# On the Extension of Argumentation Logic

Antonis Kakas[1,†], Paolo M. Mancarella[2,*,†]

[1] Department of Computer Science, University of Cyprus, Cyprus, antonis@ucy.ac.cy
[2] Department of Computer Science, University of Pisa, Italy, paolo.mancarella@unipi.it

**Abstract**

This paper shows how Argumentation Logic can be further extended to cover more fully paraconsistent forms of logical reasoning. The extension is based on the notion of non-acceptable self-defeating arguments as a generalization of the Reductio ad Absurdum principle.

**Keywords**

Propositional Logic, Argumentation Logic, Para consistency, Reductio ad Absurdum

## 1. Motivation and Background

Argumentative inference relies on the central normative condition of the acceptability of a (set of) argument(s). Informally, this condition states that "a (set of) argument(s) is acceptable only when it defends against all its counter-arguments". An acceptable argument thus forms a "case" that supports *satisfactorily* its claim and hence the claim is a possible or credulous conclusion under the argumentative reasoning. One way to formalise this notion of acceptability of arguments is via a recursive operator that first defines the more general notion of relative acceptability, $Acc(\Delta, \Delta_0)$, giving the acceptability of a set of arguments $\Delta$ with respect to a given set $\Delta_0$ and then projecting down to $Acc(\Delta, \{\})$ for the semantics.

This acceptability semantics for argumentation was first proposed in the context of Logic Programming [1, 2, 3] showing how it captures and extends the semantics of negation as failure. It was then an easy matter to apply this to abstract argumentation [4]. More recently, it was shown [5] that this type of recursive acceptability semantics for argumentation can be applied to formal logical reasoning where arguments are sets of logical formulae, e.g., propositional formulae. In this, an acceptable case of arguments corresponds to a set of formulae which can be enveloped in a model of the theory and thus a credulous conclusion corresponds to a *satisfiable* formula. We can then show that such a form of Argumentation Logic (AL) is logically equivalent to classical Propositional Logic (PL).

This equivalence holds only when reasoning under a set of given premises that are classically consistent. When the premises are inconsistent, AL does not trivialize but smoothly extends PL into a paraconsistent logic. Technically, AL does this by encompassing in its form of logical reasoning the notion of proof by contradiction in a way that prevents this from using an inconsistency in one part of the theory to derive any conclusion that may be "unrelated" to the inconsistency. In argumentation, Reduction ad absurdum is captured through the notion of *non-acceptability* of arguments, namely the contrary notion of acceptability of arguments. Non-acceptable arguments are "self-defeating" arguments. Informally, such an argument is one that either forms a counter-argument to itself or that it is a counter-argument to an argument that it necessarily needs in order to defend against some counter-argument to it. In other words, a non-acceptable or self-defeating argument invalidates its possible case of support by rendering the set of arguments in the case incompatible with each other.

In this paper, we will explore further the notion of non-acceptable arguments and study how this can give in the AL reformulation of PL new acceptable sets of arguments (under inconsistent premises) that were not recognized as such before. The main idea is that we can extend the notion of acceptability of a set of arguments $\Delta$ by requiring that any counter-argument $A$ against this is either, as before,

*CILC 2024: 39th Italian Conference on Computational Logic, June 26-28, 2024, Rome, Italy*
*Corresponding author.
†These authors contributed equally.

defended against explicitly by some other argument, or by recognizing that $A$ is by itself non-acceptable or self-defeating. In this second possibility a counter-argument is dealt with by showing, in analogy to proof by contradiction, that it is by itself invalid and hence it cannot affect the acceptability of $\Delta$. Whereas in the previous work in [5] this was used only for the limiting case of non-acceptability of a self-attacking counter-argument, in this paper we will show how more complex forms of self-defeating non-acceptable arguments can be identified and used to "neutralize" the effect of such arguments when they appear as counter-arguments to other arguments.

Section 2 reviews the acceptability semantics for general abstract argumentation frameworks under which the classical Propositional Logic is reformulated as an Argumentation logic. Section 3 defines the proposed general extension of the acceptability semantics. Section 4 applies the general theory to the specific case of AL as a reformulation of PL and shows how this extends the existing definition of AL. Section 5 concludes with a discussion of future work (further possible extensions of AL, e.g. for directly inconsistent premises, where we just extend suitably the defense relation).

## 2. Acceptability semantics of Argumentation

Let us briefly review the area of Abstract Argumentation and its semantics [4, 6] as developed and used in the area of Artificial Intelligence. In abstract argumentation we are not interested in the internal structure of arguments but only in their relative properties. An abstract argumentation framework is defined as follows.

**Definition 1.** *[Abstract Argumentation Framework]*
*An abstract argumentation framework is a triple, $\langle Arg, Att, \mathcal{D}ef\rangle$, where*

- *$Arg$ is a set (of arguments)*
- *$Att$ is a binary (partial) relation on $Arg$ (attack relation)*
- *$\mathcal{D}ef$ is a binary (partial) relation on $Arg$ (defense relation)*

Given $A, \Delta, D \subseteq Arg$, we say that $A$ **attacks** $\Delta$ (written $A \rightsquigarrow \Delta$) iff there exists $a \in A$ and $b \in \Delta$ such that $(a, b) \in Att$ and that $D$ **defends against** $A$ (written $D \twoheadrightarrow A$) iff $(d, c) \in \mathcal{D}ef$ for some $d \in D$ and $c \in A$.

This definition differs from the usual classical definition used in the area of abstract argumentation, where an argumentation framework is simply a tuple $AF^c = \langle Arg, Att^c\rangle$, consisting of arguments and an attack relation $Att^c$ between arguments. In the triple argumentation frameworks defined above, we expand the classic attack relation $Att^c$ into its two relations, i.e. whenever $(a, b) \in Att^c$ then $(a, b) \in Att$ and $(a, b) \in \mathcal{D}ef$. This transformation makes explicit the two properties captured by the classic attack relation, namely when $(a, b) \in Att^c$ then (i) the argument $a$ can form a counter-argument to $b$ and (ii) $a$ can defend against $b$ when this is a counter-argument (to any argument).

A typical realization of a triple argumentation framework in some language, $\mathcal{L}$, for constructing and comparing arguments is given by: (1) $a$ is in conflict in $\mathcal{L}$ with $b$ for $(a, b) \in Att$ to hold and (2) $a$ is at least as strong in $\mathcal{L}$ as $b$ for $(a, b) \in \mathcal{D}ef$ to hold. In such realizations, the attack relation is symmetric and the defense relation is a subset of the attack relation. The detailed study of these links is beyond the scope of this paper.

The semantics of an abstract argumentation framework is defined via subsets of arguments that satisfy an acceptability property, $\mathcal{A}cc(\Delta, \Delta_0)$, whose informal meaning is that the set of arguments $\Delta$ is acceptable in the context of a given set of arguments $\Delta_0$, only when $\Delta$ can defend against all its counter-arguments. The precise definition of the acceptability property is given as follows.

**Definition 2.** *[Acceptability property]*
*Let $AF = \langle Arg, Att, \mathcal{D}ef\rangle$ be an abstract argumentation framework and $\Delta, \Delta_0 \subseteq Arg$. Then:*

- *$\mathcal{A}cc(\Delta, \Delta_0)$ iff*

- $\Delta \subseteq \Delta_0$, *or*
- *for any* $A \subseteq Arg$ *such that* $A \rightsquigarrow \Delta$*:*
  - $A \not\subseteq \Delta \cup \Delta_0$, *and*
  - *there exists* $D \subseteq Arg$ *such that* $D \twoheadrightarrow A$ *and* $\mathcal{A}cc(D, \Delta \cup \Delta_0)$

We can thus see that for $\Delta$ to be acceptable in the context of $\Delta_0$ all its counter-arguments must be defended by arguments that are themselves acceptable in the extended context of $\Delta \cup \Delta_0$. Extending the context in this way means that (a chosen set of arguments) $\Delta$ can contribute to its own defense.

Formally, the acceptability property is defined through the least fixed point of an associated monotonic operator on the binary Cartesian product of sets of arguments

$$\mathcal{R} = 2^{Arg} \times 2^{Arg}$$

**Definition 3.** *[Acceptability operator]*
*Let* $AF = \langle Arg, Att, \mathcal{D}ef \rangle$ *be an abstract argumentation framework. The acceptability operator* $\mathcal{A} : \mathcal{R} \to \mathcal{R}$ *is defined as follows. Given* $r \in \mathcal{R}$ *and* $\Delta, \Delta_0 \subseteq Arg$, $(\Delta, \Delta_0) \in \mathcal{A}(r)$ *iff:*

- $\Delta \subseteq \Delta_0$, *or*
- *for any* $A \subseteq Arg$ *such that* $A \rightsquigarrow \Delta$*:*
  - $A \not\subseteq \Delta \cup \Delta_0$, *and*
  - *there exists* $D \subseteq Arg$ *such that* $D \twoheadrightarrow A$ *and* $(D, \Delta \cup \Delta_0) \in r$

We denote by $\mathcal{A}^{fix}$ the least fixed point of this operator. Then the semantics of an argumentation framework is given through the subsets of arguments $\Delta$ that are acceptable with respect to the empty set of arguments, i.e. such that $(\Delta, \{\}) \in \mathcal{A}^{fix}$ holds. We say that such sets of arguments are *acceptable.*

**Example 1.**
*Let* $AF = \langle Arg, Att, \mathcal{D}ef \rangle$ *be the abstract argumentation framework where*

- $Arg = \{a, b\}$
- $Att = \{(a, b), (b, a)\}$
- $\mathcal{D}ef = \{(b, a)\}$

*In this framework its two arguments attack each other but only argument* $b$ *is able to defend against its counter-argument of* $a$*, e.g., because* $b$ *is stronger than* $a$*. We can then see that the set* $\{b\}$ *is acceptable whereas the set* $\{a\}$ *is not acceptable as it cannot defend against its counter-argument* $A = \{b\}$*. Instead, if the defense relation contained also* $(a, b)$*, e.g., when the two arguments are of equal strength, then both* $\{a\}$ *and* $\{b\}$ *would be acceptable sets of arguments.*

To illustrate the basic idea of the problem for the need of extending the semantics of argumentation let us consider the following example.

**Example 2.** *[Motivating Example 1]*
*Let* $AF = \langle Arg, Att, \mathcal{D}ef \rangle$ *be the abstract argumentation framework where*

- $Arg = \{a, b\}$
- $Att = \{(a, a), (a, b)\}$
- $\mathcal{D}ef = \{\}$

*The argument set* $\{b\}$ *is not acceptable as it cannot defend against its attack by* $\{a\}$*: there are no arguments that could be used as a defense. Nevertheless,* $\{a\}$ *is itself self-attacking and hence we would expect that it is not necessary to find an explicit way of defending against it. Thus we would want the argument set* $\{b\}$ *to be acceptable.*

This idea that arguments that are themselves self-attacking or more generally, as we shall see below in this paper, self-defeating was first studied in the context of the argumentation-based semantics of Negation as Failure in Logic Programming (see [1, 2, 3] and references therein). These ideas were then lifted to the case of Abstract Argumentation in [4] showing their general applicability.

Both in the case of Logic Programming and Abstract Argumentation the approach is based on first formulating a relative notion of acceptability of arguments, as we have reviewed above in Definitions 2 and 3, from which we then extract or project to notions of absolute acceptability of arguments (recently a semantically equivalent reformulation of relative acceptability was proposed in [7]). This approach to extending the semantics of argumentation via a relative notion of acceptability is indirectly related to a variety of other approaches [8, 9, 10] most of which are studies of how to address odd loops in the attack relation of an argumentation framework. In particular, in [10] where, under the labeling semantics of argumentation, arguments can be labeled IN, OUT or UNDECIDED, there is a close link between this later label of UNDECIDED and the notion of self-defeated arguments under the relative acceptability semantics.

## 2.1. Propositional Logic as Argumentation Logic

An important application of the relative acceptability semantics is that of the reformulation of classical Propositional Logic in terms of argumentation [5]. We will briefly review this reformulation and its paraconsistent extension of Argumentation Logic as a realization of the abstract argumentation framework and its acceptability semantics.

**Definition 4.** *[Argumentation Logic Framework]*
*We denote by $\vdash_{MRA}$ the Natural Deduction direct derivation relation of propositional logic modulo Reduction ad Absudrum (MRA), i.e. without the proof rule of Reduction ad Absudrum.*

*Let $T$ be a propositional theory. The argumentation logic framework corresponding to $T$ is the triple $AF^T = \langle Args, Att, Def \rangle$ with:*

- $Arg = \{\Sigma \mid \Sigma$ *is a finite set of propositional sentences*$\}$
- *given* $\Delta, \Gamma \in Arg$, *with* $\Delta \neq \{\}$, $(\Gamma, \Delta) \in Att$ *iff* $T \cup \Gamma \cup \Delta \vdash_{MRA} \bot$
- *given* $\Delta \in Arg$, $(\{\overline{\phi}\}, \Delta) \in Def$, *where* $\overline{\phi}$ *is the complement of some sentence* $\phi \in \Delta$ *and* $(\{\}, \Delta) \in Def$ *whenever* $T \cup \Delta \vdash_{MRA} \bot$.

We see that the attack relation is symmetric, i.e. arguments are always counter-arguments of each other when together they are directly inconsistent in the context of the given premises $T$. The defense relation essentially expresses the fact that any argument can be defended against by *undermining* one of its premises. In logical terms, the defense relation expresses the property that for any formula $\phi$ we are free to choose this or its complement. The second part of the defense relation expresses the fact that if an argument is self-inconsistent with respect to the given premises, then this can be trivially defended against by the "safe" empty argument (which in turn can not be attacked). We will see below that when we extend the acceptability semantics, this second part of the defense relation will not need to be stated explicitly at this level, but will be captured at the extended acceptability semantic level.

We will denote by $\mathcal{AL}^{fix}$ (or simply by $\mathcal{AL}$) the least fixpoint of the corresponding operator $\mathcal{A}$ in the general abstract argumentation frameworks as above in definition 3. We then have a logical correspondence between propositional logic (for classically consistent premises $T$) and the argumentation acceptability semantics [5]. For any formula $\phi$: $\phi$ is acceptable, i.e., $(\{\phi\}, \{\}) \in \mathcal{AL}^{fix}$ if and only if there is a model of $T$ in which $\phi$ is true. Furthermore, for classically inconsistent premises which are directly consistent, i.e. consistent under the restricted derivation of $\vdash_{MRA}$, the argumentation semantics does not trivialize but smoothly extends the propositional deductive semantics into such cases of inconsistent premises.

The full technical details of these results can be found in [5]. For the purposes of this paper it is important to point out that the results rest on the correspondence between proofs via Reductio ad Absurdum and the non-acceptability of formulae, namely that for any formula $\phi$: $(\{\phi\}, \{\}) \notin \mathcal{AL}^{fix}$

holds if and only if there exists in Natural Deduction a restricted[1] Reductio ad Absurdum proof for the complement of $\phi$, i.e. for $\overline{\phi}$. Hence, if a posited formula $\phi$ is shown via the restricted form of Reductio ad Absurdum to lead to an inconsistency, then the argument set $\{\phi\}$ can not be acceptable. This means that some argument, $A$, that attacks $\{\phi\}$, i.e. it is directly inconsistent with it, cannot be defended against by some set of formulae $D$ that is acceptable in the context of $\{\phi\}$. In other words, the argument $\{\phi\}$ *defeats* its possible defenses, it is *self-defeating*.

## 3. Non-acceptable Arguments

In this section, we will examine further the nature of non-acceptable arguments and the relative defeatedness of such arguments in the context of a given set of arguments. In particular, we will be interested in a sub-case of non-acceptability that relates to arguments that are defeated in their own context.

Let us return to abstract argumentation and consider the following example.

**Example 3.** *[Motivating Example 2]*
*Let $AF = \langle Arg, Att, \mathcal{D}ef \rangle$ be the abstract argumentation framework where*

- $Arg = \{a, b\}$
- $Att = \{(a, b)\}$
- $\mathcal{D}ef = \{\}$

*Argument set $\{b\}$ is attacked by argument set $\{a\}$. Trivially then, $(\{b\}, \{a\}) \notin \mathcal{A}^{fix}$, i.e. $\{b\}$ is non-acceptable in the context of $\{a\}$, as $\{b\}$ is attacked by an argument that belongs to the context . We will also say that $\{b\}$ is **defeated in the context of** $\{a\}$. Similarly, if we have that $(b, a) \in Att$, i.e. $\{b\}$ also attacks $\{a\}$, then $\{a\}$ is defeated in the context of $\{b\}$. Note also that these particular contextual defeat cases hold irrespective of what is contained in the defense relation. Hence even if we had the tuple $(b, a) \in \mathcal{D}ef$ it would still be the case that $\{b\}$ is non-acceptable or defeated in the context of $\{a\}$.*

Let us return to the first motivating example.

**Example 4.** *[Motivating Example 1 cont.]*
*Let $AF = \langle Arg, Att, \mathcal{D}ef \rangle$ be the abstract argumentation framework where*

- $Arg = \{a, b\}$
- $Att = \{(a, a), (a, b)\}$
- $\mathcal{D}ef = \{\}$

*The argument set $\{a\}$ is self-attacking and hence it is non-acceptable or defeated in its own context. We consider this argument as a **self-defeating argument** exactly because it contains an (one of its) attack. Note that the property of $\{a\}$ being self-defeating is not affected by the argument $\{b\}$.*

Recognizing this property of self-defeatness of arguments, it is reasonable to require that other arguments, whose counter-arguments are such self-defeated arguments, are acceptable (wrt $\{\}$). For instance, in the above example it is reasonable to accept that argument $\{b\}$ is acceptable (in the context of $\{\}$), because its only attack is "self-defeating". In other words, it is not necessary to explicitly defend against such a self-defeating counter-argument, as this attack is an argument that invalidates itself.

The above example shows a simple (and limiting) case of a non-acceptable self-defeating argument. More complex forms of such arguments exist, as it is illustrated in the next example.

[1]This restriction requires that the posited hypothesis must be necessary for its inconsistency to be derived (see [5]). For classically consistent premises $T$ such a restricted proof always exists when a non-restricted ordinary Reductio ad Absurdum proof exists for the same posited hypothesis.

**Example 5.** *[Motivating Example 3]*
*Let $AF = \langle Arg, Att, \mathcal{D}ef \rangle$ be the abstract argumentation framework where*

- $Arg = \{a, b, a_1, d_1\}$
- $Att = \{(a, b), (a_1, a), (a, d_1), (d_1, a_1)\}$
- $\mathcal{D}ef = \{(d_1, a_1)\}$

*Argument $a$ is attacked by $a_1$ which can only be defended against by argument $d_1$. But $a$ attacks this defence of $d_1$, i.e. $d_1$ is defeated in the context of $a$. Hence, as in the example above, $a$ is non-acceptable and we can consider it as self-defeating, but now in an indirect way, because $a$ renders its necessary defending argument(s) non-acceptable or defeated in its own context. If we thus accept that $a$ is self-defeating then again, as in the example above, we would want that argument $b$, which is attacked only by $a$, to be an acceptable argument.*

These more complex forms of self-defeated arguments arise from the recursive nature of non-acceptability given by negating the recursive definition of acceptability.

**Proposition 1.** *[Non-acceptability]*
*Let $AF = \langle Arg, Att, \mathcal{D}ef \rangle$ be an abstract argumentation framework and $\Delta, \Delta_0 \subseteq Arg$. Let $non\_Acc(\Delta, \Delta_0)$ denote the statement $(\Delta, \Delta_0) \notin \mathcal{A}^{fix}$. Then the following holds directly from the definition of acceptability:*

- *$non\_Acc(\Delta, \Delta_0)$ iff $\Delta \not\subseteq \Delta_0$ and*
  $\exists A \subseteq Arg$ *such that* $A \rightsquigarrow \Delta$ *and*
  - $A \subseteq \Delta \cup \Delta_0$, *or*
  - $\forall D \subseteq Arg$ *s.t.* $D \twoheadrightarrow A$: $non\_Acc(D, \Delta \cup \Delta_0)$.

Hence, a general non-acceptable argument $A$ such that $(A, \{\}) \notin \mathcal{A}^{fix}$ holds, is one where, when we collect recursively the defenses against one of its counter-arguments and recursively the further defenses against attacks of the earlier defenses, we end up with a collection of defenses that is self-attacking.

**Definition 5.** *[Self-defeating sets of arguments]*
*Let $A, A' \subseteq Arg$. We say that $A$ is defeated in the context of $A'$ iff $non\_Acc(A, A')$ holds. When a set is defeated in the context of the empty set, i.e. $non\_Acc(A, \{\})$ holds, we say that $A$ is self-defeating and denote this by $\mathcal{SD}(A)$.*

**Proposition 2.**
*Let $AF = \langle Arg, Att, \mathcal{D}ef \rangle$ be an abstract argumentation framework.*

*(i) If $(a, b) \in Att$, then $non\_Acc(\{b\}, \{a\})$ holds;*
*(ii) If $\Delta$ is self-attacking then it is self-defeating, i.e. $\mathcal{SD}(A)$ holds.*

**Proof.**

(i) $\{a\} \rightsquigarrow \{b\}$ and $\{a\} \subseteq \{b\} \cup \{a\}$. Hence $non\_Acc(\{b\}, \{a\})$.
(ii) Obvious, since $\Delta \rightsquigarrow \Delta$ and $\Delta \subseteq \Delta \cup \{\}$.

## 4. Extended Acceptability semantics

The extension of the notion of acceptability of arguments follows the simple idea that counter-arguments that are non-acceptable or self-defeating can be dealt with without the need to explicitly defend against them. It is sufficient to recognize that such attacks are self-defeating.

**Definition 6.** *[Extended Acceptability]*
*Let $AF = \langle Arg, Att, \mathcal{D}ef \rangle$ be an abstract argumentation framework and $\Delta, \Delta_0 \subseteq Arg$. Then a set of arguments $\Delta$* **is acceptable in the context of** *$\Delta_0$, denoted by $\mathcal{A}cc^+(\Delta, \Delta_0)$, when the following holds:*

$\mathcal{A}cc^+(\Delta, \Delta_0)$ *iff*
- $\Delta \subseteq \Delta_0$, *or*
- *for any $A \subseteq Arg$ such that $A \rightsquigarrow \Delta$:*
  - $A \not\subseteq \Delta \cup \Delta_0$, *and*
  - $(A, \{\}) \notin \mathcal{A}^{fix}$, *or there exists $D \subseteq Arg$ such that $D \twoheadrightarrow A$ and $(D, \Delta \cup \Delta_0) \in \mathcal{A}^{fix}$*

**Proposition 3.**
*Let $AF = \langle Arg, Att, \mathcal{D}ef \rangle$ be an abstract argumentation framework and $\Delta, \Delta_0 \subseteq Arg$. Then*

$$(\Delta, \Delta_0) \in \mathcal{A}^{fix} \implies \mathcal{A}cc^+(\Delta, \Delta_0)$$

**Proof**. Straightforward by the definition of $\mathcal{A}^{fix}$ and $\mathcal{A}cc^+$.

**Example 6.** *[Motivating Example 1 Revisited]*
*In both of these examples $(\{b\}, \{\})$ does not belong to $\mathcal{A}^{fix}$, i.e. the argument set $\{b\}$ is not acceptable. However, $\mathcal{A}cc^+(\{b\}, \{\})$ holds because the only (minimal) attack against $\{b\}$, namely the set $\{a\}$, is self-defeating. Hence the argument set $\{b\}$ is acceptable in the extended semantics.*

### 4.1. AL$^+$: Extended Argumentation Logic

We will now revisit the reformulation of classical Propositional Logic and its paraconsistent extension by Argumentation Logic as a realization of the abstract argumentation framework. We will apply the extended acceptability semantics of the general Definition 6 to the case of Argumentation Logic. This will give an extended form of argumentation logic that takes more fully into account the presence of non-acceptable arguments. In effect, this will give a generalized use of the principle of proof by contradiction under inconsistent premises.

**Definition 7.** *[Extended Argumentation Logic]*
*Let $AF^T = \langle Args, Att, Def \rangle$ be the argumentation logic framework corresponding to a (directly consistent) propositional theory $T$. The extended argumentation logic, $\mathcal{AL}^+$, is given by:*

*$\mathcal{AL}^+(\Delta, \{\})$ holds iff for any $A \subseteq Arg$ such that $A \rightsquigarrow \Delta$:*

- $A \not\subseteq \Delta$, *and*
- $(A, \{\}) \notin \mathcal{AL}$, *or* $\exists\ D \subseteq Arg$ *such that $D \twoheadrightarrow A$ and $(D, \Delta) \in \mathcal{AL}$*

Hence a set of formulae is acceptable in $\mathcal{AL}^+$ either because its attacks could be defended acceptably, as before in the basic logic of $\mathcal{AL}$, or its attacks are non-acceptable in $\mathcal{AL}$.

The following result shows that the extended argumentation logic, $\mathcal{AL}^+$, is a "proper" extension of $\mathcal{AL}$ when the given premises $T$ are classically consistent.

**Theorem 1.** *Let $T$ be a classically consistent theory and $AF^T = \langle Arg, Att, \mathcal{D}ef \rangle$ its corresponding argumentation logic framework. Let also $\phi$ be a propositional formula such that $(\{\phi\}, \{\}) \notin \mathcal{AL}$ holds. Then $\mathcal{AL}^+(\{\phi\}, \{\})$ does not hold.*

**Proof.** This is a technical result whose proof is in the Appendix.

This means that the extension of the logic does not trivialize the original logic and specifically classical Propositional Logic for consistent premises.

**Corollary 1.** *Let $T$ be a classically consistent theory and $AF^T = \langle Arg, Att, \mathcal{D}ef \rangle$ its corresponding argumentation logic framework. Let also $\phi$ be a propositional formula such that $T \not\models \phi$. Then $T \not\models_{AL^+} \phi$*[2].

**Proof.** Follows directly from the theorem and the equivalence of propositional logic with argumentation logic, i.e. that $T \not\models \phi$ iff $T \not\models_{AL} \phi$.

We also know from proposition 3 that $\mathcal{AL}^+$ contains the original logic of $\mathcal{AL}$. Hence $\mathcal{AL}^+$, is a conservative extension of Propositional Logic and of $\mathcal{AL}$ when the given premises $T$ are classically consistent.

The following example taken from [5] clarifies the link between the extended $AL^+$ and the original $AL$ and how it gives genuinely new cases of acceptable formulae.

**Example 7.**
*Consider the following two theories of propositional logic:*

- $T_1 = \{\neg(\beta \wedge \alpha), \neg\alpha\}$ $\qquad$ $T_2 = \{\neg(\beta \wedge \alpha), \neg(\alpha \wedge \gamma), \neg(\alpha \wedge \neg\gamma)\}$

*It is easy to see that the argument $\{\beta\}$ is acceptable in $\mathcal{AL}$ relative to theory $T_1$. Its minimal attack $\{\alpha\}$ is directly self-inconsistent and hence self-attacking (i.e. $T_1 \cup \{\alpha\} \vdash_{MRA} \bot$) and so it can be defended by $\{\}$. The argument $\{\beta\}$ is also acceptable in $\mathcal{AL}$ relative to theory $T_2$ even though its attack $\alpha$ is not directly inconsistent. The defense against the attack of $\{\alpha\}$, namely $\{\neg\alpha\}$, is such that $(\{\neg\alpha\}, \{\beta\}) \in \mathcal{AL}$. Notice, however, that this attack of $\{\alpha\}$, is itself a non-acceptable self-defeating argument, as it cannot defend acceptably against its attack by $\{\gamma\}$: the only possible defense of $\{\neg\gamma\}$ in non-acceptable in the context of $\{\alpha\}$ because $\{\alpha\}$ attacks $\{\neg\gamma\}$. Therefore recognizing the non-acceptability of the attack $\{\alpha\}$ is an alternative way to enforce the acceptability of $\{\beta\}$. The extended acceptability semantics of $\mathcal{AL}^+$ uses this alternative way. Importantly, it does so in the same way for both theories $T_1, T_2$.*

The extended acceptability semantics becomes relevant when the theory of premises is inconsistent, and attacks like $\{\alpha\}$ above cannot be defended acceptably by $\{\neg\alpha\}$.

**Example 8.** *[Example 7 cnt.]*
*Consider the following theory, obtained from $T_2$ by making also $\neg\alpha$ non acceptable: $T_3 = T_2 \cup \{\neg(\neg\alpha \wedge \delta), \neg(\neg\alpha \wedge \neg\delta)\}$. The attack $\{\alpha\}$ cannot be acceptably defended because the possible defense of $\{\neg\alpha\}$ is non-acceptable in a way similar to the non-acceptability of $\{\alpha\}$ shown above (replacing $\{\gamma\}$ with $\{\delta\}$). Nevertheless, as $\{\alpha\}$ is by itself non-acceptable it is reasonable to accept $\{\beta\}$ as acceptable, as $\mathcal{AL}^+$ does.*

Finally, we point out that in the extended Argumentation Logic, $AL^+$, we can drop the second element of the defense relation on the formulation of AL in definition 4, namely that now the defence relation does not need to contain that the empty argument defends against any self-attacking argument. Indeed, this is now covered by the extended semantics, as any directly self-inconsistent arguments are non-acceptable because they are self-attacking and hence such attacks will be taken into account as harmless by the semantics of $AL^+$.

[2]Here $\models$ denotes the classical entailment of Propositional Logic. The entailment relation $\models_{AL^+}$ in $\mathcal{AL}^+$, is defined (as it is for $\models_{AL}$ in $\mathcal{AL}$) by $T \models_{AL^+} \phi$ iff $\mathcal{AL}^+(\{\phi\}, \{\})$ holds and $\mathcal{AL}^+(\{\neg\phi\}, \{\})$ does not hold.

## 5. Conclusions

We have shown how to extend Argumentation Logic to capture the intuitive idea that for attacks which are by themselves self-defeating it is not necessary to defend against. The proposed extended Argumentation Logic clarifies how reasoning to a conclusion can be achieved either by showing explicitly how the conclusion is supported or by showing that opposing conclusions are by themselves invalid. This resonates with classical logical reasoning where conclusions are either directly derived from the premises or their oppositions are shown, via Reductio ad Absurdum, to be inconsistent with the given premises.

The extended argumentation semantics is based on definition 6. We can then consider applying this definition iteratively to give possible further extensions of acceptability and study the properties of such extensions. We can also study how we can further extend the framework to allow directly inconsistent premises. For example, we can examine how we can accommodate this by simply extending the defense relation so that any two subsets of the premises which are directly inconsistent with each other are able to defend against each other.

## References


[1] A. Kakas, R. Kowalski, F. Toni, Abductive Logic Programming, Journal of Logic and Computation 2 (1992) 719–770.

[2] A. C. Kakas, P. Mancarella, P. M. Dung, The acceptability semantics for logic programs, in: Proc. of 11th Int. Conf. on Logic Programming, 1994, pp. 504–519.

[3] A. C. Kakas, F. Toni, Computing argumentation in logic programming, J. Log. Comput. 9 (1999) 515–562.

[4] A. Kakas, P. Mancarella, On the semantics of abstract argumentation., Journal of Logic and Computation [electronic only] 23 (2013) 991–1015.

[5] A. Kakas, P. Mancarella, F. Toni, On Argumentation Logic and Propositional Logic, Studia Logica 106 (2018) 237–279.

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

[7] R. Baumann, G. Brewka, M. Ulbricht, Shedding new light on the foundations of abstract argumentation: Modularization and weak admissibility, Artif. Intell. 310 (2022) 103742.

[8] P. Baroni, M. Giacomin, G. Guida, Scc-recursiveness: a general schema for argumentation semantics, Artif. Intell. 168 (2005) 162–210.

[9] G. A. Bodanza, F. A. Tohmé, Two approaches to the problems of self-attacking arguments and general odd-length cycles of attack, J. Appl. Log. 7 (2009) 403–420.

[10] M. W. A. Caminada, D. M. Gabbay, A logical account of formal argumentation, Stud Logica 93 (2009) 109–145.


## 6. Appendix: Proof of Theorem 1

We will use the following Lemma.

**Lemma 1.** *Let $T$ be a directly consistent theory and $AF^T = \langle Arg, Att, \mathcal{D}ef\rangle$ its corresponding argumentation logic framework. Let also*

$$\phi_1, \phi_2, \ldots, \phi_k, \qquad\qquad k \geq 1$$

*be a sequence of formulae such that:*

*(i)* $\phi_i \neq \phi_j$, *for each* $i \neq j$

*(ii)* $\{\phi_i\}$ *attacks* $\{\overline{\phi_{i-1}}\}$, *for each* $i = 2, \ldots, k$

*(iii)* $(\{\overline{\phi_k}\}, \{\overline{\phi_1}, \ldots, \overline{\phi_{k-1}}\}) \notin \mathcal{AL}$

*Then*

$$(\overline{\phi_1}, \{\}) \notin \mathcal{AL}$$

**Proof.** If $k = 1$ the required result is given by condition (iii). Hence let $k \geq 2$. We show

$$(\{\overline{\phi_{k-1}}\}, \{\overline{\phi_1}, \ldots, \overline{\phi_{k-2}}\}) \notin \mathcal{AL} \qquad (*)$$

Let $\Delta_0 = \{\overline{\phi_1}, \ldots, \overline{\phi_{k-2}}\}$ and $\Delta = \{\overline{\phi_{k-1}}\}$. By condition (ii), $\{\phi_k\}$ attacks $\{\overline{\phi_{k-1}}\}$. The only defence against this attack is $\{\overline{\phi_k}\}$ which is non-acceptable with respect to $\Delta \cup \Delta_0$ by (iii). Hence (*) holds. By iterating this process $k - 1$ times, we obtain the sequence of facts

$$(\{\overline{\phi_{k-2}}\}, \{\overline{\phi_1}, \ldots, \overline{\phi_{k-3}}\}) \notin \mathcal{AL}$$

$$\ldots$$

$$(\overline{\phi_1}, \{\}) \notin \mathcal{AL}$$

**Theorem 1** Let $T$ be a classically consistent theory and $AF^T = \langle Arg, Att, \mathcal{D}ef \rangle$ its corresponding argumentation logic framework. Let also $\phi$ be a propositional formula such that $(\{\phi\}, \{\}) \notin \mathcal{AL}$ holds. Then $(\{\phi\}, \{\}) \notin \mathcal{AL}^+$ also holds.

**Proof.** Theorem 2 of [5] allows us without loss of generality to restrict consideration to singleton sets of arguments both for attacks as well as defenses when examining the acceptability or non-acceptability of argument sets. Let us denote by $\Delta$ the set $\{\phi\}$. We need to show that when $(\Delta, \{\}) \notin \mathcal{AL}$ then $(\Delta, \{\}) \notin \mathcal{AL}^+$. We will show this by contradiction. Assume both $(\Delta, \{\}) \notin \mathcal{AL}$ and $(\Delta, \{\}) \in \mathcal{AL}^+$ hold. By the definition of $\mathcal{AL}$, $(\Delta, \{\}) \notin \mathcal{AL}$ implies that there exists a sequence of tuples of formulae

$$a_i = \langle \phi_i, \overline{\phi_i} \rangle \qquad i = 1, ..., k \text{ for some natural number } k \neq 0 \qquad (\dagger)$$

such that $\{\phi_1\}$ attacks $\Delta$ and for each $i = 2, ..., k$:

- $\{\phi_i\}$ attacks $\{\overline{\phi_{i-1}}\}$ (a)
- $\{\overline{\phi_i}\}$ defends against $\{\phi_i\}$ (b)
- $(\{\overline{\phi_i}\}, \Delta \cup \{\overline{\phi_j} \mid j = 1, \ldots, i-1\}) \notin \mathcal{AL}$ (c)
- $\Delta \cup \{\overline{\phi_j} \mid j = 1, ..., k-1\}$ attacks $\{\overline{\phi_k}\}$ (d)

This is a sequence of consecutive defenses and attacks, starting from $\Delta$ as the first defense, such that all defenses in the sequence are non-acceptable with respect to the union of defenses that are prior to them. By the definition of non-acceptability, the third condition above also means that all these defences are not equal to each other, i.e. $\overline{\phi_n} \neq \overline{\phi_m}$ for any $n \neq m$, and hence also $\phi_n \neq \phi_m$ for any $n \neq m$.

Let us also assume that $k$ is the smallest number for which such a sequence exists for the non-acceptability of $\Delta = \{\phi\}$. If $k = 1$, then the non-acceptability of $\Delta$ comes from the fact that $\Delta$ is self-attacking (i.e., $\{\phi_1\} = \Delta$). By the definition of $\mathcal{AL}^+$, this is in direct contradiction with $(\Delta, \{\}) \in \mathcal{AL}^+$.

Hence $k > 1$. Notice also that, for any $n = 2, ..., k-1$, $\phi_n \neq \phi$, as otherwise $k$ would not be the least value for such a sequence.

we have two cases:

- Case 1: $\overline{\phi_j}$ attacks $\overline{\phi_k}$ for some $j \leq k$
  i.e. some previous defense $\overline{\phi_j}$ or itself attacks $\overline{\phi_k}$, or
- Case 2: $\Delta$ attacks $\overline{\phi_k}$.

**Case (1):** We can show that $(\{\overline{\phi_1}\}, \{\}) \notin \mathcal{AL}$ holds as follows. First notice that $\overline{\phi_j}$ attacks $\overline{\phi_k}$ implies $(\{\overline{\phi_k}\}, \{\overline{\phi_j}\}) \notin \mathcal{AL}$, by definition of $\mathcal{AL}$. Since all $\phi_i$ are different between them, we have that

$$(\{\overline{\phi_k}\}, \{\overline{\phi_1}, \overline{\phi_2}, ..., \overline{\phi_{j-1}}, \overline{\phi_j}, \overline{\phi_{j+1}}, ..., \overline{\phi_{k-1}}\}) \notin \mathcal{AL}$$

This, along with (a) above and, again, the fact that all $\phi_i$ are different between them, allow us to apply Lemma 1 and to conclude

$$(\{\overline{\phi_1}\}, \{\}) \notin \mathcal{AL}.$$

We also know from $(\Delta, \{\}) \in \mathcal{AL}^+$ that the attack of $\phi_1$ against $\Delta$ must also be non-acceptable, i.e., we also have $(\{\phi_1\}, \{\}) \notin \mathcal{AL}$ holds. These two facts lead (via Theorem 2 in [5]) to $T \models \phi_1$ and $T \models \overline{\phi_1}$ contradicting the classical consistency of $T$.

**Case 2: $\Delta$ attacks $\overline{\phi_k}$**

By the symmetry of the attack we also have that $\overline{\phi_k}$ attacks $\Delta$. This is another attack against $\Delta$. We will show that its defense $\phi_k$ is non-acceptable both in the empty context, i.e., $(\{\phi_k\}, \{\}) \notin \mathcal{AL}$ holds, and is also non-acceptable with respect to $\Delta$, i.e., also $(\{\phi_k\}, \Delta) \notin \mathcal{AL}$ holds. The latter means that the attack of $\overline{\phi_k}$ cannot be acceptably defended against and so by $(\Delta, \{\}) \in \mathcal{AL}^+$ this attack must be non-acceptable, i.e., $(\{\overline{\phi_k}\}, \{\}) \notin \mathcal{AL}$. These two results, as above in case 1, using Theorem 2 in [5]) give $T \models \phi_k$ and $T \models \overline{\phi_k}$ leading to a contradiction of the classical consistency of the premises $T$.

To show these two non-acceptability results we consider the sequence of formulae, $\psi_1, ..., \psi_k$ where $\psi_i = \phi_{k+1-i}$ for $i = 1, ..., k$. The sequence starts with $\psi_1 = \phi_k$ that defends against the attack of $\overline{\phi_k}$ on $\Delta$. From the way that the formulae $\phi_i$ have been constructed above and by the symmetry of attacks and the definition of defense in argumentation the formulae $\psi_i$ are different and the following hold:

(i) $\overline{\psi_i} \neq \overline{\psi_j}$, for each $i \neq j$
(ii) $\{\overline{\psi_i}\}$ attacks $\{\psi_{i-1}\}$, for each $i = 2, \ldots, k$
(iii) $(\{\psi_k\}, \{\}) \notin \mathcal{AL}$

We also have the following two conditions:

(iii) $(\{\psi_k\}, \{\}) \notin \mathcal{AL}$
(iv) $\Delta$ attacks $\psi_k$

These two conditions hold by the construction of the sequence of $\phi_i$ and $\psi_k = \phi_1$ ($\phi_1$ attacks $\Delta$ so the reverse also holds and $\phi_1$ cannot be acceptably defended by $\Delta$ so this attack must be non-acceptable for $(\Delta, \{\}) \in \mathcal{AL}^+$ hold to hold).

From the forth condition it follows trivially that $(\{\psi_k\}, \Delta \cup \{\psi_1, ..., \psi_{k-1}\}) \notin \mathcal{AL}$. Then as in Lemma 1 it follows that $(\{\psi_1\}, \Delta) \notin \mathcal{AL}$ (i.e., $(\{\phi_k\}, \Delta) \notin \mathcal{AL}$ as required) since also the formula $\phi$ in $\Delta$ is different from all the $\phi_i$ and hence the $\psi_i$.

Finally, we consider the third condition (iii) above and extend the sequence of $\psi_i$ with the sequence of attacks and defenses that the non-acceptability of $\psi_k$ entails. This is a sequence of formulae $\sigma_1, ..., \sigma_m$ such that $\sigma_1$ attacks $\psi_k$ and $\sigma_j$ attacks $\overline{\sigma_{j-1}}$ for $j = 2, \ldots, m$. All the formulae $\sigma_j$ are different between them but some may be equal to some of the $\psi_i$ formulae. If this is the case we can iterative replace such formulae $\psi_i$, starting with the one that appears deepest in $\sigma_j$ sequence, with the subsequence of the sequence of $\sigma_j$ formulae until we have a sequence of formuale which are all different. We can then apply again Lemma 1 to arrive at the result of $(\{\psi_1\}, \{\}) \notin \mathcal{AL}$, i.e., that $(\{\phi_k\}, \{\}) \notin \mathcal{AL}$, as required.