frameworks.

Given an argumentation framework $AF = \langle \mathcal{A}, \mathcal{R} \rangle$ we say that an argument $a \in \mathcal{A}$ attacks an argument $b \in \mathcal{A}$ if and only if $(a,b) \in \mathcal{R}$. Given an argumentation framework $\langle \mathcal{A}, \mathcal{R} \rangle$ and an argument $x \in \mathcal{A}$ we denote by x^- the set of arguments attacking x and by x^+ the set of arguments attacked by x. Given a set $B \subseteq \mathcal{A}$ we denote by B^- the set of arguments attacking some $x \in B$ and by B^+ the set of arguments attacked by some $x \in B$.

A labeling-based semantics maps every argumentation framework to a set of labelings, which are functions that map every argument of an argumentation framework to a label. All the labeling-based semantics considered in this paper are defined using three possible labels: I indicates that the argument is accepted, O that the argument is rejected, and U that the acceptance of the argument is undecided.

Definition 2.2 [labeling [6]] A *labeling* of an argumentation framework $AF = \langle \mathcal{A}, \mathcal{R} \rangle$ is a function L: $\mathcal{A} \to \{I, O, U\}$. We denote by $\mathcal{L}(AF)$ the set of all labelings of AF. We also denote a labeling L by the set of pairs $\{(x_1, L(x_1)), ..., (x_n, L(x_n))\}$ where $\mathcal{A} = \{x_1, ...x_n\}$.

Definition 2.3 [labeling-based semantics] A labeling-based semantics σ defines a function $\mathcal{L}_\sigma$ that associates every $AF \in \mathcal{AF}$ with a set $\mathcal{L}_\sigma(AF) \subseteq \mathcal{L}(AF)$.

Definition 2.4 [Complete labeling] Let $AF = \langle \mathcal{A}, \mathcal{R} \rangle$ be an argumentation framework. A labeling $L \in \mathcal{L}(AF)$ is complete if and only if, for all $x \in \mathcal{A}$:

- $L(x) = I$ if and only if, for all $y \in x^-$, $L(y) = O$.
- $L(x) = O$ if and only if, for some $y \in x^-$, $L(y) = I$.
- $L(x) = U$ if and only if, not for all $y \in x^-$, $L(y) = O$ and no $y \in x^-$, $L(y) = I$.

Example 2.5 [Four arguments] The argumentation framework visualized on the left hand side of Figure 1 is defined by $AF = \langle \{a,b,c,d\}, \{(a,b),(b,a),(c,d),(d,c)\} \rangle$. There are nine complete labelings: $\{(a,U),(b,U),(c,U),(d,U)\}$, $\{(a,I),(b,O),(c,U),(d,U)\}, \{(a,O),(b,I),(c,U),(d,U)\}, \{(a,U),(b,U),(c,I),(d,O)\}$,

$\{(a,U),(b,U),(c,O),(d,I)\},\{(a,I),(b,O),(c,I),(d,O)\},\{(a,I),(b,O),(c,O),(d,I)\}$, $\{(a,O),(b,I),(c,I),(d,O)\}$, $\{(a,O),(b,I),(c,O),(d,I)\}$.

Fig. 1. An argumentation framework (AF) and a bipolar argumentation framework (BAF)

A bipolar argumentation framework is an extension of Dung's framework. It is based on a binary attack relation between arguments and a binary support relation over the set of arguments.

Definition 2.6 [Bipolar argumentation framework [8]] A *bipolar argumentation framework* (BAF, for short) is a triple $\langle \mathcal{A}, \mathcal{R}, \mathcal{S} \rangle$ consists of: a set $\mathcal{A}$ of arguments, a binary relation $\mathcal{R}$ on $\mathcal{A}$ called attack relation, and another binary relation $\mathcal{S}$ on $\mathcal{A}$ called support relation, and $\mathcal{R} \cap \mathcal{S} = \emptyset$.

An AF is a special BAF with the form $\langle \mathcal{A}, \mathcal{R}, \emptyset \rangle$. We denote by $\mathcal{BAF}$ the set of all bipolar argumentation frameworks. A *BAF* can be represented as a directed graph. Given $a,b,c \in \mathcal{A}$, $(a,b) \in \mathcal{R}$ means a attacks b, noted as $a \rightarrow b$; $(b,c) \in \mathcal{S}$ means b supports c, noted as $b \dashrightarrow c$.

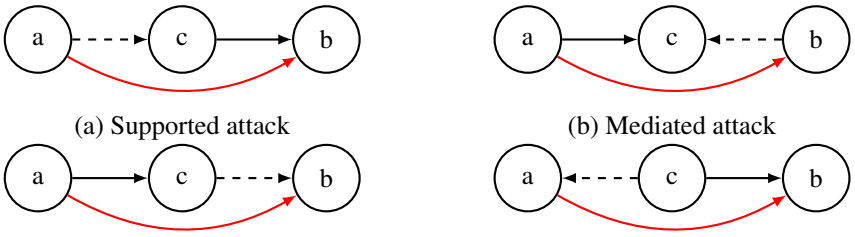
Example 2.7 [Four arguments, continued] The bipolar argumentation framework visualized at the right hand side of Figure 1 extends the argumentation framework in Example 1 so that a supports d.

Support relations only influence the semantics when there are also attacks, which leads to the study of the interactions between attack and support. In the literature, the different kinds of relations between support and attack have been studied as different notions of indirect attack.

Definition 2.8 [Four indirect attacks [24]] Let $BAF = \langle \mathcal{A}, \mathcal{R}, \mathcal{S} \rangle$ be a bipolar argumentation framework and $a,b \in \mathcal{A}$, there is:

- a supported attack from a to b in *BAF* iff there exists an argument c s.t. there is a sequence of supports from a to c and c attacks b, represented as $(a,b) \in \mathcal{R}^{sup}$.
- a mediated attack from a to b in *BAF* iff there exists an argument c s.t. there is a sequence of supports from b to c and a attacks c, represented as $(a,b) \in \mathcal{R}^{med}$.
- a secondary attack from a to b in *BAF* iff there exists an argument c s.t. there is a sequence of supports from c to b and a attacks c, $(a,b) \in \mathcal{R}^{sec}$.
- an extended attack from a to b in *BAF* iff there exists an argument c s.t. there is a sequence of supports from c to a and c attacks b, $(a,b) \in \mathcal{R}^{ext}$.

Definition 2.9 [Super-mediated attack [10]] Let $BAF = \langle \mathcal{A}, \mathcal{R}, \mathcal{S} \rangle$ be a bipolar argumentation framework and $a,b \in \mathcal{A}$, there is a super-mediated attack from a to b iff



(a) Supported attack (b) Mediated attack

(c) Secondary attack (d) Extended attack

Fig. 2. Four kinds of indirect attack

there exists an argument c such that there is a sequence of supports from b to c and a directly attacks c or supported-attacks c, represented as $(a,b) \in \mathcal{R}^{med}_{R^{sup}}$.

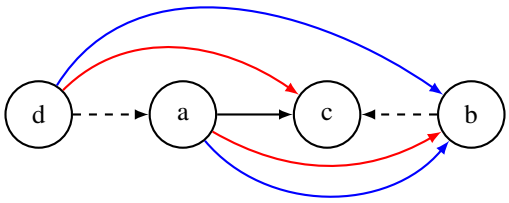


Fig. 3. Super-mediated attack

We can obtain various kinds of indirect attacks according to different interpretations of support relation. These indirect attacks were built from the combination of direct attacks and the supports. Then from the obtained indirect attacks and the support we can build additional indirect attacks and so on.

Definition 2.10 [Tiered indirect attacks [24]] Given a $BAF = \langle \mathcal{A}, \mathcal{R}, \mathcal{S} \rangle$, the tiered indirect attacks of BAF are as follows :

- $R_0^{ind} = \emptyset$
- $R_1^{ind} = \{R_\emptyset^{sup}, R_\emptyset^{sec}, R_\emptyset^{med}, R_\emptyset^{ext}\}$
- $R_i^{ind} = \{R_E^{sup}, R_E^{sec}, R_E^{med}, R_E^{ext} \mid E \subseteq R_{i-1}^{ind}\}$ for $i > 1$, where:
  - $R_E^{sup} = \{(a,b) \mid$ there exists an argument c s.t. there is a sequence of supports from a to c and $(c,b) \in R \cup \bigcup E\}$
  - $R_E^{sec} = \{(a,b) \mid$ there exists an argument c s.t. there is a sequence of supports from c to b and $(a,c) \in R \cup \bigcup E\}$
  - $R_E^{med} = \{(a,b) \mid$ there exists an argument c s.t. there is a sequence of supports from b to c and $(a,c) \in R \cup \bigcup E\}$
  - $R_E^{ext} = \{(a,b) \mid$ there exists an argument c s.t. there is a sequence of supports from c to a and $(c,b) \in R \cup \bigcup E\}$

With R^{ind} we denote the collection of all sets of indirect attacks $\bigcup_{i=0}^{\infty} R_i^{ind}$.

Definition 2.11 [Existing reductions of BAF to AF] Given a $BAF = \langle \mathcal{A}, \mathcal{R}, \mathcal{S} \rangle, \forall a,b,c \in \mathcal{A}$:

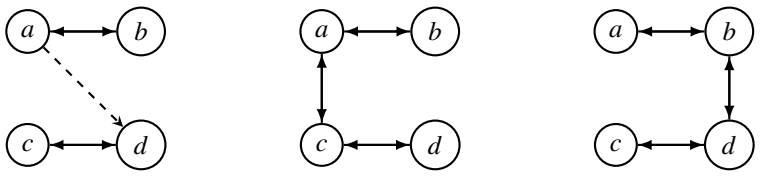
- SupportedReduction [10] (RS for short): $(a,b) \in \mathcal{R}^{sup}$ is the collection of supported attacks iff $(a,c) \in \mathcal{S}$ and $(c,b) \in \mathcal{R}$, $RS(BAF) = (\mathcal{A}, \mathcal{R} \cup \mathcal{R}^{sup})$.
- MediatedReduction [10] (RM for short): $(a,b) \in \mathcal{R}^{med}$ is the collection of mediated attacks iff $(b,c) \in \mathcal{S}$ and $(a,c) \in \mathcal{R}$, $RM(BAF) = \big(\mathcal{A}, \mathcal{R} \cup \mathcal{R}^{med}\big)$.
- SecondaryReduction [10] (R2 for short): $(a,b) \in \mathcal{R}^{sec}$ is the collection of secondary attacks iff $(c,b) \in \mathcal{S}$ and $(a,c) \in \mathcal{R}$, $R2(BAF) = (\mathcal{A}, \mathcal{R} \cup \mathcal{R}^{sec})$.
- ExtendedReduction [10] (RE for short): $(a,b) \in \mathcal{R}^{ext}$ is the collection of extended attacks, iff $(c,a) \in \mathcal{S}$ and $(c,b) \in \mathcal{R}$, $RE(BAF) = (\mathcal{A}, \mathcal{R} \cup \mathcal{R}^{ext})$.
- DeductiveReduction [24] (RD for short): Let $\mathcal{R}' = \{\mathcal{R}^{sup}, \mathcal{R}^{med}_{\mathcal{R}^{sup}}\} \subseteq \mathcal{R}^{ind}$ be the collection of supported and super-mediated attacks in BAF, $RD(BAF) = (\mathcal{A}, \mathcal{R} \cup \bigcup \mathcal{R}')$.
- NecessaryReduction [24] (RN for short): Let $\mathcal{R}' = \{\mathcal{R}^{sec}, \mathcal{R}^{ext}\} \subseteq \mathcal{R}^{ind}$ be the collection of secondary and extended attacks in BAF, $RN(BAF) = (\mathcal{A}, \mathcal{R} \cup \bigcup \mathcal{R}')$.

Next, we define the labeling-based semantics of BAF.

Definition 2.12 [labeling of BAF] A labeling of a $BAF = \langle \mathcal{A}, \mathcal{R}, \mathcal{S} \rangle$ is a function L: $\mathcal{A} \to \{I, O, U\}$. We denote by $\mathcal{L}(BAF)$ the set of all labelings of BAF.

Definition 2.13 [Complete labeling of BAF] Let $\omega \in \{RS, RM, R2, RE, RD, RN\}$ be a reduction of BAF to AF. A labeling L of a $BAF = \langle \mathcal{A}, \mathcal{R}, \mathcal{S} \rangle$ is complete under ω, iff L is a complete labeling of $\omega(BAF)$.

Example 2.14 [Labeling-based semantics of BAF with RD and RN] Consider the bipolar argumentation framework in Figure 4.1. If the interpretation of support from a to d is deductive, a supported-attacks c, c mediated-attacks a. We have $RD(\mathcal{F}) = \langle Ar, att \cup \{(a,c),(c,a)\} \rangle$ as visualized in Figure 4.2, there are five complete labelings: $\{(a,O),(b,I),(c,U),(d,U)\}, \{(a,U),(b,U),(c,O),(d,I)\}, \{(a,I),(b,O),(c,O),(d,I)\}$, $\{(a,O),(b,I),(c,O),(d,I)\}, \{(a,O),(b,I),(c,I),(d,O)\}$. If the interpretation of support from a to d is necessary, then b secondary-attacks d, and d extended-attacks b. We have $RN(\mathcal{F}) = \langle Ar, att \cup \{(b,d),(d,b)\} \rangle$ as visualized in Figure 4.3, there are five complete labelings: $\{(a,I),(b,O),(c,U),(d,U)\}$, $\{(a,U),(b,U),(c,I),(d,O)\}$, $\{(a,I),(b,O),(c,O),(d,I)\}$, $\{(a,I),(b,O),(c,I),(d,O)\}, \{(a,O),(b,I),(c,I),(d,O)\}$.



4.1 A BAF $\mathcal{F}$ 4.2 RD($\mathcal{F}$) 4.3 RN($\mathcal{F}$)

Fig. 4. Deductive and necessary interpretations give different corresponding AFs

3 Evidential support

BAF with evidential support has been studied by N. Oren and T. J. Norman [23]. They analyse the importance of introducing evidential support into argumentation framework and proposed the traditional extension-based semantics of BAF with evidential support. Besides, they add moreover that elements of evidential support are unique, that support is minimal, and so on.

To keep our presentation uniform and to compare evidential support to deductive and necessary support, we only consider the fragment of bipolar argumentation frameworks where individual arguments attack or support other arguments. This also simplifies the following definitions.

Moreover, evidential support contains special arguments which do not need to be supported by other arguments. Such arguments may have to satisfy other constraints, for example that they cannot be attacked by ordinary arguments, or that they cannot attack ordinary arguments. To keep our analysis uniform, we do not explicitly distinguish such special arguments, but encode them implicitly: if an argument supports itself, then it is such a special argument. This leads to the following definition of an evidential sequence for an argument.

Definition 3.1 [Evidential sequence] Given a $BAF = \langle \mathcal{A}, \mathcal{R}, \mathcal{S} \rangle$. A sequence $(a_0, \ldots, a_n)$ of elements of $\mathcal{A}$ is an evidential sequence for argument a_n iff $(a_0, a_0) \in \mathcal{S}$, and for $0 \leq i < n$ we have $(a_i, a_{i+1}) \in \mathcal{S}$.

We give our labeling-based semantics as follows.

Definition 3.2 [Complete labeling of BAF with evidential support] Let $BAF = \langle \mathcal{A}, \mathcal{R}, \mathcal{S} \rangle$ be a bipolar argumentation framework. A labeling $L \in \mathcal{L}(BAF)$ is complete under evidential support iff for all $a \in \mathcal{A}$:

(i) $L(a) = I$ iff, there is an evidential sequence $(a_0, \ldots, a)$ for a, s.t. $\forall a' \in \{a_0, \ldots, a\}^-$, it holds that $L(a') = O$.

(ii) $L(a) = O$ iff, for all evidential sequence $(a_0, \ldots, a)$ for a, $\exists a' \in \mathcal{A}$ s.t. $L(a') = I$ and $a' \in \{a_0, \ldots, a\}^-$.

(iii) $L(a) = U$ iff, both of the conditions in (i) and (ii) are not satisfied.

Example 3.3 illustrates the complete semantics of BAF with evidential support.

Example 3.3 Assume a $BAF = \langle \mathcal{A}, \mathcal{R}, \mathcal{S} \rangle$ with evidential support, in which $\mathcal{A} = \{a, b, c, d\}$, $\mathcal{R} = \{(d, b)\}$, $\mathcal{S} = \{(a, a), (a, b), (b, c), (d, d)\}$, as depicted in Figure 5. The only complete labeling of BAF is $\{(a, I), (b, O), (c, O), (d, I)\}$.

4 A principle-based robustness analysis

In this section, we present a principle-based robustness analysis of bipolar argumentation. Due to space limitations, we provide some proofs of results; the remaining all other proofs are available in the supplemental material. [1] Principles 1 to 4 extend the robustness principles introduced by Baroni and Giacomin [2] and further developed

[1] `https://drive.google.com/file/d/1c1B2iuAGWokm4FFIAEoM8VUSOZuWO1kt/view?usp=sharing`

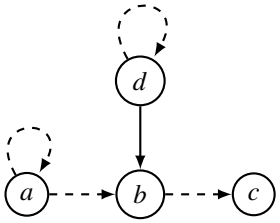


Fig. 5. A BAF with evidential support

by Rienstra et al. [28]. These principles characterize how semantics behave when an argumentation framework is modified by adding or removing an attack relation. In this work, we adapt and apply these principles to bipolar argumentation frameworks, allowing for structural changes involving both attack and support relations.

Principle 1 says that adding an attack between any two arguments does not change the original semantics of the framework.

Principle 1 (Attack addition persistence) *Let σ be a semantics and let $X,Y \in \{O,I,U\}$. We say that σ satisfies XY addition persistence if and only if for all $\langle \mathcal{A},\mathcal{R},\mathcal{S}\rangle \in \mathcal{BAF}$ and $x,y \in \mathcal{A}$, if $L \in \mathcal{L}_\sigma(\langle \mathcal{A},\mathcal{R},\mathcal{S}\rangle)$, $L(x) = X$ and $L(y) = Y$, then $L \in \mathcal{L}_\sigma((\mathcal{A},\mathcal{R}\cup\{(x,y)\},\mathcal{S}))$.*

The results for Principle 1 are summarized in Table 1. It shows that all seven semantics preserve labelings when the attacked argument is labeled Out (OO, UO, IO), and all fail when an In-labeled argument is attacked or attacking (UI, IU, II). For the remaining cases: OU, UU, and OI, only RM, RD, and REv preserve the labeling across the board, making them the most robust to added attacks. In contrast, RS, RE, and RN are least robust, failing in all three of these cases.

Table 1
Attack Addition persistence for labeling-based complete semantics. Note that in Tables 1–4 each cell shows whether a semantics *persists* (✓) or *does not persist* (×) after adding or removing an attack/support (x,y). Columns are labeled by the ordered pair **XY**, where **X** is the *original* label of the source argument x and **Y** that of the target y. For example, *IO* in Table 1 denotes that we are adding the attack from an In-labeled argument to an Out-labeled argument.

	OO	OU	UO	UU	OI	UI	IO	IU	II
RS	✓	×	✓	×	×	×	✓	×	×
RM	✓	✓	✓	✓	✓	×	✓	×	×
R2	✓	✓	✓	✓	×	×	✓	×	×
RE	✓	×	✓	×	×	×	✓	×	×
RD	✓	✓	✓	✓	✓	×	✓	×	×
RN	✓	×	✓	×	×	×	✓	×	×
REv	✓	✓	✓	✓	✓	×	✓	×	×

Principle 2 says that removing an attack between any two arguments does not change the original semantics of the framework.

Principle 2 (Attack Removal persistence) *Let σ be a semantics and let $X,Y \in$*

$\{O,I,U\}$. *We say that* σ *satisfies XY removal persistence if and only if for all* $\langle \mathcal{A},\mathcal{R},\mathcal{S}\rangle \in \mathcal{BAF}$ *and* $x,y \in \mathcal{A}$*, if* $L \in \mathcal{L}_\sigma(\langle \mathcal{A},\mathcal{R},\mathcal{S}\rangle)$*,* $L(x) = X$ *and* $L(y) = Y$*, then* $L \in \mathcal{L}_\sigma((\mathcal{A},\mathcal{R}\setminus\{(x,y)\},\mathcal{S}))$.

Table 2
Attack Removal persistence for labeling-based complete semantics

	OO	OU	UO	UU	OI	UI	IO	IU	II
RS	×	×	×	×	×	✓	✓	✓	✓
RM	✓	✓	✓	×	✓	✓	×	✓	✓
R2	✓	✓	✓	×	✓	✓	×	✓	✓
RE	×	×	×	×	✓	✓	×	✓	✓
RD	✓	×	×	×	×	✓	×	✓	✓
RN	×	×	×	×	✓	✓	×	✓	✓
REv	✓	✓	✓	×	✓	✓	×	✓	✓

The results for Principle 2 are summarized in Table 2. It shows universal satisfaction of UI, IU, and II removal persistence, and universal failure on UU. RM, R2, and REv are the most stable, preserving all other cases, while RS fails nearly all. RN, RE, and RD show mixed sensitivity, especially when an Out argument previously attacked an In or Undecided one. These differences highlight how some semantics tightly couple rejections to attack structure, while others are more relaxed.

Principle 3 says that adding a support relation between any two arguments does not change the original semantics of the framework.

Principle 3 (Support Addition persistence) *Let* σ *be a semantics and let* $X,Y \in \{O,I,U\}$. *We say that* σ *satisfies XY addition persistence if and only if for all* $\langle \mathcal{A},\mathcal{R},\mathcal{S}\rangle \in \mathcal{BAF}$ *and* $x,y \in \mathcal{A}$*, if* $L \in \mathcal{L}_\sigma(\langle \mathcal{A},\mathcal{R},\mathcal{S}\rangle)$*,* $L(x) = X$ *and* $L(y) = Y$*, then* $L \in \mathcal{L}_\sigma((\mathcal{A},\mathcal{R},\mathcal{S}\cup\{(x,y)\}))$.

The results for Principle 3 are summarized in Table 3. It shows that only the II case is universally preserved, and UO universally fails. RM, RD, and REv again stand out, preserving five of the remaining seven configurations. R2, RE, and RN are the most sensitive, only preserving IO, IU, and II. RS sits in between. The results reflect that semantics differ in how they treat added supports from non-accepted sources.

Proposition 4.1 *The complete semantics satisfy OI, UI and II support addition persistence under supported reduction.*

Proof. For any bipolar argumentation framework $\mathcal{F} = \langle \mathcal{A},\mathcal{R},\mathcal{S}\rangle$ such that there exists the set of arguments $\{l,x,y,z\} \subseteq \mathcal{A}$ where there exists a sequence of supports from l to x and an attack from y to z, let $\mathcal{F}' = \langle \mathcal{A},\mathcal{R}'\rangle$ be the argumentation framework obtained by applying the supported reduction to $\mathcal{F}$. Let $\mathcal{F}_1 = \langle \mathcal{A},\mathcal{R},\mathcal{S}\cup\{(x,y)\}\rangle$, and let $\mathcal{F}'_1 = \langle \mathcal{A},\mathcal{R}'_1\rangle$ be the argumentation framework obtained by applying the supported reduction to $\mathcal{F}_1$. We represent the four argumentation frameworks in Figure 6. Let L be a complete labeling of $\mathcal{F}'$.

- $L(x) \in \{O,I,U\}$ and $L(y) = I$: $L(z) = O$ since $L(y) = I$, $L(l) \in \{O,I,U\}$, from [28],

Table 3
Support Addition persistence for labeling-based complete semantics

	OO	OU	UO	UU	OI	UI	IO	IU	II
RS	×	×	×	×	✓	✓	×	×	✓
RM	✓	✓	×	✓	✓	✓	×	×	✓
R2	×	×	×	×	×	×	✓	✓	✓
RE	×	×	×	×	×	×	✓	✓	✓
RD	✓	✓	×	✓	✓	✓	×	×	✓
RN	×	×	×	×	×	×	✓	✓	✓
REv	✓	✓	×	✓	✓	✓	×	×	✓

the complete semantics satisfy OO, UO and IO attack addition persistence, which means that L is a complete labeling of $\mathcal{F}'_1$.

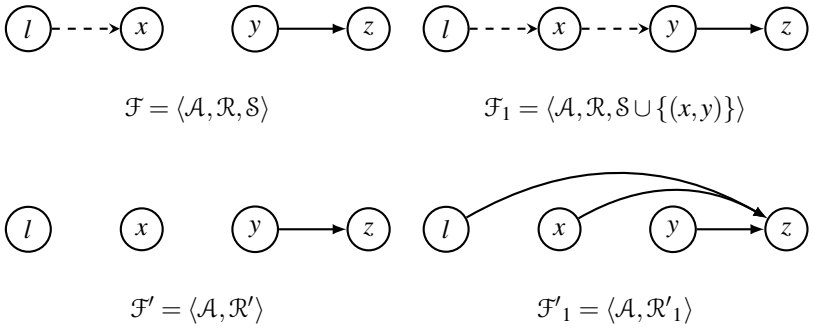


Fig. 6. The complete semantics satisfy OI, UI and II support addition persistence under supported reduction.

□

Proposition 4.2 *The complete semantics violates IO support addition persistence under supported reduction.*

Proof. Consider the following counterexample in Figure 7. The complete labeling of $\mathcal{F}'$ $\{(l,I),(x,I),(b,I),(y,O),(z,I)\}$ is no longer a complete labeling after adding the attacks from l and x to z in $\mathcal{F}'$. The new complete labeling (of $\mathcal{F}'_1$) is $\{(l,I),(x,I),(b,I),(y,O),(z,O)\}$. □

Proposition 4.3 *The complete semantics violates OU support addition persistence under supported reduction.*

Proof. Consider the following counterexample in Figure 8. The complete labeling of $\mathcal{F}'$ $\{(l,I),(x,O),(h,I),(b,U),(y,U),(z,U)\}$ is no longer a complete labeling after adding the attacks from l and x to z in $\mathcal{F}'$. The new complete labeling (of $\mathcal{F}'_1$) is $\{(l,I),(x,O),(h,I),(b,I),(y,O),(z,O)\}$. □

Proposition 4.4 *The complete semantics violates UO support addition persistence under supported reduction.*

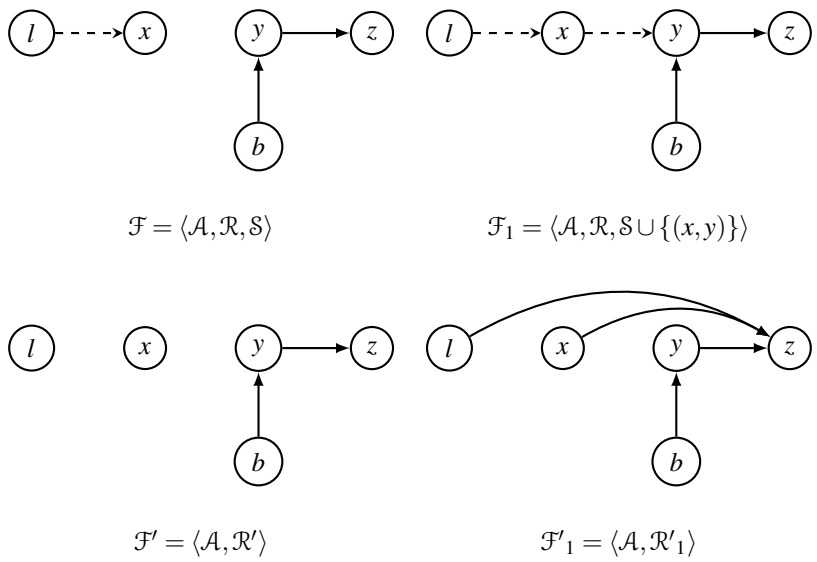


Fig. 7. The complete semantics violates IO support addition persistence under supported reduction.

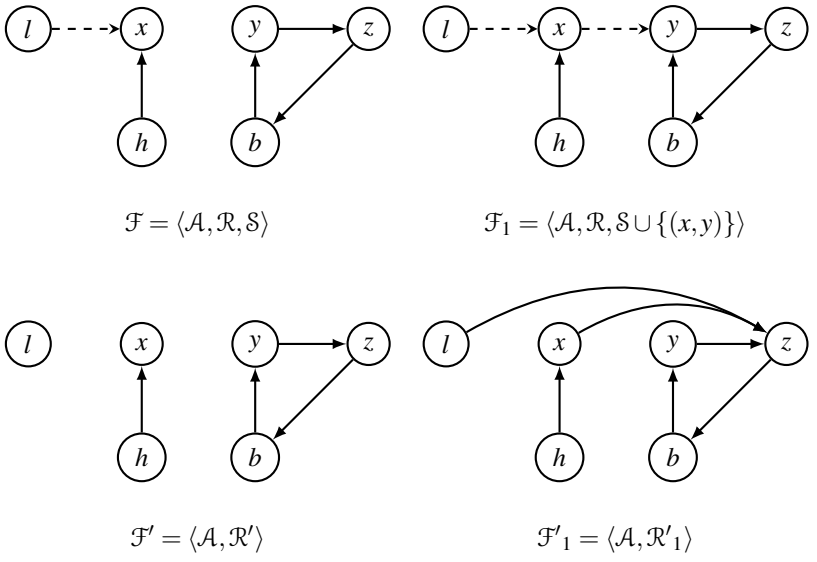


Fig. 8. The complete semantics violates OU support addition persistence under supported reduction.

Proof. Consider the following counterexample in Figure 9. The complete labeling of $\mathcal{F}'$ $\{(f,I),(l,O),(x,U),(b,I),(y,O),(z,I)\}$ is no longer a complete labeling after adding the attacks from l and x to z in $\mathcal{F}'$. The new complete labeling (of $\mathcal{F}'_1$) is $\{(f,I),(l,O),(x,U),(b,I),(y,O),(z,U)\}$. □

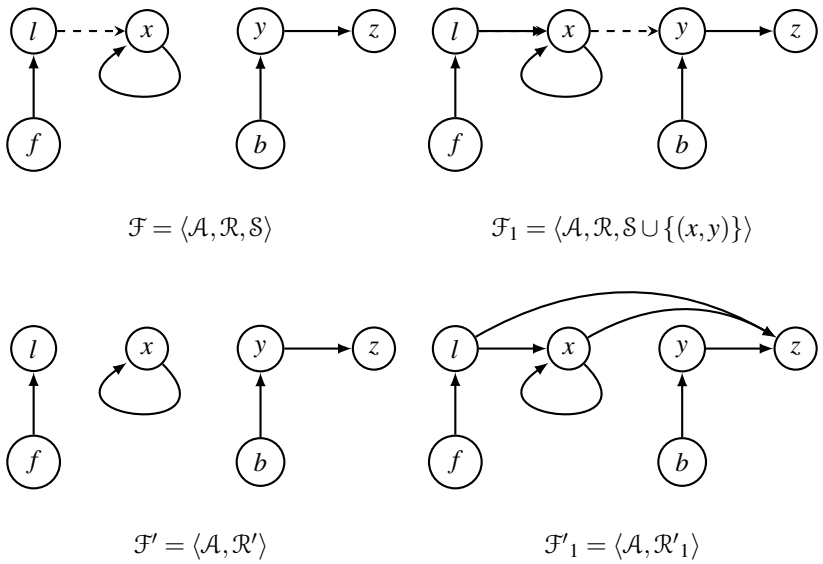


Fig. 9. The complete semantics violates UO support addition persistence under supported reduction.

Principle 4 says that removing a support between any two arguments does not change the original semantics of the framework.

Principle 4 (Support Removal persistence) *Let σ be a semantics and let $X,Y \in \{O,I,U\}$. We say that σ satisfies XY removal persistence if and only if for all $\langle \mathcal{A},\mathcal{R},\mathcal{S} \rangle \in \mathcal{BAF}$ and $x,y \in \mathcal{A}$, if $L \in \mathcal{L}_\sigma(\langle \mathcal{A},\mathcal{R},\mathcal{S} \rangle)$, $L(x) = X$ and $L(y) = Y$, then $L \in \mathcal{L}_\sigma((\mathcal{A},\mathcal{R},\mathcal{S} \setminus \{(x,y)\}))$.*

The results for Principle 4 are summarized in Table 4. It shows that all semantics preserve UI, and most preserve IO and II—with the exception of REv, which fails both due to its direct reliance on evidential chains. RM and R2 are the most robust, failing only in OO and UU. RN, RS, and RE are more fragile when supports between Out-labeled or Undecided arguments are removed, which calls for reevaluation.

Table 4
Support Removal persistence for labeling-based complete semantics

	OO	OU	UO	UU	OI	UI	IO	IU	II
RS	✓	✓	×	✓	✓	✓	×	×	✓
RM	×	✓	✓	×	✓	✓	✓	✓	✓
R2	×	✓	✓	×	✓	✓	✓	✓	✓
RE	✓	×	✓	×	×	✓	✓	✓	✓
RD	×	✓	✓	×	×	✓	✓	✓	✓
RN	×	×	✓	×	×	✓	✓	✓	✓
REv	✓	✓	✓	×	✓	✓	✓	×	×

5 Related work

Support relations, unlike the attack relation, remain controversial in the literature. The evaluation of bipolar argumentation semantics through a principle-based lens is relatively recent but growing. For example, there are studies analyzing bipolar argumentation semantics by focusing on different interpretations of support [11,33]. This line of research has been further extended to domains such as legal reasoning, where multiple interpretations of support correspond to different legal interpretations [32]. Principle-based analyses have also been applied to new semantics defined using novel notions of defense and to social choice-based approaches to argument evaluation [31]. Doder et al. [13] specifically investigate principle-based characterizations of ranking semantics for frameworks with necessities.

Applications of BAFs have been particularly studied in fields like explainable AI (XAI) and argumentation-based chatbots, providing a motivation for our research. In the context of XAI, Kampik et al. explore the changes in quantitative bipolar argumentation frameworks to provide sufficient, necessary, and counterfactual explanations in response to updates within these frameworks [18]. It underscores the importance of studying the dynamic aspects of BAFs, highlighting their crucial role in understanding operational dynamics. In the realm of chatbots, as Federico Castagna et al noted [7]: "Speaking of the underlying argumentation framework of argumentation-based chatbots, when embedding a knowledge base into an AF, the Bipolar framework (and its variants QBAF and WBAF) turns out to be the most common option. This choice is related to the additional information provided by BAFs which encompass support relations rather than just attacks, allowing for an intuitive formalisation of both endorsements and conflicts between pieces of data." For instance, the interactive recommender systems developed by Rago et al. [27,26] utilize a BAF and tripolar argumentation framework to embed their underlying knowledge bases, thereby enhancing the clarity of their recommendations through rich, argument-based explanations. In a similar vein, Cocarascu et al. describe argumentative dialogical agents that construct a quantitative bipolar argumentation framework to facilitate structured dialogues based on movie reviews [12]. The integration of BAFs with advances in generative AI and hybrid models has fostered innovations such as ArguBot [4], developed using Google DialogFlow [30]. This system employs ASPARTIX [15] to compute arguments from an underlying BAF to support (pro-bot) or challenge (con-bot) the user's opinion about the topic of dialogue. Lastly, the conversational agent designed by Fazzinga et al. incorporates BAFs to manage dialogues and argumentation effectively, showcasing the versatility and extensive applicability of BAFs in contemporary AI applications [16]. All these developments underline the significance of ongoing studies into the dynamics and robustness of bipolar argumentation semantics.

6 Summary

In this paper we analysed robustness properties for seven variants of the complete semantics for bipolar argumentation frameworks. Six variants arise from reduction-based approaches that interpret support as necessary or deductive, while the seventh is defined directly for evidential support. We use four robustness principles—the two

attack-oriented principles of Rienstra et al. [28] together with two support-oriented principles—and compare the variants exhaustively. Tables 1–4 make the impact of adding or removing attacks or supports explicit, allowing practitioners (1) to select a semantics that remains stable under the specific updates their application performs, and (2) to identify precisely the situations in which recomputation of labels is unavoidable.

Beyond these results, our analysis situates naturally within the A-BDI metamodel (Argumentation as Balancing, Dialogue, and Inference) [34]. A-BDI views the three conceptualizations as complementary rather than exclusive, and principles as a means to select among existing methods or to define new ones. This perspective aligns with the reasoning alignment view [29]: Reasoning Alignment Diagrams (RADs) are commutative reasoning representations that align a source specification with an argumentation-based explanation path; they compose an "assert (inference)" RAD with a "listen (revision)" RAD to model dialogue—agents that can say (argumentation/inference) and hear (belief revision) while preserving alignment. Our robustness classification supports this agenda by indicating when local inference within a BAF remains stable under dialogue-driven updates, and how changes to relations or to the interpretation of support affect balancing in dynamic contexts.

Future work. A natural next step is to generalise reasoning alignment between argumentation as inference and argumentation as balancing, by studying bipolar argumentation within structured argumentation [19]. In parallel, the rise of large language models foregrounds the dialogue perspective: in multi-agent settings, each agent can maintain a local BAF as its individual reasoning state while the dialogue protocol drives assert, question, and revise moves; our robustness results then indicate when edits prompted by these moves leave the agent's labels stable and when recomputation is required. On the technical side, we plan to extend the robustness analysis beyond complete semantics (e.g., grounded, preferred, stable); to investigate settings where interpretations of support vary over time or across agents and use our tables to anticipate when acceptability persists or changes; and to formalise optimisation heuristics suggested by Tables 1–4 (for instance, "no-recompute" cases under specific edit patterns) to support incremental solvers and enforcement procedures in dynamic BAF-based systems.

Acknowledgments

We thank the anonymous reviewer for their comments. This work is supported by the Luxembourg National Research Fund (FNR) through the following projects: The Epistemology of AI Systems (EAI) (C22/SC/17111440), DJ4ME – A DJ for Machine Ethics: the Dialogue Jiminy (O24/18989918/DJ4ME), Logical Methods for Deontic Explanations (LoDEx) (INTER/DFG/23/17415164/LoDEx), Symbolic and Explainable Regulatory AI for Finance Innovation (SERAFIN) (C24/19003061/SERAFIN), and the University of Luxembourg for the Marie Speyer Excellence Grant for the project Formal Analysis of Discretionary Reasoning (MSE-DISCREASON).

References

[1] Baroni, P., D. Gabbay, M. Giacomin and L. van der Torre, editors, **1**, College Publications, 2018.

[2] Baroni, P. and M. Giacomin, *On principle-based evaluation of extension-based argumentation semantics*, Artificial Intelligence **171** (2007), pp. 675–700.

[3] Baumann, R., S. Doutre, J.-G. Mailly and J. P. Wallner, *Enforcement in formal argumentation*, IfColog Journal of Logics and their Applications (FLAP) **8** (2021), pp. 1623–1678.

[4] Bistarelli, S., C. Taticchi and F. Santini, *A chatbot extended with argumentation*, in: M. D'Agostino, F. A. D'Asaro and C. Larese, editors, *Proceedings of the 5th Workshop on Advances in Argumentation in Artificial Intelligence 2021 co-located with the 20th International Conference of the Italian Association for Artificial Intelligence (AIxIA 2021), Milan, Italy, November 29th, 2021*, CEUR Workshop Proceedings **3086** (2021).

[5] Boella, G., D. M. Gabbay, L. van der Torre and S. Villata, *Support in abstract argumentation*, in: *Proceedings of the Third International Conference on Computational Models of Argument (COMMA'10)*, Frontiers in Artificial Intelligence and Applications, IOS Press, 2010, pp. 40–51.

[6] Caminada, M., *On the issue of reinstatement in argumentation*, in: *European Workshop on Logics in Artificial Intelligence*, Springer, 2006, pp. 111–123.

[7] Castagna, F., N. Kokciyan, I. Sassoon, S. Parsons and E. Sklar, *Computational argumentation-based chatbots: a survey*, arXiv preprint arXiv:2401.03454 (2024).

[8] Cayrol, C. and M.-C. Lagasquie-Schiex, *On the acceptability of arguments in bipolar argumentation frameworks*, in: *European Conference on Symbolic and Quantitative Approaches to Reasoning and Uncertainty*, Springer, 2005, pp. 378–389.

[9] Cayrol, C. and M.-C. Lagasquie-Schiex, *Bipolar abstract argumentation systems*, in: *Argumentation in Artificial Intelligence*, Springer, 2009 pp. 65–84.

[10] Cayrol, C. and M.-C. Lagasquie-Schiex, *Bipolarity in argumentation graphs: Towards a better understanding*, International Journal of Approximate Reasoning **54** (2013), pp. 876–899.

[11] Cayrol, C. and M.-C. Lagasquie-Schiex, *An axiomatic approach to support in argumentation*, in: *International Workshop on Theory and Applications of Formal Argumentation*, Springer, 2015, pp. 74–91.

[12] Cocarascu, O., A. Rago and F. Toni, *Extracting dialogical explanations for review aggregations with argumentative dialogical agents.*, in: *AAMAS*, 2019, pp. 1261–1269.

[13] Doder, D., S. Vesic and M. Croitoru, *Ranking semantics for argumentation systems with necessities*, in: *IJCAI 2020-29th International Joint Conference on Artificial Intelligence*, International Joint Conferences on Artificial Intelligence Organization, 2021, pp. 1912–1918.

[14] Dung, P. M., *On the Acceptability of Arguments and Its Fundamental Role in Nonmonotonic Reasoning, Logic Programming, and n-Person Games*, Artificial Intelligence **77** (1995), pp. 321–357.

[15] Egly, U., S. A. Gaggl and S. Woltran, *Aspartix: Implementing argumentation frameworks using answer-set programming*, in: *International Conference on Logic Programming*, Springer, 2008, pp. 734–738.

[16] Fazzinga, B., A. Galassi and P. Torroni, *An argumentative dialogue system for covid-19 vaccine information*, in: *International Conference on Logic and Argumentation*, Springer, 2021, pp. 477–485.

[17] Gabbay, D., G. Kern-Isberner, G. Simari and M. Thimm, editors, **3**, College Publications, 2024.

[18] Kampik, T., K. Čyras and J. R. Alarcón, *Change in quantitative bipolar argumentation: Sufficient, necessary, and counterfactual explanations*, International Journal of Approximate Reasoning **164** (2024), p. 109066.

[19] Müller, M. A., S. Vesic and B. Yun, *Interpreting preferred semantics in structured bipolar argumentation* (2025).

[20] Niskanen, A., D. Neugebauer, M. Järvisalo and J. Rothe, *Deciding acceptance in incomplete argumentation frameworks*, in: *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI-20)* (2020), pp. 2942–2949.

[21] Niskanen, A., J. P. Wallner and M. Järvisalo, *Extension enforcement under grounded semantics in abstract argumentation*, in: *Sixteenth International Conference on Principles of Knowledge Representation and Reasoning*, 2018.

[22] Nouioua, F. and V. Risch, *Argumentation frameworks with necessities*, in: *International Conference on Scalable Uncertainty Management*, Springer, 2011, pp. 163–176.

[23] Oren, N. and T. J. Norman, *Semantics for evidence-based argumentation*, in: *Computational Models of Argument*, IOS Press, 2008 pp. 276–284.
[24] Polberg, S., *Intertranslatability of abstract argumentation frameworks*, Technical Report DBAI-TR-2017-104, Institute for Information Systems, Technical University of Vienna (2017).
[25] Potyka, N., *Continuous dynamical systems for weighted bipolar argumentation*, in: M. Thielscher, F. Toni and F. Wolter, editors, *Principles of Knowledge Representation and Reasoning: Proceedings of the Sixteenth International Conference, KR 2018, Tempe, Arizona, 30 October - 2 November 2018* (2018), pp. 148–157.
[26] Rago, A., O. Cocarascu, C. Bechlivanidis, D. Lagnado and F. Toni, *Argumentative explanations for interactive recommendations*, Artificial Intelligence **296** (2021), p. 103506.
[27] Rago, A., O. Cocarascu and F. Toni, *Argumentation-based recommendations: Fantastic explanations and how to find them.*, , **18**, 2018, pp. 1949–1955.
[28] Rienstra, T., C. Sakama, L. van der Torre and B. Liao, *A principle-based robustness analysis of admissibility-based argumentation semantics*, Argument & Computation (2020), pp. 1–35.
[29] Rienstra, T., L. van der Torre and L. Yu, *Reasoning alignment for Agentic AI: Argumentation, belief revision, and dialogue*, Journal of Applied Logics - IfCoLog Journal **12** (2025), pp. 1683–1712.
[30] Sabharwal, N. and A. Agrawal, *Introduction to google dialogflow*, Cognitive virtual assistants using google dialogflow: develop complex cognitive bots using the google dialogflow platform (2020), pp. 13–54.
[31] Yu, L., C. Al Anaissy, S. Vesic, X. Li and L. van der Torre, *A principle-based analysis of bipolar argumentation semantics*, in: *European Conference on Logics in Artificial Intelligence*, Springer, 2023, pp. 209–224.
[32] Yu, L., R. Markovich and L. van der Torre, *Interpretations of support among arguments*, in: *Legal Knowledge and Information Systems*, IOS Press, 2020 pp. 194–203.
[33] Yu, L. and L. van der Torre, *A principle-based approach to bipolar argumentation*, in: *NMR 2020: Non-Monotonic Reasoning Workshop Notes*, 2020, CEUR Workshop Proceedings, Vol. 2672.
[34] Yu, L. and L. van der Torre, *The A-BDI metamodel for human-level AI: Argumentation as balancing, dialogue and inference*, in: *International Conference on Logic and Argumentation*, Springer, 2025, pp. 361–379.
[35] Yu, L., L. van der Torre and R. Markovich, *Thirteen challenges of formal and computational argumentation*, in: M. Thimm and G. R. Simari, editors, *Handbook of Formal Argumentation, Volume 3*, forthcoming .