Argumentation Logic

Antonis KAKAS [a,1], Francesca TONI [b] and Paolo MANCARELLA [c]
[a] *University of Cyprus*
[b] *Imperial College London, UK*
[c] *Università di Pisa, Italy*

Abstract. We propose a novel logic-based argumentation framework, called Argumentation Logic (AL), built upon a restriction of classical Propositional Logic (PL) as its underlying logic. This allows us to control the application of Reduction ad Absurdum (RA). In the case of classically consistent theories, AL and PL are equivalent, and RA is recovered through a notion of (non-)acceptability of arguments. In the case of classically inconsistent theories, AL is an extension of PL that does not trivialize, enjoying good logic-based argumentation and general logical properties.

Keywords. Logic, Argumentation, Semantics, Inconsistency

1. Introduction

The argumentation community has developed several approaches to modelling logic-based argumentation (e.g. see [3] for an overview). Existing logic-based argumentation frameworks typically start with an underlying logic (equipped with a logical language, a notion of entailment, and a notion of inconsistency) and define arguments from sentences in the language and attacks between arguments using entailment and inconsistency. Then (semantic or procedural) notions of 'acceptable arguments' give a way to reason with inconsistent theories in the given logic, and resolve inconsistencies dialectically. Many logic-based argumentation frameworks use (or include) classical logic as their underlying logic. Thus, these approaches take classical logic as their starting point and build argumentation frameworks on top.

We present a logic-based argumentation framework that goes inside classical logic and re-interprets classical entailment via argumentation before even considering reasoning in the presence of inconsistencies. Our approach brings together two main aspects. Firstly, it uses a notion of *acceptability of arguments* to re-construct classical entailment with consistent theories as well as to support reasoning with inconsistent theories. This notion goes beyond standard semantic notions in many forms of argumentation (e.g. admissibility in [7]) and is inspired by early work in logic programming [6,11]. Secondly, it recognizes the special role of *Reductio ad Absurdum* (RA) in Natural Deduction (ND) and pushes this into the dialectical level and out of the underlying logic.

[1] Corresponding Author: Department of Computer Science, University of Cyprus, 75 Kallipoleos Str., P.O. Box 537 CY-1678 Nicosia, Cyprus; E-mail: antonis@cs.ucy.ac.cy

In our logic-based argumentation framework, that we call *Argumentation Logic* (AL), arguments are built from the given theory and sets of propositional formulae. The acceptability of an argument ensures that any other argument that attacks it can be defended against. AL separates proofs into *direct* and *indirect*, the former being without the use of RA, and defines attack in terms of direct proof of inconsistency.

AL can be shown to be equivalent to classical Propositional Logic (PL) when the given theory is classically consistent, but it does not trivialize when the theory is classically inconsistent. AL thus provides an approach for logic-based argumentation that extends PL by handling classical inconsistency in a natural way.

The paper is organised as follows. Section 2 provides some motivation. Section 3 gives preliminaries. Section 4 defines argumentation frameworks in AL, drawn from a propositional theory. Section 5 gives AL in the case of *directly consistent* theories, namely where inconsistency cannot be derived by direct proofs, and section 6 proves some properties thereof. Section 7 shows how AL extends propositional logic. Section 8 gives AL in the case of *directly inconsistent* theories, as a generalisation of the directly consistent case. Section 9 puts AL in the context of related work and section 10 concludes.

2. Motivation

Let us consider an example, adapted from [3]. We are given the propositional logic theory $T = \{s, \neg(r \wedge s), \neg r \rightarrow u, \neg u\}$, where propositions r, s, u represent the statements:

- r: *EC is too liberal wrt workers rights*
- s: *EC is not sufficiently liberal wrt workers rights*
- u: *European unemployment rises*

T is classically inconsistent. In order to derive the inconsistency in a proof system like that of Natural Deduction, Reduction ad Absurdum (RA, or proof by contradiction) needs to be applied, e.g. for proving $\neg r$, by assuming r, which together with s in T gives $r \wedge s$ thus leading to a contradiction with $\neg(r \wedge s)$ in T; then $\neg r$ leads to inconsistency *directly*, without using RA.

Given this theory, can s be supported by an acceptable *argument*? In order to answer this question we need to decide what we want to consider as supporting arguments and how to determine the acceptability of arguments. For example, does the sub-theory $\{s\}$ of T form an acceptable argument for its (direct) consequence s? This argument may be deemed to be *attacked*, for example, by the rest of the theory, $T_1 = \{\neg(r \wedge s), \neg r \rightarrow u, \neg u\}$, since $T = \{s\} \cup T_1$ is inconsistent. It may also be deemed to be attacked by $T_2 = \{r\}$, since $T \cup \{s\} \cup T_2$ is also inconsistent, and directly so, namely without requiring RA.

No matter which attacks we choose to consider, in order to deem $\{s\}$ acceptable, all these attacks need to be *defended* against, possibly by means of additional arguments that may be also deemed acceptable: a *dialectical process* is required to ascertain this acceptability, as for example indicated in [3].

How do we define attack and defence between arguments? Clearly, inconsistency needs to be at the heart of both. If we choose defence to coincide with attack, then, since inconsistency is symmetric, each argument can defend against any attack trivially by simply attacking back. The trivialization of logical inconsistent reasoning means that

either we need to tune the dialectical process carefully, e.g. as in [3], or we need to separate attack and defence. We will follow the latter option.

Within the required dialectical process, should we allow attacks that are inconsistent on their own? E.g., in the earlier example, should the full theory T be allowed to attack $\{s\}$, since $\{s\} \cup T = T$ is inconsistent? One could argue that such attacks should not be allowed in the first place since they are inconsistent on their own, and in classical logic we can logically derive anything from an inconsistent set of sentences, by applying RA. RA is essential to ensure completeness, and thus needs to be retained but controlled to avoid spurious attacks in the dialectical process. We choose to push RA in the dialectical process, by equating its application to determine the acceptability of arguments, and disallow its use to determine attacks and defences.

3. Preliminaries on Natural Deduction

Let $\mathscr{L}$ be a Propositional Logic (PL) language and $\vdash$ denote the provability relation of Natural Deduction (ND) in PL.[2] Throughout the paper, theories and sentences will always refer to theories and sentences wrt $\mathscr{L}$.

Definition 1 *Let T be a theory and ϕ a sentence. A* direct derivation *for ϕ (from T) is a ND derivation of ϕ (from T) without any application of RA. If there is a direct derivation for ϕ (from T) we say that ϕ is* directly derived *from T, denoted $T \vdash_{MRA} \phi$.*

Example 1 *Let $T = \{\alpha \wedge \beta \rightarrow \bot, \neg\beta \rightarrow \bot\}$. The following is a ND derivation for $\neg\alpha$ (from T) that is* not *direct:*

	$\lceil\alpha$		*hypothesis*
		$\lceil\neg\beta$	*hypothesis*
		$\neg\beta \rightarrow \bot$	*from T*
		$\bot\rfloor$	$\rightarrow E$
	$\neg\neg\beta$		$\neg I$ *(RA)*
	β		$\neg E$
	$\alpha \wedge \beta$		$\wedge I$
	$\alpha \wedge \beta \rightarrow \bot$		*from T*
	$\bot\rfloor$		$\rightarrow E$
$\neg\alpha$			$\neg I$ *(RA)*

Definition 2 *A theory T is* classically inconsistent *iff $T \vdash \bot$. A theory T is* directly inconsistent *iff $T \vdash_{MRA} \bot$. A theory T is* classically/directly consistent *iff it is not classically/directly inconsistent, respectively.*

Trivially, if a theory is classically consistent then it is directly consistent, e.g. as in the case of T in example 1. However, a directly consistent theory may be classically inconsistent, e.g. as in the case of $T = \{\alpha \rightarrow \bot, \neg\alpha \rightarrow \bot\}$.

We will use a special kind of ND derivations, that we call *Reduction ad Absurdum Natural Deduction* (RAND). These are ND derivations with an outermost application of RA. Example 1 shows a RAND derivation. RAND derivations of $\neg\phi$ will sometimes be

[2] See appendix A for a review of the ND rules we use, including $\neg I$/Reduction ad Absurdum (RA).

denoted by $\lceil \phi \dots \bot \rfloor$ (where ϕ is the hypothesis). The RAND derivation in example 1 can be denoted by $\lceil \alpha \dots \lceil \neg\beta \dots \bot \rfloor \dots \bot \rfloor$.

ND, with its syllogistic roots, has a natural argumentative interpretation: a direct derivation of a formula can be interpreted as an argument supporting the formula. However, this argumentative interpretation cannot be given naturally to RA, as the premise of this rule is an argument for rejecting the complement of the conclusion of the rule rather than an argument for directly supporting this conclusion.

4. Argumentation Logic Frameworks

Definition 3 *The* argumentation logic (AL) framework corresponding to *a theory T is the triple $\langle Args^T, Att^T, Def^T \rangle$ with:*

- *$Args^T = \{T \cup \Sigma | \Sigma$ is a set of sentences $\}$ is the set of all expansions of T by sets of sentences wrt $\mathscr{L}$;*
- *given $a, b \in Args^T$, with $a = T \cup \Delta$, $b = T \cup \Gamma$, such that $\Delta \neq \{\}$, $(b,a) \in Att^T$ iff $a \cup b \vdash_{MRA} \bot$;*
- *given $a, d \in Args^T$, with $a = T \cup \Delta$, $(d,a) \in Def^T$ iff*
  1. $d = T \cup \{\neg\phi\}$ ($d = T \cup \{\phi\}$) for some $\phi \in \Delta$ (respectively $\neg\phi \in \Delta$), or
  2. $d = T \cup \{\}$ and $a \vdash_{MRA} \bot$.

In the remainder, b attacks *a* (wrt T) *stands for $(b,a) \in Att^T$ and d* defends *or* is a defence against *a (wrt T) stands for $(d,a) \in Def^T$.*

Note that, since T is fixed, we will often *equate arguments $T \cup \Sigma$ to sets of sentences Σ*. So, for example, we will refer to $T \cup \{\} = T$ as the *empty argument*. Similarly, we will often equate a defence to a set of sentences. In particular, when $d = T \cup D$ defends/is a defence against $a = T \cup \Delta$ we will say that *D defends/is a defence against Δ (wrt T)*.

The attack relation between arguments is defined in terms of a direct derivation of inconsistency. Note that, trivially, for $a = T \cup \Delta$, $b = T \cup \Gamma$, $(b,a) \in Att^T$ iff $T \cup \Delta \cup \Gamma \vdash_{MRA} \bot$. The following example illustrates the notion of attack:

Example 2 *Given T in example 1, $\{\}$ attacks $\{\neg\beta\}$, $\{\beta\}$ attacks $\{\alpha\}$ (and vice-versa), $\{\neg\beta\}$ attacks $\{\beta\}$ (and vice-versa).*

Note that the attack relation is symmetric except for the case of the empty argument. Indeed, for a, b both non-empty, it is always the case that a attacks b iff b attacks a. However, the empty argument cannot be attacked by any argument (as the attacked argument is required to be non-empty), but the empty argument can attack an argument. As an additional example, given $T = \{\alpha, \neg\alpha\}$, $\{\}$ attacks $\{\alpha\}$ and $\{\}$ attacks $\{\beta\}$ (for any sentence β), since $T \vdash_{MRA} \bot$. Finally, note that our notion of attack includes the special case of attack between a sentence and its negation, since, for any theory T, $\{\phi\}$ attacks $\{\neg\phi\}$ (and vice-versa), for any sentence ϕ.

The notion of defence is a subset of the attack relation. In the first case of the definition we defend against an argument by adopting the complement[3] of some sentence in the argument, whereas in the second case we defend against any directly inconsistent

[3]The complement of a sentence ϕ is $\neg\phi$ and the complement of a sentence $\neg\phi$ is ϕ.

set using the empty argument. Then, in example 2, $\{\neg\beta\}$ defends against the attack $\{\beta\}$. Note that the empty argument cannot be defended against if T is directly consistent.

5. AL for Directly Consistent Theories

In this section we assume that T is *directly consistent*. As conventional in argumentation, we define a notion of acceptability of sets of arguments to determine which conclusions can be justified (or not) from the given theory. The intuition is that "an argument is acceptable iff all its counter-arguments are not". This has existed since the early nineties (see [6,11]) and is studied more recently in [12]. Our definition of acceptability and non-acceptability is formalised in terms of the least fix point of (monotonic) operators on the cartesian product of the set of arguments, as follows:

Definition 4 *Let $\langle Args^T, Att^T, Def^T\rangle$ be the AL framework corresponding to a directly consistent theory T, and $\mathscr{R}$ the set of binary relations over $Args^T$.*

- *The* acceptability operator $\mathscr{A}_T$:$\mathscr{R}\rightarrow\mathscr{R}$ *is defined as follows: for any $acc \in \mathscr{R}$ and $a, a_0 \in Args^T$:* $\quad (a,a_0) \in \mathscr{A}_T(acc)$ *iff*
  - $a \subseteq a_0$, *or*
  - *for any $b \in Args^T$ such that b attacks a wrt T,*
    - $b \not\subseteq a_0 \cup a$, and
    - there is $d \in Args^T$ that defends against b wrt T such that $(d, a_0 \cup a) \in acc$.
- *The* non-acceptability operator $\mathscr{N}_T : \mathscr{R} \rightarrow \mathscr{R}$ *is defined as follows: for any $nacc \in \mathscr{R}$ and $a, a_0 \in Args^T$:* $\quad (a,a_0) \in \mathscr{N}_T(nacc)$ *iff*
  - $a \not\subseteq a_0$, *and*
  - *there is $b \in Args^T$ such that b attacks a wrt T and*
    - $b \subseteq a_0 \cup a$, or
    - for any $d \in Args^T$ that defends against b wrt T, $(d, a_0 \cup a) \in nacc$.

These $\mathscr{A}_T$ and $\mathscr{N}_T$ operators are monotonic wrt set inclusion and hence their repeated application starting from the empty binary relation will have a least fixed point.

Definition 5 *ACC^T and $NACC^T$ denote the least fixed points of $\mathscr{A}_T$ and $\mathscr{N}_T$ respectively. We say that a* is acceptable wrt a_0 in T *iff $ACC^T(a,a_0)$, and a* is not acceptable wrt a_0 in T *iff $NACC^T(a,a_0)$. We also say that a* is acceptable in T *iff $ACC^T(a,\{\})$, and a* is not acceptable in T *iff $NACC^T(a,\{\})$.*

Note that non-acceptability, $NACC^T(a,a_0)$, is the same as the classical negation of $ACC^T(a,a_0)$, i.e. $NACC^T(a,a_0) = \neg ACC^T(a,a_0)$. We will use these two versions of non-acceptability interchangeably. The following examples illustrate non-acceptability.

Example 3 *Consider T in example 1. $T \cup \{\neg\beta\}$ is classically and directly inconsistent, and $T \cup \{\alpha\}$ is classically inconsistent but directly consistent. It is easy to see that $NACC^T(\{\neg\beta\},\{\})$ holds, as illustrated in figure 1 (left)*[4]*, since $\{\neg\beta\} \not\subseteq \{\}$, $b = \{\}$ attacks $\{\neg\beta\}$ and $\{\} \subseteq \{\neg\beta\}$. Also, $NACC^T(\{\alpha\},\{\})$ holds, as illustrated in figure 1 (right). Indeed:*

[4]Here and throughout the paper, $\uparrow$ denotes an attack and $\Uparrow$ denotes a defence.

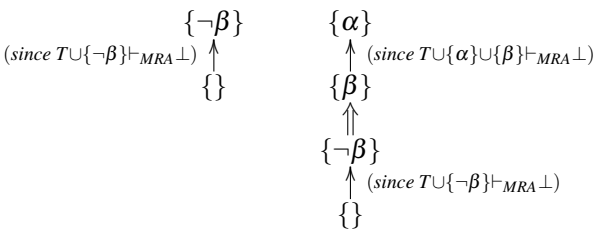

Figure 1. Illustration of $NACC^T(\{\neg\beta\},\{\})$ (left) and $NACC^T(\{\alpha\},\{\})$ (right), for example 3.

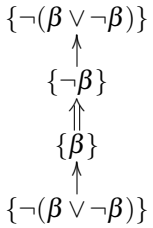

Figure 2. Illustration of $NACC^T(\{\neg(\beta\vee\neg\beta)\},\{\})$ for example 5.

- *since* $\{\alpha\}\not\subseteq\{\}$, $b=\{\beta\}$ *attacks* $\{\alpha\}$ *and* $\{\neg\beta\}$ *is the only defence against b, to prove that* $NACC^T(\{\alpha\},\{\})$ *it suffices to prove that* $NACC^T(\{\neg\beta\},\{\alpha\})$*;*
- *since* $\{\neg\beta\}\not\subseteq\{\alpha\}$, $b=\{\}$ *attacks* $\{\neg\beta\}$ *and* $\{\}\subseteq\{\alpha,\neg\beta\}$, $NACC^T(\{\neg\beta\},\{\alpha\})$ *holds as required.*

Note that if an argument a is attacked by the empty argument, then it is acceptable wrt any a_0 iff $a\subseteq a_0$, since there is no defence against the empty argument. This observation is used in the following example.

Example 4 *Given* $T=\{\alpha\rightarrow\bot,\neg\alpha\rightarrow\bot\}$, $NACC^T(\{\alpha\},\{\})$ *and* $NACC^T(\{\neg\alpha\},\{\})$ *both hold:* $NACC^T(\{\alpha\},\{\})$ *holds as* $\{\alpha\}$ *is attacked by* $\{\}$*;* $NACC^T(\{\neg\alpha\},\{\})$ *holds as* $\{\neg\alpha\}$ *is attacked by* $\{\}$.

The following example illustrates non-acceptability in the case of an empty theory.

Example 5 *For* $T=\{\}$, $NACC^T(\{\neg(\beta\vee\neg\beta)\},\{\})$ *holds, as illustrated in figure 2. Also, trivially,* $NACC^T(\{\beta\wedge\neg\beta\},\{\})$ *holds, since it is attacked by the empty argument.*

A novel, alternative notion of *entailment* can be defined for theories that are directly consistent in terms of the (non-) acceptability semantics for AL frameworks, as follows:

Definition 6 *Let* T *be a directly consistent theory and* ϕ *a sentence. Then* ϕ *is* AL-entailed by T *(denoted* $T\models_{AL}\phi$*) iff* $ACC^T(\{\phi\},\{\})$ *and* $NACC^T(\{\neg\phi\},\{\})$.

This is motivated by the argumentation perspective, where an argument is held if it can be successfully defended and it cannot be successfully objected against.

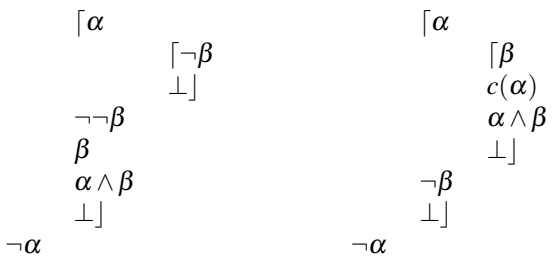

Figure 3. Two RAND derivations of $\neg\alpha$ in example 3: $\mathtt{d}_1$ (left) and $\mathtt{d}_2$ (right).

6. Properties of AL for Directly Consistent Theories

The following result gives a core property of the notion of AL-entailment wrt the notion of direct derivation in PL, for *directly consistent* theories.

Proposition 1 *Let T be a* directly consistent *theory and ϕ a sentence such that $T \vdash_{MRA} \phi$. Then $T \models_{AL} \phi$.*

Proof: Let $a = T \cup \Delta$ be any attack against $\{\phi\}$, i.e. $T \cup \{\phi\} \cup \Delta \vdash_{MRA} \bot$. Since $T \vdash_{MRA} \phi$ then $T \cup \Delta \vdash_{MRA} \bot$. Since T is directly consistent, $\Delta \neq \{\}$. Hence any such a can be defended against by the empty argument. Since $ACC^T(\{\}, \Sigma)$, for any set of sentences Σ, then $ACC^T(\{\phi\}, \{\})$ holds. Moreover, since $T \vdash_{MRA} \phi$, necessarily $T \cup \{\neg\phi\} \vdash_{MRA} \bot$. Hence the empty argument attacks $\{\neg\phi\}$ and thus $NACC^T(\{\neg\phi\}, \{\})$ holds. QED

The following theorem (proven in appendix B) shows how RA, deleted from the ND proof system within $\vdash_{MRA}$, is brought back through the notion of non-acceptability. This theorem will be used, in section 7, to prove (one half of) the link between AL and PL.

Theorem 1 *Let T be a* directly consistent *theory and ϕ a sentence. If $NACC^T(\{\phi\}, \{\})$ holds then there exists a RAND derivation of $\neg\phi$ from T.*

For example, the RAND derivation corresponding to the proof of $NACC^T(\{\alpha\}, \{\})$ in figure 1 is $\mathtt{d}_1$ in figure 3. Here, the inner RAND derivation in $\mathtt{d}_1$ corresponds to the non-acceptability of the defence $\{\neg\beta\}$ against the attack $\{\beta\}$ against $\{\alpha\}$. Derivation $\mathtt{d}_2$ in figure 1 is an alternative RAND of $\neg\alpha$, but this cannot be obtained from any proof of $NACC^T(\{\alpha\}, \{\})$, because there is a defence against the attack $\{\neg\beta\}$ given by the empty set (in other words, $\mathtt{d}_2$ does not identify a useful attack, that cannot be defenced against, for proving non-acceptability).

The following result gives a 'cut rule' for AL wrt the undelyind direct logic given by $\vdash_{MRA}$ (see appendix C for the proof).

Proposition 2 *Let T be a* directly consistent *theory and ϕ a sentence such that $T \vdash_{MRA} \phi$ and $T \cup \{\phi\} \models_{AL} \psi$. Then $T \models_{AL} \psi$.*

7. From AL to PL and Back

The following result gives a core property of the notion of non-acceptability for *classically consistent* theories.

Proposition 3 *Let T be* classically consistent *and ϕ a sentence. If $NACC^T(\{\neg\phi\},\{\})$ holds then $ACC^T(\{\phi\},\{\})$ holds.*

Proof: By theorem 1, since $NACC^T(\{\neg\phi\},\{\})$, then $T \vdash \phi$. Suppose, by contradiction, that $ACC^T(\{\phi\},\{\})$ does not hold. Then $NACC^T(\{\phi\},\{\})$ holds (since $NACC^T(\{\phi\},\{\})$ = $\neg ACC^T(\{\phi\},\{\})$) and by theorem 1 there is a RAND derivation of $\neg\phi$ from T and thus $T \vdash \neg\phi$. This implies that T is classically inconsistent: contradiction. QED

Thus, in PL, trivially AL-entailment reduces to the notion on non-acceptability:

Corollary 1 *Let T be a* classically consistent *theory and ϕ a sentence. Then $T \models_{AL} \phi$ iff $NACC^T(\{\neg\phi\},\{\})$.*

The following property sanctions that AL-entailment implies classical derivability:

Corollary 2 *Let T be a* classically consistent *theory and ϕ a sentence. If $T \models_{AL} \phi$ then $T \vdash \phi$.*

Proof: If $NACC^T(\{\neg\phi\},\{\})$, then, by theorem 1, there is a RAND derivation of $\neg\neg\phi$ from T and thus $T \vdash \phi$. QED

This corollary gives that consequences of a classically consistent theory under $\models_{AL}$ are classical consequences too. Although proposition 1 sanctions that all *direct* consequences are retrieved by $\models_{AL}$, in general not all classical consequences are retrieved by $\models_{AL}$, namely the converse of corollary 2 does not hold. For example, $\{\neg\alpha\} \not\models_{AL} \alpha \rightarrow \beta$, i.e., under $\models_{AL}$, implication is not material. However, if we restrict attention to theories *expressed using connectives $\wedge$ and $\neg$ only* (without loss of generality wrt PL), thus forcing implication to be interpreted as material implication, then all classical consequences of classically consistent theories are retrieved by $\models_{AL}$ and the two logics, AL and PL, are equivalent for classically consistent theories. To show this we first show that AL can retrieve classical consequences obtained by a special kind of derivations. These are RAND derivations satisfying the *genuine absurdity property*: this is satisfied by a RAND (sub-)derivation when its hypothesis ϕ is necessary for its direct derivation of $\perp$. This property is illustrated by example 3: $\mathtt{d}_1$ and $\mathtt{d}_2$ in figure 3 are both RAND derivations of $\neg\alpha$, but only $\mathtt{d}_1$ satisfies the genuine absurdity property (wrt T). Indeed, in $\mathtt{d}_2$, α is not necessary in the outer RAND direct derivation of $\perp$. In the case of directly consistent theories T expressed using connectives $\wedge$ and $\neg$ only, if there exists a RAND derivation of $\neg\phi$ from T that fully satisfies the genuine absurdity property then $T \models_{AL} \phi$ [9]. Since, in the case of classically consistent theories, for every RAND derivation there exists a RAND derivation of the same conclusion that fully satisfies the genuine absurdity property [10], AL retrieves fully PL and non-acceptability brings back RA.

8. AL beyond Directly Consistent Theories

We extend the notion of AL-entailment for directly inconsistent theories, T. The earlier result of proposition 2, ensuring that AL-entailment remains closed under direct consequences, suggests a natural way to extend AL: we can consider the various maximal

directly consistent subsets of the direct closure of any such T as a way to separate the dichotomy of the direct inconsistency of T and define notions of entailment wrt to these subsets, thus generalizing the earlier notion for directly consistent theories.

Definition 7 *Let T be any theory, and $Cn(T) = \{\phi | T \vdash_{MRA} \phi\}$ be all direct consequences of T. Then a sentence ϕ is sceptically AL-entailed (s-AL-entailed in short) or credulously AL-entailed (c-AL-entailed in short) by T (denoted $T \models^s_{AL} \phi$/$T \models^c_{AL} \phi$, respectively) iff $T' \models_{AL} \phi$ for all/some maximally (wrt $\subseteq$) directly consistent $T' \subseteq Cn(T)$, respectively.*

Example 6 *Consider the directly inconsistent theory $T = \{\alpha \wedge \beta, \neg\beta\}$. The maximally directly consistent sub-theories of $Cn(T)$ are $Cn(T_1)$ and $Cn(T_2)$ for $T_1 = \{\alpha \wedge \beta\}$ and $T_2 = \{\alpha, \neg\beta\}$. Therefore, $T \models^s_{AL} \alpha$ and $T \not\models^s_{AL} \beta$, $T \not\models^s_{AL} \neg\beta$.*

Consider now the (directly inconsistent) theory $T' = \{\alpha \wedge \beta, \neg\beta, \alpha\}$. Since $Cn(T') = Cn(T)$, T' is equivalent, under $\models^s_{AL}$, to T.

This example shows why the closure $Cn(T)$ rather than T is appropriate in the definition of s-AL-entailment. Indeed, the maximally directly consistent sub-theories of T in the example are $T_1 = \{\alpha \wedge \beta\}$ and $T_2^* = \{\neg\beta\}$, where $T_2^* \not\models_{AL} \alpha$. Instead, the maximally directly consistent sub-theories of T' in this example are $T_3 = \{\alpha \wedge \beta, \alpha\}$ and $T_4 = \{\neg\beta, \alpha\}$, where $T_3 \models_{AL} \alpha$, $T_4 \models_{AL} \alpha$. Therefore, counter-intuitively, T and T' in this example would not be equivalent under $\models^s_{AL}$ without the closure.

The need for closing the given theory under $\vdash_{MRA}$ can be understood from the argumentation perspective as including in the argumentation framework all arguments explicitly given, as members of the theory, or implicitly given, as direct consequences thereof. In other words, working with the closure ensures, given proposition 2, that the $\models^s_{AL}$ semantics remains invariant under equivalent re-writings (under $\vdash_{MRA}$) of the given theory.

The following example illustrates the difference between $\models^s_{AL}$ and $\models^c_{AL}$.

Example 7 *Let $T = \{\alpha, \alpha \rightarrow \bot, \neg\alpha \rightarrow \bot\}$. T is directly inconsistent. The maximally directly consistent subsets of the closure of T are $Cn(\{\alpha \rightarrow \bot, \neg\alpha \rightarrow \bot\})$ and $Cn(\{\alpha, \neg\alpha \rightarrow \bot\})$. Then $T \models^c_{AL} \alpha$ but $T \not\models^s_{AL} \alpha$.*

The notions of $\models^s_{AL}$/$\models^c_{AL}$are extensions of the notion of $\models_{AL}$, in the sense of the following property, that follows directly from the definitions:

Proposition 4 *Let T be a directly consistent theory and $\phi \in \mathscr{L}$. Then $T \models^s_{AL} \phi$ iff $T \models^c_{AL} \phi$ iff $T \models_{AL} \phi$.*

The following example illustrates the "paraconsistency" of the notion of $\models^s_{AL}$ / $\models^c_{AL}$, i.e. how it avoids trivialization and how parts of the theory that do not contribute to the direct inconsistency are sceptically entailed:

Example 8 *Consider the directly inconsistent theory $T = \{\alpha, \neg\alpha, \beta\}$. The maximally directly consistent sub-theories of $Cn(T)$ are $Cn(\{\alpha, \beta\})$ and $Cn(\{\neg\alpha, \beta\})$. Thus*

- *$T \models^s_{AL} \beta$, $T \models^s_{AL} \beta \vee \neg\beta$, $T \models^s_{AL} \alpha \vee \neg\alpha$, but*
- *$T \not\models^s_{AL} \alpha$ and $T \not\models^s_{AL} \neg\alpha$*

9. Related Work

Besnard and Hunter [3] proposed an argumentation framework based upon classical logic with the aim (that we share) to use argumentation to reason with possibly inconsistent classical theories, beyond the realms of classical logic. In their approach, arguments are defined in terms of sub-theories of a given (typically inconsistent) theory and they have minimal and consistent supports (wrt the full classical consequence relation). Attacks are defined in terms of a notion of canonical undercut that relies on arguments for the negation of the support of attacked argument. Further, the evaluation of arguments is given through a related tree structure of defeated or undefeated nodes. To illustrate how their approach differes from ours, consider the two classically (and directly) inconsistent theories $T_1 = \{\alpha, \beta, \neg\alpha \vee \neg\beta\}$ and $T_2 = \{\alpha, \beta, \alpha \wedge \beta, \neg\alpha \vee \neg\beta\}$. For neither theories AL entails α, whereas the approach of [3] gives α in the case of T_2.

Several properties for logic-based argumentation systems have been suggested, e.g. in [4,2,8,1], requiring that their semantics follow a desired behaviour. They typically refer to the extensions of the argumentation framework imposing, in an axiomatic way, certain properties on them. Often these properties are called rationality postulates to emphasize their link to the logical nature of the frameworks. Given the nature of our approach, they can be applied either at the level of the extensions of the argumentation framework, i.e. at the level of acceptable arguments, $ACC^T(a, \{\})$, or at the full logical level of AL itself, i.e. at the level of the logical entailment $\models_{AL}$, where these postulates refer to properties often required from a logical system.

For example, the various consistency postulates formulated in [8,1] require that extensions of the argumentation framework are consistent with respect to the underlying logic, either in the sense that they do not imply a contradiction or that they are not trivial implying everything in the language. As we have seen acceptable arguments in *AL* have such consistency properties and at the full logical level *AL* does not trivialize under any form of inconsistency, classical or direct, wrt $\vdash_{MRA}$.

Similarly, closure postulates, e.g. as in [4,1], state that extensions should be closed under the underlying logic. In our case this would require that given an argument $a = T \cup \Delta$ that is acceptable, i.e. $ACC^T(\Delta, \{\})$ holds, and a set of formulae Γ such that $T \cup \Delta \vdash_{MRA} \Gamma$ then $ACC^T(\Delta \cup \Gamma, \{\})$ should also hold. This follows directly from the fact that any attack against $\Delta \cup \Gamma$ is also an attack against Δ and that this can be defended in the same way. Furthermore, the meaning of the closure postulates at the full logical level is that of the logic satisfying cut-rules in their entailment, which indeed is the fact that motivates the formulation of these postulates in logic-based argumentation. As we have seen in proposition 2 such properties are enjoyed by AL.

10. Conclusions and Future Work

We have presented a logic-based argumentation framework, called Argumentation Logic (AL), as an extension of classical Propositional Logic (PL) through a formulation of PL itself as a logic of arguments. *AL* handles classical inconsistency in a natural and foundational way and as such it enjoys desirable properties, at the full logical level of AL entailment.

In using logic-based argumentation for various application problems, e.g. to capture common sense default reasoning or legal reasoning, it is well known that priorities

amongst the logical sentences that form the premises of arguments can be very useful leading to what is usually called preference based argumentation, see e.g. [13]. AL has recently been used in the formalization of psychological theories of story comprehension [5] so that this can include contrapositive reasoning with default rules. A more general study of the extension of *AL* as a preference-based argumentation framework with priorities on the sentences of the given theory, in this context of achieving a more general synthesis of defeasible and strict classical reasoning, forms an important part of our future work.

A system to visualize the argumentative reasoning of *AL* is currently under development with the ultimate aim to use this as a tool to support a dialectical process for resolving conflicts in the context of applications.

References

[1] L. Amgoud. Postulates for logic-based argumentation systems. *International Journal of Approximate Reasoning*, 2014. To appear.

[2] L. Amgoud and P. Besnard. Bridging the gap between abstract argumentation systems and logic. In *SUM*, pages 12–27, 2009.

[3] P. Besnard and A. Hunter. *Elements of Argumentation*. MIT Press, 2008.

[4] M. Caminada and L. Amgoud. An axiomatic account of formal argumentation. In *BNAIC*, pages 327–328, 2005.

[5] I. Diakidou, A. Kakas, L. Michael, and R. Miller. A psychology-inspired approach to automated narrative text comprehension. In *KR*, 2014. To appear.

[6] P. M. Dung, A. C. Kakas, and P. Mancarella. Negation as failure revisited. Technical report, University of Pisa, 1992.

[7] P.M. Dung. On the acceptability of arguments and its fundamental role in non-monotonic reasoning, logic programming and n-person games. *Artificial Intelligence*, 77(1–2):321–357, 1995.

[8] N. Gorogiannis and A. Hunter. Instantiating abstract argumentation with classical logic arguments: Postulates and properties. *Artificial Intelligence*, 175(9-10):1479–1497, 2011.

[9] A. Kakas, F. Toni, and P. Mancarella. Argumentation logic. Technical report, Department of Computer Science, University of Cyprus, Cyprus, April 2012.

[10] A. Kakas, F. Toni, and P. Mancarella. On reductio ad absurdum in propositional logic. Technical report, Department of Computer Science, University of Cyprus, Cyprus, December 2013.

[11] A. C. Kakas, P. Mancarella, and P. M. Dung. The acceptability semantics for logic programs. In *ICLP*, pages 504–519, 1994.

[12] Antonis Kakas and Paolo Mancarella. On the semantics of abstract argumentation. *Journal of Logic and Computation*, 23:991–1015, 2013.

[13] S. Modgil and H. Prakken. A general account of argumentation with preferences. *Artificial Intelligence*, 195:361–397, 2013.

A. Appendix: Natural Deduction

We use the following rules, for any propositional formulae ϕ, ψ, χ in $\mathscr{L}$:

$$\wedge I : \frac{\phi, \psi}{\phi \wedge \psi} \quad \wedge E : \frac{\phi \wedge \psi}{\phi} \quad \wedge E : \frac{\phi \wedge \psi}{\psi} \quad \vee I : \frac{\phi}{\phi \vee \psi} \quad \vee I : \frac{\psi}{\phi \vee \psi} \quad \rightarrow I : \frac{\lceil \phi \ldots \psi \rfloor}{\phi \rightarrow \psi} \quad \perp I : \frac{\phi, \neg\phi}{\perp}$$

$$\neg E : \frac{\neg\neg\phi}{\phi} \quad \neg I/RA : \frac{\lceil \phi \ldots \perp \rfloor}{\neg\phi} \quad \vee E : \frac{\phi \vee \psi, \lceil \phi \ldots \chi \rfloor, \lceil \psi \ldots \chi \rfloor}{\chi} \quad \rightarrow E : \frac{\phi, \phi \rightarrow \psi}{\psi}$$

where $\lceil \zeta, \ldots \rfloor$ is a (sub-)derivation with ζ referred to as the *hypothesis*. $\perp$ stands for inconsistency.

B. Appendix: Proof of theorem 1

We will use the following lemma:

Lemma 1 *For any theory $T \subseteq \mathscr{L}$ and for any set of sentences $\Delta \subseteq \mathscr{L}$ such that $T \cup \Delta$ is directly consistent, if $NACC^T(\{\phi\}, \Delta)$ holds then there exists a RAND derivation of $\neg\phi$ from $T \cup \Delta$.*

Proof of lemma 1: We use induction on the number of iterations of the $\mathscr{N}_T$ operator whose least fixed point defines $NACC^T$ (see definition 4).

Base Case: $NACC^T(\{\phi\}, \Delta)$ holds at the first iteration of $\mathscr{N}_T$. Then, there exists A such that A attacks $\{\phi\}$ (namely $T \cup A \cup \{\phi\} \vdash_{MRA} \bot$) and $A \subseteq \Delta \cup \{\phi\}$. Thus, $T \cup \Delta \cup \{\phi\} \vdash_{MRA} \bot$ and, trivially, there exists a RAND derivation $\lceil \phi \ldots \bot \rfloor$ (with no RAND sub-derivations) of $\neg\phi$ from $T \cup \Delta$.

Induction Hypothesis: For any $\psi \in \mathscr{L}$, for any $\mathscr{E}$ such that $T \cup \mathscr{E}$ is directly consistent, if $NACC^T(\{\psi\}, \mathscr{E})$ holds after k iterations of $\mathscr{N}_T$, then there exists a RAND derivation of $\neg\psi$ from $T \cup \mathscr{E}$.

Inductive Step: Assume $NACC^T(\{\phi\}, \Delta)$ holds after $k+1$ iterations of $\mathscr{N}_T$, for some Δ such that $T \cup \Delta$ is directly consistent. Then there exists A such that
(i) A attacks $\{\phi\}$ (namely $T \cup A \cup \{\phi\} \vdash_{MRA} \bot$), but $A \not\subseteq \Delta \cup \{\phi\}$; and
(ii) for each defence D against A, $NACC^T(D, \Delta \cup \{\phi\})$ holds after k iterations of $\mathscr{N}_T$.
Since $A \not\subseteq \Delta \cup \{\phi\}$, $A \neq \{\}$. Also, by compactness of $\vdash_{MRA}$ (holding by compactness of $\vdash$), we can assume that A is finite. Let $A = \{\psi_1, \ldots, \psi_n\}$. Then, $D_i = \{\neg\psi_i\}$, for any $i = 1, \ldots, n$, is a defence against A and hence satisfies property (ii) above, i.e. $NACC^T(D_i, \Delta \cup \{\phi\})$ holds after k iterations. Note that $T \cup \Delta \cup \{\phi\}$ is directly consistent, as otherwise Δ attacks $\{\phi\}$ wrt T and $NACC^T(\{\phi\}, \Delta)$ would hold at the first iteration.
Hence, by the induction hypothesis, there exists a RAND derivation of $\neg\neg\psi_i$, for any $i = 1, \ldots, n$. from $T \cup \Delta \cup \{\phi\}$. We can construct a RAND derivation, $\mathtt{d}$, of $\neg\phi$ from $T \cup \Delta$, with top derivation $\mathtt{d} : \lceil \phi \ldots \bot \rfloor$ using the RAND derivations of $\neg\neg\psi_i$ from $T \cup \Delta \cup \{\phi\}$ as child sub-derivations. Note that in the top derivation we can use the $\neg E$ rule to derive ψ_i from each $\neg\neg\psi_i$, and hence, by definition of the attack A, the derivation $\mathtt{d}$ indeed leads directly to inconsistency from $T \cup \Delta$.
The resulting $\mathtt{d}$ is a RAND of $\neg\phi$ from $T \cup \Delta$ as any use of ϕ in the sub-derivations of $\neg\neg\psi_i$ from $T \cup \Delta \cup \{\phi\}$ can now be replicated using the copy operation of ϕ from $\mathtt{d}$. QED

To prove the theorem, assume now that $NACC^T(\{\phi\}, \{\})$ holds. Directly from lemma 1 with $\Delta = \{\}$, if T is directly consistent then there is a RAND derivation $\mathtt{d}$ of $\neg\phi$ from T.

C. Appendix: Proof of proposition 2 (Sketch)

We need to show (1) $ACC^T(\{\psi\}, \{\})$ and (2) $NACC^T(\{\neg\psi\}, \{\})$.

(1) By induction on the length of the branches of the $ACC^{T \cup \{\phi\}}(\Delta, \Delta_0)$ (or the number of iterations of the $\mathscr{A}_T$ operator) we show $ACC^{T \cup \{\phi\}}(\Delta, \Delta_0)$ implies $ACC^T(\Delta, \Delta_0)$, for any sets of sentences Δ, Δ_0. The base case is trivial as $\Delta \subseteq \Delta_0$ does not depend on the given theory. Let $ACC^{T \cup \{\phi\}}(\Delta, \Delta_0)$ hold and consider an attack A on Δ w.r.t. T, i.e. $T \cup \Delta \cup A \vdash_{MRA} \bot$. Then A is also an attack on Δ w.r.t. $T \cup \{\phi\}$, i.e. $(T \cup \{\phi\}) \cup \Delta \cup A \vdash_{MRA} \bot$. Hence from $ACC^{T \cup \{\phi\}}(\Delta, \Delta_0)$ we know that $A \not\subseteq \Delta \cup \Delta_0$ holds and that there is a defence D against A. There are two cases for this defence. $D = \{\}$ when $(T \cup \{\phi\}) \cup A \vdash_{MRA} \bot$. But then $T \cup A \vdash_{MRA} \bot$ also holds since $T \vdash_{MRA} \phi$ and thus $D = \{\}$ is also a defence on the attack A w.r.t. T. Otherwise, there exists $\chi \in A$ such that $D = \{\chi^c\}$ is a defence (for χ^c the complement of χ), i.e. $ACC^{T \cup \phi}(D, \Delta \cup \Delta_0)$. By the induction hypothesis $ACC^T(D, \Delta \cup \Delta_0)$ also holds and hence this D is also a defence against A w.r.t. T.

(2) By induction on the length of the branches of the $NACC^{T \cup \{\phi\}}(\Delta, \Delta_0)$ (or the number of iterations of the $\mathscr{N}_T$ operator) we show $NACC^{T \cup \{\phi\}}(\Delta, \Delta_0)$ implies $NACC^T(\Delta, \Delta_0)$. The base case follows from the fact that an attack A against Δ w.r.t. $T \cup \{\phi\}$ is also against Δ w.r.t. T since $T \vdash_{MRA} \phi$ and hence $A \subseteq \Delta \cup \Delta_0$ does not depend on the given theory. From the same observation that any attack A w.r.t. $T \cup \{\phi\}$ is also an attack w.r.t. T the inductive step follows straightforwardly as the branches of the non-acceptability wrt $T \cup \{\phi\}$ of the possible defences against an attack will be of strictly smaller length than those of $NACC^{T \cup \{\phi\}}(\Delta, \Delta_0)$.