Metalevel Argumentation

Sanjay Modgil*, Trevor Bench-Capon†

Abstract

The abstract nature of Dung's theory of argumentation accounts for its widespread application as a general framework for various species of non-monotonic reasoning, and, more generally, reasoning in the presence of conflict. In this paper we formalise reasoning *about* argumentation within the Dung argumentation paradigm itself. A metalevel Dung argumentation framework is itself instantiated by arguments that make statements *about* arguments, their interactions, and their evaluation in an object-level argumentation framework. We show how Dung's theory, and object level extensions of Dung's theory, such as those intended to accommodate preferences, can then be uniformly characterised by metalevel argumentation in a Dung framework. We then discuss how this provides for application of the full range of theoretical and practical developments of Dung's theory, to extensions of Dung's theory, and provides for integration and further augmentation of these extensions.

Keywords: Argumentation, Dung, Metalevel, Preferences, Values

1 Introduction

Formal models [5] of argumentation have been extensively studied and applied within Artificial Intelligence [13, 47]. Many developments build on Dung's seminal theory of argumentation [26]. A Dung *argumentation framework* is a directed graph consisting of a set of arguments $\mathcal{A}$ and a binary conflict based *attack* relation $\mathcal{R}$ on $\mathcal{A}$. The extensions, and so justified arguments of a framework are then defined under different semantics. Extensions are defined through application of an 'acceptability calculus', whereby an argument $x \in \mathcal{A}$ is said to be *acceptable* with respect to $S \subseteq \mathcal{A}$, iff any argument y that attacks x is itself attacked by some argument z in S. For example, if S is a maximal (under set inclusion) set such that all its contained arguments are acceptable with respect to S, then S is said to be an extension under the *preferred* semantics.

Dung's theory has been developed in a number of directions. Some works formalise collective attacks between *sets* of arguments [16, 44]. In other works, the success of an attack from x to y is contingent on y not being preferred to x according to some given preference relation on $\mathcal{A}$ [1], or the value promoted by y not being ranked higher than the value promoted by x, according to a given ordering on values [11]. More recently, [38, 35]'s *Extended Argumentation Framework* (EAF) extends Dung's framework to

*Corresponding author: Sanjay Modgil (sanjaymodgil@yahoo.co.uk), Department of Informatics, King's College London, Strand, London WC2R 2LS, UK (+44 (0)788 307 5206)

†T.J.M Bench-Capon (tbc@csc.liv.ac.uk) Department of Computer Science, University of Liverpool, Liverpool L69 7ZF, UK

include arguments that attack attacks. Thus, if x and y attack each other, then an argument z justifying a preference for y over x, *attacks the attack* from x to y. EAFs thus accommodate argumentation based reasoning *about* possibly conflicting preference information, within the argumentation framework itself. [6, 7] generalise this idea to allow recursive attacks on attacks, while a number of works also augment Dung's framework to include a *support* relation on arguments [2, 45]. Finally, [9] was the first to model recursive attacks on attacks, and support, albeit without providing Dung style semantics.

The continuing development, influence and application of Dung's theory can be attributed to its abstract nature, and to the encoding of intuitive generic principles of commonsense reasoning in the acceptability calculus. Its abstract nature allows for instantiation by various logical formalisms; one is free to choose a logic $\mathcal{L}$ and define what constitutes an argument and attack between arguments defined by a theory in $\mathcal{L}$. A theory's inferences can then be defined in terms of the claims of the justified arguments constructed from the theory (an argument essentially being a proof of a candidate inference — the argument's claim — in the underlying logic). Indeed, many logic programming formalisms and non-monotonic logics have been shown to conform to Dung's semantics [21, 26, 30], thus testifying to the general applicability of the principles encoded in the acceptability calculus. Dung's theory can therefore be understood as a *semantics* for non-monotonic reasoning. In this view, what appropriately accounts for the correctness of an inference is that an argument for the inference can be shown to rationally prevail in the face of arguments for opposing inferences, where, one can claim that: *it is application of the acceptability calculus that encodes logic neutral, rational means for establishing such standards of correctness*.

In this paper we further substantiate the above claim, by formalising reasoning *about* argumentation within the Dung argumentation paradigm itself. The basic idea is that given an object-level argumentation framework, one can consider metalevel arguments that can be explicitly categorised according to the types of claim made about the arguments and their relations in the object level framework. These metalevel arguments can then themselves be related by an attack relation in a Dung framework, where this metalevel attack relation satisfies constraints imposed by the claim based categorisation. One can then show a correspondence between the object level framework and its metalevel formulation, such that the justified arguments of the object level framework can be computed directly from its metalevel formulation.

For example, given an object level Dung framework $(\mathcal{A},\mathcal{R})$, one can consider metalevel arguments that make claims such as 'x is justified', 'x is rejected', 'x attacks y', about arguments $x, y \in \mathcal{A}$. These metalevel arguments can themselves be organised into a Dung framework such that an argument claiming 'x is justified' is a justified argument of the metalevel framework, iff x is a justified argument of the object level framework. Thus, the acceptability calculus applied at the metalevel characterises the use of the acceptability calculus at the object level.

The remainder of this paper is organised as follows. Section 2 reviews Dung's theory, and the above mentioned developments of the theory. In Section 3 we augment a Dung argumentation framework $(\mathcal{A},\mathcal{R})$ to obtain a 5 tuple *Structured Argumentation Framework* (SAF) that includes a function, a language and a set of constraints, such that the function maps arguments in $\mathcal{A}$ to claims in the language, and given this claim based categorisation of arguments, a set of constraints on $\mathcal{R}$ is specified. We then define a specific language in which one can express claims about object level frameworks, and so identify *Metalevel Argumentation Frameworks* (MAFs) as a special class of

SAFs. We then show how Dung frameworks, their generalisation to accommodate collective attacks [44], their extensions to accommodate preferences [1] and values [11], and a special, but widely applicable, class of [38]'s extended framework, can all be formulated as instances of metalevel argumentation.

In Section 4 we discuss some implications and applications of metalevel argumentation. Firstly, since MAFs formalise Dung argumentation and many of its developments, within the Dung paradigm itself, one can bring the full range of theoretical results and techniques for Dung argumentation to bear on developments of Dung argumentation. Secondly, in the spirit of Dung's original theory, MAFs adopt a level of abstraction that makes limited commitments to the instantiating logics. Thus metalevel arguments can be instantiated based on the existence of arguments in different object level frameworks, which in turn may be constructed from different underlying logics. This not only allows for integration and further extension of different forms of abstract argumentation, but also provides principled means for instantiation by, and integration of arguments constructed from different underlying logics.

The contents of this paper first appeared as a technical report [41], and builds on and substantially extends previous work of ours [39]. The latter was the first work to formalise the insight that attacks can themselves be treated as arguments in a Dung framework. This insight has been described in a number of subsequent works [6, 17, 19, 29] that are discussed in Section 5's review of related work. We conclude in Section 6 in which we also point to future work.

To summarise, the contributions of this paper are as follows:

1. We formalise abstract metalevel argumentation frameworks that adopt the same basic machinery of a Dung framework, but overlay more structure by identifying classes of claims about arguments in object level frameworks, and thus constraints on the attack relation.

2. Dung's abstract argumentation theory can be said to identify general dialectical principles that underpin common-sense reasoning as encoded in a range of non-monotonic reasoning formalisms. We promote and substantiate this view by showing how a number of developments of Dung's theory can be uniformly characterised in terms of these dialectical principles.

3. We discuss how by formalising Dung argumentation and its developments, within the Dung paradigm itself, the full range of theoretical and practical results and techniques for Dung's work can be applied to its developments. We also discuss how metalevel argumentation frameworks provide a unifying formalism in which to integrate and further extend the various developments of Dung argumentation, and allow for instantiation by, and integration of arguments constructed from different underlying logics and theories.

2 Abstract Argumentation Theories

In this section we review Dung's theory [26] and its various developments to accommodate collective attacks [44], preference based [1] and value based argumentation [11], and attacks on attacks [38].

Definition 1 A *Dung argumentation framework* (AF) is a tuple $(\mathcal{A}, \mathcal{R})$, where $\mathcal{A}$ is a set of arguments, and $\mathcal{R} \subseteq (\mathcal{A} \times \mathcal{A})$ is a binary attack relation on $\mathcal{A}$.

Definition 2 Let $(\mathcal{A}, \mathcal{R})$ be an AF, and $S \subseteq \mathcal{A}$. Then $x \in \mathcal{A}$ is acceptable w.r.t. S iff for all $y \in \mathcal{A}$ such that $y\mathcal{R}x$, there exists a $z \in S$ such that $z\mathcal{R}y$

The acceptability of arguments underpins evaluation of the status of arguments:

Definition 3 Let $\Delta = (\mathcal{A}, \mathcal{R})$. Let $S \subseteq \mathcal{A}$ such that $\forall x, y \in S$, it is not the case that $x\mathcal{R}y$, in which case S is said to be *conflict free*. Then S is an extension of Δ that is:

admissible iff each argument in S is acceptable w.r.t S;
complete iff S is *admissible*, and every argument acceptable w.r.t. S is in S;
grounded iff S is the minimal (w.r.t. set inclusion) *complete* extension;
preferred iff S is a maximal (w.r.t. set inclusion) *complete* extension;
stable iff S is *admissible* and every argument not in S is attacked by an argument in S.

For $s \in \{$complete, preferred, stable, grounded$\}$:

- If $x \in \mathcal{A}$ is in at least one s extension of Δ then x is said to be credulously justified under the s semantics.
- If $x \in \mathcal{A}$ is in all s extensions of Δ then x is said to be sceptically justified under the s semantics.
- If $x \in \mathcal{A}$ is not in any s extension of Δ then x is said to be rejected under the s semantics.

We introduce some notation that will be of use in the remainder of this paper:

Notation 1 *Let $(\mathcal{A}, \mathcal{R})$ be an AF, and $E \subseteq \mathcal{A}$.*

- $\overrightarrow{E+}$ *denotes the set of attacks originating from arguments in E:*
  $\overrightarrow{E+} = \{(x, y) \mid x \in E, x\mathcal{R}y\}$
- $E+$ *denotes the set of arguments attacked by arguments in E:*
  $E+ = \{y \mid x \in E, x\mathcal{R}y\}$
- $E-$ *denotes the set of arguments that attack arguments in E:*
  $E- = \{y \mid y\mathcal{R}x, x \in E\}$

The extensions and status of arguments in an AF can be defined in terms of labellings [22, 31, 49]. Here, we review the labelling approach presented in [22, 23].

Definition 4 A labelling is a total function $\mathcal{L}$ that assigns a label `IN`, `OUT` or `UNDEC` to each argument $x \in \mathcal{A}$ in an AF $(\mathcal{A}, \mathcal{R})$. Henceforth, we say that: $\texttt{in}(\mathcal{L}) = \{x | \mathcal{L}(x) = \texttt{IN}\}$; $\texttt{out}(\mathcal{L}) = \{x | \mathcal{L}(x) = \texttt{OUT}\}$; $\texttt{undec}(\mathcal{L}) = \{x | \mathcal{L}(x) = \texttt{UNDEC}\}$.

Legal labellings of arguments are then defined, and used as a basis for characterising the extensions of an AF.

Definition 5 Let $\mathcal{L}$ be a labelling for $(\mathcal{A}, \mathcal{R})$ and $x \in \mathcal{A}$.

- x is legally `IN` iff x is labelled `IN` and every y that attacks x is labelled `OUT`.
- x is legally `OUT` iff x is labelled `OUT` and there is at least one y that attacks x and y is labelled `IN`.
- x is legally `UNDEC` iff there is no y that attacks x such that y is labelled `IN`, and it is not the case that: for all $y \in \mathcal{A}$ if y attacks x, then y is labelled `OUT`.

Definition 6 For $\mathtt{l} \in \{\mathtt{IN}, \mathtt{OUT}, \mathtt{UNDEC}\}$ an argument x is said to be illegally $\mathtt{l}$ iff x is labelled $\mathtt{l}$, and it is not legally $\mathtt{l}$. Then:

an admissible labelling $\mathcal{L}$ is a labelling without arguments that are illegally $\mathtt{IN}$ and without arguments that are illegally $\mathtt{OUT}$;
a complete labelling $\mathcal{L}$ is an admissible labelling without arguments that are illegally $\mathtt{UNDEC}$.

Let $\mathcal{L}$ be a complete labelling. Then $\mathcal{L}$ is :

a grounded labelling iff there does not exist a complete labelling $\mathcal{L}'$ such that $\mathtt{in}(\mathcal{L}') \subset \mathtt{in}(\mathcal{L})$;
a preferred labelling iff there does not exist a complete labelling $\mathcal{L}'$ such that $\mathtt{in}(\mathcal{L}') \supset \mathtt{in}(\mathcal{L})$;
a stable labelling iff $\mathtt{undec}(\mathcal{L}) = \emptyset$.

In [23], the following is shown to hold:

Theorem 1 Let $\Delta = (\mathcal{A}, \mathcal{R})$. For $s \in \{$admissible, complete, grounded, preferred, stable$\}$: E is an s extension of Δ iff there exists an s labelling $\mathcal{L}$ with $\mathtt{in}(\mathcal{L}) = E$.

Notice the extra expressivity compared with the extension based approach. An s labelling $\mathcal{L}$ not only identifies the arguments in an s extension E (the arguments labelled $\mathtt{IN}$), but also the arguments in $E+$ and $E-$ (the union of which are the arguments labelled $\mathtt{OUT}$). Note that since each s extension is admissible, it follows that it is always the case that $E- \subseteq E+$. Also note that those arguments labelled $\mathtt{UNDEC}$ are neither in E, $E+$ or $E-$, and so are the arguments that are neither in the s extension E identified by the $\mathtt{IN}$ arguments, or attacked by or attack (an argument in) E. Observe that the following follows straightforwardly from Definition 5 and Theorem 1.

Proposition 1[1] Let $\Delta = (\mathcal{A}, \mathcal{R})$, and for $s \in \{$admissible, complete, grounded, preferred, stable$\}$, let E be an s extension of Δ. Then there exists an s labelling $\mathcal{L}$ of Δ such that $\mathtt{in}(\mathcal{L}) = E$, and $\mathtt{out}(\mathcal{L}) = (E+) \cup (E-)$.

Finally note that we can say that the arguments labelled $\mathtt{OUT}$ or $\mathtt{UNDEC}$ are *potentially* rejected arguments in the sense that if these arguments are $\mathtt{OUT}$ or $\mathtt{UNDEC}$ in all other s labellings, then they are said to be *rejected* as defined in Definition 3.

Recent works generalise binary attacks to allow for attacks between sets of arguments [16, 44]. Here, we briefly review [44], in which the attack relation is defined from sets of arguments to single arguments [2]. [44] motivate the need for such a generalisation by illustrating cases where a number of arguments can interact and constitute a stronger attack on any given argument, and where requiring that the logical contents of the attacking arguments be combined into a single attacking argument may not only be unnatural (by artificially requiring the combination of separate but orthogonal reasons), but may also lead to unwanted results.

Definition 7 A *DungC* argumentation framework is a tuple $(\mathcal{A}, \mathcal{R})$, where $\mathcal{A}$ is a set of arguments, and $\mathcal{R} \subseteq (2^{\mathcal{A}} \backslash \{\emptyset\}) \times \mathcal{A}$.

[1] All proofs of results in this paper appear in the Appendix.
[2] [44] argues that allowing sets of arguments to attack other sets does not provide greater flexibiilty

- $x \in \mathcal{A}$ is acceptable w.r.t. $S \subseteq \mathcal{A}$ iff for all $B \subseteq \mathcal{A}$ such that $B\mathcal{R}x$, there exists a $C \subseteq S$ such that $C\mathcal{R}y$ for some $y \in B$.
- $S \subseteq \mathcal{A}$ is conflict free iff $\forall S' \subseteq S, \forall x \in S$ it is not the case that $S'\mathcal{R}x$

Given the above definitions of conflict free and acceptability for DungC frameworks, the extension-based semantics and status of arguments are defined in the same way as for a standard Dung framework (i.e., as in Definition 3). Hence, one can straightforwardly see that a standard Dung framework is simply a special case of a DungC framework. If each attack originates from a singleton set of arguments, then the definitions of acceptability and conflict-free in Definition 7 coincide with those given in Definitions 2 and 3. As one would therefore expect, the fundamental results that hold for Dung frameworks are also shown to hold for DungC frameworks [44].

We now review approaches that augment Dung's theory so as to formalise the role of the relative strengths of arguments. In these approaches, an attack by x on y succeeds as a *defeat* only if y is not stronger than x. The status of arguments is then determined on the basis of the derived defeat relation, rather than the original attack relation.

In Preference based Argumentation [1], an AF is augmented with a preference ordering on $\mathcal{A}$, so that an attack by x on y succeeds as a defeat, only if y is not strictly preferred to x.

Definition 8 A *Preference based Argumentation Framework* (PAF) is a tuple $(\mathcal{A}, \mathcal{R}, \mathcal{P})$, where $\mathcal{A}$ is a set of arguments, $\mathcal{R} \subseteq \mathcal{A} \times \mathcal{A}$, and $\mathcal{P}$ is a preordering on $\mathcal{A} \times \mathcal{A}$.

Let $\gg_\mathcal{P}$ denote the strict ordering associated with $\mathcal{P}$, i.e., $y \gg_\mathcal{P} x$ iff $(y, x) \in \mathcal{P}$ and $(x, y) \notin \mathcal{P}$. Then:

- $\forall x, y \in \mathcal{A}$, x *defeats* y iff $x\mathcal{R}y$ and not $(y \gg_\mathcal{P} x)$.
- For $s \in \{$admissible, complete, preferred, stable, grounded$\}$, E is an s extension of $(\mathcal{A}, \mathcal{R}, \mathcal{P})$ iff E is an s extension of the Dung framework ($\mathcal{A}$, *defeat*).
- The justified arguments of $(\mathcal{A}, \mathcal{R}, \mathcal{P})$ are the justified arguments of ($\mathcal{A}$, *defeat*).

The preference relation in PAFs is entirely abstract. Value based Argumentation [11] gives more content to the notion of preferences, by relating the strength of arguments to the values promoted by accepting them. Note that preferences over values are subjective, and depend on the person or persons, i.e., the *audience*, to whom the argument is addressed. Hence, a Value based Argumentation Framework (VAF) extends Dung's framework to include a set of values, a function mapping arguments to these values and a set of audiences (i.e., a set of total orderings on these values). An argument x *defeats* y w.r.t. an audience a, if x attacks y, and a does not rank the value promoted by y higher than the value promoted by x.

Definition 9 A *Value based Argumentation Framework* is a 5-tuple $(\mathcal{A},\mathcal{R},V,val,P)$ where val is a function from $\mathcal{A}$ to a non-empty set of values V, and P is a set $\{\mathsf{a}_1, \ldots, \mathsf{a}_n\}$, where each a_i names a total ordering (audience) $>_{\mathsf{a}_i}$ on $V \times V$.

- An *audience specific* VAF ($aVAF$) is a 5-tuple $(\mathcal{A},\mathcal{R},V,val,\mathsf{a})$ where $\mathsf{a} \in P$.

- Given an $aVAF$ $(\mathcal{A},\mathcal{R},V,val,\mathsf{a})$, $\forall x, y \in \mathcal{A}$:
x *defeats*$_\mathsf{a}$ y iff $x\mathcal{R}y$, and it is not the case that $val(y) >_\mathsf{a} val(x)$.

- For $s \in \{$admissible, complete, preferred, stable, grounded$\}$, E is an s extension of $(\mathcal{A},\mathcal{R},V,val,\mathsf{a})$ iff E is an s extension of the Dung framework ($\mathcal{A}$, *defeat*$_\mathsf{a}$).

- The justified arguments of $(\mathcal{A},\mathcal{R},V,val,\mathsf{a})$ are the justified arguments of ($\mathcal{A}$, *defeat*$_\mathsf{a}$).

In PAFs and VAFs, preference orderings and values are applied to generate a defeat relation which is a subset of the attack relation containing only those attacks that are successful. In *Extended Argumentation* [38], this is achieved by directly attacking attacks with arguments, so that if x attacks y, and z attacks the attack from x to y, then z is interpreted as claiming that y is stronger than x. The rationale for concluding that y is stronger than x is thus itself now part of the domain of discourse, and is encoded as an argument z in the object-level framework. One can therefore account for reasoning and indeed arguing *about*, as well as *with*, defeasible and possibly conflicting information about the relative strengths of arguments.

Extended Argumentation Frameworks (*EAF*s) thus extend Dung frameworks with a second attack relation $\mathcal{D}$ from arguments to attacks.

Definition 10 An *Extended Argumentation Framework* is a tuple ($\mathcal{A}$, $\mathcal{R}$, $\mathcal{D}$), where $\mathcal{A}$ is a set of arguments, $\mathcal{R} \subseteq \mathcal{A} \times \mathcal{A}$, $\mathcal{D} \subseteq \mathcal{A} \times \mathcal{R}$, and if $(z,(x,y))$, $(z',(y,x)) \in \mathcal{D}$ then (z,z'), $(z',z) \in \mathcal{R}$

Note the constraint that if z attacks x's attack on y, and z' attacks y's attack on x, then z and z' are required to attack each other (since they express contradictory preferences).

The notion of a successful attack, henceforth referred to as a *defeat*, is then parameterised w.r.t. preferences specified by some given set S of arguments:

Definition 11 y defeats$_S$ x, denoted $y \rightarrow^s x$, iff $(y,x) \in \mathcal{R}$ and $\neg\exists z \in S$ s.t. (z,(y,x)) $\in S$.

A conflict free set of arguments is then defined to account for the case where y *asymmetrically* attacks x, but given a preference for x over y, both may appear in a conflict free set and hence an extension (as in [11]) (notice that a conflict free set does not admit arguments that symmetrically attack).

Definition 12 S is conflict free iff $\forall x,y \in S$: if $(y,x) \in \mathcal{R}$ then $(x,y) \notin \mathcal{R}$, and $\exists z \in S$ s.t. (z,(y,x)) $\in \mathcal{D}$.

The acceptability of an argument x w.r.t. a set S is now defined for an *EAF*. The basic idea is that for any attacker y of x, a reinstating attack $z \rightharpoonup y$ from $z \in S$ must itself be reinstated against $\mathcal{D}$ attacks on $z \rightharpoonup y$. The definition is motivated in more detail in [38] and requires the notion of a *reinstatement set* for a defeat.

Definition 13 Let $S \subseteq \mathcal{A}$ in ($\mathcal{A}$, $\mathcal{R}$, $\mathcal{D}$). Let $R_S = \{x_1 \rightarrow^S y_1, \ldots, x_n \rightarrow^S y_n\}$ where for $i = 1 \ldots n$, $x_i \in S$. Then R_S is a reinstatement set for $a \rightarrow^S b$, iff $a \rightarrow^S b \in R_S$ and:

$$\forall x \rightarrow^S y \in R_S, \forall y' \text{ s.t. } (y',(x,y)) \in \mathcal{D}, \exists x' \rightarrow^S y' \in R_S$$

Definition 14 x is acceptable w.r.t. $S \subseteq \mathcal{A}$ iff $\forall y$ s.t. $y \rightarrow^S x$, $\exists z \in S$ s.t. $z \rightarrow^S y$ and there is a *reinstatement set* for $z \rightarrow^S y$.

In Figure 1, a is acceptable w.r.t. S. We have $b \rightarrow^S a$, $c \rightarrow^S b$, and there is a reinstatement set $\{c \rightarrow^S b, d \rightarrow^S e\}$ for $c \rightarrow^S b$. Note that if we had $f \twoheadrightarrow (d \rightharpoonup e)$, and no argument in S defeating f, there would be no reinstatement set, and a would not be acceptable w.r.t. S.

Given the definitions of conflict free and acceptability for *EAF*s, admissible, complete, preferred and stable semantics for *EAF*s are now defined as for AFs in Definition

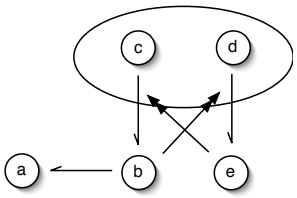

Figure 1: a is EAF acceptable w.r.t. S (single headed arrows denote ordinary attacks, and double headed arrows denote attacks on attacks).

3 (except that x defeats$_S$ y replaces $(x, y) \in \mathcal{R}$ in the definition of stable extensions). In [38] it is shown that *EAF*s inherit many of the results that hold for Dung frameworks. However, since an *EAF*'s characteristic function F (which takes as input a set S of arguments and returns the arguments acceptable w.r.t. S) is not in general monotonic, the grounded extension cannot be characterised as the least (under set inclusion) complete extension. However, starting with the empty set, iterating F does yield a monotonically increasing sequence. Let $G^0 = \emptyset$, $G^{i+1} = F(G^i)$. The grounded extension of an *EAF* is then defined as $\bigcup_{i=0}^{\infty} G^i$.

3 Formalising Abstract Argumentation in Metalevel Argumentation Frameworks

In this section we formalise Metalevel Argumentation Frameworks which essentially augment a Dung framework $(\mathcal{A}, \mathcal{R})$ with a language for representing the claims of arguments in $\mathcal{A}$, and constraints on the attack relation $\mathcal{R}$ that account for the arguments' claims. The arguments in $\mathcal{A}$ are arguments claiming statements *about* object level abstract argumentation frameworks, and the constraints on $\mathcal{R}$ essentially characterise the reasoning by which one determines the justified arguments of the object level framework. We then show how the varieties of abstract argumentation reviewed in Section 2 can be formalised as instances of metalevel argumentation.

3.1 Introducing Metalevel Argumentation

Section 2's review of abstract argumentation described how, in general, establishing that an argument x is justified under a semantics s is based on evaluation of the acceptability of x w.r.t. sets of arguments. The acceptability of x w.r.t. some subset S of $\mathcal{A}$, hinges on whether attacks of the form $y\mathcal{R}x$ succeed as defeats. If for each such $y\mathcal{R}x$, y is successfully attacked (defeated) by some $z \in S$, then z effectively undermines the success of the attack from y to x; z can be said to *reinstate* x. The rules defining legal labelling assignments for Dung frameworks correspond intuitively to the extension-based use of this reinstatement principle:

R1 x is legally `IN` (i.e., x is in an admissible extension) in an s labelling (credulously justified under the semantics s) iff every attack $y\mathcal{R}x$ on x fails.

R2 An attack $y\mathcal{R}x$ fails if y is `OUT` (i.e., y is attacked by a reinstating z that is `IN`).

R3 x is legally `OUT` (potentially rejected under the semantics s) if at least one attack $y\mathcal{R}x$ succeeds (i.e., y in `IN`).

Given a Dung framework $\Delta = (\mathcal{A}, \mathcal{R})$, a statement asserting the existence of an argument $x \in \mathcal{A}$, and its purported membership of an admissible extension of Δ, constitutes a metalevel argument ξ claiming 'x is justified'. That is to say, ξ is an argument of the form 'there is an $x \in \mathcal{A}$ that is in an admissible extension of Δ, and so x is justified'. The existence of an attack $y\mathcal{R}x$ constitutes a metalevel argument α, that claims that 'y *successfully attacks* and so *defeats* x' (note that logics for asserting statements of this kind have been proposed in [18, 33, 53] and will be briefly reviewed in Section 5). Corresponding to R1, R2 and R3, we have the following metalevel reasoning (MR) principles:

MR1 The justified status of x is challenged by any defeat on x, and so we have a metalevel attack from α to ξ.

MR2 y does not defeat x if y is rejected. Hence, in the metalevel, α is attacked by an argument τ claiming that 'y is rejected'. Thus τ reinstates ξ (characterising the object level reinstatement of x.)

MR3 y defeats x if y is justified. Hence, in the metalevel, the argument ψ claiming that 'y is justified' attacks the argument τ claiming that 'y is rejected', so reinstating the argument α that claims that 'y defeats x'.

The metalevel arguments and their attacks are illustrated in Figure 2a. Figure 2b shows the metalevel argumentation corresponding to the object level reinstatement of x, by some z that attacks y. Finally, Figure 2c shows the metalevel argumentation corresponding to an object level symmetric attack between arguments x and y ($x\mathcal{R}y$ and $y\mathcal{R}x$). In this case the metalevel argumentation characterises the object level reinstatement of x by x itself.

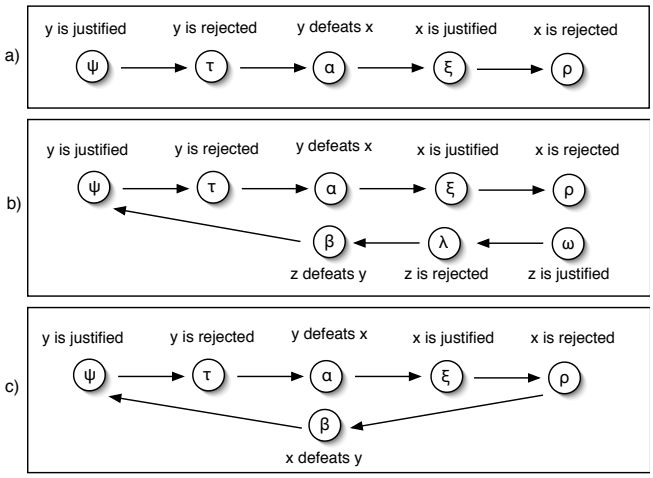

Figure 2: Meta-level arguments and their attacks

Our description of metalevel arguments assumes their classification according to the nature of the claims they make about an object level framework, and attacks amongst these metalevel arguments based on this classification. This suggests the more general notion of a structured argumentation framework, whereby one augments a Dung framework $(\mathcal{A}, \mathcal{R})$ to include a mapping $\mathcal{C}$ from arguments in $\mathcal{A}$ to claims specified in some language $\mathcal{L}$, and based on these claims a set of constraints that are thus imposed on $\mathcal{R}$.

Definition 15 A *Structured Argumentation Framework* (SAF) is a tuple Δ = ($\mathcal{A}$, $\mathcal{R}$, $\mathcal{C}$, $\mathcal{L}, \mathcal{D}$), where:

- ($\mathcal{A}$, $\mathcal{R}$) is a Dung argumentation framework
- $\mathcal{L}$ is a claim language
- $\mathcal{C}$ is a claim function mapping arguments in $\mathcal{A}$ to sets of wff in $\mathcal{L}$
- $\mathcal{R}$ satisfies a set of constraints $\mathcal{D}$, where each constraint is a rule of the form:

  if $l \in \mathcal{C}(\alpha)$ and $l' \in \mathcal{C}(\beta)$ then $(\alpha, \beta) \in \mathcal{R}$

  where $\alpha, \beta \in \mathcal{A}$, and l, l' are wff of $\mathcal{L}$.

  We say that *an attack relation $\mathcal{R}$ is defined by $\mathcal{D}$* iff $(\alpha, \beta) \in \mathcal{R}$ implies that the claims of α and β satisfy the antecedent of some constraint rule in $\mathcal{D}$.

For $s \in \{$admissible, complete, grounded, preferred, stable$\}$, we say that E is an s extension of Δ iff E is an s extension of ($\mathcal{A}$, $\mathcal{R}$), and $\alpha \in \mathcal{A}$ is a justified argument of Δ iff α is a justified argument of ($\mathcal{A}$, $\mathcal{R}$).

We now identify a special class of SAFs — Metalevel Argumentation Frameworks (MAFs) — by specifying a language $\mathcal{L}$ whose wff are built from constants, sets of constants, sets of pairs of constants (where these constants may name arguments and or values), and predicates of the form $justified$, $rejected$, $defeat$ e.t.c. Claims in this language will thus refer to the properties of arguments, values and preferences in an object level framework.

Definition 16 A *Metalevel Argumentation Framework* (MAF) is a SAF $\Delta_{\mathcal{M}}$ = ($\mathcal{A}$, $\mathcal{R}$, $\mathcal{C}$, $\mathcal{L}_{\mathcal{M}}, \mathcal{D}$), where $\mathcal{L}_{\mathcal{M}}$ consists of a countable set of constant symbols and the set of predicates[3]:

$$\{justified, defeat, rejected, preferred, val, val\_pref, audience\}.$$

The set of wff of $\mathcal{L}_{\mathcal{M}}$ is defined by the following BNF:

$\mathcal{L}_{\mathcal{M}} : X ::= x, \{x_1, \ldots, x_n\}, \{(x_1, x_2), \ldots, (x_m, x_n)\} \mid justified(X) \mid$
$rejected(X) \mid defeat(X, X') \mid preferred(X, X') \mid val(X, X') \mid val\_pref(X, X')$
$\mid audience(X)$

where x, x_i ranges over the constant symbols.

In the following sections we show how the varieties of abstract argumentation reviewed in Section 2 can be formalised as instances of argumentation in a MAF. In each case, the constraints in $\mathcal{D}$ characterise evaluation of the justified arguments in the object level framework.

3.2 Formalising Dung's abstract argumentation theory in Metalevel Argumentation Frameworks

A *MAF*'s formalisation of Dung argumentation consists of: i) metalevel arguments whose construction is based on the existence of object level attacks, and so claim defeats between object level arguments; ii) metalevel arguments whose construction is

[3] $\mathcal{L}_{\mathcal{M}}$ is extensible: here we include the predicate and constant symbols required in this paper.

based on the existence of object level arguments and their purported membership or non-membership of admissible extensions, and that claim that the object level arguments are justified, respectively rejected; iii) constraints on the metalevel attack relation that capture MR1 – MR3 in Section 3.1:

Definition 17 A Dung MAF is a tuple ($\mathcal{A}_\mathcal{M}$, $\mathcal{R}_\mathcal{M}$, $\mathcal{C}$, $\mathcal{L}_M, \mathcal{D}_d$), where $\mathcal{D}_d$ =

$\{D1$: if $defeat(Y, X) \in \mathcal{C}(\alpha)$ and $justified(X) \in \mathcal{C}(\beta)$ then $(\alpha, \beta) \in \mathcal{R}_\mathcal{M}$

$D2$: if $defeat(Y, X) \in \mathcal{C}(\alpha)$ and $rejected(Y) \in \mathcal{C}(\beta)$ then $(\beta, \alpha) \in \mathcal{R}_\mathcal{M}$

$D3$: if $justified(X) \in \mathcal{C}(\alpha)$ and $rejected(X) \in \mathcal{C}(\beta)$ then $(\alpha, \beta) \in \mathcal{R}_\mathcal{M}$ $\}$

Definition 18 A Dung MAF Δ_M = ($\mathcal{A}_\mathcal{M}$, $\mathcal{R}_\mathcal{M}$, $\mathcal{C}$, $\mathcal{L}_\mathcal{M}, \mathcal{D}_d$) is said to be a metalevel formulation of the Dung AF Δ = ($\mathcal{A}, \mathcal{R}$) iff:

- $\lceil x \rceil$ is a constant in $\mathcal{L}_M$ iff $x \in \mathcal{A}$ [4]
- $\mathcal{A}_\mathcal{M}$ is the union of the disjoint sets $\mathcal{A}_{\mathcal{M}1}$, $\mathcal{A}_{\mathcal{M}2}$, $\mathcal{A}_{\mathcal{M}3}$, where:
  1. $\alpha \in \mathcal{A}_{\mathcal{M}1}$, $\mathcal{C}(\alpha) = \{justified(\lceil x \rceil)\}$ iff $x \in \mathcal{A}$
  2. $\beta \in \mathcal{A}_{\mathcal{M}2}$, $\mathcal{C}(\beta) = \{rejected(\lceil x \rceil)\}$ iff $x \in \mathcal{A}$
  3. $\gamma \in \mathcal{A}_{\mathcal{M}3}$, $\mathcal{C}(\gamma) = \{defeat(\lceil y \rceil, \lceil x \rceil)\}$ iff $(y, x) \in \mathcal{R}$
- $\mathcal{R}_\mathcal{M}$ is defined by $\mathcal{D}_d$.

Notation 2 In general, a metalevel argument may be mapped by $\mathcal{C}$ to more than one claim. However, from hereon we only consider arguments that make single claims, and so as an abuse of notation may denote such arguments by the claims they make:

- If $\mathcal{C}(\alpha) = \{justified(\lceil x \rceil)\}$ we write $(j - x)$ to refer to α.
- If $\mathcal{C}(\alpha) = \{rejected(\lceil x \rceil)\}$ we write $(r - x)$ to refer to α.
- If $\mathcal{C}(\alpha) = \{defeat(\lceil y \rceil, \lceil x \rceil)\}$ we write $(y$ def $x)$ to refer to α.

Given a Dung framework Δ = ($\mathcal{A}, \mathcal{R}$), one can evaluate the justified arguments of Δ by evaluating the justified arguments of the Dung *MAF* Δ_M.

Theorem 2 Let Δ_M = ($\mathcal{A}_\mathcal{M}$, $\mathcal{R}_\mathcal{M}$, $\mathcal{C}$, $\mathcal{L}_\mathcal{M}, \mathcal{D}_d$) be the MAF of a Dung framework $(\mathcal{A}, \mathcal{R})$. Then for $s \in \{$complete, grounded, preferred, stable$\}$, $(j - x) \in \mathcal{A}_\mathcal{M}$ is a credulously, respectively sceptically, justified argument of Δ_M under the s semantics, iff $x \in \mathcal{A}$ is a credulously, respectively sceptically, justified argument of Δ under the s semantics.

Note the extra expressivity that results from the metalevel formulation of a Dung framework. Given an extension E of Δ, the corresponding extension of Δ_M identifies the arguments labelled `IN` and OUT by a corresponding labelling for E.

Proposition 2 Let $\Delta_\mathcal{M}$ = ($\mathcal{A}_\mathcal{M}$, $\mathcal{R}_\mathcal{M}$, $\mathcal{C}$, $\mathcal{L}_\mathcal{M}, \mathcal{D}_d$) be the MAF of a Dung framework Δ = ($\mathcal{A}, \mathcal{R}$). For $s \in \{$admissible, complete, grounded, preferred, stable$\}$:
There exists an s labelling $\mathcal{L}$ of Δ iff there exists an s extension E of $\Delta_\mathcal{M}$ such that:
1) $x \in$ `in`($\mathcal{L}$) iff $(j - x) \in E$; 2) $y \in$ `out`($\mathcal{L}$) iff $(r - y) \in E$.

[4] Sense (Gödel) quotes $\lceil \rceil$ standardly abbreviate metalevel representations of object level formulae.

To illustrate, consider Figure 3's object level Dung framework Δ, and its metalevel formulation Δ_M. Observe that $\{d, c, b\}$ is the grounded, preferred and stable extension of Δ, corresponding to $\{(d \text{ def } e), (j-d), (r-e), (c \text{ def } e), (j-c), (r-a), (b \text{ def } a), (j-b)\}$ being the grounded, preferred and stable extension of Δ_M.

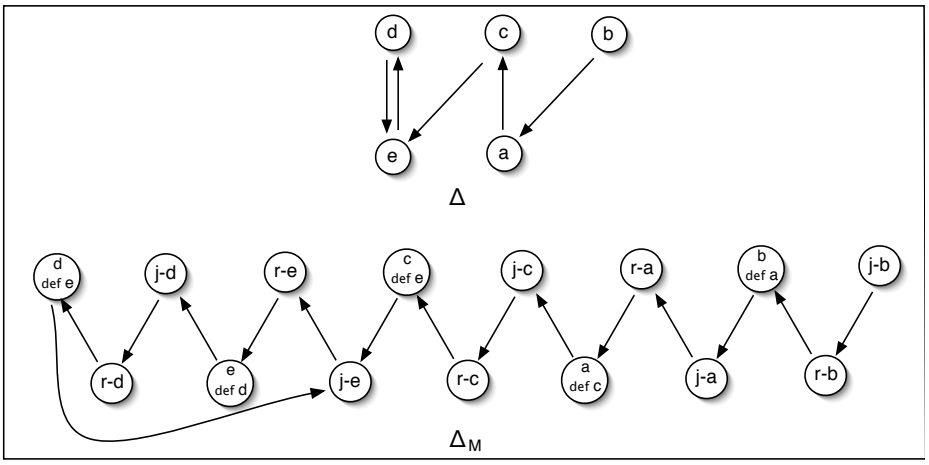

Figure 3: A Dung argumentation framework and its metalevel formulation

3.3 Formalising Collective Attacks in Meta-level Argumentation Frameworks

The existence of a collective attack $B\mathcal{R}x$ constitutes an argument α claiming 'B defeats x', and for each $y \in B$, an argument claiming 'y is rejected' attacks α.

Definition 19 A DungC *MAF* is a tuple ($\mathcal{A}_\mathcal{M}$, $\mathcal{R}_\mathcal{M}$, $\mathcal{C}$, $\mathcal{L}_\mathcal{M}, \mathcal{D}_{dc}$) , where $\mathcal{D}_{dc}$ =

$\{D1$: if $defeat(Y, X) \in \mathcal{C}(\alpha)$ and $justified(X) \in \mathcal{C}(\beta)$ then $(\alpha, \beta) \in \mathcal{R}_\mathcal{M}$

$D2$: if $defeat(Y, X) \in \mathcal{C}(\alpha)$, $rejected(Z) \in \mathcal{C}(\beta)$ and $Z \in Y$, then $(\beta, \alpha) \in \mathcal{R}_\mathcal{M}$

$D3$: if $justified(X) \in \mathcal{C}(\alpha)$ and $rejected(X) \in \mathcal{C}(\beta)$ then $(\alpha, \beta) \in \mathcal{R}_\mathcal{M}$ $\}$

Definition 20 A DungC MAF Δ_M = ($\mathcal{A}_\mathcal{M}$, $\mathcal{R}_\mathcal{M}$, $\mathcal{C}$, $\mathcal{L}_\mathcal{M}, \mathcal{D}_{dc}$) is said to be a metalevel formulation of the DungC framework Δ = ($\mathcal{A}, \mathcal{R}$) iff:

- $\lceil x \rceil$ is a constant in $\mathcal{L}_M$ iff $x \in \mathcal{A}$
- $\mathcal{A}_\mathcal{M}$ is the union of the disjoint sets $\mathcal{A}_{\mathcal{M}1}$, $\mathcal{A}_{\mathcal{M}2}$, $\mathcal{A}_{\mathcal{M}3}$, where:
  1. $\alpha \in \mathcal{A}_{\mathcal{M}1}, \mathcal{C}_\mathcal{M}(\alpha) = \{justified(\lceil x \rceil)\}$ iff $x \in \mathcal{A}$
  2. $\beta \in \mathcal{A}_{\mathcal{M}2}, \mathcal{C}_\mathcal{M}(\beta) = \{rejected(\lceil x \rceil)\}$ iff $x \in \mathcal{A}$
  3. $\gamma \in \mathcal{A}_{\mathcal{M}3}, \mathcal{C}_\mathcal{M}(\gamma) = \{defeat(\{\lceil y_1 \rceil, \ldots, \lceil y_n \rceil\}, \lceil x \rceil)\}$ iff $(\{y_1, \ldots, y_n\}, \{x\}) \in \mathcal{R}$
- $\mathcal{R}_\mathcal{M}$ is defined by $\mathcal{D}_{dc}$.

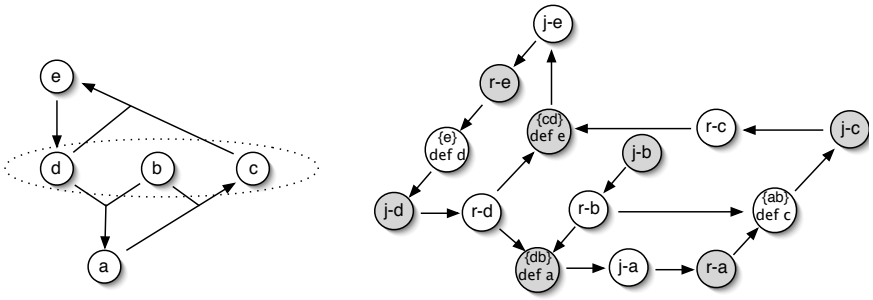

Figure 4: A DungC framework with collective attacks, and its metalevel formulation

We employ a similar abuse of notation as in Notation 2, writing $(\{y_1, \ldots, y_n\}$ def $x)$ to refer to an argument claiming $defeat(\{\lceil y_1 \rceil, \ldots, \lceil y_n \rceil\}, \lceil x \rceil)$.
Consider Figure 4's object level DungC framework with the following attacks:

$\{a, b\} \rightarrow c$, $\{d, b\} \rightarrow a$, $\{e\} \rightarrow d$, $\{d, c\} \rightarrow e$,

One can easily verify that $E1 = \{d, b, c\}$ (the encircled arguments) and $E2 = \{a, b, e\}$ are the preferred and stable extensions, and $\{b\}$ the grounded extension. Figure 4 shows the object level framework's MAF, and the arguments (shaded) in the preferred and stable extension $E1'$ corresponding to $E1$. In general:

Theorem 3 Let $\Delta_M = (\mathcal{A}_\mathcal{M}, \mathcal{R}_\mathcal{M}, \mathcal{C}, \mathcal{L}_\mathcal{M}, \mathcal{D}_{dc})$ be the MAF of a DungC framework $\Delta = (\mathcal{A}, \mathcal{R})$. Then for $s \in \{$complete, grounded, preferred, stable$\}$, $(j - x) \in \mathcal{A}_\mathcal{M}$ is a credulously, respectively sceptically, justified argument of Δ_M under the s semantics, iff $x \in \mathcal{A}$ is a credulously, respectively sceptically, justified argument of Δ under the s semantics.

3.4 Formalising Preference Based Argumentation in Meta-level Argumentation Frameworks

As discussed in Section 2, PAFs, VAFs and *EAF*s provide *additional* information which enable, when evaluating the framework, to say that an attack fails to succeed as a defeat, even though the attacking argument is justified. In terms of metalevel argumentation, this additional information is the source of additional *arguments* which can be used to attack arguments of the form (x def y).

Given an object level PAF, the existence of a strict preference $x \gg_\mathcal{P} y$ constitutes a metalevel argument claiming that 'x is strictly preferred to y'. In addition to MR2 in Section 3.1, we thus have the additional following metalevel characterisation of the object level reasoning that determines the justified arguments of a PAF:

> MR2$'$: y does not defeat x (i.e., y's attack on x fails) if x is strictly preferred to y. Hence, in the metalevel, α claiming 'y defeats x' is attacked by an argument ρ claiming that 'x is strictly preferred to y'. Thus ρ reinstates ξ claiming 'x is justified'.

Definition 21 A P-MAF is a tuple $(\mathcal{A}_\mathcal{M}, \mathcal{R}_\mathcal{M}, \mathcal{C}, \mathcal{L}_\mathcal{M}, \mathcal{D}_p)$, where $\mathcal{D}_p = \mathcal{D}_d \cup$ $\{$ D4 : if $defeat(Y, X) \in \mathcal{C}(\alpha)$ and $s\_preferred(X, Y) \in \mathcal{C}(\beta)$ then $(\beta, \alpha) \in \mathcal{R}_\mathcal{M}\}$

Definition 22 A $P\text{-}MAF$ Δ_M = ($\mathcal{A}_\mathcal{M}$, $\mathcal{R}_\mathcal{M}$, $\mathcal{C}$, $\mathcal{L}_\mathcal{M}, \mathcal{D}_p$) is said to be a metalevel formulation of the PAF Δ = ($\mathcal{A}, \mathcal{R}, \mathcal{P}$) iff:

- $\lceil x \rceil$ is a constant in $\mathcal{L}_\mathcal{M}$ iff $x \in \mathcal{A}$
- $\mathcal{A}_\mathcal{M}$ is the union of the disjoint sets $\mathcal{A}_{\mathcal{M}1}$, $\mathcal{A}_{\mathcal{M}2}$, $\mathcal{A}_{\mathcal{M}3}$, $\mathcal{A}_{\mathcal{M}4}$, where $\mathcal{A}_{\mathcal{M}1}$, $\mathcal{A}_{\mathcal{M}2}$ and $\mathcal{A}_{\mathcal{M}3}$ are defined as in Definition 18, and:
  4. $\delta \in \mathcal{A}_{\mathcal{M}4}$, $\mathcal{C}(\delta) = \{s\_preferred(\lceil x \rceil, \lceil y \rceil)\}$ iff $x \gg_\mathcal{P} y$
     (from hereon, we may write (xPy) to refer to an argument of the form δ).
- $\mathcal{R}_\mathcal{M}$ is defined by $\mathcal{D}_p$.

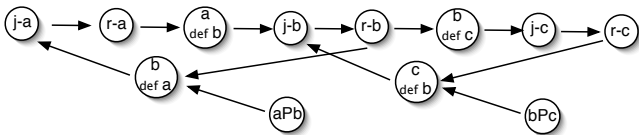

Figure 5: A PAF formulated as a metalevel framework

Figure 5 shows the metalevel formulation of the PAF ($\mathcal{A} = \{a, b, c\}$, $\mathcal{R} = \{(a, b), (b, a), (b, c), (c, b)\}$, $\mathcal{P} = \{(a, b), (b, c)\}$). $E' = \{(j - a), (a \text{ def } b), (r - b), (j - c), (aPb), (bPc)\}$ is the single grounded/preferred/stable extension of the $P\text{-}MAF$, corresponding to the single grounded/preferred/stable extension $\{a, c\}$ of the object level PAF.

Theorem 4 Let Δ_M = ($\mathcal{A}_\mathcal{M}$, $\mathcal{R}_\mathcal{M}$, $\mathcal{C}$, $\mathcal{L}_\mathcal{M}, \mathcal{D}_p$) be the $P\text{-}MAF$ of a PAF Δ = ($\mathcal{A}, \mathcal{R}, \mathcal{P}$). Then for $s \in$ {complete, grounded, preferred, stable}, $(j - x) \in \mathcal{A}_\mathcal{M}$ is a credulously, respectively sceptically, justified argument of Δ_M under the s semantics, iff $x \in \mathcal{A}$ is a credulously, respectively sceptically, justified argument of Δ under the s semantics.

Note that the above example illustrates *resolution* of a framework obtained by replacing symmetric attacks with asymmetric attacks. Properties relating frameworks and their resolutions have been studied in [8] and [33]. Our metalevel formulation of PAFs provides a general setting for further formal study of *resolution semantics* [8], whereby the reasoning by which an object-level framework's resolutions are obtained is now modelled in terms of argumentation within its metalevel formulation.

3.5 Formalising Value Based Argumentation in Meta-level Argumentation Frameworks

In VAFs we have additional information: a set of values, a function mapping arguments to values, and a set of audiences representing totally ordered preference relations on values. The metalevel characterisation of the object level reasoning applied to determine the justified arguments of an audience specific VAF ($aVAF$), augments MR2 in Section 3.1 as follows:

MR2.1 y does not defeat x (i.e., y's attack on x fails) if x's value is preferred to y's value. Hence, in the metalevel, α claiming 'y defeats x' is attacked by *a value*

preference argument ν claiming that 'x's value is preferred to y's value'. Thus ν reinstates ξ claiming 'x is justified'.

MR2.2 The preference of x's value over y's value is challenged by the contrary preference. Hence, in the metalevel each ν is symmetrically attacked by the metalevel argument ν' claiming that 'y's value is preferred to x's value'.

MR2.3 The $aVAF$'s choice of audience (total ordering on values) endorses the pairwise value preferences specified by the total ordering. Thus, an audience constitutes a metalevel argument that claims a total ordering, and that attacks any value preference arguments that contradict the endorsed value preferences. Hence, if $val(x) >_{\mathsf{a}} val(y)$, then a constitutes an *audience argument* that attacks ν' and thus reinstates ν.

We can also represent all audiences in a VAF, where given $P = \{\mathsf{a}_1, \ldots, \mathsf{a}_n\}$, then each a_i constitutes an *audience argument* that symmetrically attacks every other audience argument, since by definition, for all i, j such that $i \neq j$, a_i and a_j contradict each other on at least one value preference.

Definition 23 A V-MAF is a tuple ($\mathcal{A}_\mathcal{M}$, $\mathcal{R}_\mathcal{M}$, $\mathcal{C}$, $\mathcal{L}_\mathcal{M}, \mathcal{D}_v$) , where $\mathcal{D}_v = \mathcal{D}_d \cup$

{D4 : if $defeat(Y, X) \in \mathcal{C}(\alpha)$ and $val\_pref(val(X, V), val(Y, V')) \in \mathcal{C}(\beta)$ then $(\beta, \alpha) \in \mathcal{R}_\mathcal{M}$

D5 : if $val\_pref(val(Y, V'), val(X, V)) \in \mathcal{C}(\alpha)$ and $val\_pref(val(X, V), val(Y, V'))$ $\in \mathcal{C}(\beta)$ then $(\beta, \alpha) \in \mathcal{R}_\mathcal{M}$

D6 : if $audience(Z) \in \mathcal{C}(\alpha)$ and $val\_pref(val(Y, V'), val(X, V)) \in \mathcal{C}(\beta)$, where Z is a set of pairs of constants such that $(V, V') \in Z$, then $(\alpha, \beta) \in \mathcal{R}_\mathcal{M}$

D7 : if $audience(Z) \in \mathcal{C}(\beta)$ and $audience(Z') \in \mathcal{C}(\alpha)$, where Z and Z' are sets of pairs of constants such that $(V, V') \in Z$, $(V', V) \in Z'$, then $(\beta, \alpha) \in \mathcal{R}_\mathcal{M}$}

Definition 24 A V-MAF Δ_M = ($\mathcal{A}_\mathcal{M}$, $\mathcal{R}_\mathcal{M}$, $\mathcal{C}$, $\mathcal{L}_\mathcal{M}, \mathcal{D}_v$) is said to be a metalevel formulation of the VAF Δ = ($\mathcal{A}$, $\mathcal{R}$, V, val, P) iff:

- $\lceil x \rceil$ is a constant in $\mathcal{L}_\mathcal{M}$ iff $x \in \mathcal{A}$ or $x \in V$
- $\mathcal{A}_\mathcal{M}$ is the union of the disjoint sets $\mathcal{A}_{\mathcal{M}1}$, $\mathcal{A}_{\mathcal{M}2}$, $\mathcal{A}_{\mathcal{M}3}$, $\mathcal{A}_{\mathcal{M}4}$, $\mathcal{A}_{\mathcal{M}5}$ where $\mathcal{A}_{\mathcal{M}1} \ldots \mathcal{A}_{\mathcal{M}3}$ are defined as in Definition 18, and:
  4. $\nu \in \mathcal{A}_{\mathcal{M}4}$, $\mathcal{C}(\nu) = \{val\_pref(val(\lceil x \rceil, \lceil v \rceil), val(\lceil y \rceil, \lceil v' \rceil))\}$ iff $val(x) = v$, $val(y) = v'$, and $v \neq v'$ (henceforth we may write $(x_v P y_{v'})$ to refer to an argument of the form ν)
  5. $\epsilon \in \mathcal{A}_{\mathcal{M}5}$, $\mathcal{C}(\epsilon) = \{ \{(v_1, v_2) \ldots (v_m, v_n))\} \}$ iff $\mathsf{a} \in P$, $\mathsf{a} = v_1 >_{\mathsf{a}} v_2 \ldots v_m >_{\mathsf{a}} v_n$ (henceforth we may write $(>_{\mathsf{a}})$ to refer to an argument of the form ϵ)
- $\mathcal{R}_\mathcal{M}$ is defined by $\mathcal{D}_v$.

The V-MAF of an $aVAF$ ($\mathcal{A}$,$\mathcal{R}$,V,val,a) is defined as above, where $P = \{\mathsf{a}\}$.

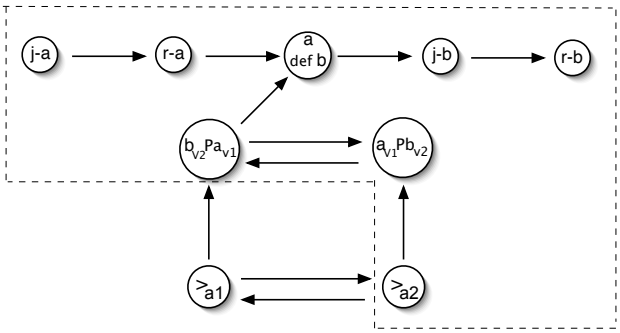

Figure 6: A VAF and aVAF (within the dashed line) formulated as metalevel frameworks

Example 1 Figure 6 shows the V-MAF formulation of the VAF:
($\mathcal{A} = \{a, b\}$, $\mathcal{R} = \{(a, b)\}$, $V = \{v1, v2\}$, $val(a) = v1, val(b) = v2$, $P = \{\mathsf{a}_1 = \{(v_1, v_2)\}, \mathsf{a}_2 = \{(v_2, v_1)\}\}$)

We have two preferred extensions of the V-MAF formulation, one for each audience:

$E1 = \{(>_{a1}), (a_{v1}Pb_{v2}), (j - a), (a \text{ def } b), (r - b)\}$

$E2 = \{(>_{a2}), (b_{v2}Pa_{v1}), (j - a), (j - b)\}$

$E2$ is then the single preferred extension of the V-MAF formulation of the $aVAF$ for audience *a2* (shown outlined by the dashed line in Figure 6).

Theorem 5 Let Δ_M = ($\mathcal{A}_\mathcal{M}$, $\mathcal{R}_\mathcal{M}$, $\mathcal{C}$, $\mathcal{L}_\mathcal{M}, \mathcal{D}_v$) be the V-MAF of an $aVAF$ Δ = ($\mathcal{A}$,$\mathcal{R}$,V,val,a). Then for $s \in \{$complete, grounded, preferred, stable$\}$, $(j - x) \in \mathcal{A}_\mathcal{M}$ is a credulously, respectively sceptically, justified argument of Δ_M under the s semantics, iff $x \in \mathcal{A}$ is a credulously, respectively sceptically, justified argument of Δ under the s semantics.

In [11], the arguments in every preferred extension for every audience in a VAF are referred to as *objectively* acceptable. The arguments that appear in at least one preferred extension for at least one audience in a VAF are referred to as *subjectively* acceptable. These notions correspond to the sceptically, respectively credulously justified arguments (under the preferred semantics) of the VAF's metalevel formulation.

Definition 25 Given a VAF Δ = $(\mathcal{A}, \mathcal{R}, V, val, P)$, and an argument $x \in \mathcal{A}$:

- x is objectively acceptable iff $\forall \mathsf{a} \in P$, x is a sceptically justified argument of the $aVAF$ ($\mathcal{A}$,$\mathcal{R}$,V,val,a) under the preferred semantics.
- x is subjectively acceptable iff $\exists \mathsf{a} \in P$, x is a credulously justified argument of the $aVAF$ ($\mathcal{A}$,$\mathcal{R}$,V,val,a) under the preferred semantics.

Theorem 6 Let Δ_M = ($\mathcal{A}_\mathcal{M}$, $\mathcal{R}_\mathcal{M}$, $\mathcal{C}$, $\mathcal{L}_\mathcal{M}, \mathcal{D}_v$) be the V-MAF of a VAF Δ = ($\mathcal{A}$,$\mathcal{R}$,V,val,P). Then for any $x \in \mathcal{A}$, $(j - x) \in \mathcal{A}_\mathcal{M}$:

1. x is an objectively acceptable argument of Δ iff $(j - x)$ is a sceptically justified argument of Δ_M under the preferred semantics.

2. x is a subjectively acceptable argument of Δ iff $(j - x)$ is a credulously justified argument of Δ_M under the preferred semantics.

3.6 Formalising Extended Argumentation in Meta-level Argumentation Frameworks

In both preference and value based argumentation, information assumed to be exogenous to the domain of argumentation based reasoning is used to undermine the success of attack as defeats. In *EAF*s, such information is part of the object level domain of argumentation, and in keeping with the abstract nature of Dung's approach no commitments are made to the nature of this information. Rather, the use of the information to undermine the success of attacks is abstractly characterised; by defining a new attack relation that originates from an argument, and that attacks an attack.

Intuitively, the metalevel characterisation of the object level reasoning in an *EAF*, extends that in a Dung *MAF* so that arguments of the form $(j - x)$ and arguments of the form $(q$ def $r)$ are respectively attacked by arguments of the form $(y$ def $x)$ and $(p$ def $(q$ def $r))$. Just as $(y$ def $x)$ is attacked by $(r - y)$ and $(r - y)$ is attacked by $(j - y)$, so $(p$ def $(q$ def $r))$ is attacked by $(r - p)$ and $(r - p)$ is attacked by $(j - p)$.

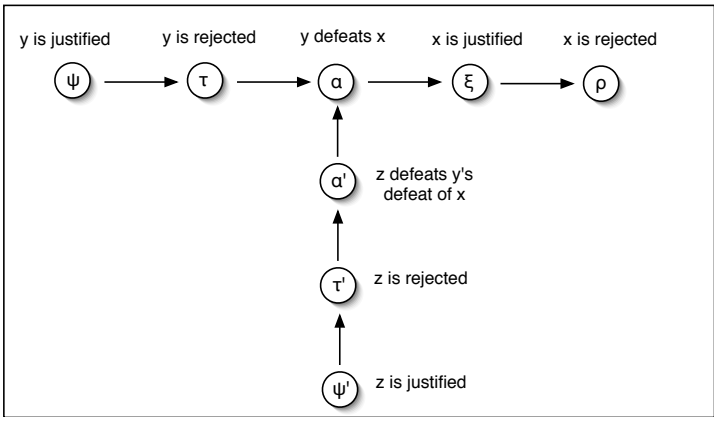

Figure 7: Meta-level formulation of an attack on an attack

Hence, in addition to MR1, MR2 and MR3 in Section 3.1 we have the following analogous metalevel reasoning principles:

MR1′ The existence of an attack $(z, (y, x))$, constitutes a metalevel argument α' claiming 'z *successfully attacks* and so *defeats* y's defeat of x'; hence α' attacks α claiming 'y *defeats* x' (see Figure 7). Any metalevel attack on α', challenging z's defeat of y's defeat of x, reinstates α, and characterises the object level reinstatement of y's attack on x.

MR2′ If z is rejected, then z does not defeat y's defeat of x. Hence, α' is attacked by τ' claiming that 'z is rejected'. Thus τ' reinstates α.

MR3′ If z is justified, then z defeats y's defeat of x. Thus, ψ' claiming that 'z is justified' attacks τ' claiming 'z is rejected', so reinstating α'.

Definition 26 An E-MAF is a tuple ($\mathcal{A}_\mathcal{M}$, $\mathcal{R}_\mathcal{M}$, $\mathcal{C}$, $\mathcal{L}_\mathcal{M}, \mathcal{D}_e$), where:
$\mathcal{D}_e$ = $\mathcal{D}_d \cup \{$ $D4$: if $defeat(Z, (defeat(Y, X)) \in \mathcal{C}(\alpha)$ and $defeat(Y, X) \in \mathcal{C}(\beta)$ then $(\alpha, \beta) \in \mathcal{R}_\mathcal{M}\}$.

Definition 27 An E-MAF Δ_M = ($\mathcal{A}_\mathcal{M}$, $\mathcal{R}_\mathcal{M}$, $\mathcal{C}$, $\mathcal{L}_\mathcal{M}, \mathcal{D}_e$) is said to be a metalevel formulation of the EAF Δ = ($\mathcal{A}, \mathcal{R}, \mathcal{D}$) iff:

- $\lceil x \rceil$ is a constant in $\mathcal{L}_\mathcal{M}$ iff $x \in \mathcal{A}$
- $\mathcal{A}_\mathcal{M}$ is the union of the disjoint sets $\mathcal{A}_{\mathcal{M}1}$, $\mathcal{A}_{\mathcal{M}2}$, $\mathcal{A}_{\mathcal{M}3}$, $\mathcal{A}_{\mathcal{M}4}$, where $\mathcal{A}_{\mathcal{M}1}$, $\mathcal{A}_{\mathcal{M}2}$ and $\mathcal{A}_{\mathcal{M}3}$ are defined as in Definition 18, and:
  4. $\delta \in \mathcal{A}_{\mathcal{M}4}$, $\mathcal{C}(\delta)$ = $\{defeat(\lceil z \rceil, (defeat(\lceil y \rceil, \lceil x \rceil))\}$ iff (z,(y, x)) $\in \mathcal{D}$ (from hereon, we write $(zD(yDx))$ to refer to an argument of the form δ).
- $\mathcal{R}_\mathcal{M}$ is defined by $\mathcal{D}_e$.

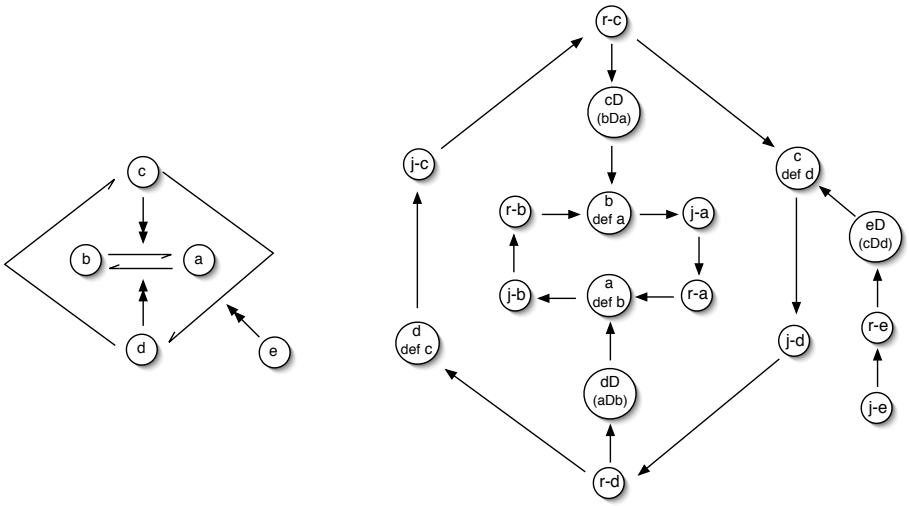

Figure 8: An EAF and its meta-level formulation

Figure 8 shows an *EAF*[5] and its metalevel formulation. In general, a correspondence does not hold between the justified arguments of *EAF*s and their metalevel formulations. Consider the *EAF* Δ in Figure 9, in which a attacks b, and b itself attacks the attack from a [6]. Δ's single preferred extension is E = $\{a, b\}$, since E is conflict free according to Definition 12, a is obviously acceptable w.r.t. E, and b is acceptable w.r.t. E since no argument $defeats_E$ b. However, there exist two preferred extensions of Δ's metalevel formulation Δ_M:

$$\{(j - a), (a \text{ def } b), (r - b)\} \text{ and } \{(j - a), (j - b), (bD(aDb))\}$$

Hence, a and b are sceptically justified arguments of Δ, whereas only $(j - a)$ is a sceptically justified argument of Δ_M.

[5]The *EAF* shown is the 'weather example' used to motivate extended argumentation in [38]

[6]The example demonstrates the kind of self-reference exhibited by the *liar paradox* ("this sentence is false") in that b is interpreted as asserting a conclusion about itself, viz. that "b is preferred to a"

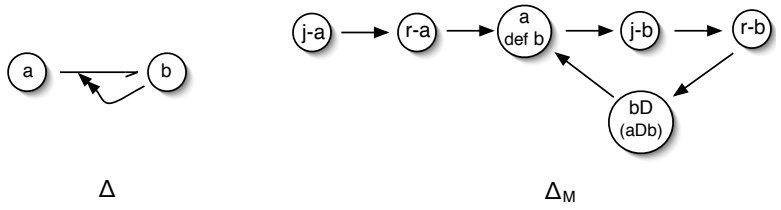

Figure 9: A non-hierarchical *EAF* and its metalevel formulation Δ_M

The example reveals a distinction between the ontological status ascribed to $\mathcal{R}$ attacks in EAFs and their metalevel formulation as arguments. *If* one were to ascribe to attacks the same status as arguments, then one might justifiably consider $\{a, b\}$ and $\{a\}$ as distinct preferred extensions of the EAF Δ, where the latter preferred extension implicitly contains the attack $a \rightharpoonup b$. However, EAFs treat $\mathcal{R}$ attacks as second class citizens, and with some justification given that the existence of an attack is contingent on the existence of arguments but not vice versa (i.e., one can consider the existence of arguments independently of attacks, but not vice versa). In this view it is legitimate to identify $\{a, b\}$ as the set inclusion maximal admissible set of *arguments*.

Our metalevel formulation does not discriminate between metalevel arguments constituted by the existence of $\mathcal{R}$ attacks and arguments in $\mathcal{A}$; these attacks and arguments are effectively assumed to be on a par. This in turn means that unlike attacks at the object level, attacks formalised as metalevel arguments can indirectly reinstate themselves via the rejection of the very same argument that is the target of the attack. This is illustrated in Δ_M in Figure 9 in which $(a$ def $b)$ is part of a 4 cycle, so that $(a$ def $b)$ reinstates $(r - b)$, which in turn reinstates $(a$ def $b)$ (hence we say $(a$ def $b)$ *indirectly* reinstates itself) in the preferred extension $\{(j - a), (a \text{ def } b), (r - b)\}$. Now, notice that the above implications of the distinct ontological treatment of arguments and attacks would not manifest if we focussed on EAFs that maintain a strict separation between object level arguments and the arguments that attack attacks between object level arguments. Indeed, [38] study a special class of *hierarchical EAF* in which the argumentation is 'stratified' into levels so that, intuitively, each level is a Dung framework $(\mathcal{A}, \mathcal{R})$ in which all $\mathcal{R}$ attacks are restricted to arguments within the framework. These $\mathcal{R}$ attacks are then attacked by $\mathcal{D}$ attacks that exclusively originate from arguments in the immediate metalevel. It is interesting to note that the characteristic functions of hierarchical *EAF*s are monotonic, so enabling characterisation of their grounded extensions as the least fixed point of their characteristic functions.

Definition 28 Δ = $(\mathcal{A}, \mathcal{R}, \mathcal{D})$ is a *hierarchical EAF* iff there exists a partition Δ_H = ($((\mathcal{A}_1, \mathcal{R}_1), \mathcal{D}_1)$, ..., $((\mathcal{A}_j, \mathcal{R}_j), \mathcal{D}_j)$, ...) such that:

- $\mathcal{A}$ = $\bigcup_{i=1}^{\infty} \mathcal{A}_i$, $\mathcal{R}$ = $\bigcup_{i=1}^{\infty} \mathcal{R}_i$, $\mathcal{D}$ = $\bigcup_{i=1}^{\infty} \mathcal{D}_i$, and for $i = 1 \ldots \infty$, $(\mathcal{A}_i, \mathcal{R}_i)$ is a Dung argumentation framework.
- $(C,(A, B)) \in \mathcal{D}_i$ implies $(A, B) \in \mathcal{R}_i$, $C \in \mathcal{A}_{i+1}$

Δ is a *bounded hierarchical EAF* iff its partition Δ_H is of the form ($((\mathcal{A}_1, \mathcal{R}_1), \mathcal{D}_1)$, ..., $((\mathcal{A}_n, \mathcal{R}_n), \mathcal{D}_n)$), where $\mathcal{D}_n$ = $\emptyset$

A correspondence can then be shown between bounded hierarchical *EAF*s and their metalevel formulations:

Theorem 7 Let Δ_M = ($\mathcal{A}_\mathcal{M}$, $\mathcal{R}_\mathcal{M}$, $\mathcal{C}$, $\mathcal{L}_\mathcal{M}, \mathcal{D}_e$) be the E-MAF of a bounded hierarchical *EAF* $\Delta = (\mathcal{A}, \mathcal{R}, \mathcal{D})$. Then for $s \in \{$complete, grounded, preferred, stable$\}$, $(j - x) \in \mathcal{A}_\mathcal{M}$ is a credulously, respectively sceptically, justified argument of Δ_M under the s semantics, iff $x \in \mathcal{A}$ is a credulously, respectively sceptically, justified argument of Δ under the s semantics.

The EAF in Figure 7 is not hierarchical, whereas the EAF in Figure 8 is. Its single preferred extension $\{e, d, b\}$ corresponds to the single preferred extension

$\{(j - e), (eD(cDd)), (j - d), (d \text{ def } c), (r - c), (dD(aDb)), (j - b), (b \text{ def } a), (r - a)\}$

of its metalevel formulation in Figure 8.

Although [38] discusses and illustrates requirements for EAFs that do not conform to the hierarchical restriction, we observe that many applications of extended argumentation can be naturally accommodated under the hierarchical restriction; in particular application of extended argumentation to agent reasoning over beliefs, goals and actions [36, 43], and to the modelling of case law [14]. Note that in Section 6 we comment further on the lack of correspondence between non-hierarchical EAFs and their metalevel formulations.

Prior to discussing applications of metalevel argumentation, we conclude with a brief comment on a recent critique [3] of PAFs, VAFs and EAFs. The critique claims that these frameworks may yield unintended results, and so could be seen as undermining the value of our metalevel formulations. In [3], a single motivating example is used to substantiate the critique. An expert's assessment that a given violin is produced by Stradivari (s), and that Stradivari produced violins are expensive ($s \rightarrow e$), is used to construct an argument $a1$ claiming that the violin is expensive. $a1$ is then attacked (on its premise s) by $a2$ claiming that the violin is not expensive ($\neg s$), where the source of $a2$ is less reliable than the expert, so that $a1$ is stronger than $a2$. Hence the attack from $a2$ to $a1$ fails, so that, as claimed in [3], one obtains an inconsistent extension containing $a1$ and $a2$. However, the example as presented in [3] fails to acknowledge that the extensions of a framework are evaluated on the basis of *all* the arguments and attacks defined by the instantiating theory. Hence, in the example there will also be an expert's argument $a3$ simply claiming s, where $a3$ and $a2$ symmetrically attack, and where the greater reliability of the expert means that $a3$ asymmetrically defeats $a2$, so yielding the single consistent preferred extension $\{a1, a3\}$.

4 Applications of Metalevel Argumentation

4.1 Applying Results and Techniques for Dung Argumentation to Developments of Dung Argumentation

We have described metalevel formulations of various object level developments of Dung's argumentation theory, and shown that the justified arguments of the object level frameworks can be characterised in terms of the justified arguments of their metalevel formulations. Since MAFs adopt the same basic machinery of Dung frameworks, these correspondences allow one to transition the full range of theoretical and practical results and techniques defined for Dung argumentation, to these various developments.

Consider, for example, the use of labellings for characterising the extensions of a Dung framework. Algorithms for computing labellings, and therefore extensions, have also been proposed [23, 42, 50, 51]. However, there has been little, if any work on labellings for the various developments of Dung's framework reviewed in Section 2[7]. Given the previous section's correspondences between the extensions of object level frameworks and their metalevel formulations, one can now use the labelling approach and algorithms at the metalevel in order to characterise and compute the object level extensions. For example, suppose Δ is a hierarchical EAF, and let Δ_M = ($\mathcal{A}_\mathcal{M}$, $\mathcal{R}_\mathcal{M}$, $\mathcal{C}$, $\mathcal{L}_\mathcal{M}, \mathcal{D}_e$) be its E-MAF as defined in Definition 27. We can then apply labelling algorithms to compute the labellings and so extensions of the Dung framework ($\mathcal{A}_\mathcal{M}$, $\mathcal{R}_\mathcal{M}$) under all of Dung's semantics. Given Theorem 7, we thus obtain the extensions of Δ.

Consider also the substantial body of work on argument game proof theories for Dung frameworks (AFs) (e.g., [24, 27, 42, 52]). These games are defined for establishing the justified status of a given argument x under each of Dung's semantics, and take the form of dialogues between a proponent and an opponent of a given argument. The proponent starts with an argument to be tested, after which each player must attack the other player's arguments with a counterargument. The initial argument is provable if the proponent has a winning strategy, i.e., if he can make the opponent run out of moves however the opponent chooses to attack. The precise rules of the argument game depend on the semantics which the proof theory is meant to capture.

While these games work well for AFs, there is relatively little work on defining argument games for the various developments of AFs described in Section 2. Furthermore, the few games that have been defined (e.g., [10, 12, 37]) are considerably more complex than those defined for AFs. However, one can now apply argument games for AFs, to metalevel frameworks, so that given the previous section's correspondences shown, one can establish the justified status of arguments of the form $(j - x)$ in order to establish the status of x in the corresponding object level framework.

For example, consider Two Player Immediate Response (TPI) games for demonstrating the justified status under credulous preferred and sceptical preferred semantics, for *coherent*[8] AFs [27, 52]. In [27], three moves are used: COUNTER, BACKUP and RETRACT. Either player can use COUNTER to attack the last argument played by their counterpart, whereas BACKUP and RETRACT are used respectively by opponent and proponent to back-track when no argument is available to attack the argument last played. A game for VAFs was then proposed in [10], in which an additional VALUE move is made available to both players. The VALUE move is used when there is no argument available to allow a COUNTER move, and allows the player to defend an argument by claiming an audience advocating that its value is preferred to its attacker. A record of such moves must be kept so as to block a VALUE move expressing an audience whose value preference contradicts a previous VALUE move's audience advocated preference. Although this game is effective for some frameworks, it is more complicated than the original TPI in that it has this additional move and requires maintenance of additional structures to record the currently expressed value preferences. Additionally, certain types of framework present problems.

A desirable feature of VAFs is that rather than the audience being fixed in advance,

[7]The only exception we know of is the labellings defined for EAFs for the admissible, preferred and stable semantics [37]. Algorithms for computing these extensions are not defined in this work

[8]Every preferred extension of a *coherent* Dung framework is also stable, and since in general stable extensions are preferred, the stable and preferred extensions of a coherent AF coincide.

it should emerge from the reasoning process itself[9]. This issue is explored in [12], in which given a VAF, the proponent wishes to establish a position in which certain arguments are to be accepted, certain other arguments to be rejected, and the proponent is indifferent to the inclusion or exclusion of the remaining arguments. [12] presents a game in which an audience for which the position is acceptable is determined, or it is shown that no such audience exists. The moves of the game are, however, rather complicated, and lack the clear intuitions of the TPI game moves.

However, we can now apply the standard TPI games to our metalevel formulations of $aVAF$s and VAFs. Given Theorem 5, we can then evaluate the justified status of an argument $(j - x)$ in the metalevel formulation, so evaluating the status of x in the object level VAF. No additional moves or structures recording the currently expressed value preferences are required. Furthermore, since the value preferences and audiences appear as arguments in the admissible set constructed during the course of a TPI game, we can play the standard TPI game to identify the audience needed to make a proponent's position acceptable.

4.2 Extending and Integrating Abstract Argumentation in Metalevel Frameworks

In the spirit of Dung's original theory, MAFs adopt a level of abstraction that makes limited commitments to the instantiating logics. Thus metalevel arguments can be built from statements about arguments, preferences, attacks etc. in different object level frameworks, which in turn may be constructed from different underlying logics. We now discuss how this provides for integration and further extension of the various developments of abstract argumentation, and for instantiation by, and integration of arguments constructed from different theories in different underlying logics, where one theory may encode metalevel reasoning about the arguments defined by another theory.

Recall that in VAFs audiences serve as oracles in that it is not possible to debate the merits of belonging to one audience rather than another. At the metalevel, however, the audiences are arguments within the framework, and as such are open to attack like any other argument, allowing one to advance arguments for and against particular audiences. Consider Definition 23's metalevel formulation of value based argumentation. Given an object level VAF we can construct the arguments and attacks instantiating a V-MAF, and *additionally* include in the V-MAF metalevel arguments that refer to object level arguments and attacks defined by argumentation-based reasoning about what the audience should be.

One possible source of such arguments may be moral principles. In a debate on fox hunting, for example, one might find an argument along the lines of *fox hunting is enjoyed by many people* in conflict with an argument such as *fox hunting causes animal suffering*. Resolving this conflict requires choosing between the audience preferring the value of human enjoyment to the value of animal welfare, and the audience endorsing the contrary preference. Opponents of hunting could now appeal to some moral standards, claiming that it is not a legitimate choice to prefer human enjoyment to animal welfare, using an argument such as *no rational person could promote enjoyment at the expense of animal welfare*. Such an argument would attack any audience endorsing the preference for human enjoyment. In turn this argument could be subject to attack, so that the debate shifts to what preferences are legitimate.

[9] Echoing Searle's view [48] that value preferences are the product of reasoning rather than an input to it.

This kind of argumentation is particularly common in the legal domain, especially when considering common law and the role of precedent cases. Often a case will turn on how the court chooses to resolve a conflict between possible purposes that the law can serve. Consider the well known property law case of *Pierson v Post* that has been the subject of much discussion in AI and Law. In this case there is a conflict between the value of encouraging a socially useful activity, and the need to have clear law to minimise disputes. While the minority opinion in *Pierson* favoured the first value, the majority preferred clear law to social utility. In subsequent cases where this choice is presented, *Pierson* can be cited as an argument against adopting a preference for social utility.

This additional expressiveness is important for representing such domains. Whereas object level frameworks, such as that produced for *Pierson* and related cases in [15] could identify the choices confronting the courts, they could capture neither the choice actually made, nor the rationale for the choice, both of which are essential for a proper representation of precedential reasoning. For a recent representation of case law which uses metalevel frameworks to allow such argumentation, see [14].

Consider also that our definition of E-MAFs (Definition 26) admits (given the BNF specification of $\mathcal{L}_{\mathcal{M}}$ in Definition 16) arguments with claims of the form $defeat(Z_n, defeat(Z_{n-1}, defeat(Z_{n-2}, \ldots)))$, where each such argument attacks an argument with claim $defeat(Z_{n-1}, defeat(Z_{n-2}, \ldots))$. That is to say we can model recursive attacks on attacks on attacks on ... etc (see Figure 10). Indeed, such a metalevel formulation might even provide a basis for defining an object level framework extending EAFs to accommodate such recursive attacks, in the sense that one might verify the correctness of such an object level formalisation by showing a correspondence with the metalevel formulation. Recently, [6, 7] have proposed just such an object level formalisation of recursive attacks, and we will discuss this work in Section 5.

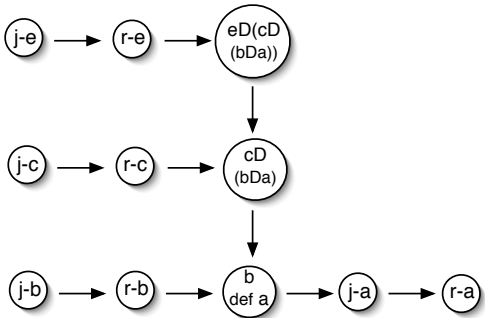

Figure 10: An E-MAF with metalevel formulations of recursive attacks on attacks.

In light of the above discussion, the question naturally arises as to whether there are requirements for formalising argumentation in metalevel frameworks independently of correspondences with object level frameworks. Our recent work on integrating accrual and dialectical argumentation [40] suggests such requirements. In [40] we argue that accrual should be modelled in terms of reasoning about the application of preferences to sets of arguments, and describe how this can be formalised in MAFs in which the constraints on the metalevel attack relation augment those for Dung MAFs ($D1$, $D2$

and $D3$ in Definition 17), with constraints that encode and so ensure satisfaction of the three principles of accrual identified by Prakken in [46].

A number of works (e.g., [32, 34]) have described requirements for extending value based argumentation so that criteria other than the audience based ranking of values can be used to undermine the success of attacks. For example, consider that two arguments a and b symmetrically attack, where a and b promote the same value. One may then wish to arbitrate between a and b based on other criteria, such as the relative trustworthiness of the distinct advocates (or sources) of each argument, or the degree to which each argument promotes a value. We can formalise a metalevel integration of value and preference argumentation by straightforwardly combining the P-MAFs and V-MAFs of Sections 3.4 and 3.5, by adding an additional constraint specifying attacks between contrary pairwise value preferences and strict preferences given by the preference relation.

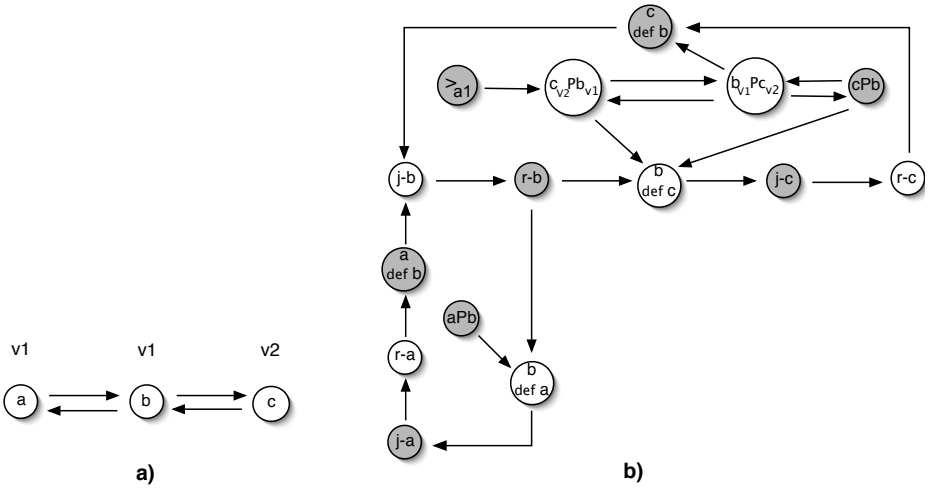

Figure 11: A VAF (a) and preference ordering formalised as a VP-MAF (b) in which the arguments in one of its two extensions are shaded

Definition 29 A VP-MAF is a tuple ($\mathcal{A}_\mathcal{M}$, $\mathcal{R}_\mathcal{M}$, $\mathcal{C}$, $\mathcal{L}_\mathcal{M}$, $\mathcal{D}_{vp}$), where $\mathcal{D}_{vp} = \mathcal{D}_v \cup \mathcal{D}_p \cup \{D' :$ if $\mathcal{C}(\alpha) = val\_pref(val(Y,V'), val(X,V))$ and $\mathcal{C}(\beta) = s\_preferred(X,Y)$ then $(\beta,\alpha),(\alpha,\beta) \in \mathcal{R}_\mathcal{M}\}$.

Consider the $aVAF$ in Figure 11 where the single audience a1 orders value $v1$ over $v2$. Consider also a separately defined preference ordering yielding the strict preferences $c \gg_\mathcal{P} b$, $a \gg_\mathcal{P} b$, where the latter strict preference resolves the choice between the symmetrically attacking a and b, each of which promote the same value. Notice that there are two preferred extensions of the VP-MAF; the one containing the shaded arguments as shown in Figure 11b), and the second containing the same arguments except that $b_{v1}Pc_{v2}$ replaces $(c$ def $b)$ and (cPb). Both preferred extensions contain the sceptically justified arguments $(j-a)$ and $(j-c)$, where c is justified at the object level, is in the latter case based exclusively on its reinstatement by a, and in the former case additionally results from c's preference over b.

Consider also that one can instantiate a VP-MAF by reference to arguments constructed from different underlying logics, where one logic may encode metalevel reasoning about the arguments and attacks defined by a theory in another logic. For example, one might instantiate a VP-MAF with:

1. 'action' arguments constructed in a BDI logic, instantiating schemes and critical questions for practical reasoning as described in [4];
2. 'value' arguments that refer to the values promoted by 'action' arguments, as described in [34] in which arguments are built from first order theories consisting of facts that assign values to constants naming the above action arguments, and orderings on values;
3. 'preference' arguments, whose construction from first order theories encoding clinical trial conclusions as to the relative efficacy of drugs is described in [34] (preference arguments thus arbitrate between action arguments for alternative medical actions that promote the same value (health) but do so to differing degrees, as concluded by clinical trials).

5 Related Work

This paper builds on and substantially extends previous work of ours [39] in which bounded hierarchical EAFs are rewritten as Dung frameworks. In [39] we 'expand' $\mathcal{R}$ attacks $x \rightharpoonup y$ in an EAF, to obtain attacks $x \rightarrow \overline{x} \rightarrow \overrightarrow{xy} \rightarrow y$. A $\mathcal{D}$ attack $(z,(x,y))$ is then rewritten as an attack $z \rightarrow \overrightarrow{xy}$ in the rewrite. The rewrites presented in [39] have also been described in subsequent works by other authors [6, 7, 17, 19, 20, 29].

Of particular interest is [6, 7]'s extension of EAFs – $AFRA$ (*Argumentation Framework with Recursive Attacks*) – to accommodate recursive attacks on attacks, and their rewrite as Dung frameworks. It is instructive to examine how $AFRAs$ accommodate attacks on attacks and formalise these as a Dung framework. Consider arguments a, b, c, d, and attacks (a, b), (c, d) and $(b,(c, d))$ ($AFRAs$ also allow attacks on attacks on attacks etc.). The notion of direct and indirect defeats from attacks to arguments and attacks is defined, so that for the given example:

(a, b) *directly defeats* b, (c, d) *directly defeats* d, $(b,(c, d))$ *directly defeats* (c, d) and (a, b) *indirectly defeats* $(b,(c, d))$

Based on these notions of defeat, notions of conflict free, acceptability and admissible and preferred extensions are defined. A Dung argumentation framework rewrite is also defined. Figure 12a) shows [6]'s rewrite for the above example, and by way of comparison the metalevel E-MAF formulation is also shown in Figure 12b). Intuitively, [6] effectively model attacks as arguments, in a manner similar to our metalevel formulation. Given that [6]'s motivation is to obtain a rewrite rather than formalise metalevel argumentation, the rewrite does not include arguments corresponding to metalevel arguments of the form $(r - x)$. Intuitively, however, a correspondence obtains between [6]'s rewrite and our metalevel formulation, since an argument $(r - b)$ will be acceptable iff $(a$ def $b)$ is acceptable, so that one can formulate an attack directly from $(a$ def $b)$ to $(bD(cDd))$. These observations suggest therefore that the metalevel formulation of recursive attacks described in Section 4.2 can be viewed as a metalevel formulation of [6]'s object level formalisation of recursive attacks. A formal demonstration of this is a topic for future work. For the moment, recall Section 3.6's discussion on the lack of correspondence for metalevel formulations of *non-hierarchical*

EAFs, and how this reflects the distinction in ontological status that [38] ascribes to arguments and attacks; a distinction that fails to be preserved in the metalevel formulation. It is this very same distinction that is responsible for the characteristic function of EAFs failing to satisfy monotonicity (although *hierarchical* EAFs do satisfy monotonicity)[10]. However, [6, 7] do not make such a distinction (and so obtain characteristic functions that are in general monotonic). Thus, one would expect a full correspondence for the metalevel formulation of $AFRA$s.

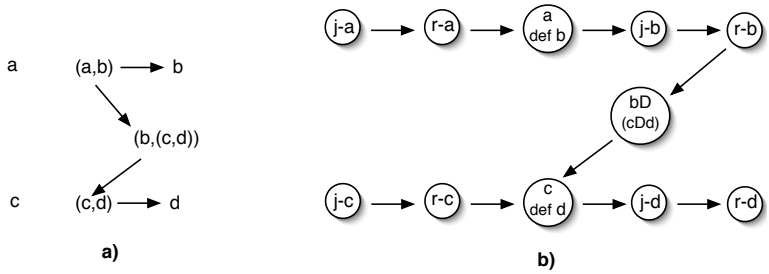

Figure 12: Comparing rewrite in [6] (a) and the metalevel formulation of an *EAF* (b)

The contents of this paper first appeared as a technical report [41], subsequent to which a special issue of *Studia Logica* published a number of interesting papers on *New Ideas in Argumentation*. Specifically, in [29], attacks are expanded in a similar manner to that described by our rewrites in [39]. The expansions in [29] also model recursive attacks on attacks, additionally allowing for the case where although an argument x can be rejected, this does not mean rejection of attacks that emanate from x.

The above works on recursive attacks [6, 7, 29] are not concerned with this paper's primary goal of establishing a methodology for metalevel argumentation and showing how varieties of object level argumentation can be uniformly characterised using this methodology. In this regard, the work of Boella et.al. [17, 20] is more closely related. Boella et.al. also describe use of the rewrites first presented in [39] to suggest a methodology for meta-argumentation, and outline similar aspirations to ours; to instantiate Dung's argumentation framework with meta-arguments so as to model Dung argumentation and its various extensions. In [17], an object level attack $a \rightarrow b$ is represented by the meta-arguments and attacks: $accept(a) \rightarrow X_{a,b} \rightarrow Y_{a,b} \rightarrow accept(b)$. [17] then describes how attacks on arguments of the form $Y_{a,b}$ can be used to model attacks on attacks and recursive attacks on attacks. However, [17] and [20] do not formally prove correspondences between the varieties of object level argumentation reviewed in Section 2, and their metalevel formulations. The importance of the correspondences shown in this paper, is that they offer a unifying perspective on a range of developments of Dungs original framework, and facilitate application of results and techniques for Dung's framework to their various developments.

We also mention the meta-argumentation frameworks of [25]. This work does not propose a methodology of metalevel argumentation in which arguments make claims about object level frameworks, but rather proposes a specific meta-argumentation framework for object level *bipolar* argumentation frameworks [2]. The object-level support

[10] [38]'s counter-example to monotonicity is illustrated by Figure 1: A is acceptable w.r.t. $S = \{c, d\}$, but is not acceptable with respect to the conflict free $S' = \{c, b, d, e\}$. However, note that the attacks (c, b), (d, e) can be assumed to be contained within S but not S' (given the reinstatement set $\{c \rightarrow^S b, d \rightarrow^S e\}$) if one were to ascribe them the same status as arguments. Under this assumption, $S' \not\supseteq S$.

relation is used to identify 'coalitions' (sets of conflict free arguments related by the support relation) that constitute meta-arguments in a Dung framework, and that are related by an meta-attack relation based on the object-level attack relation. The justified arguments of the object level framework can then be identified based on the metalevel Dung framework.

Finally, a number of works formalise logics that explicitly refer to arguments and their relations and properties [18, 33, 53]. In particular, [53] are motivated by the claim that "*argumentation and formal dialogue is necessarily a meta-logical process*", and so formalise metalogics in which one can reason about what constitutes an argument in the object level, attack and defeat relations between these arguments, their properties (e.g. the values they promote), and their status in the object level. While [53] (and [18, 33]) provide meta-logics for reasoning about arguments, and argument construction based on these meta-logical statements, they do not relate these to argumentation frameworks, and so the argumentation at the metalevel is not considered using this single powerful abstraction. Indeed, one can see how these meta-logics can be used to define arguments, attacks and metalevel claims for instantiating the metalevel argumentation frameworks described in this paper.

6 Conclusions

In this paper we have proposed a general methodology of metalevel argumentation, whereby: 1) meta-arguments are constructed based on statements about an object-level framework; 2) constraints on the metalevel attack relation refer to the claims that the metalevel arguments make about the object level framework, and so characterise the reasoning applied to evaluate the justified arguments of the object level framework; 3) the justified arguments of the metalevel framework are evaluated under the standard Dung semantics.

We have shown correspondences between the object level frameworks and their metalevel formulations; correspondences that we not only suggest yield a number of practical benefits, but also support a rhetorical aim of this paper: to support the view that Dung's acceptability calculus identifies general and widely applicable principles of commonsense reasoning. Specifically, we have shown how collective attacks, preference based, value based, and hierarchical extended argumentation can all be formulated as instances of Dung argumentation.

Future work will investigate metalevel formulations of other developments of Dung's theory. For example, frameworks augmented with support relations [2, 45]. For the moment, we observe that if x supports y, then an attack on x propagates to y. At the metalevel this would be formalised in terms of a metalevel attack from $(r-x)$ to $(j-y)$, so that if the argument $(r-x)$ claiming 'x is rejected' is in an admissible extension E (and is therefore reinstated by some $(z \text{ def } x) \in E$), then it cannot be that $(j-y) \in E$. As discussed in Section 5, future work will also develop metalevel formulations of recent frameworks accommodating recursive attacks on attacks [6, 7, 29].

In Section 4 we proposed a number of benefits of metalevel argumentation. The correspondences shown in Section 3 allow one to transition the full range of theoretical and practical results and techniques for Dung argumentation, to developments of Dung argumentation. In particular, we discussed how the labelling approach and argument game proof theories developed for Dung's framework, can now be applied to compute the extensions and justified status of arguments in the object-level frameworks formulated in the metalevel frameworks (thus avoiding requirements for development

of more complex algorithms and proof theories for the object-level frameworks). Since the metalevel formulations introduce extra arguments, future work will focus on how efficiency gains can be obtained. For example, when applying argument games to metalevel frameworks, one could allow a player to play more than one argument in a single move. In particular a player could move an argument of the form $(x \text{ def } y)$, followed by an argument of the form $(j - x)$, given that the player's counterpart will always be able to play $(r - x)$ in response to $(x \text{ def } y)$, which in turn can always be countered by $(j - x)$. If these two moves were played together the counterpart would then have the choice of attacking either $(x \text{ def } y)$ or $(j - x)$. Changes of this sort would eliminate some unnecessary rounds, but not otherwise impact on the game [11].

Argument game proof theories for Dung frameworks have informed formalisation of argumentation-based dialogue systems (see [47] for a review), where, for example, one agent seeks to persuade another to adopt a belief it does not already hold to be true, or when agents deliberate about what actions to execute, or negotiate over resources. Another direction for future work will be to formalise similar such dialogues for metalevel frameworks, allowing, for example, agents to debate value preferences in value based deliberation over actions, and preferences in negotiation.

In Section 4.2 we discussed how MAFs provide for extending and integrating various forms of abstract argumentation. In particular, we suggested extending the metalevel formulation of $V - MAF$s to incorporate arguments expressing constraints on audiences, and pointed to a recent work in which MAFs integrate accrual with Dung's dialectical mode of argumentation [40]. We also described how arguments for preference orderings and value preferences, built from different underlying theories encoded in different logics, can be integrated in metalevel integrations of value and preference based argumentation. These examples by no means exhaust the possible extensions and integrations, and there is much scope for future work in these areas. For example, one might look to integrate preferences with collective attacks, or model the strength of attacks [9] in terms of weights assigned to metalevel arguments of the form $(x \text{ def } y)$, where an attack's weight may need to exceed a certain threshold (as recently described in [28]), and the failure to do so may be modelled as a metalevel attack on $(x \text{ def } y)$.

7 Appendix

The following lemmas are used for the proofs of the main results in this paper. We first define the expansion of DungC and Dung frameworks, obtained by substituting some subset of the attacks (X, y) in $\mathcal{R}$ (where in the case of a Dung framework X is a single argument) as shown in Figure 13.

Definition 30 Let $\Delta = (\mathcal{A}, \mathcal{R})$ be a DungC framework. Then $\Delta' = (\mathcal{A}', \mathcal{R}')$ is said to be an expansion of Δ iff:

- $\mathcal{R}'$ is any set of attacks $(\mathcal{R}^* \subseteq \mathcal{R}) \cup \{expand((X, y))|(X, y) \in (\mathcal{R} - \mathcal{R}^*)\}$ where $expand((X, y)) = \{(\{x\}, \overline{x}), (\{\overline{x}\}, \overrightarrow{Xy})|x \in X\} \cup (\{\overrightarrow{Xy}\}, y)\}$
- $\mathcal{A}' = \mathcal{A} \cup \{\overline{x}, \overrightarrow{Xy}|(\{\overline{x}\}, \overrightarrow{Xy}) \in \mathcal{R}')\}$

[11] Note also that this will enhance the naturalness of such dialogical games, in that it is more natural for a player to assert in one move that 'x is justified and x attacks y' rather than simply assert that 'x attacks y' and then only assert 'x is justified' in response to the opponent asserting that 'x is rejected'

Let $\Delta = (\mathcal{A}, \mathcal{R})$ be a Dung framework. Then $\Delta' = (\mathcal{A}', \mathcal{R}')$ is said to be an expansion of Δ iff Δ' is the expansion of the DungC framework $(\mathcal{A}, \mathcal{R}^s)$ where $\mathcal{R}^s = \{(\{x\}, y) | (x, y) \in \mathcal{R}\}$.

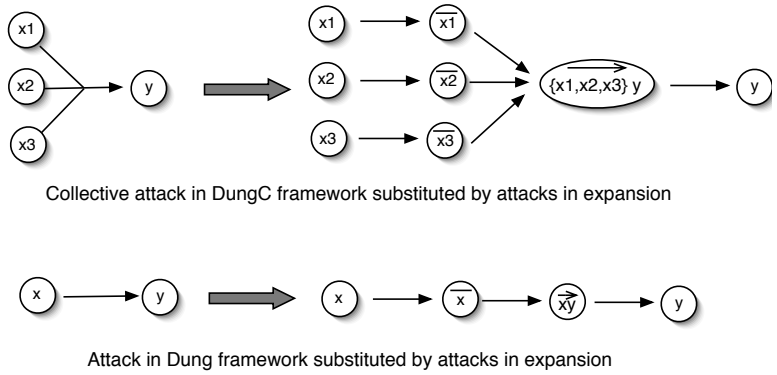

Figure 13: Expansions of DungC collective attacks and Dung attacks

Lemmas 1 and 2 prove an iff correspondence between admissible extensions of DungC frameworks and their expansions.

Lemma 1 Let $(\mathcal{A}', \mathcal{R}')$ be an expansion of the DungC framework $(\mathcal{A}, \mathcal{R})$. Let E be an admissible extension of $(\mathcal{A}, \mathcal{R})$. Then E' is an admissible extension of $(\mathcal{A}', \mathcal{R}')$ such that:

1. $E' = E \cup \{\overrightarrow{Xy} | X \subseteq E, (X, y) \in \mathcal{R}, expand((X, y)) \subseteq \mathcal{R}'\} \cup \{\overline{y} | X \subseteq E, (X, y) \in \mathcal{R}, \overline{y} \in \mathcal{A}'\}$
2. $\forall \alpha \in \mathcal{A}$, if α is acceptable w.r.t. E then α is acceptable w.r.t. E'

Proof: Suppose $\alpha \in \mathcal{A}$, α acceptable w.r.t. E, and $\Omega \mathcal{R}' \alpha$.
– If $\Omega \mathcal{R} \alpha$, then $\exists \Gamma \subseteq E$ s.t. $\Gamma \mathcal{R} \delta$, $\delta \in \Omega$. Suppose it is not the case that $\Gamma \subseteq E'$, $\Gamma \mathcal{R}' \delta$. Then, $\Gamma \mathcal{R} \delta$ is some $\mathcal{R}$ attack (X, y) such that $expand((X, y)) \subseteq \mathcal{R}'$. By definition of E', $\overrightarrow{Xy} \in E'$, and since $\{\overrightarrow{Xy}\} \mathcal{R}' y$, α is acceptable w.r.t. E'.
– Suppose it is not the case that $\Omega \mathcal{R} \alpha$. Then Ω is some $\overrightarrow{Yz}$, α some z, where $\forall y \in Y$, $\{y\} \mathcal{R}' \overline{y}$, $\{\overline{y}\} \mathcal{R}' \overrightarrow{Yz}$. Since $\overrightarrow{Yz}$ is obtained by expanding $Y \mathcal{R} z$, and by admissibility of E, $\exists X \subseteq E$ s.t. $X \mathcal{R} y'$ for some $y' \in Y$, then by definition of E', $\overline{y}' \in E'$, and so α (= z) is acceptable w.r.t E'.

To show E' is admissible it remains to show that arguments $\overrightarrow{Xy}, \overline{y} \in E'$ are acceptable w.r.t. E'. If $\overrightarrow{Xy} \in E'$ then $\forall x \in X$, $\{\overline{x}\} \mathcal{R}' \overrightarrow{Xy}$ and $\{x\} \mathcal{R}' \overline{x}$, and since $X \subseteq E$ and so $X \subseteq E'$, then $\overrightarrow{Xy}$ is acceptable w.r.t. E'. If $\overline{y} \in E'$ then $(y, \overline{y}) \in \mathcal{R}'$, $(X, y) \in \mathcal{R}$ and $X \subseteq E$ hence $X \subseteq E'$. If $expand((X, y)) \nsubseteq \mathcal{R}'$ then $X \mathcal{R}' y$, and so $\overline{y}$ is acceptable w.r.t. E'. If $expand((X, y)) \subseteq \mathcal{R}'$, then $\overrightarrow{Xy} \in E'$, where $\overrightarrow{Xy} \mathcal{R}' y$, hence $\overline{y}$ is acceptable w.r.t. E'.

Lemma 2 Let $(\mathcal{A}', \mathcal{R}')$ be an expansion of the DungC framework $(\mathcal{A}, \mathcal{R})$. Let E' be an admissible extension of $(\mathcal{A}', \mathcal{R}')$. Then $E = (E' \cap \mathcal{A})$ is an admissible extension of $(\mathcal{A}, \mathcal{R})$ such that:

- $\overrightarrow{Xy} \in E'$ implies $X \subseteq E$, $(X, y) \in \mathcal{R}$, and $\overline{y} \in E'$ implies $\exists X$, $X \subseteq E$, $(X, y) \in \mathcal{R}$
- $\forall \alpha \in \mathcal{A}$, α is acceptable w.r.t. E' implies α is acceptable w.r.t. E

Proof: Suppose $\alpha \in \mathcal{A}$, α acceptable w.r.t. E', $\Omega \mathcal{R} \alpha$.
– Suppose $\Omega \mathcal{R}' \alpha$. Hence, $\exists \Gamma \subseteq E'$ s.t. $\Gamma \mathcal{R}' \delta$, $\delta \in \Omega$. Suppose it is not the case that $\Gamma \subseteq E$, $\Gamma \mathcal{R} \delta$. Then Γ is of the form $\{\overrightarrow{Xy}\}$ (and so $(X, y) \in \mathcal{R}$), δ of the form y, and $\forall x \in X$, $\{\overline{x}\} \mathcal{R}' \overrightarrow{Xy}$, $\{x\} \mathcal{R}' \overline{x}$, and by admissibility of E', $\forall x \in X$, $x \in E'$. Since $E = (E' \cap \mathcal{A})$, $X \subseteq E$, and since $(X, y) \in \mathcal{R}$ then α is acceptable w.r.t. E. Note that we have also shown that $\overrightarrow{Xy} \in E'$ implies $X \subseteq E$, $(X, y) \in \mathcal{R}$.
– Suppose it is not the case that $\Omega \mathcal{R}' \alpha$. Then Ω is some Y, α some z, $expand((Y, z)) \subseteq \mathcal{R}'$. Hence $\{\overrightarrow{Yz}\} \mathcal{R}' z$, and by assumption of acceptability of z w.r.t. E', $\exists \overline{y} \in E'$ s.t. $\{\overline{y}\} \mathcal{R}' \overrightarrow{Yz}$, where $y \in Y$. Since $\{y\} \mathcal{R}' \overline{y}$, then by the admissibility of E' either:
1) $\exists X \subseteq E'$, $X \mathcal{R}' y$, $(X, y) \in \mathcal{R}$, $X \subseteq E$, hence α (= z) is acceptable w.r.t. E, or;
2) $\exists \overrightarrow{Xy} \in E'$ s.t. $\{\overrightarrow{Xy}\} \mathcal{R}' y$, where $expand((X, y)) \subseteq \mathcal{R}'$ and $\forall x \in X$, $\{\overline{x}\} \mathcal{R}' \overrightarrow{Xy}$, $\{x\} \mathcal{R}' \overline{x}$, and so by admissibility of E', $X \subseteq E'$, hence $X \subseteq E$, α (= z) is acceptable w.r.t. E. Note we have also shown $\overline{y} \in E'$ implies $\exists X$, $X \subseteq E$, $(X, y) \in \mathcal{R}$.

Lemma 3 Let $\Delta' = (\mathcal{A}', \mathcal{R}')$ be an expansion of the DungC framework $\Delta = (\mathcal{A}, \mathcal{R})$. Then for $s \in \{$admissible, complete, preferred, grounded, stable $\}$, E is an s extension of $(\mathcal{A}, \mathcal{R})$ iff E' is an s extension of $(\mathcal{A}', \mathcal{R}')$, where:

1. $\forall \alpha \in \mathcal{A}$, $\alpha \in E$ iff $\alpha \in E'$
2. $\exists X \subseteq E$, $(X, y) \in \mathcal{R}$ iff $\overrightarrow{Xy} \in E'$, where $expand((X, y)) \subseteq \mathcal{R}'$
3. $\exists X \subseteq E$, $(X, y) \in \mathcal{R}$ iff $\overline{y} \in E'$, where $\overline{y} \in \mathcal{A}'$

Proof:
1. s **= admissible**. **1.1** Left to right half follows from Lemma 1. **1.2**. Right to left half follows from Lemma 2.

Let us define functions h and g s.t.

> For any admissible extension E of Δ, $E' = h(E)$ as defined above.
> For any admissible extension E' of Δ', $E = g(E')$ as defined above.

We show that:
a) h is monotonically strictly increasing in the sense that $\forall E, F$ s.t. E and F are admissible extensions of Δ and $E \subset F$, then $h(E) \subset h(F)$
Suppose E and by **1.1** the corresponding admissible $E' = h(E)$. Suppose $E \subset F$ and by **1.1** the corresponding admissible $F' = h(F)$. It is obvious to see that $E' \subset F'$
b) g is monotonically strictly increasing in the sense that $\forall E', F'$ s.t. E' and F' are admissible extensions of Δ' and $E' \subset F'$, then $g(E') \subset g(F')$
Suppose E' and by **1.2** the corresponding admissible $E = g(E')$. Suppose $E' \subset F'$. Then:

> $\forall \alpha \in (F' - E')$, if $\alpha \in \mathcal{A}$ or α is of the form $\overline{y}$ or $\overrightarrow{Xy}$, then $\alpha \notin E$, respectively $\neg \exists X \subseteq E$ s.t. $(X, y) \in \mathcal{R}$, since otherwise, by application of **1.1** to E, we would have $\alpha \in E'$, respectively $\overline{y}$ or $\overrightarrow{Xy} \in E'$. **(i)**

By **1.2**, let F be the corresponding admissible extension of Δ. Given **i)**, $E \subset F$.

2 s **= complete**.
2.1 Left to right half: Suppose E is complete. Applying **1.1**, E' is an admissible extension of Δ', where $E = g(E')$. Suppose E' is not complete. Then $\exists \alpha \notin E'$, α acceptable w.r.t. E', and so by Dung's fundamental lemma [26], $F' = E' \cup \{\alpha\}$ is admissible where $F' \supset E'$. Applying **1.2**, $F = g(F')$ is an admissible extension of Δ, where by **b)**, $E \subset F$, contradicting E is complete.
2.2 Right to left half: Suppose E' is complete. Applying **1.2**, E is an admissible extension of Δ, where $E' = h(E)$. Suppose E is not complete. Then $\exists \alpha \notin E$, α acceptable w.r.t. E, and so by Dung's fundamental lemma, $F = E \cup \{\alpha\}$ is admissible where $F \supset E$. Applying **1.1**, $F' = h(F)$ is an admissible extension of Δ', where by **a)**, $E' \subset F'$, contradicting E' is complete.

3 s **= preferred**.
3.1 Left to right half: Suppose E is preferred. The proof now proceeds in the same way as **2.1**, except that supposing E' is not preferred immediately implies $\exists F' \supset E'$ s.t. F' is admissible.
3.2 Right to left half: Suppose E' is preferred. The proof now proceeds in the same way as **2.2**, except that supposing E is not preferred immediately implies $\exists F \supset E$ s.t. F is admissible.

4 s **= grounded**.
4.1 Left to right half: Suppose E is grounded. Applying **2.1**, E' is a complete extension of Δ', where $E = g(E')$. Suppose E' is not grounded. Then $\exists F' \subset E'$, F' is complete. Applying **2.2**, $F = g(F')$ is a complete and so admissible extension of Δ, where by **b)**, $F \subset E$, contradicting E is grounded.
4.2 Right to left half: Suppose E' is grounded. Applying **2.2**, E is a complete extension of Δ, where $E' = h(E)$. Suppose E is not grounded. Then $\exists F \subset E$, F is complete. Applying **2.1**, $F' = h(F)$ is a complete and so admissible extension of Δ', where by **a)**, $F' \subset E'$, contradicting E' is grounded.

5 s **= stable**:
5.1 Left to right half: Suppose E is stable. Applying **2.1**, E' is complete. Suppose $\alpha \in \mathcal{A}$, $\alpha \notin E'$. Then $\alpha \notin E$, $\exists \Gamma \subseteq E$ s.t. $(\Gamma, \alpha) \in \mathcal{R}$. Since $E \subseteq E'$, $\Gamma \subseteq E'$. Suppose $(\Gamma, \alpha) \notin \mathcal{R}'$. Then $(\Gamma, \alpha) = (X, y)$, $expand((X, y)) \subseteq \mathcal{R}'$ and so $\overrightarrow{Xy} \in E'$, $(\overrightarrow{Xy}, y) \in \mathcal{R}'$. Suppose $\exists \alpha \in (\mathcal{A}' - \mathcal{A})$, $\alpha \notin E'$, $\neg \exists \Gamma \subseteq E'$ s.t. $(\Gamma, \alpha) \in \mathcal{R}'$. Then either:
α is of the form $\overline{y}$, in which case $\{y\}\mathcal{R}'\overline{y}$, $y \notin E'$. But then we have already shown that $y \in \mathcal{A}$ is attacked by some subset of E', and so $\overline{y}$ is acceptable w.r.t. E', contradicting E' is complete;
α is of the form $\overrightarrow{Xy}$, in which case $\forall x \in X$, $\{\overline{x}\}\mathcal{R}'\overrightarrow{Xy}$, $\overline{x} \notin E'$. For any such $\overline{x}$, $x \in E'$, and since $\{x\}\mathcal{R}'\overline{x}$, then $\overrightarrow{Xy}$ is acceptable w.r.t. E', contradicting E' is complete.
5.2 Right to left half: Suppose E' is a stable extension. Applying **2.2**, E is complete. Suppose some $\alpha \in \mathcal{A}$, $\alpha \notin E$. Then $\alpha \notin E'$, $\exists \Gamma \subseteq E'$ s.t. $(\Gamma, \alpha) \in \mathcal{R}'$. Suppose it is not the case that $\Gamma \subseteq E$ and $(\Gamma, \alpha) \in \mathcal{R}$. Then Γ is some $\{\overrightarrow{Xy}\}$, α some y, and by **2.2**, $\exists X \subseteq E$, $(X, y) \in \mathcal{R}$.

Notice that since the definitions of conflict free, acceptability and the extensions of a standard Dung framework are a special case of DungC frameworks (i.e., the case where every attack originates from a singleton set of arguments), then corollaries of the above

results establish the same correspondences for Dung frameworks and their expansions.

Corollary 1 Let $\Delta' = (\mathcal{A}', \mathcal{R}')$ be an expansion of the Dung framework $\Delta = (\mathcal{A}, \mathcal{R})$. Then for $s \in \{$admissible, complete, preferred, grounded, stable$\}$, E is an s extension of $(\mathcal{A}, \mathcal{R})$ iff E' is an s extension of $(\mathcal{A}', \mathcal{R}')$, where:

1. $\forall \alpha \in \mathcal{A}$, $\alpha \in E$ iff $\alpha \in E'$
2. $\exists x \in E$, $(x, y) \in \mathcal{R}$ iff $\overrightarrow{xy} \in E'$, where $expand((x, y)) \in \mathcal{R}'$
3. $\exists x \in E$, $(x, y) \in \mathcal{R}$ iff $\overline{y} \in E'$, where $\overline{y} \in \mathcal{A}'$

7.1 Proofs for Section 2

Proposition 1 Let $\Delta = (\mathcal{A}, \mathcal{R})$, and for $s \in \{$admissible, complete, grounded, preferred, stable$\}$, let E be an s extension of Δ. Then there exists an s labelling $\mathcal{L}$ of Δ such that $\mathtt{in}(\mathcal{L}) = E$, and $\mathtt{out}(\mathcal{L}) = (E+) \cup (E-)$.

Proof: Obvious, given Theorem 1 and Definition 5.

7.2 Proofs for Sections 3.2 and 3.3

Since DungC frameworks with collective attacks are a straightforward generalisation of Dung frameworks, we establish results for the former, and then state Section 3.2's results as corollaries of these results. In what follows we will use the following generalisation of Notation 1.

Notation 3 Let $(\mathcal{A}, \mathcal{R})$ be a DungC framework, and $E \subseteq \mathcal{A}$.

- $\overrightarrow{E+}$ denotes the set of attacks originating from *sets of* arguments in E: $\overrightarrow{E+} = \{(B, y) \mid B \subseteq E, B\mathcal{R}y\}$
- $E+$ denotes the set of arguments attacked by *sets of* arguments $B \subseteq E$: $E+ = \{y \mid B \subseteq E, B\mathcal{R}y\}$
- $E-$ denotes the set of *sets of* arguments that attack arguments in E: $E- = \{B \mid B\mathcal{R}x, x \in E\}$

Lemma 4 Let $\Delta_M = (\mathcal{A}_\mathcal{M}, \mathcal{R}_\mathcal{M}, \mathcal{C}, \mathcal{L}_\mathcal{M}, \mathcal{D}_{dc})$ be the MAF of a DungC framework $\Delta = (\mathcal{A}, \mathcal{R})$. Then for $s \in \{$admissible, complete, grounded, preferred, stable$\}$, E is an s extension of Δ iff E' is an s extension of Δ_M, where:

1. $X\mathcal{R}y \in \overrightarrow{E+}$ iff $(X \text{ def } y) \in E'$
2. $y \in E+$ iff $(r - y) \in E'$
3. $x \in E$ iff $(j - x) \in E'$

Proof: Let $\Delta^* = (\mathcal{A}^*, \mathcal{R}^*)$ be the expansion of $\Delta = (\mathcal{A}, \mathcal{R})$ such that $\forall (X, y) \in \mathcal{R}$, $expand((X, y)) \subseteq \mathcal{R}^*$. Let $\Delta^{**} = (\mathcal{A}^{**}, \mathcal{R}^{**})$ be Δ^*'s *augmentation*, defined as follows:

$$\mathcal{A}^{**} = \mathcal{A}^* \cup \{\overline{y} | y \in \mathcal{A}\} \text{ and } \mathcal{R}^{**} = \mathcal{R}^* \cup \{(\{y\}, \overline{y}) | y \in \mathcal{A}\}$$

In other words Δ^{**} is defined by additionally including attacks $(\{y\}, \overline{y})$ for those y that are not a member of some set Y s.t. $(Y, x) \in \mathcal{R}$, and are thus are not obtained by expanding the attacks in Δ. The following holds:

> For $s \in \{$admissible, complete, grounded, preferred, stable$\}$, E^* is an s extension of Δ^* iff E^{**} is an s extension of Δ^{**}, where $E^* \subseteq E^{**}$, and $(E^{**} - E^*) = \{\overline{y} | \overrightarrow{Xy} \in E^*, \overline{y} \notin \mathcal{A}^*\}$ **(i)**

(i) follows given that no $\mathcal{R}^{**}$ attacks originate from the extra $\overline{y}$ arguments in $(\mathcal{A}^{**} - \mathcal{A}^*)$, and so $\forall \alpha \in E^* \cap E^{**}$, α is acceptable w.r.t. E^* iff α is acceptable w.r.t. E^{**}, and since $\overrightarrow{Xy} \in E^*$ iff $\overrightarrow{Xy} \in E^{**}$, $\{\overrightarrow{Xy}\}\mathcal{R}^* y$ iff $\{\overrightarrow{Xy}\}\mathcal{R}^{**} y$, and $\forall y$, $\{y\}\mathcal{R}^{**}\overline{y}$, then each $\overline{y} \in (E^{**} - E^*)$ is acceptable w.r.t. E^{**}.

It should now be obvious to see that the argument graphs Δ_M and Δ^{**} are isomorphic, where f is a bijective function from $\mathcal{A}_\mathcal{M}$ to $\mathcal{A}^{**}$ s.t.

- $f((j - x))$ = x (where $x \in \mathcal{A}, \mathcal{A}^*, \mathcal{A}^{**}$)
- $f((r - y))$ = $\overline{y}$
- $f((X \text{ def } y))$ = $\overrightarrow{Xy}$

and g is a bijective function from $\mathcal{R}_\mathcal{M}$ to $\mathcal{R}^{**}$ s.t. $g(\alpha, \beta) = (f(\alpha), f(\beta))$.

To see that this is so, observe that $(\mathcal{A}_\mathcal{M}, \mathcal{R}_\mathcal{M})$ is effectively obtained by replacing every $x \in \mathcal{A}$ by $(j - x)$, and then every attack (X, y) is expanded, interspersing $(j - x)\mathcal{R}_\mathcal{M}(r - x)$, $(r - x)\mathcal{R}_\mathcal{M}(X \text{ def } y)$ for all $x \in X$, and $(X \text{ def } y)\mathcal{R}_\mathcal{M}(j - y)$, and for *every* $(j - x)$, the attack $(j - x)\mathcal{R}_\mathcal{M}(r - x)$ is added.

We now prove the main result:

• By lemma 3, E is an s extension of Δ iff E^* is an s extension of the expansion Δ^*, where $\forall \alpha \in \mathcal{A}$, $\alpha \in E$ iff $\alpha \in E^*$, $X \subseteq E$ and $(X, y) \in \mathcal{R}$ iff $\overrightarrow{Xy} \in E^*$ (recall that every attack (X, y) is expanded in Δ^*) and $\overline{y} \in E^*$ s.t. $\overline{y} \in \mathcal{A}^*$.

• Given **(i)**, E is an s extension of Δ iff E^{**} is an s extension of the augmentation Δ^{**} of Δ^*, where $\forall \alpha \in \mathcal{A}$, $\alpha \in E$ iff $\alpha \in E^{**}$, $X \subseteq E$ and $(X, y) \in \mathcal{R}$ iff $\overrightarrow{Xy}, \overline{y} \in E^{**}$, (given that $\overrightarrow{Xy} \in E^{**}$ implies $\overline{y} \in E^{**}$).

• By the isomorphism of Δ_M and Δ^{**}: E is an s extension of Δ iff E' is an s extension of Δ_M, where $\forall x \in \mathcal{A}$, $x \in E$ iff $(j - x) \in E'$, $X \subseteq E$, $(X, y) \in \mathcal{R}$ (i.e., $X\mathcal{R}y \in \overrightarrow{E+}$ and $y \in E+$) iff $((X \text{ def } y))$, $(r - y) \in E'$.

Corollary 2 Let Δ_M = ($\mathcal{A}_\mathcal{M}$, $\mathcal{R}_\mathcal{M}$, $\mathcal{C}$, $\mathcal{L}_\mathcal{M}, \mathcal{D}_d$) be the MAF of a Dung framework Δ = $(\mathcal{A}, \mathcal{R})$. Then for $s \in \{$admissible, complete, grounded, preferred, stable$\}$, E is an s extension of Δ iff E' is an s extension of Δ_M, where:

1. $x\mathcal{R}y \in \overrightarrow{E+}$ iff $(x \text{ def } y) \in E'$
2. $y \in E+$ iff $(r - y) \in E'$
3. $x \in E$ iff $(j - x) \in E'$

Theorem 2 Let Δ_M = ($\mathcal{A}_\mathcal{M}$, $\mathcal{R}_\mathcal{M}$, $\mathcal{C}$, $\mathcal{L}_\mathcal{M}, \mathcal{D}_d$) be the MAF of a Dung framework $(\mathcal{A}, \mathcal{R})$. Then for $s \in \{$complete, grounded, preferred, stable$\}$, $(j - x) \in \mathcal{A}_\mathcal{M}$ is a credulously, respectively sceptically, justified argument of Δ_M under the s semantics, iff $x \in \mathcal{A}$ is a credulously, respectively sceptically, justified argument of Δ under the s semantics.

Proof Follows from Corollary 2.

Proposition 2 Let $\Delta_{\mathcal{M}}$ = ($\mathcal{A}_{\mathcal{M}}$, $\mathcal{R}_{\mathcal{M}}$, $\mathcal{C}$, $\mathcal{L}_{\mathcal{M}}, \mathcal{D}_d$) be the MAF of a Dung framework Δ = $(\mathcal{A}, \mathcal{R})$. For $s \in$ {admissible, complete, grounded, preferred, stable}:
There exists an s labelling $\mathcal{L}$ of Δ iff there exists an s extension E of $\Delta_{\mathcal{M}}$ such that: 1) $x \in \mathtt{in}(\mathcal{L})$ iff $(j-x) \in E$; 2) $y \in \mathtt{out}(\mathcal{L})$ iff $(r-y) \in E$.

Proof:
Left to right: Let $\mathcal{L}$ be an s labelling of Δ. By Theorem 1, E' = $\mathtt{in}(\mathcal{L})$ is an s extension of Δ. By Corollary 2, there is an s extension E of $\Delta_{\mathcal{M}}$, where E = $\{(j-x)|x \in E'\}$ $\cup$ $\{(x \text{ def } y)|\ x\mathcal{R}y \in \overrightarrow{E'+}\} \cup \{(r-y)|y \in E'+\}$. Hence, $x \in \mathtt{in}(\mathcal{L})$ implies $(j-x) \in E$, and since by Definition 5, $y \in \mathtt{out}(\mathcal{L})$ implies $y \in E'+$, then $y \in \mathtt{out}(\mathcal{L})$ implies $(r-y) \in E$.
Right to left: Let E be an s extension of $\Delta_{\mathcal{M}}$. Let E' = $\{x|(j-x) \in E\}$ be the s extension of Δ as defined in Corollary 2, where if $(r-y) \in E$ then $y \in E'+$. By Proposition 1 there is an s labelling $\mathcal{L}$ where $\mathtt{in}(\mathcal{L})$ = E', $\mathtt{out}(\mathcal{L})$ = $E'+$ (recall that $E'- \subseteq E'+$).

Theorem 3 Let Δ_M = ($\mathcal{A}_{\mathcal{M}}$, $\mathcal{R}_{\mathcal{M}}$, $\mathcal{C}$, $\mathcal{L}_{\mathcal{M}}, \mathcal{D}_{dc}$) be the MAF of a DungC framework Δ = $(\mathcal{A}, \mathcal{R})$. Then for $s \in$ {complete, grounded, preferred, stable}, $(j-x) \in \mathcal{A}_{\mathcal{M}}$ is a credulously, respectively sceptically, justified argument of Δ_M under the s semantics, iff $x \in \mathcal{A}$ is a credulously, respectively sceptically, justified argument of Δ under the s semantics.

Proof: Follows from Lemma 4.

7.3 Proofs for Sections 3.4 and 3.5

In what follows we make use of the following notation:

Notation 4 Let $(\mathcal{A}, defeat)$ be defined on the basis of $(\mathcal{A}, \mathcal{R}, \mathcal{P})$ as in Definition 8, and let $E \subseteq \mathcal{A}$.

- $\overrightarrow{E+}$ denotes the set of defeats from arguments in E:
  $\overrightarrow{E+}$ = $\{(x,y) \mid (x,y) \in defeat\ ,\ x \in E\ \}$
- $E+$ denotes the set of arguments defeated by arguments in E:
  $E+$ = $\{y \mid (x,y) \in defeat, x \in E\ \}$
- $E-$ denotes the set of arguments that defeat arguments in E:
  $E-$ = $\{y \mid (y,x) \in defeat, x \in E\ \}$

Lemma 5 Let ($\mathcal{A}_{\mathcal{P}}$, $\mathcal{R}_{\mathcal{P}}$, $\mathcal{C}$, $\mathcal{L}_{\mathcal{M}}, \mathcal{D}_p$) be the P-MAF of a PAF $(\mathcal{A}, \mathcal{R}, \mathcal{P})$. For $s \in$ {admissible, complete, preferred, stable, grounded}:
E is an s extension of $(\mathcal{A}, \mathcal{R}, \mathcal{P})$ iff $E' \cup \{(xPy)|x \gg_{\mathcal{P}} y\}$ is an s extension of $(\mathcal{A}_{\mathcal{P}}, \mathcal{R}_{\mathcal{P}})$, where:

1. $(x,y) \in \overrightarrow{E+}$ iff $(x \text{ def } y) \in E'$
2. $y \in E+$ iff $(r-y) \in E'$
3. $x \in E$ iff $(j-x) \in E'$

Proof By Definition 8, E is an s extension of $(\mathcal{A}, \mathcal{R}, \mathcal{P})$ iff E is an s extension of the Dung framework ($\mathcal{A}$, *defeat*), where $(x,y) \in$ *defeat* iff $(x,y) \in \mathcal{R}$ and $\neg(y \gg_{\mathcal{P}} x)$. Let Δ_M = ($\mathcal{A}_{\mathcal{M}}$, $\mathcal{R}_{\mathcal{M}}$, $\mathcal{C}$, $\mathcal{L}_{\mathcal{M}}, \mathcal{D}_d$) be the MAF of ($\mathcal{A}$, *defeat*). By Corollary 2, E is an s extension of ($\mathcal{A}$, *defeat*) iff E^M is an s extension of ($\mathcal{A}_{\mathcal{M}}$, $\mathcal{R}_{\mathcal{M}}$), where:

1. $(x, y) \in \overrightarrow{E+}$ iff $(x \text{ def } y) \in E^M$
2. $y \in E+$ iff $(r - y) \in E^M$
3. $x \in E$ iff $(j - x) \in E^M$

Hence it suffices to show the following result:

> E^M is an s extension of ($\mathcal{A}_\mathcal{M}$, $\mathcal{R}_\mathcal{M}$) iff $E^P = E^M \cup \{(xPy) | x \gg_\mathcal{P} y\}$ is an s extension of $(\mathcal{A}_\mathcal{P}, \mathcal{R}_\mathcal{P})$

Firstly, note that it is straightforward to show that:
i) $\mathcal{A}_\mathcal{M} \subseteq \mathcal{A}_\mathcal{P}$, where $\mathcal{A}_\mathcal{P} - \mathcal{A}_\mathcal{M} = \{(xPy) | x \gg_\mathcal{P} y\} \cup \{(y \text{ def } x) | x \gg_\mathcal{P} y, y\mathcal{R}x\}$
ii) $\mathcal{R}_\mathcal{M} \subseteq \mathcal{R}_\mathcal{P}$, where $\mathcal{R}_\mathcal{P} - \mathcal{R}_\mathcal{M} =$
$\{((xPy), (y \text{ def } x)), ((r - y), (y \text{ def } x)), ((y \text{ def } x), (j - x)) | x \gg_\mathcal{P} y, y\mathcal{R}x\}$

1.1 *Left to right half for s = admissible*: Assume an admissible extension E^M of ($\mathcal{A}_\mathcal{M}$, $\mathcal{R}_\mathcal{M}$) and E^P defined as above. We show that every $\alpha \in E^P$ is acceptable w.r.t. E^P:

1.1.1 Since each $(xPy) \in \mathcal{A}_\mathcal{P}$ is not attacked by any argument, then each $(xPy) \in \mathcal{A}_\mathcal{P}$ is acceptable w.r.t. E^P.

1.1.2 Suppose $\alpha \in E^M$ and so $\alpha \in E^P$.
– Suppose $(\beta, \alpha) \in \mathcal{R}_\mathcal{P}$ and $(\beta, \alpha) \in \mathcal{R}_\mathcal{M}$. Hence, $\exists \gamma \in E^M$, $(\gamma, \beta) \in \mathcal{R}_\mathcal{M}$, and since $\mathcal{R}_\mathcal{M} \subseteq \mathcal{R}_\mathcal{P}$, $(\gamma, \beta) \in \mathcal{R}_\mathcal{P}$, where $\gamma \in E^P$ by definition of E^P.
– Suppose $(\beta, \alpha) \in \mathcal{R}_\mathcal{P}$ and $(\beta, \alpha) \notin \mathcal{R}_\mathcal{M}$. Since $\alpha \in E^M$, then by i) and ii) it must be the case that α is of the form $(j - x)$, β is of the form $(y \text{ def } x)$, and $(xPy)\mathcal{R}_\mathcal{P}(y \text{ def } x)$, where $(xPy) \in E^P$.

1.2 *Right to left half for s = admissible*: Assume E^P is an admissible extension of ($\mathcal{A}_\mathcal{P}$, $\mathcal{R}_\mathcal{P}$) and E^M defined as above. We show that every $\alpha \in E^M$ is acceptable w.r.t. E^M. Suppose some $(\beta, \alpha) \in \mathcal{R}_\mathcal{M}$. Hence $(\beta, \alpha) \in \mathcal{R}_\mathcal{P}$ and $\exists \gamma \in E^P$, $\gamma \mathcal{R}_\mathcal{P} \beta$.

1.2.1 Suppose $\gamma \in E^M$, $(\gamma, \beta) \notin \mathcal{R}_\mathcal{M}$. By i) and ii), γ must be of the form $(r - y)$, β of the form $(y \text{ def } x)$, and $(y \text{ def } x) \notin \mathcal{A}_\mathcal{M}$, contradicting $(\beta, \alpha) \in \mathcal{R}_\mathcal{M}$.

1.2.2 Suppose $\gamma \notin E^M$, in which case γ is of the form (xPy), β is of the form $(y \text{ def } x)$, and by i), $(y \text{ def } x) \notin \mathcal{A}_\mathcal{M}$, contradicting $(\beta, \alpha) \in \mathcal{R}_\mathcal{M}$.

• $s \in \{$*complete, grounded, preferred*$\}$. We define functions f and g s.t.

> For any admissible extension E^M of $\Delta^M = (\mathcal{A}_\mathcal{M}, \mathcal{R}_\mathcal{M})$, $E^P = h(E^M)$ as defined above.
> For any admissible extension E^P of $\Delta^P = (\mathcal{A}_\mathcal{P}, \mathcal{R}_\mathcal{P})$, $E^M = g(E^P)$ as defined above.

We show that:
a) h is monotonically strictly increasing in the sense that $\forall E^M, F^M$ s.t. E^M and F^M are admissible extensions of Δ^M and $E^M \subset F^M$, then $h(E^M) \subset h(F^M)$.
Suppose E^M and by **1.1** the corresponding admissible $E^P = h(E^M)$. Suppose $E^M \subset F^M$ and by **1.1** the corresponding admissible $F^P = h(F^M)$. It is obvious to see that $E^P \subset F^P$
b) g is monotonically strictly increasing in the sense that $\forall E^P, F^P$ s.t. E^P and F^P are

admissible extensions of Δ^P and $E^P \subset F^P$, then $g(E^P) \subset g(F^P)$.
Suppose E^P and by **1.2** the corresponding admissible $E^M = g(E^P)$. Suppose $E^P \subset F^P$, where by 1.1.1, $\forall \alpha \in (F^P - E^P)$, α is not of the form (xPy). Hence, by **1.2**, F^M is the corresponding admissible extension of Δ, where $E^M \subset F^M$.

Given **a)** and **b)**, the result is shown to hold for $s \in \{$complete, grounded, preferred$\}$ in exactly the same way as in Lemma 3.

• *Left to right half for s = stable*: Assume E^M is stable, and E^P defined as above. Suppose $\exists \alpha \notin E^P$ s.t. no argument in E^P $\mathcal{R}_\mathcal{P}$ attacks α. Then $\alpha \in \mathcal{A}_\mathcal{M}$, since otherwise α is of the form (xPy), contradicting $(xPy) \in E^P$, or α is of the form $(y\text{def}x)$, where $(xPy)\mathcal{R}_\mathcal{P}(y \text{ def } x)$, contradicting no argument in E^P $\mathcal{R}_\mathcal{P}$ attacks α. Hence, $\alpha \notin E^M$ (since otherwise $\alpha \in E^P$ by definition of E^P), and since $\mathcal{R}_\mathcal{M} \subseteq \mathcal{R}_\mathcal{P}$, α is not $\mathcal{R}_\mathcal{M}$ attacked by any argument in E^M, contradicting E^M is stable.

Right to left half for s = stable: Assume E^P is stable. If E^M is not stable then $\exists \beta \in \mathcal{A}_\mathcal{M}$, $\beta \notin E^M$ s.t. no argument in E^M $\mathcal{R}_\mathcal{M}$ attacks β. Since $\mathcal{A}_\mathcal{M} \subseteq \mathcal{A}_\mathcal{P}$, then $\beta \in \mathcal{A}_\mathcal{P}$. Since $E^M \subseteq E^P$, no argument in $E^P - E^M$ is in $\mathcal{A}_\mathcal{M}$, and E^P is stable, then $\beta \notin E^P$ and there is an argument α in E^P that $\mathcal{R}_\mathcal{P}$ attacks β. $\alpha \in E^P$ is either:
- an argument of the form (xPy), in which case β is of the form $(y\text{def}x)$. But then by i) and ii), $(y\text{def}x) \notin \mathcal{A}_\mathcal{M}$, contradicting $\beta \in \mathcal{A}_\mathcal{M}$.
- not of the form (xPy), in which case $\alpha \in E^M$. By assumption that no argument in E^M $\mathcal{R}_\mathcal{M}$ attacks β, and by i) and ii), $\alpha\mathcal{R}_\mathcal{P}\beta = (y\text{def}x)\mathcal{R}_\mathcal{P}(j-x)$ or $(r-x)\mathcal{R}_\mathcal{P}(y\text{def}x)$, where $(y \text{ def } x) \notin \mathcal{A}_\mathcal{M}$, contradicting $\alpha \in \mathcal{A}_\mathcal{M}$ and $\beta \in \mathcal{A}_\mathcal{M}$ respectively.

Theorem 4 Let Δ_M = ($\mathcal{A}_\mathcal{M}$, $\mathcal{R}_\mathcal{M}$, $\mathcal{C}$, $\mathcal{L}_\mathcal{M}, \mathcal{D}_p$) be the P-MAF of a PAF Δ = $(\mathcal{A}, \mathcal{R}, \mathcal{P})$. Then for $s \in \{$complete, grounded, preferred, stable$\}$, $(j - x) \in \mathcal{A}_\mathcal{M}$ is a credulously, respectively sceptically, justified argument of Δ_M under the s semantics, iff $x \in \mathcal{A}$ is a credulously, respectively sceptically, justified argument of Δ under the s semantics.

Proof Since every complete (and so grounded, preferred and stable) extension of ($\mathcal{A}_\mathcal{M}$, $\mathcal{R}_\mathcal{M}$) contains the set $\{(xPy) | x \gg_\mathcal{P} y\}$, then the theorem follows from Lemma 5.

Lemma 6 Let Δ_V = ($\mathcal{A}_\mathcal{V}$, $\mathcal{R}_\mathcal{V}$, $\mathcal{C}$, $\mathcal{L}_\mathcal{M}, \mathcal{D}_v$) be the V-MAF of an $aVAF$ Δ = ($\mathcal{A}$,$\mathcal{R}$,V, val,a). Let the extensions of Δ be the extensions of ($\mathcal{A}$, *defeat*$_\mathsf{a}$), where x *defeats*$_\mathsf{a}$ y iff $x\mathcal{R}y$, and $\neg(val(y) >_\mathsf{a} val(x))$ (as defined in Definition 9). Let $\overrightarrow{E+}$, $E+$ be defined as in Notation 4. Then, for $s \in \{$admissible, complete, preferred, stable, grounded$\}$:
E is an s extension of Δ iff $E' \cup \{(x_v P y_{v'}) | v >_\mathsf{a} v'\} \cup \{(>_\mathsf{a})\}$ is an s extension of $(\mathcal{A}_\mathcal{V}, \mathcal{R}_\mathcal{V})$, where:

1. $(x, y) \in \overrightarrow{E+}$ iff $(x \text{ def } y) \in E'$
2. $y \in E+$ iff $(r - y) \in E'$
3. $x \in E$ iff $(j - x) \in E'$

Proof Let ($\mathcal{A}_\mathcal{M}$, $\mathcal{R}_\mathcal{M}$, $\mathcal{C}$, $\mathcal{L}_\mathcal{M}, \mathcal{D}_d$) be the Dung MAF of ($\mathcal{A}$, *defeat*$_\mathsf{a}$). By Corollary 2, E is an s extension of ($\mathcal{A}$, *defeat*$_\mathsf{a}$) iff E^M is an s extension of ($\mathcal{A}_\mathcal{M}$, $\mathcal{R}_\mathcal{M}$), where:

1. $(x, y) \in \overrightarrow{E+}$ iff $(x \text{ def } y) \in E^M$
2. $y \in E+$ iff $(r - y) \in E^M$
3. $x \in E$ iff $(j - x) \in E^M$

Hence, letting $E^M = E'$, it suffices to show that E^M is an s extension of ($\mathcal{A}_\mathcal{M}$, $\mathcal{R}_\mathcal{M}$) iff $E^V = E^M \cup \{(x_v P y_{v'}) | v >_\mathsf{a} v'\} \cup \{(>_\mathsf{a})\}$ is an s extension of $(\mathcal{A}_\mathcal{V}, \mathcal{R}_\mathcal{V})$.

Firstly, it is straightforward to show that:

i) $\mathcal{A}_\mathcal{M} \subseteq \mathcal{A}_\mathcal{V}$, where:
$\mathcal{A}_\mathcal{V} - \mathcal{A}_\mathcal{M} = \{>_\mathsf{a}\} \cup \{(x_v P y_{v'}) | (x_v P y_{v'}) \in \mathcal{A}_\mathcal{V}\} \cup \{(y \text{ def } x) | val(x) >_\mathsf{a} val(y)\}$
i.e., the set of arguments $\mathcal{A}_\mathcal{V}$ extends $\mathcal{A}_\mathcal{M}$ with the audience argument $>_\mathsf{a}$, all value preference arguments, and the arguments $(y \text{ def } x)$ that by definition of Δ_V are attacked by value preference arguments endorsed by a.

ii) $\mathcal{R}_\mathcal{M} \subseteq \mathcal{R}_\mathcal{V}$, where $\mathcal{R}_\mathcal{V} - \mathcal{R}_\mathcal{M}$ is the set of attacks $((x_v P y_{v'}), (y_{v'} P x_v))$ between value preference arguments, all attacks $((>_\mathsf{a}), (y_{v'} P x_v))$ from the audience argument to value preference arguments, all attacks $((x_v P y_{v'}), (y \text{ def } x))$ from value preference to attack arguments, and incoming and outgoing attacks to and from arguments $(y \text{ def } x)$ that do not appear in $\mathcal{A}_\mathcal{M}$ given that $val(x) >_\mathsf{a} val(y)$), i.e.:

$$\{((r - y), (y \text{ def } x)), ((y \text{ def } x), (j - x)) | val(x) >_\mathsf{a} val(y)\}$$

iii) E^V contains the audience argument $(>_\mathsf{a})$ and partitions the value preference arguments, so that for every $(y_{v'} P x_v) \notin E^V$, $(y_{v'} P x_v)$ is attacked by $(>_\mathsf{a})$ (it is not the case that $v' >_\mathsf{a} v$) thus reinstating every audience *endorsed* $(x_v P y_{v'}) \in E^V$ against the attack by $(y_{v'} P x_v)$.

Left to right half for s = admissible: Assume E^M is an admissible extension of ($\mathcal{A}_\mathcal{M}$, $\mathcal{R}_\mathcal{M}$) and E^V defined as above. We show that every $\alpha \in E^V$ is acceptable w.r.t. E^V:

• Let $\alpha = (>_\mathsf{a})$. Then $(>_\mathsf{a})$ is acceptable w.r.t. E^V since $>_\mathsf{a}$ is not attacked by any argument, and by iii), all $(x_v P y_{v'}) \in E^V$ are acceptable w.r.t. E^V.

• Suppose $\alpha \in E^M \cap E^V$.
– Suppose $(\beta, \alpha) \in \mathcal{R}_\mathcal{V}$ and $(\beta, \alpha) \in \mathcal{R}_\mathcal{M}$. Hence, $\exists \gamma \in E^M$, $(\gamma, \beta) \in \mathcal{R}_\mathcal{M}$, and since $\mathcal{R}_\mathcal{M} \subseteq \mathcal{R}_\mathcal{V}$, $(\gamma, \beta) \in \mathcal{R}_\mathcal{V}$, where $\gamma \in E^V$ by definition of E^V.
– Suppose $(\beta, \alpha) \in \mathcal{R}_\mathcal{V}$ and $(\beta, \alpha) \notin \mathcal{R}_\mathcal{M}$. Then by i) and ii) it must be the case that α is of the form $(j - x)$, β is of the form $(y \text{ def } x)$, where $val(x) >_\mathsf{a} val(y)$, and $(j - x)$ is acceptable w.r.t. E^V given $(x_v P y_{v'}) \mathcal{R}_\mathcal{V} (y \text{ def } x)$, $(x_v P y_{v'}) \in E^V$.

Right to left half for s = admissible: Assume E^V is an admissible extension of ($\mathcal{A}_\mathcal{V}$, $\mathcal{R}_\mathcal{V}$) and E^M defined as above. We show that every $\alpha \in E^M$ is acceptable w.r.t. E^M. Suppose $(\beta, \alpha) \in \mathcal{R}_\mathcal{M}$. Hence $\beta \mathcal{R}_\mathcal{V} \alpha$ and $\exists \gamma \in E^V$, $\gamma \mathcal{R}_\mathcal{V} \beta$.

• Suppose $\gamma \in E^M$, $(\gamma, \beta) \notin \mathcal{R}_\mathcal{M}$. By i) and ii), γ must be of the form $(r - y)$, β of the form $(y \text{ def } x)$, and $(y \text{ def } x) \notin \mathcal{A}_\mathcal{M}$, contradicting $(\beta, \alpha) \in \mathcal{R}_\mathcal{M}$.

• Suppose $\gamma \notin E^M$. Then γ is of the form $(x_v P y_{v'})$ or $(>_\mathsf{a})$, and β is of the form $(y \text{ def } x)$ or $(y_{v'} P x_v)$. By i), in either case any such β is not in $\mathcal{A}_\mathcal{M}$, contradicting $(\beta, \alpha) \in \mathcal{R}_\mathcal{M}$.

Left to right and right to left half for $s \in$ {complete, grounded, preferred}: Let us define functions f and g s.t. for any admissible extension E^M of $\Delta^M = (\mathcal{A}_\mathcal{M}, \mathcal{R}_\mathcal{M})$, $E^V = h(E^M)$ as defined above, and for any admissible extension E^V of Δ^V = ($\mathcal{A}_\mathcal{V}$, $\mathcal{R}_\mathcal{V}$), $E^M = g(E^V)$ as defined above.

We show that **a)** h is monotonically strictly increasing, and **b)** g is monotonically strictly increasing, in the same way as in Lemma 5, substituting the superscript V for P, and in the proof of **b)** noting that $\forall \alpha \in (F^V - E^V)$, by definition of E^V and F^V, α is not a value preference argument $(x_v P y_{v'})$ or the audience argument $(>_\mathsf{a})$. Given **a)** and **b)**, the result is shown to hold for $s \in$ {complete, grounded, preferred} in exactly the same way as in Lemma 3.

Left to right half for s = stable: The proof proceeds in the same way as for the left to right half for s = stable in Lemma 5 (substituting $\mathcal{R}_\mathcal{V}$ for $\mathcal{R}_\mathcal{P}$), except we show $\alpha \in \mathcal{A}_\mathcal{M}$ as follows. Suppose otherwise. Then:
– α is the audience argument $(>_\mathsf{a})$, contradicting $\alpha \notin E^V$, or;
– by iii), α is a value preference argument $(y_{v'} P x_v)$ attacked by $(>_\mathsf{a})$, contradicting no argument in E^V $\mathcal{R}_\mathcal{V}$ attacks α, or α is a value preference argument $(x_v P y_{v'})$ endorsed by $(>_\mathsf{a})$, contradicting $\alpha \notin E^V$, or;
– α is of the form $(y$ def $x)$ s.t. $val(x) >_\mathsf{a} val(y)$, and so $(y$ def $x)$ is attacked by some $(x_v P y_{v'}) \in E^V$ that is endorsed by $(>_\mathsf{a})$, contradicting no argument in E^V $\mathcal{R}_\mathcal{V}$ attacks α.

Right to left half for s = stable: Assume E^V is stable and E^M defined as above. By the right to left for s = complete, E^M is complete. If E^M is not stable then $\exists \beta \in \mathcal{A}_\mathcal{M}$, $\beta \notin E^M$ s.t. no argument in E^M $\mathcal{R}_\mathcal{M}$ attacks β. Since $\mathcal{A}_\mathcal{M} \subseteq \mathcal{A}_\mathcal{V}$, then $\beta \in \mathcal{A}_\mathcal{V}$, and since $E^M \subseteq E^V$ and E^V is stable, there is a α in E^V that $\mathcal{R}_\mathcal{V}$ attacks β. Suppose $\alpha \in E^V - E^M$. Then α is of the form $(x_v P y_{v'})$ or $(>_\mathsf{a})$, and β is of the form $(y$ def $x)$ or $(y_{v'} P x_v)$, in either case contradicting $\beta \in \mathcal{A}_\mathcal{M}$. Suppose $\alpha \in E^V$, $\alpha \in E^M$, where by assumption that no argument in E^M $\mathcal{R}_\mathcal{M}$ attacks β, and by i) and ii), $\alpha \mathcal{R}_\mathcal{V} \beta$ = $(y$ def $x) \mathcal{R}_\mathcal{V} (j - x)$ or $(r - x) \mathcal{R}_\mathcal{V} (y$ def $x)$, and $(y$ def $x) \notin \mathcal{A}_\mathcal{M}$. Hence, the first case contradicts $\alpha \in \mathcal{A}_\mathcal{M}$, and the second case contradicts $\beta \in \mathcal{A}_\mathcal{M}$.

Theorem 5 Let Δ_M = ($\mathcal{A}_\mathcal{M}$, $\mathcal{R}_\mathcal{M}$, $\mathcal{C}$, $\mathcal{L}_\mathcal{M}, \mathcal{D}_v$) be the V-MAF of an $aVAF$ Δ = ($\mathcal{A}$,$\mathcal{R}$,V,val,a). Then for $s \in$ {complete, grounded, preferred, stable}, $(j - x) \in \mathcal{A}_\mathcal{M}$ is a credulously, respectively sceptically, justified argument of Δ_M under the s semantics, iff $x \in \mathcal{A}$ is a credulously, respectively sceptically, justified argument of Δ under the s semantics

Proof Given that every complete (and so grounded, preferred and stable) extension of ($\mathcal{A}_\mathcal{M}$, $\mathcal{R}_\mathcal{M}$) contains the set $\{(x_v P y_{v'}) | v >_\mathsf{a} v'\} \cup \{(>_\mathsf{a})\}$, then the theorem follows immediately from Lemma 6.

Lemma 7 Let Δ_V = ($\mathcal{A}_\mathcal{V}$, $\mathcal{R}_\mathcal{V}$, $\mathcal{C}$, $\mathcal{L}_\mathcal{M}, \mathcal{D}_v$) be the V-MAF of a VAF Δ = ($\mathcal{A}$,$\mathcal{R}$,V,val, P). Then E is a preferred extension of ($\mathcal{A}$,$\mathcal{R}$,V,val,a), where $\mathsf{a} \in P$, iff $E' \cup \{(x_v P y_{v'}) | v >_\mathsf{a} v'\} \cup \{(>_\mathsf{a})\}$ is a preferred extension of Δ_V, where:

1. $(x, y) \in \overrightarrow{E+}$ iff $(x$ def $y) \in E'$
2. $y \in E+$ iff $(r - y) \in E'$
3. $x \in E$ iff $(j - x) \in E'$

Proof Given Lemma 6 it suffices to show that:

E^* = $E' \cup \{(x_v P y_{v'}) | v >_\mathsf{a} v'\} \cup \{(>_\mathsf{a})\}$ is a preferred extension of the metalevel formulation Δ_a = ($\mathcal{A}_\mathsf{a}$, $\mathcal{R}_\mathsf{a}$, $\mathcal{C}, \mathcal{L}_\mathcal{M}, \mathcal{D}_v$) of ($\mathcal{A}$,$\mathcal{R}$,$V$,$val$,$\mathsf{a}$), where $\mathsf{a} \in P$, iff E^* is a preferred extension of Δ_V.
Firstly, note that it is straightforward to show that for each Δ_a:

1. $\mathcal{A}_\mathsf{a} \subseteq \mathcal{A}_V$, where $\mathcal{A}_V - \mathcal{A}_\mathsf{a}$ = $\{(>_{\mathsf{a}'}) | \mathsf{a}' \neq \mathsf{a}, \mathsf{a}' \in P\}$
2. $\mathcal{R}_\mathsf{a} \subseteq \mathcal{R}_V$, where $\mathcal{R}_V - \mathcal{R}_\mathsf{a}$ = $\{((>_{\mathsf{a}'}), (x_v P y_{v'})), ((>_{\mathsf{a}'}), (>_\mathsf{a})), ((>_\mathsf{a}), (>_{\mathsf{a}'}))\ | \mathsf{a}' \neq \mathsf{a}, \mathsf{a}' \in P, v >_\mathsf{a} v'\}$

a) Let E^* be an admissible extension of some Δ_a. We show that E^* is an admissible extension of Δ_V. Suppose $(\beta, \alpha) \in \mathcal{R}_V$, $\alpha \in E^*$, and:

– $(\beta, \alpha) \in \mathcal{R}_{\mathsf{a}}$, in which case $\exists\gamma \in E^*$, $\gamma\mathcal{R}_{\mathsf{a}}\beta$, and by 2, $\gamma\mathcal{R}_V\beta$;
– $(\beta, \alpha) \notin \mathcal{R}_{\mathsf{a}}$, in which case by 1 and 2, β must be some $(>_{\mathsf{a}'})$, $\mathsf{a}' \neq \mathsf{a}$, α is either some $(x_v P y_{v'})$ or $(>_{\mathsf{a}})$. But then $(>_{\mathsf{a}})\mathcal{R}_V(>_{\mathsf{a}'})$ where $(>_{\mathsf{a}}) \in E^*$.

b) Let E^* be an admissible extension of Δ_V. We show that E^* is an admissible extension of Δ_{a}. Suppose $\beta\mathcal{R}_{\mathsf{a}}\alpha$, $\alpha \in E^*$. By 2), $\beta\mathcal{R}_V\alpha$, and by the admissibility of E^*, $\exists\gamma \in E^*$, $\gamma\mathcal{R}_V\beta$. Suppose $\neg(\gamma\mathcal{R}_{\mathsf{a}}\beta)$. Since $\beta \in \mathcal{A}_{\mathsf{a}}$, then by 1 and 2, it must be that γ is some $(>_{\mathsf{a}'})$ s.t. $\mathsf{a}' \neq \mathsf{a}$, and β is a value preference argument $(x_v P y_{v'})$. But this contradicts E^* is a conflict free subset of $\mathcal{A}_V$, given that $(>_{\mathsf{a}})\mathcal{R}_V(>_{\mathsf{a}'})$ and $(>_{\mathsf{a}}) \in E^*$

Suppose E^* is a preferred extension of some Δ_{a}. By **a)**, E^* is an admissible extension of Δ_V. Suppose E^* is not a preferred extension of Δ_V. Then, $\exists E^{**} \supset E^*$ s.t. E^{**} is an admissible extension of Δ_V. Suppose $\alpha \in (E^{**} - E^*)$, where $\alpha \in (\mathcal{A}_{\mathcal{V}} - \mathcal{A}_{\mathsf{a}})$. But then this contradicts E^{**} is conflict free, given that α is either a value preference or audience argument, and $(>_{\mathsf{a}}) \in E^*$, $(>_{\mathsf{a}})\mathcal{R}_V(y_{v'} P x_v)$ for every $(y_{v'} P x_v) \notin E^*$ (see iii) in Lemma 6), and $(>_{\mathsf{a}})\mathcal{R}_V(>_{\mathsf{a}'})$ for every $\mathsf{a}' \neq \mathsf{a}$. Suppose $\alpha \in (E^{**} - E^*)$, where $\alpha \in \mathcal{A}_{\mathsf{a}}$. By **b)**, E^{**} is an admissible extension of Δ_{a}, contradicting E^* is a preferred extension of Δ_{a}.

Suppose E^* is a preferred extension of Δ_V. By **b)**, E^* is an admissible extension of Δ_{a}. Suppose E^* is not a preferred extension of Δ_{a}. Then, $\exists E^{**} \supset E^*$ s.t. E^{**} is an admissible extension of Δ_{a}. But then by **a)**, E^{**} is an admissible extension of Δ_V, contradicting E^* is a preferred extension of Δ_V.

Theorem 6 Let Δ_M = ($\mathcal{A}_{\mathcal{M}}$, $\mathcal{R}_{\mathcal{M}}$, $\mathcal{C}$, $\mathcal{L}_{\mathcal{M}}, \mathcal{D}_v$) be the V-MAF of a VAF Δ = ($\mathcal{A}$,$\mathcal{R}$,V,val,P). Then for any $x \in \mathcal{A}$, $(j - x) \in \mathcal{A}_{\mathcal{M}}$:

1. x is an objectively acceptable argument of Δ iff $(j - x)$ is a sceptically justified argument of Δ_M under the preferred semantics.

2. x is a subjectively acceptable argument of Δ iff $(j - x)$ is a credulously justified argument of Δ_M under the preferred semantics.

Proof
1) *Left to right*: Let x be an objectively acceptable argument of Δ. Suppose any preferred extension E' of Δ_M. Since E' is complete, it must contain some set $\{(x_v P y_{v'}) \mid v >_{\mathsf{a}} v'\} \cup \{(>_{\mathsf{a}})\}$ where $(>_{\mathsf{a}}) \in \mathcal{A}_{\mathcal{M}}$. Suppose $(j - x) \notin E'$. But then by the right to left half of Lemma 7, there is a corresponding preferred extension E of some ($\mathcal{A}$,$\mathcal{R}$,V,val,a) that does not contain x, contradicting x is an objectively acceptable argument of Δ.
Right to left: Let E' be any preferred extension of Δ_M, where E' must contain some set $\{(x_v P y_{v'}) | v >_{\mathsf{a}} v'\} \cup \{(>_{\mathsf{a}})\}$. We have $(j - x) \in E'$, and by the right to left half of Lemma 7, there is a corresponding preferred extension E of some ($\mathcal{A}$,$\mathcal{R}$,V,val,a) that contains x.

2) *Left to right*: Let E be the preferred extension of some ($\mathcal{A}$,$\mathcal{R}$,V,val,a), $x \in E$. By the left to right half of Lemma 7, there is a corresponding preferred extension E' of Δ_M s.t. $(j - x) \in E'$.
Right to left: Let E' be any preferred extension of Δ_M, where E' must contain some set $\{(x_v P y_{v'}) | v >_{\mathsf{a}} v'\} \cup \{(>_{\mathsf{a}})\}$. Let $(j - x) \in E'$. By the right to left half of Lemma 7, there is a corresponding preferred extension E of ($\mathcal{A}$,$\mathcal{R}$,V,val,a) that contains x.

7.4 Proofs for Section 3.6

Lemmas 8, 9 and 10 are used in the proof of Theorem 7.

Lemma 8 Let Δ_H = ($((\mathcal{A}_1, \mathcal{R}_1), \mathcal{D}_1)$, ..., $((\mathcal{A}_n, \mathcal{R}_n), \mathcal{D}_n)$) be the partition of the bounded hierarchical $\Delta = (\mathcal{A}, \mathcal{R}, \mathcal{D})$.
Let Δ_M = ($\mathcal{A}_\mathcal{M}$, $\mathcal{R}_\mathcal{M}$, $\mathcal{C}$, $\mathcal{L}_\mathcal{M}, \mathcal{D}_e$) be the metalevel formulation of Δ as defined in Definition 27.
Then there exists a partition Δ_{MH} of Δ_M such that:
Δ_{MH} = ($((\mathcal{A}'_1, \mathcal{R}'_1), (\mathcal{A}'_{1-\mathcal{D}}, \mathcal{R}'_{1-\mathcal{D}}))$, ..., $((\mathcal{A}'_n, \mathcal{R}'_n), (\mathcal{A}'_{n-\mathcal{D}}, \mathcal{R}'_{n-\mathcal{D}}))$), where:

1. $\mathcal{A}_\mathcal{M}$ = $\bigcup_{i=1}^{n}(\mathcal{A}'_i \cup \mathcal{A}'_{i-\mathcal{D}})$ and $\mathcal{R}_M$ = $\bigcup_{i=1}^{n}$ ($\mathcal{R}'_i \cup \mathcal{R}'_{i-\mathcal{D}}$)
2. for $i = 1 \ldots n$, $(\mathcal{A}'_i, \mathcal{R}'_i)$ are the arguments and attacks in the Dung MAF formulation of $(\mathcal{A}_i, \mathcal{R}_i)$
3. for $i = 1 \ldots n$: $(z, (y, x)) \in \mathcal{D}_i$ **iff**

   $(\{(j - z), (r - z), (zD(yDx)),(y\text{def}x)\} \subseteq \mathcal{A}'_{i-\mathcal{D}}$,

   $\{((j-z),(r-z)), ((r-z),(zD(yDx))), ((zD(yDx)),(y\text{def}x)) \subseteq \mathcal{R}'_{i-\mathcal{D}}$
4. $(\mathcal{A}'_{n-\mathcal{D}}, \mathcal{R}'_{n-\mathcal{D}})$ = $(\emptyset, \emptyset)$, $\mathcal{D}_n$ = $\emptyset$

Proof Proof is obvious. Intuitively, the partition Δ_{MH} corresponds to the partition Δ_H, where each Dung framework $(\mathcal{A}_i, \mathcal{R}_i)$ is formulated as its Dung MAF with arguments and attacks $(\mathcal{A}'_i, \mathcal{R}'_i)$, and the $\mathcal{D}_i$ attacks are formulated as the metalevel attacks in $(\mathcal{A}'_{i-\mathcal{D}}, \mathcal{R}'_{i-\mathcal{D}})$. We illustrate with the example in Figure 14.

Lemma 9 Let Δ_H, Δ_M and its partition Δ_{MH} = ($((\mathcal{A}'_1, \mathcal{R}'_1), (\mathcal{A}'_{1-\mathcal{D}}, \mathcal{R}'_{1-\mathcal{D}}))$, ..., $((\mathcal{A}'_n, \mathcal{R}'_n), (\mathcal{A}'_{n-\mathcal{D}}, \mathcal{R}'_{n-\mathcal{D}}))$) be defined as in Lemma 8. Let us define the tuple:

($((\mathcal{A}'_1, \mathcal{R}'_1), \mathcal{R}'_{1-\mathcal{D}r})$, ..., $((\mathcal{A}'_n, \mathcal{R}'_n), \mathcal{R}'_{n-\mathcal{D}r})$)

where for $i = 1 \ldots n-1$, the Dung framework $(\mathcal{A}'_{i-\mathcal{D}}, \mathcal{R}'_{i-\mathcal{D}})$) with attacks ($(j - z)$,$(r - z)$), ($(r - z)$,$(zD(yDx))$), ($(zD(yDx))$,$(y\text{def}x)$), is replaced by the singleton set of attacks $\mathcal{R}'_{i-\mathcal{D}r}$ = $\{((j - z), (y\text{def}x))\}$ (notice that for $i = 1 \ldots n-1$, $(y\text{def}x)$ $\in \mathcal{A}'_i$ and $(j - z) \in \mathcal{A}'_{i+1}$)
Let the *reduction* Δ_{MHr} of Δ_{MH} be defined as $(\mathcal{A}_{MHr}, \mathcal{R}_{MHr})$ where [12]:

$\mathcal{A}_{MHr}$ = $\bigcup_{i=1}^{n} \mathcal{A}'_i$ and $\mathcal{R}_{MHr}$ = $\bigcup_{i=1}^{n}$ ($\mathcal{R}'_i \cup \mathcal{R}'_{i-\mathcal{D}r}$).

Then $\forall \alpha \in \mathcal{A}_{MHr} \cap \mathcal{A}_\mathcal{M}$, for $s \in \{$complete, grounded, preferred, stable$\}$, α is a credulously, respectively sceptically justified argument of Δ_{MHr} under the s semantics, iff α is a credulously, respectively sceptically justified argument of Δ_M under the s semantics.

Proof: Δ_{MH} is an expansion of Δ_{MHr} such that each $((j - z), (y\text{def}x)) \in \mathcal{R}_{MHr}$ is expanded to obtain the set $\{((j-z),(r-z)), ((r-z),(zD(yDx))), ((zD(yDx)),(y\text{def}x))\}$ in $\mathcal{R}_{MHr}$. The result therefore follows from Corollary 1.

Lemma 10 Let E be an admissible extension of a bounded hierarchical EAF Δ = $(\mathcal{A}, \mathcal{R}, \mathcal{D})$. Let $x \in E$, $x \rightarrow^E y$ and Rs a reinstatement set for $x \rightarrow^E y$. Then $\forall F \supset E$ s.t. F is admissible, $x \rightarrow^F y$ and Rs is a reinstatement set for $x \rightarrow^F y$.

Proof Given the partition ($((\mathcal{A}_1, \mathcal{R}_1), \mathcal{D}_1)$, ..., $((\mathcal{A}_n, \mathcal{R}_n), \mathcal{D}_n)$) of Δ, we can partition

[12] See Figure 14 for an example of Δ_{MH} and its reduction Δ_{MHr}

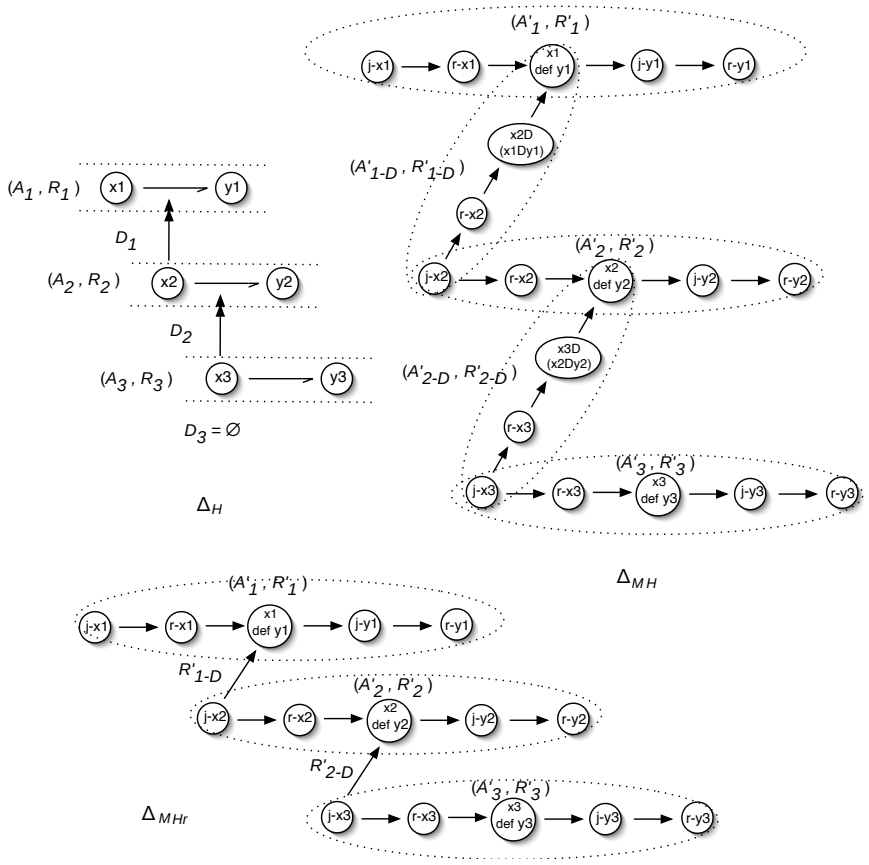

Figure 14: Δ_H is the hierarchical partition of the EAF Δ, and Δ_{MH} is the hierarchical partition of Δ's metalevel formulation Δ_M. Δ_{MHr} is the reduction of Δ_{MH}.

E into $E_1 \cup \cdots \cup E_n$ where $x \in E_i$ iff $x \in \mathcal{A}_i$ and $(y, x) \in \mathcal{R}$ implies $(y, x) \in \mathcal{R}_i$, $(x, y) \in \mathcal{R}$ implies $(x, y) \in \mathcal{R}_i$, and $(y', (x, y)) \in \mathcal{D}$ implies $(y', (x, y)) \in \mathcal{D}_i$, $y' \in \mathcal{A}_{i+1}$. We now prove the result by induction on i:

Base case: Suppose $x \in E_{n-1}$, $x \rightarrow^E y$, and $y'_1, \ldots, y'_m$ s.t. for $k = 1 \ldots m$, $(y'_k, (x, y)) \in \mathcal{D}_{n-1}$. Since $\mathcal{D}_n = \emptyset$, then the reinstatement set Rs for $x \rightarrow^E y$ is of the form $\{x \rightarrow^E y, x'_1 \rightarrow^E y'_1, \ldots, x'_m \rightarrow^E y'_m\}$, since for $k = 1 \ldots m$ $\neg\exists(z, (x'_k, y'_k)) \in \mathcal{D}$. The latter also implies that for any admissible F s.t. $F \supset E$, $y'_k \notin F$ (since otherwise F would not be conflict free), $x \rightarrow^F y$ and $Rs = \{x \rightarrow^F y, x'_1 \rightarrow^F y'_1, \ldots, x'_m \rightarrow^F y'_m\}$ is a reinstatement set for $x \rightarrow^F y$.

Inductive hypothesis: The result holds for $x \in E_j$, $j > i$.

General case: Suppose $x \in E_i$, $x \rightarrow^E y$ and a reinstatement set Rs for $x \rightarrow^E y$. Suppose $\{y'_1, \ldots, y'_m\} = \{y' | (y', (x, y)) \in \mathcal{D}\}$. By assumption of Rs, for $k = 1 \ldots m$, $\exists x'_k \in E_{i+1}$ s.t. $x'_k \rightarrow^E y'_k$. Hence, $Rs = \{x \rightarrow^E y\} \cup \bigcup_{k=1}^m Rs_k$ where Rs_k is a reinstatement set for $x'_k \rightarrow^E y'_k$. By inductive hypothesis, for k = $1 \ldots m$, $x'_k \rightarrow^F y'_k$ and Rs_k is a reinstatement set for $x'_k \rightarrow^F y'_k$. Hence, $x \rightarrow^F y$ (since for $k = 1 \ldots m$, $y'_k \notin F$, given that by Proposition 2 in [38] no two arguments defeat$_F$ each other in a conflict free F) and $\bigcup_{k=1}^m Rs_k \cup \{x \rightarrow^F y\}$ is a reinstatement set for $x \rightarrow^F y$.

Theorem 7 Let Δ_M = ($\mathcal{A}_\mathcal{M}$, $\mathcal{R}_\mathcal{M}$, $\mathcal{C}$, $\mathcal{L}_\mathcal{M}, \mathcal{D}_e$) be the E-MAF of a bounded hierarchical *EAF* Δ = $(\mathcal{A}, \mathcal{R}, \mathcal{D})$. Then for $s \in \{$complete, grounded, preferred, stable$\}$, $(j - x) \in \mathcal{A}_\mathcal{M}$ is a credulously, respectively sceptically, justified argument of Δ_M under the s semantics, iff $x \in \mathcal{A}$ is a credulously, respectively sceptically, justified argument of Δ under the s semantics.

Proof Let Δ_H = ($((\mathcal{A}_1, \mathcal{R}_1), \mathcal{D}_1)$, ..., $((\mathcal{A}_n, \mathcal{R}_n), \mathcal{D}_n)$) be the partition of Δ. Let Δ_{MH} be the partition of Δ_M. Let Δ_{MHr} = $(\mathcal{A}_{MHr}, \mathcal{R}_{MHr})$ be the reduction of Δ_{MH}, as defined by Lemma 9 on the basis of

$$(((\mathcal{A}'_1, \mathcal{R}'_1), \mathcal{R}'_{1-\mathcal{D}r}), \ldots, ((\mathcal{A}'_n, \mathcal{R}'_n), \mathcal{R}'_{n-\mathcal{D}r}))$$

Given Lemmas 8 and 9, it suffices to show that:

> $(j - x) \in \mathcal{A}_{MHr}$ is a credulously, respectively sceptically, justified argument of Δ_{MHr} under the s semantics, iff $x \in \mathcal{A}$ is a credulously, respectively sceptically, justified argument of Δ under the s semantics.

To show the above we show that: E is an s extension of Δ iff E' is an s extension of Δ_{MHr}, where:

1. $x \in E$, x defeats$_E$ y and there is a reinstatement set for the defeat $x \rightarrow_E y$ iff $(x \text{ def } y), (r - y) \in E'$
2. $x \in E$ iff $(j - x) \in E'$

Observe that:

O1 Referring to the EAF Δ and its hierarchical partition Δ_H: $\forall(\beta, \alpha) \in \mathcal{R}$, $\forall(\gamma, (\beta, \alpha)) \in \mathcal{D}$, $(\beta, \alpha) \in \mathcal{R}_i$ iff $(\gamma, (\beta, \alpha)) \in \mathcal{D}_i$, $\gamma \in \mathcal{A}_{i+1}$

O2 For $i = 1 \ldots n$: $(\mathcal{A}'_i, \mathcal{R}'_i)$ are the arguments and attacks in the Dung MAF formulation of $(\mathcal{A}_i, \mathcal{R}_i)$ in the partition Δ_H of Δ

O3 For $i = 1 \ldots n - 1$, $(\gamma, \delta) \in \mathcal{R}'_{i-\mathcal{D}r}$ implies $\gamma \in \mathcal{A}'_{i+1}$, $\delta \in \mathcal{A}'_i$, and γ is an argument of the form $(j - z)$, δ is an argument of the form $(y\text{def}x)$.

O4 For $i = 1 \ldots n$, $((j - z), (y\text{def}x)) \in \mathcal{R}'_{i-\mathcal{D}r}$ iff $(z, (y, x)) \in \mathcal{D}_i$.

$O1 - O4$ imply that E can be partitioned into $E_1, \ldots, E_n$, and E' into $E'_1, \ldots, E'_n$, and that the theorem is shown by proving by induction on i, the following result:
$E^i = (E_i \cup \ldots \cup E_n)$ is an s extension of $(\mathcal{A}^i, \mathcal{R}^i, \mathcal{D}^i)$ = $(\mathcal{A}_i \cup \cdots \cup \mathcal{A}_n, \mathcal{R}_i \cup \cdots \cup \mathcal{R}_n, \mathcal{D}_i \cup \cdots \cup \mathcal{D}_n)$ iff $E^{i'} = (E'_i \cup \ldots \cup E'_n)$ is an s extension of $(\mathcal{A}^{i'}, \mathcal{R}^{i'})$ = $(\mathcal{A}'_i \cup \cdots \cup \mathcal{A}'_n, \mathcal{R}'_i \cup \mathcal{R}'_{i-\mathcal{D}r} \cup \cdots \cup \mathcal{R}'_n \cup \mathcal{R}'_{n-\mathcal{D}r})$, where:

1. $z \in E^i$, z defeats$_{E^i}$ y and there is a reinstatement set for the defeat $z \rightarrow_{E^i} y$ iff $(z \text{ def } y), (r - y) \in E^{i'}$
2. $x \in E^i$ iff $(j - x) \in E^{i'}$

1) s = *admissible*. Firstly, note that:

R1 $E^i = (E_i \cup \ldots \cup E_n)$ is an admissible extension of $(\mathcal{A}^i, \mathcal{R}^i, \mathcal{D}^i)$ implies $\forall j > i$, $E^j = (E_j \cup \ldots \cup E_n)$ is an admissible extension of $(\mathcal{A}^j, \mathcal{R}^j, \mathcal{D}^j)$.
To show the above, assume $\alpha \in E^j$, $\beta \rightarrow^{E^j} \alpha$, where given the partition of Δ,

if $(\epsilon,(\beta,\alpha)) \in \mathcal{D}$ then $(\epsilon,(\beta,\alpha)) \in \mathcal{D}^j$, $\epsilon \in \mathcal{A}_{j+1}$. Hence $\beta \rightarrow^{E^i} \alpha$, and by the admissibility of E^i, $\exists\gamma \in E^i$, $\gamma \rightarrow^{E^i} \beta$ and there is a reinstatement set RS_i for $\gamma \rightarrow^{E^i} \beta$. Given the partition of Δ, $\gamma \in \mathcal{A}^j$ and if $(\delta,(\gamma,\beta)) \in \mathcal{D}$ then $(\delta,(\gamma,\beta)) \in \mathcal{D}^j$, $\delta \in \mathcal{A}_{j+1}$. Hence, it is straightforward to show that $\gamma \rightarrow^{E^j} \beta$ and there is a reinstatement set RS_j for $\gamma \rightarrow^{E^j} \beta$.

R2 $E^{i'}$ is an admissible extension of $(\mathcal{A}^{i'},\mathcal{R}^{i'})$ implies $\forall j > i$, $E^{j'}$ is an admissible extension of $(\mathcal{A}^{j'},\mathcal{R}^{j'})$.
This follows given Δ's partition, which implies that $\forall j > i$, no argument in $(\mathcal{A}^{i'} - \mathcal{A}^{j'})\ \mathcal{R}^i$ attacks an argument in $\mathcal{A}^{j'}$, and so for any $\alpha \in E^{j'}$, $(\beta,\alpha) \in \mathcal{R}^{i'}$ iff $(\beta,\alpha) \in \mathcal{R}^{j'}$, and $\exists\gamma \in E^{i'}$ s.t. $(\gamma,\beta) \in \mathcal{R}^{i'}$ iff $\gamma \in E^{j'}$ and $(\gamma,\beta) \in \mathcal{R}^{j'}$.

Base case (i = n): Since $\mathcal{D}_n = \emptyset$, z defeats$_{E^n}$ y iff $z\mathcal{R}_n y$, and trivially there is a reinstatement set for $z \rightarrow^{E^n} y$. Also, $\mathcal{R}_{n-\mathcal{D}r}$ = $\emptyset$, and $(\mathcal{A}'_n,\mathcal{R}'_n)$ are the arguments and attacks in the Dung MAF formulation of $(\mathcal{A}_n,\mathcal{R}_n)$. Hence, for $i = n$, the result follows immediately from Corollary 2.

Inductive hypothesis (IH): The result holds for $j > i$.

General Case:
Left to right half: Let $E^i = (E_i \cup \ldots \cup E_n)$ be an admissible extension of $(\mathcal{A}^i,\mathcal{R}^i,\mathcal{D}^i)$. We show that $E^{i'}$ as defined above is an admissible extension of $(\mathcal{A}^{i'},\mathcal{R}^{i'})$. By R1, $E^{i+1} = (E_{i+1} \cup \ldots \cup E_n)$ is an admissible extension of $(\mathcal{A}^{i+1},\mathcal{R}^{i+1},\mathcal{D}^{i+1})$, and by *IH*, $E^{i+1'} = (E'_{i+1} \cup \ldots \cup E'_n)$ is an admissible extension of $(\mathcal{A}^{i+1'},\mathcal{R}^{i+1'})$. We show that $E^{i'}$, where $E^{i'} \supset E^{i+1'}$, is an admissible extension of $(\mathcal{A}^{i'}\mathcal{R}^{i'})$.
Suppose $x \in E_i$. Then $(j-x) \in E_{i'}$. We show $(j-x)$ is acceptable w.r.t. $E^{i'}$:

By definition of Δ and Δ_{MHr}, $\exists(y,x) \in \mathcal{R}^i$ iff $\exists((y\text{def}x),(j-x)) \in \mathcal{R}^{i'}$. Suppose $y\mathcal{R}^i x$. By assumption of x acceptable w.r.t. E^i, $\exists z \in E_i$ s.t. $z \rightarrow^{E^i} y$ and there is a reinstatement set for $z \rightarrow^{E^i} y$. By definition of $E^{i'}$, $(j-z),(z\text{def}y),(r-y) \in E^{i'}$, where $(r-y)\mathcal{R}^{i'}(y\text{def}x)$, and so $(j-x)$ is acceptable w.r.t. $E^{i'}$.

The result is shown in full by showing that $(r-y)$ and $(z\text{def}y)$ are acceptable w.r.t. $E^{i'}$:

Firstly, $(z\text{def}y) \in E'_i$ reinstates $(r-y)$ against the attack $(j-y)\mathcal{R}^{i'}(r-y)$.
Secondly, $(j-z) \in E'_i$ reinstates $(z\text{def}y)$ against the attack $(r-z)\mathcal{R}^{i'}(z\text{def}y)$.
However, suppose $\exists(j-y') \in E'_{i+1}$ s.t. $((j-y'),(z\text{def}y)) \in \mathcal{R}'_{i-\mathcal{D}r}$. Hence, $(y',(z,y)) \in \mathcal{D}_i$, and by assumption of a reinstatement set for $z \rightarrow^{E^i} y$, $\exists z' \in E_{i+1}$, $z' \rightarrow^{E^i} y'$, and there is a reinstatement set for $z' \rightarrow^{E^i} y'$. By R1 and *IH* above, $(j-z'),(z'\text{def}y') \in E'_{i+1}$, and given $(z'\text{def}y')\mathcal{R}^{i+1'}(j-y')$ and so $(z'\text{def}y')\mathcal{R}^{i'}(j-y')$, $(z\text{def}y)$ is acceptable w.r.t. $E^{i'}$.

Right to left half: Let $E^{i'} = (E'_i \cup \ldots \cup E'_n)$ be an admissible extension of $(\mathcal{A}^{i'},\mathcal{R}^{i'})$. We show that E^i as defined above is an admissible extension of $(\mathcal{A}^i,\mathcal{R}^i,\mathcal{D}^i)$. By R2, $E^{i+1'} = (E'_{i+1} \cup \ldots \cup E'_n)$ is an admissible extension of $(\mathcal{A}^{i+1'},\mathcal{R}^{i+1'})$, and by *IH*, $E^{i+1} = (E_{i+1} \cup \ldots \cup E_n)$ is an admissible extension of $(\mathcal{A}^{i+1},\mathcal{R}^{i+1},\mathcal{D}^{i+1})$. We show that E^i, where $E^i \supset E^{i+1}$, is an admissible extension of $(\mathcal{A}^i,\mathcal{R}^i,\mathcal{D}^i)$.
Suppose $(j-x) \in E'_i$. Then $x \in E_i$. We show x is acceptable w.r.t. E^i:

By definition of Δ and Δ_{MHr}, $\exists(y,x) \in \mathcal{R}^i$ iff $\exists((y\text{def}x),(j-x)) \in \mathcal{R}^{i'}$. Suppose $((y\text{def}x),(j-x)) \in \mathcal{R}^{i'}$. Since $E^{i'}$ is admissible, either:
a) $\exists(j-x') \in E'_{i+1}$ s.t. $(j-x')\mathcal{R}'_{i-\mathcal{D}r}(y\text{def}x)$, and so $(x',(y,x)) \in \mathcal{D}_i$. By R2 and *IH*, $x' \in E_{i+1}$, and so $y \nrightarrow^{E^i} x$,
or;
b) $\exists(r-y) \in E'_i$ s.t. $(r-y)\mathcal{R}^{i'}(y\text{def}x)$, and since $(j-y)\mathcal{R}^{i'}(r-y)$, $\exists(z\text{def}y) \in E'_i$ s.t. $(z\text{def}y)\mathcal{R}^{i'}(j-y)$ (and so $z\mathcal{R}^i y$) and since $(r-z)\mathcal{R}^{i'}(z\text{def}y)$, $\exists(j-z) \in E'_i$ s.t. $(j-z)\mathcal{R}^{i'}(z\text{def}y)$. By definition, $z \in E_i$.
Suppose $(j-y'_1)\ldots(j-y'_m) \in \mathcal{A}'_{i+1}$ s.t. for $k=1\ldots m$, $(j-y'_k)\mathcal{R}^{i'}(z\text{def}y)$, in which case for $k=1\ldots m$, $(y'_k,(z,y)) \in \mathcal{D}i$. For each such $(j-y'_k)$, by admissibility of $E^{i'}$, $\exists(z'\text{def}y'_k) \in E'_{i+1}$ s.t. $(z'\text{def}y'_k)\mathcal{R}^{i'}(j-y'_k)$, and since $(r-z')\mathcal{R}^{i'}(z'\text{def}y'_k)$, $\exists(j-z') \in E'_{i+1}$ s.t. $(j-z')\mathcal{R}^{i'}(r-z')$. By R2 and *IH*, $z' \in E_{i+1}$, $z' \rightarrow^{E^{i+1}} y'_k$, and there is a reinstatement set Rs_k for $z' \rightarrow^{E^{i+1}} y'_k$.
Hence, we have $z \in E_i$, $z \rightarrow^{E^i} y$, and there is a reinstatement set $\bigcup_{k=1}^{m} Rs_k \cup \{z \rightarrow^{E^i} y\}$ for $z \rightarrow^{E^i} y$. Hence x is acceptable w.r.t. E^i.

We have shown: **1.1** = the left to right half for s = admissible, and; **1.2** = the right to left half for s = admissible. We define functions h and g s.t.

> For any admissible extension E of Δ, E' = $h(E)$.
> For any admissible extension E' of Δ_{MHr}, E = $g(E')$.

From hereon, we will let Δ' = $(\mathcal{A}',\mathcal{R}')$ denote Δ_{MHr}. We show that:

a) h is monotonically strictly increasing.
Suppose E and by **1.1** the corresponding admissible E' = $h(E)$. Suppose $E \subset F$ and by **1.1** the corresponding admissible F' = $h(F)$. Suppose $\exists x \in E$ s.t. $x \rightarrow^E y$ and there is a reinstatement set for $x \rightarrow^E y$. By lemma 10, $x \rightarrow^F y$ and there is a reinstatement set for $x \rightarrow^F y$, and so it must be that $E' \subset F'$.

b) g is monotonically strictly increasing.
Suppose E' and by **1.2** the corresponding admissible E = $g(E')$. Suppose $E' \subset F'$, where:

> $\forall \alpha \in (F'-E')$, if α is of the form $(j-x)$, or α is of the form $(r-y)$ or $(x\text{def}y)$, then $x \notin E$, respectively $\neg\exists x \in E$ s.t. $x \rightarrow^E y$ and there is a reinstatement set for $x \rightarrow^E y$, since otherwise, by **1.1**, we would have $(j-x) \in E'$, respectively $(r-y)$ or $(x\text{def}y) \in E'$. **(i)**

By **1.2**, let F be the corresponding admissible extension of Δ. If $\exists(j-x) \in (F'-E')$ then $x \in F$, and by **i)**, $E \subset F$. If $\exists(r-y) \in (F'-E')$ or $\exists(x\text{def}y) \in (F'-E')$, then $\exists x \in F$ s.t. $x \rightarrow^F y$ and there is a reinstatement set for $x \rightarrow^F y$. Given **i)**, there are three cases to consider: 1) Suppose $x \in E$ and $x \nrightarrow^E y$. Then given $E \subseteq F$ it cannot be that $x \rightarrow^F y$; 2) Suppose $x \in E$, $x \rightarrow^E y$ and there is no reinstatement set for $x \rightarrow^E y$. But then since there is a reinstatement set for $x \rightarrow^F y$, it must be that $E \subset F$; 3) Suppose $x \notin E$. Then $E \subset F$.

Left to right and right to left half for $s \in$ {complete, grounded, preferred}. Given **a)** and **b)**, the theorem is shown to hold in exactly the same way as in Lemma 3.

Left to right half for s = stable: Suppose E is stable. Hence E is complete [13]. By the

[13] Suppose otherwise. Then $x \notin E$ and x acceptable w.r.t. E. Since E is stable, $\exists y \in E$ s.t. $y \rightarrow^E x$, and by acceptability of x, $\exists z \in E$ s.t. $z \rightarrow^E y$, contradicting Proposition 2 in [38] which states that no two arguments defeat$_E$ each other in a conflict free E.

left to right for s = complete, E' is complete. Suppose $\alpha \notin E'$, α is not $\mathcal{R}'$ attacked by an argument in E'. There are three cases to consider:
a) α is some $(j-x) \notin E'$. Then $x \notin E$, $\exists y \in E$ s.t. $y \rightarrow^E x$. Notice that there must be a reinstatement set for $y \rightarrow^E x$, since to suppose otherwise means that for some $y' \in E, x' \notin E$ s.t. $y' \rightarrow^E x'$, then $\exists(x'', (y', x')) \in \mathcal{D}$ s.t. $x'' \notin E$, and $\neg\exists y'' \in E$ s.t. $y'' \rightarrow^E x''$, contradicting E is stable. Hence, $(y\text{def}x) \in E'$, where $(y\text{def}x)\mathcal{R}'(j-x)$.
b) Suppose some $(r-x) \notin E'$, and so given $(j-x)\mathcal{R}'(r-x)$, $(j-x) \notin E'$. But then we have shown in a) that $(y\text{def}x) \in E'$, where $(y\text{def}x)\mathcal{R}'(j-x)$, and so $(r-x)$ is acceptable w.r.t. E', contradicting E' is complete.
c) Suppose some $(y\text{def}x) \notin E'$, and so given $(r-y)\mathcal{R}'(y\text{def}x)$, $(r-y) \notin E'$. But then we have shown in b) that $(j-y) \in E'$, $(j-y)\mathcal{R}'(r-y)$. Suppose some $(j-x')$ s.t. $(j-x')\mathcal{R}'(y\text{def}x)$, $(j-x') \notin E'$. We have shown in a) that $(y'\text{def}x') \in E'$, where $(y'\text{def}x')\mathcal{R}'(j-x')$. Hence $(y\text{def}x)$ is acceptable w.r.t. E', contradicting E' is complete.

Right to left half for s = stable: Suppose E' is a stable extension. By the right to left for s = complete, E is complete. Suppose some $x \notin E$. Then $(j-x) \notin E'$, $\exists(y\text{def}x) \in E'$ s.t. $(y\text{def}x)\mathcal{R}'(j-x)$, and by the right to left half for s = complete, $y \in E$, $y \rightarrow^E x$.

References

[1] L. Amgoud and C. Cayrol. A reasoning model based on the production of acceptable arguments. *Annals of Mathematics and Artificial Intelligence*, 34(1-3):197–215, 2002.

[2] L. Amgoud, C. Cayrol, M. Lagasquie-Schiex, and P. Livet. On bipolarity in argumentation frameworks. *International Journal of Intelligent Systems*, 23(10):1062–1093, 2008.

[3] L. Amgoud and S. Vesic. Repairing preference-based argumentation frameworks. In *Proceedings of the 21st international jont conference on Artifical intelligence*, pages 665–670, 2009.

[4] K. M. Atkinson, T. J. M. Bench-Capon, and P. McBurney. Computational representation of practical argument. *Synthese*, 152(2):157–206, 2006.

[5] Katie Atkinson and Trevor J. M. Bench-Capon. Action-based alternating transition systems for arguments about action. In *Proc. Twenty-Second AAAI Conference on Artificial Intelligence (AAAI'07)*, pages 24–29, 2007.

[6] P. Baroni, F. Cerutti, M. Giacomin, and G. Guida. Encompassing attacks to attacks in abstract argumentation frameworks. In *Proc. 10th European Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty*, pages 83–94, 2009.

[7] P. Baroni, F. Cerutti, M. Giacomin, and G. Guida. Afra: Argumentation framework with recursive attacks. *International Journal of Approximate Reasoning*, to appear.

[8] P. Baroni and M. Giacomin. Resolution-based argumentation semantics. In *Computational Models of Argument: Proceedings of COMMA 2008*, pages 25–36, 2008.

[9] H. Barringer, D. M. Gabbay, and J. Woods. Temporal dynamics of support and attack networks: From argumentation to zoology. *Mechanizing Mathematical Reasoning*, pages 59–98, 2005.

[10] T. J. M. Bench-Capon. Agreeing to differ: modelling persuasive dialogue between parties with different values. *Informal Logic*, 22(3):231–245, 2003.

[11] T. J. M. Bench-Capon. Persuasion in practical argument using value-based argumentation frameworks. *Journal of Logic and Computation*, 13(3):429–448, 2003.

[12] T. J. M. Bench-Capon, S. Doutre, and P. E. Dunne. Audiences in argumentation frameworks. *Artificial Intelligence*, 171(1):42–71, 2007.

[13] T. J. M. Bench-Capon and P. E. Dunne. Argumentation in artificial intelligence. *Artificial Intelligence*, 171:10–15, 2007.

[14] T. J. M. Bench-Capon and S. Modgil. Case law in extended argumentation frameworks. In *Proc. 12th International Conference on Artificial Intelligence in Law (ICAIL'09)*, pages 118–127, 2009.

[15] T.J.M. Bench-Capon. Representation of case law as an argumentation framework. In *Legal Knowledge and Information Systems. Proceedings of JURIX 2002*, pages 103–112, 2002.

[16] A. Bochman. Collective argumentation and disjunctive programming. *Journal of Logic and Computation*, 13 (3):405–428, 2003.

[17] G. Boella, D. M. Gabbay, L. van der Torre, and S. Villata. Meta-argumentation modelling 1: Methodology and techniques. *Studia Logica*, 93(2-3):297–355, December 2009.

[18] G. Boella, J. Hulstijn, and L. van der Torre. A logic of abstract argumentation. In *Proc. 2nd International Workshop on Argumentation in Multi-Agent Systems*, pages 29–41, 2005.

[19] G. Boella, L. van der Torre, and S. Villata. Social viewpoints for arguing about coalitions. In *Proc. 11th Pacific Rim International Conference on Multi-agents*, pages 66–77, 2008.

[20] G. Boella, L. van der Torre, and S. Villata. On the acceptability of meta-arguments. In *Proc. IEEE/WIC/ACM International Conference on Intelligent Agent Technology*, pages 259–262, 2009.

[21] A. Bondarenko, P.M. Dung, R.A. Kowalski, and F. Toni. An abstract, argumentation-theoretic approach to default reasoning. *Artificial Intelligence*, 93:63–101, 1997.

[22] M. Caminada. On the issue of reinstatement in argumentation. In *10th European Conference on Logic in Artificial Intelligence (JELIA)*, pages 111–123, 2006.

[23] M. Caminada. An algorithm for computing semi-stable semantics. In *9th European Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty (ECSQARU'09)*, pages 222–234, 2007.

[24] C. Cayrol, S. Doutre, and J. Mengin. On Decision Problems related to the preferred semantics for argumentation frameworks. *Journal of Logic and Computation*, 13(3):377–403, 2003.

[25] C. Cayrol and M.-Ch. Lagasquie-Schiex. Coalitions of arguments: A tool for handling bipolar argumentation frameworks. *International Journal of Intelligent Systems*, 25(1):83–109, 2010.

[26] P. M. Dung. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial Intelligence*, 77:321–357, 1995.

[27] P. E. Dunne and T. J. M. Bench-Capon. Two party immediate response dispute: Properties and efficiency. *Artificial Intelligence*, 149:221–250, 2003.

[28] P. E. Dunne, A. Hunter, P. McBurney, S. Parsons, and M. Wooldridge. Inconsistency tolerance in weighted argument systems. In *Proc. 8th Int. Joint Conference on Autonomous Agents and Multiagent Systems*, pages 851–858, 2009.

[29] D. M. Gabbay. Semantics for higher level attacks in extended argumentation frames part 1: Overview. *Studia Logica*, 93(2-3):357–381, December 2009.

[30] G. Governatori and M. J. Maher. An argumentation-theoretic characterization of defeasible logic. In *Proceedings of the Fourteenth European Conference on Artificial Intelligence*, pages 469–473, 2000.

[31] H. Jakobovits and D. Vermeir. Robust semantics for argumentation frameworks. *Journal of logic and computation*, 9(2):215–261, 1999.

[32] S. Kaci and L. van der Torre. Preference-based argumentation: Arguments supporting multiple values. *International Journal of Approximate Reasoning*, 48(3):730–751, 2008.

[33] S. Modgil. Hierarchical argumentation. In *Proc. 10th European Conference on Logics in Artificial Intelligence (JELIA)*, pages 319–332, 2006.

[34] S. Modgil. Value based argumentation in hierarchical argumentation frameworks. In *Computational Models of Argument: Proceedings of COMMA 2006*, pages 297–308, 2006.

[35] S. Modgil. An abstract theory of argumentation that accommodates defeasible reasoning about preferences. In *Proc. 9th European Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty*, pages 648–659, 2007.

[36] S. Modgil. An argumentation based semantics for agent reasoning. In *Proc. Workshop on Languages, methodologies and development tools for multi-agent systems (LADS 07)*, pages 37–53, 2007.

[37] S. Modgil. Labellings and games for extended argumentation frameworks. In *Twenty-first International Joint Conference on Artificial Intelligence (IJCAI-09)*, pages 873–878, 2009.

[38] S. Modgil. Reasoning about preferences in argumentation frameworks. *Artificial Intelligence*, 173(9-10):901–934, 2009.

[39] S. Modgil and T. J. M. Bench-Capon. Integrating object and meta-level value based argumentation. In *Computational Models of Argument: Proceedings of COMMA 2008*, pages 240–251, 2008.

[40] S. Modgil and T. J. M. Bench-Capon. Integrating dialectical and accrual modes of argumentation. In *3rd International Conference on Computational Models of Argument (COMMA 2010)*, 2010 (to appear).

[41] S. Modgil and T.J.M Bench-Capon. Metalevel argumentation. *Technical Report*, page www.csc.liv.ac.uk/research/ techreports/techreports.html, September, 2009.

[42] S. Modgil and M. Caminada. Proof theories and algorithms for abstract argumentation frameworks. In I. Rahwan and G. Simari, editors, *Argumentation in AI*, pages 105–129. Springer-Verlag, 2009.

[43] S. Modgil and M. Luck. Argumentation based resolution of conflicts between desires and normative goals. In *Proc. 5th Int. Workshop on Argumentation in Multi-Agent Systems*, pages 252–263, 2008.

[44] S. H. Nielsen and S. Parsons. A generalization of dung's abstract framework for argumentation: Arguing with sets of attacking arguments. In *Proc. 3rd Int. Workshop on Argumentation in Multi-agent Systems*, pages 54–73, 2006.

[45] N. Oren and T. J. Norman. Semantics for evidence-based argumentation. In *Computational Models of Argument: Proceedings of COMMA 2008*, pages 276–284, 2008.

[46] H. Prakken. A study of accrual of arguments, with applications to evidential reasoning. In *ICAIL '05: Proc. 10th Int. Conf. on Artificial intelligence and law*, pages 85–94, 2005.

[47] I. Rahwan and G. Simari, editors. *Argumentation in AI*. Springer-Verlag, 2009.

[48] J.R. Searle. *Rationality in Action*. MIT Press, Cambridge, MA, 2001.

[49] B. Verheij. Two approaches to dialectical argumentation: Admissible sets and argumentation stages. In *Proc. of the biannual International Conference on Formal and Applied Practical Reasoning (FAPR) workshop*, pages 357–368, 1996.

[50] B. Verheij. A labeling approach to the computation of credulous acceptance in argumentation. In *Proc. 12th International Joint Conference on Artificial Intelligence*, pages 623–628, 2007.

[51] G. Vreeswijk. An algorithm to compute minimally grounded and admissible defence sets in argument systems. In *Computational Models of Argument: Proceedings of COMMA 2006*, pages 109–120, UK, 2006.

[52] G. A. W. Vreeswijk and H. Prakken. Credulous and sceptical argument games for preferred semantics. In *Proc. 7th European Workshop on Logic for Artificial Intelligence*, pages 239–253, 2000.

[53] M. Wooldridge, P. McBurney, and S. Parsons. On the meta-logic of arguments. In *Proc. Fourth international joint conference on Autonomous agents and multiagent systems*, pages 560–567, 2005.