

# Integrating Object and Meta-Level Value Based Argumentation

Sanjay Modgil<sup>a</sup> Trevor Bench-Capon<sup>b</sup>

<sup>a</sup>*Department of Computer Science, Kings College London*

<sup>b</sup>*Department of Computer Science, University of Liverpool*

**Abstract.** A recent extension to Dung’s argumentation framework allows for arguments to express preferences between other arguments. Value based argumentation can be formalised in this extended framework, enabling meta-level argumentation about the values that arguments promote, and the orderings on these values. In this paper, we show how extended frameworks integrating meta-level reasoning about values can be rewritten as Dung frameworks, and show a soundness and completeness result with respect to the rewrites. We then describe how value orderings can emerge, or be ‘formed’, as a result of dialogue games based on the rewritten frameworks, and illustrate the advantages of this approach over existing dialogue games for value based argumentation frameworks.

## 1. Introduction

A Dung argumentation framework [5] consists of a set of arguments related by a binary conflict based attack relation. A ‘calculus of opposition’ is then applied to determine the sets of acceptable arguments under different extensional semantics. The framework abstracts from the underlying logic in which the arguments and attack relation are defined. Dung’s theory has thus become established as a general framework for various species of non-monotonic reasoning, and, more generally, reasoning in the presence of conflict.

The extensional semantics may yield multiple sets of acceptable arguments (extensions). The sceptically justified arguments are those that appear in every extension. However, one may then be faced with the problem of how to choose between arguments that belong to at least one, but not all extensions (the credulously justified arguments), when they conflict. One solution is to provide some means for preferring one argument to another so that one can then determine whether attacks succeed and defeat the arguments they attack. For example, Value Based Argumentation Frameworks (VAFs) [2] associate each argument with a social value which it promotes, and this property determines the strength of arguments by reference to an ordering on these social values. Given such a preference ordering, one obtains a defeat relation with respect to that ordering. It has been shown that if a framework does not contain cycles comprising only arguments of equal strength, then on the basis of the defined defeat relation obtained with respect to a given ordering, one can obtain a unique, non-empty extension under the *preferred* semantics.

However, in general, preference information is often itself defeasible, conflicting and so may itself be subject to argumentation based reasoning. Hence, Dung’s framework has

recently been extended [7] to include arguments that claim preferences between other arguments. Specifically, the framework is extended to include a second attack relation such that an argument expressing a preference between two other arguments, attacks the binary attack between these two conflicting arguments, thus determining which attacks succeed as defeats. Arguments expressing contradictory preferences attack each other, and one can then argue over which of these ‘preference arguments’ is preferred to, and so defeats, the other. The justified arguments of an *Extended Argumentation Framework (EAF)* can then be evaluated under the full range of Dung’s extensional semantics. Examples of value based argumentation in the extended semantics have informally been described in [7], whereby different value orderings may yield contradictory preferences, requiring meta-level reasoning *about* values and value orderings to determine a unique set of justified arguments.

In this paper the extended semantics are reviewed in section 2. Sections 3 and 4 then describe the main contributions of this paper:

- 1) We formalise *EAFs* integrating meta-level reasoning about values and value orderings, and then show that such *EAFs* can be rewritten as Dung argumentation frameworks. We show a soundness and completeness result for the rewrite.
- 2) Given 1), we can then exploit results and techniques applied to Dung argumentation frameworks. In particular, we show how value orderings can emerge from dialogue games based on the above rewrites, and demonstrate the advantages of this approach over games proposed specifically for *VAFs* [3].

## 2. Extended Argumentation Frameworks

A Dung argumentation framework (*AF*) [5] is of the form  $(Args, \mathcal{R})$  where  $\mathcal{R} \subseteq (Args \times Args)$  can denote either attack or defeat. An argument  $A \in Args$  is defined as acceptable w.r.t. some  $S \subseteq Args$ , if for every  $B$  such that  $(B, A) \in \mathcal{R}$ , there exists a  $C \in S$  such that  $(C, B) \in \mathcal{R}$ . Intuitively,  $C$  ‘reinstates’  $A$ . In [5], the acceptability of a **set** of arguments under different extensional semantics is then defined. The definition of admissible and preferred semantics are given here, in which  $S \subseteq Args$  is conflict free if no two arguments in  $S$  are related by  $\mathcal{R}$ .

**Definition 1** Let  $S \subseteq Args$  be a conflict free set.

- $S$  is admissible iff each argument in  $S$  is acceptable w.r.t.  $S$
- $S$  is a preferred extension iff  $S$  is a set inclusion maximal admissible extension

From hereon, an argument is said to be credulously, respectively sceptically, justified, iff it belongs to at least one, respectively all, preferred extensions. We now present the extended argumentation semantics described in [7]. By way of motivation, consider two individuals **P** and **O** exchanging arguments  $A, B \dots$  about the weather forecast:

**P** : “Today will be dry in London since the BBC forecast sunshine” =  $A$

**O** : “Today will be wet in London since CNN forecast rain” =  $B$

**P** : “But the BBC are more trustworthy than CNN” =  $C$

**O** : “However, statistics show that CNN are more accurate than the BBC” =  $C'$

**O** : “And basing a comparison on statistics is more rigorous and rational than basing a comparison on your instincts about their relative trustworthiness” =  $E$

Arguments  $A$  and  $B$  symmetrically attack, i.e.,  $(A, B), (B, A) \in \mathcal{R}$ .  $\{A\}$  and  $\{B\}$  are admissible. We then have argument  $C$  claiming that  $A$  is preferred to  $B$ . Hence  $B$  does not successfully attack (defeat)  $A$ , but  $A$  does defeat  $B$ . Intuitively,  $C$  is an argument for

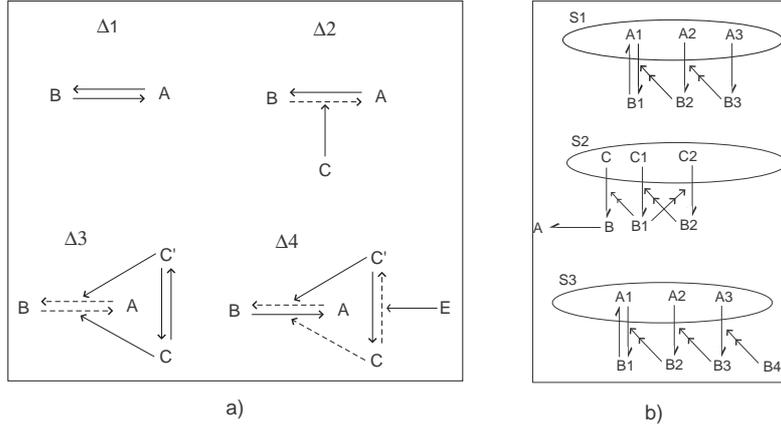


Figure 1.

$A$ 's repulsion of, or defence against,  $B$ 's attack on  $A$ , i.e.,  $C$  defence attacks  $B$ 's attack on  $A$  ( $\Delta 2$  in figure 1a)) so that  $B$ 's attack on  $A$  does not succeed as a defeat.  $B$ 's attack on  $A$  is 'cancelled out', and we are left with  $A$  defeating  $B$ . Only  $\{A\}$  is now admissible. Of course, given  $C'$  claiming a preference for  $B$  over  $A$  and so defence ( $d$ ) attacking  $A$ 's attack on  $B$ , then we will have that  $\{A\}$  and  $\{B\}$  are now both admissible, since neither defeats the other.  $C$  and  $C'$  claim contradictory preferences and so attack each other ( $\Delta 3$  in figure 1a)). These attacks can themselves be subject to  $d$  attacks in order to determine the defeat relation between  $C$  and  $C'$  and so  $A$  and  $B$ .  $E$   $d$  attacks the attack from  $C$  to  $C'$  ( $\Delta 4$  in figure 1a)), and so determines that  $C'$  defeats  $C$ ,  $B$  defeats  $A$ , and the discussion concludes in favour of  $O$ 's argument that it will be a wet day in London. We now formally define the elements of an *Extended Argumentation Framework*, and the defeat relation that is now parameterised w.r.t. some set  $S$  of arguments.

**Definition 2** An *Extended Argumentation Framework (EAF)* is a tuple  $(Args, \mathcal{R}, \mathcal{D})$  such that  $Args$  is a set of arguments, and:

- $\mathcal{R} \subseteq Args \times Args$
- $\mathcal{D} \subseteq (Args \times \mathcal{R})$
- If  $(C, (A, B)), (C', (B, A)) \in \mathcal{D}$  then  $(C, C'), (C', C) \in \mathcal{R}$

**Notation 1** We may write  $A \rightarrow B$  to denote  $(A, B) \in \mathcal{R}$ . If in addition  $(B, A) \in \mathcal{R}$ , we may write  $A \rightleftharpoons B$ . We may also write  $C \rightarrow (A \rightarrow B)$  to denote  $(C, (A, B)) \in \mathcal{D}$

**Definition 3**  $A$  defeats <sub>$S$</sub>   $B$ , denoted by  $A \rightarrow^S B$ , iff  $(A, B) \in \mathcal{R}$  and  $\neg \exists C \in S$  s.t.  $(C, (A, B)) \in \mathcal{D}$ .

Referring to the weather forecast example,  $A$  defeats <sub>$\emptyset$</sub>   $B$  but  $A$  does not defeat <sub>$\{C'\}$</sub>   $B$  ( $A \not\rightarrow^{\{C'\}} B$ ). The notion of a conflict free set  $S$  of arguments is now defined so as to

account for the case where an argument  $A$  asymmetrically attacks  $B$ , but given a preference for  $B$  over  $A$ , both may appear in a conflict free set and hence an extension (as in the case of value based argumentation).

**Definition 4**  $S$  is conflict free iff  $\forall A, B \in S$ : if  $(B, A) \in \mathcal{R}$  then  $(A, B) \notin \mathcal{R}$ , and  $\exists C \in S$  s.t.  $(C, (B, A)) \in \mathcal{D}$ .

We now define the acceptability of an argument  $A$  w.r.t. a set  $S$  for an *EAF*. The definition is motivated in more detail in [7] and relates to an intuitive requirement (captured by Dungs fundamental lemma in [5]) on what it means for an argument to be acceptable w.r.t. an admissible set  $S$  of arguments: *if  $A$  is acceptable with respect to  $S$ , then  $S \cup \{A\}$  is admissible*. To ensure satisfaction of this requirement, acceptability for *EAFs* requires the notion of a *reinstatement set* for a defeat.

**Definition 5** Let  $S \subseteq \text{Args}$  in  $(\text{Args}, \mathcal{R}, \mathcal{D})$ . Let  $R_S = \{X_1 \rightarrow^S Y_1, \dots, X_n \rightarrow^S Y_n\}$  where for  $i = 1 \dots n$ ,  $X_i \in S$ . Then  $R_S$  is a reinstatement set for  $C \rightarrow^S B$ , iff:

- $C \rightarrow^S B \in R_S$ , 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 6**  $A$  is acceptable w.r.t.  $S \subseteq \text{Args}$  iff  $\forall B$  s.t.  $B \rightarrow^S A, \exists C \in S$  s.t.  $C \rightarrow^S B$  and there is a *reinstatement set* for  $C \rightarrow^S B$ .

In figure1b),  $A1$  is acceptable w.r.t.  $S1$ . We have  $B1 \rightarrow^{S1} A1$  and  $A1 \rightarrow^{S1} B1$ . The latter is based on an attack that is  $d$ -attacked by  $B2$ . However,  $A2 \rightarrow^{S1} B2$ , which in turn is challenged by  $B3$ . But then,  $A3 \rightarrow^{S1} B3$ . We have the reinstatement set  $\{A1 \rightarrow^{S1} B1, A2 \rightarrow^{S1} B2, A3 \rightarrow^{S1} B3\}$  for  $A1 \rightarrow^{S1} B1$ . Note that  $A$  is acceptable w.r.t.  $S2$  given the reinstatement set  $\{C \rightarrow^{S2} B, C1 \rightarrow^{S2} B1, C2 \rightarrow^{S2} B2\}$  for  $C \rightarrow^{S2} B$ . Finally  $A1$  is not acceptable w.r.t  $S3$  since no argument in  $S3$  defeats $_{S3}$   $B4$ .

Admissible and preferred semantics for *EAFs* are now given by definition 1, where conflict free is defined as in definition 4. (Dung's definition of complete, stable and grounded semantics also apply to *EAFs* [7]). In our weather example,  $\{B, D, E\}$  is the single preferred extension. In [7] we show that *EAFs* inherit many of the fundamental results holding for extensions of a Dung framework. In particular: a) If  $S$  is admissible and arguments  $A$  and  $A'$  are acceptable w.r.t.  $S$ , then  $S \cup \{A\}$  is admissible and  $A'$  is acceptable w.r.t.  $S \cup \{A\}$ ; b) the set of all admissible extensions of an *EAF* form a complete partial order w.r.t. set inclusion; c) for each admissible  $S$  there exists a preferred extension  $S'$  such that  $S \subseteq S'$ ; d) Every *EAF* possesses at least one preferred extension.

### 3. Value Based Argumentation in Extended Argumentation Frameworks

In this section we show how meta-level argumentation about values and value orderings can be captured in a special class of *EAFs*. We then show that these *EAFs* can be rewritten as Dung argumentation frameworks, and go on to show a soundness and completeness result with the original *EAFs*.

**Definition 7** A value-based argumentation framework (*VAF*) is a 5-tuple  $\langle \text{Args}, \mathcal{R}, V, val, P \rangle$  where  $val$  is a function from  $\text{Args}$  to a non-empty set of values  $V$ , and  $P$  is a set  $\{a_1, \dots, a_n\}$ , where each  $a_i$  names a total ordering (audience)  $>_a$  on  $V \times V$ .

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

Given an  $aVAF$   $\Gamma = \langle Args, \mathcal{R}, V, val, a \rangle$ , one can then say that  $A \in Args$  defeats <sub>$a$</sub>   $B \in Args$ , if  $(A, B) \in \mathcal{R}$  and it is not the case that  $val(B) >_a val(A)$ . Letting  $\mathcal{R}_a$  denote the binary relation defeats <sub>$a$</sub> , then the extensions and justified arguments of  $\Gamma$  are now those of the framework  $(Args, \mathcal{R}_a)$  (as defined in definition 1. If for every  $(A, B) \in \mathcal{R}$  either  $val(A) >_a val(B)$  or  $val(B) >_a val(A)$ , and assuming no cycles in the same value in  $\Gamma$ , then there is guaranteed to be a unique, non-empty preferred extension of  $\Gamma$ , and a polynomial time algorithm to find it [2].

Pairwise orderings on values can be interpreted as *value preference arguments* in an *EAF*. Consider the mutually attacking arguments  $A$  and  $B$ , respectively promoting values  $v_1$  and  $v_2$ . Then  $v_1 > v_2$  can be interpreted as a *value preference argument* expressing that  $A$  is preferred to  $B$  -  $(v_1 > v_2) \rightarrow (B \dashv A)$  - and  $v_2 > v_1$  as expressing the contrary preference (see figure 2a). The choice of audience can then be expressed as an *audience argument* that attacks the attacks between the value preference arguments. Suppose the argument  $v_1|v_2$  denoting the chosen audience that orders  $v_1 > v_2$ . Then  $v_1|v_2 \rightarrow ((v_2 > v_1) \dashv (v_1 > v_2))$ . Now the unique preferred extension of the *EAF* in figure 2a) is  $\{A, v_1 > v_2, v_1|v_2\}$ . In this way, we can represent the meta-level reasoning required to find the preferred extension of an  $aVAF$ .

**Definition 8** Let  $\Gamma$  be an  $aVAF$   $\langle Args, \mathcal{R}, V, val, a \rangle$ . Then the *EAF*  $\Delta = (Args1 \cup Args2 \cup Args3, \mathcal{R}1 \cup \mathcal{R}2 \cup \mathcal{R}3, \mathcal{D}1 \cup \mathcal{D}2 \cup \mathcal{D}3)$  is defined as follows:

1.  $Args1 = Args, \mathcal{R}1 = \mathcal{R}$
2.  $\{v > v' | v, v' \in V, v \neq v'\} \subseteq Args2^1$ ,  
 $\{(v > v', v' > v) | v > v', v' > v \in Args2\} \subseteq \mathcal{R}2$
3.  $\{a\} \subseteq Args3, \emptyset \subseteq \mathcal{R}3$
4.  $\{(v > v', (A, B)) | (A, B) \in \mathcal{R}1, val(B) = v, val(A) = v'\} \subseteq \mathcal{D}1$   
 $\{(a, (v > v', v' > v)) | a \in Args3, (v > v', v' > v) \in \mathcal{R}2, v' >_a v\} \subseteq \mathcal{D}2$   
 $\mathcal{D}3 = \emptyset$

If in 2,3 and 4, the  $\subseteq$  relation is replaced by  $=$ , then  $\Gamma$  and  $\Delta$  are said to be *equivalent*.

If an  $aVAF$   $\Gamma$  and the defined *EAF*  $\Delta$  are *equivalent*, one can straightforwardly show that for any  $A \in Args$  in  $\Gamma$ ,  $A$  is a sceptically, respectively credulously, justified argument of  $\Gamma$  iff  $A$  is a sceptically, respectively credulously, justified argument of  $\Delta$ .

Notice that  $\Delta$  is defined so that one could additionally consider other arguments and attacks in levels 2 and 3. For example, arguments in level 2 that directly attack *value preference arguments*, or arguments in level 3 representing different audiences. Notice also the hierarchical nature of the defined *EAF*  $\Delta$ . It is stratified into three levels such that binary attacks are between arguments within a given level, and defence attacks originate from arguments in the immediate meta-level. In general then, incorporating meta-level argumentation about values and value orderings can be modelled in hierarchical *EAFs* <sup>2</sup>

<sup>1</sup>Note that this adds an additional  $|V|(|V| - 1)$  arguments: this, however, is acceptable, since it is only polynomial in the number of values.

<sup>2</sup>See [7] for examples illustrating requirements for *EAFs* that do not 'stratify' the argumentation in this way.

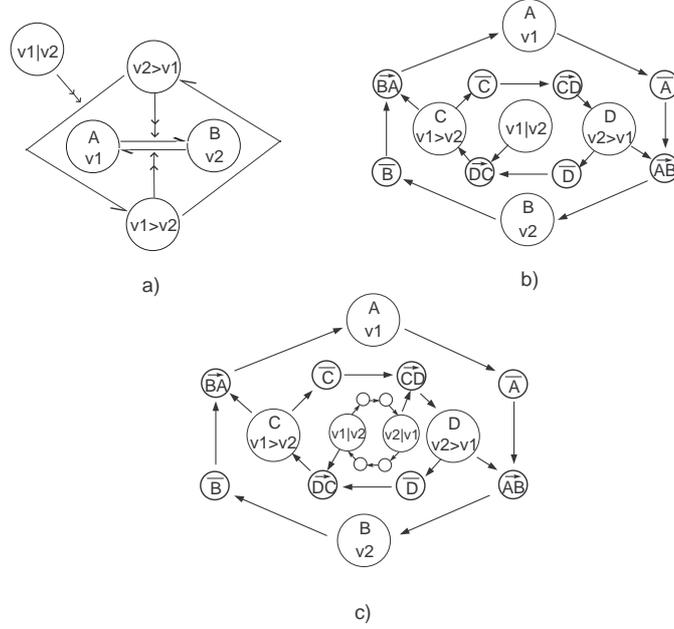


Figure 2.

**Definition 9**  $\Delta = (Args, \mathcal{R}, \mathcal{D})$  is a hierarchical *EAF* iff there exists a partition  $\Delta_H = ((Args_1, \mathcal{R}_1), \mathcal{D}_1), \dots, ((Args_n, \mathcal{R}_n), \mathcal{D}_n)$  such that:

- $Args = \bigcup_{i=1}^n Args_i$ ,  $\mathcal{R} = \bigcup_{i=1}^n \mathcal{R}_i$ ,  $\mathcal{D} = \bigcup_{i=1}^n \mathcal{D}_i$ , and for  $i = 1 \dots n$ ,  $(Args_i, \mathcal{R}_i)$  is a Dung argumentation framework.
- $\mathcal{D}_n = \emptyset$ , and for  $i = 1 \dots n - 1$ ,  $(C, (A, B)) \in \mathcal{D}_i$  implies  $(A, B) \in \mathcal{R}_i$ ,  $C \in Args_{i+1}$

We now show that it is possible to rewrite a hierarchical *EAF*  $\Delta$  as a Dung argumentation framework  $AF_\Delta$ , such that the preferred extensions of  $\Delta$  and  $AF_\Delta$  are equivalent modulo the additional arguments included in the rewrite. Firstly, we define an *expansion* of a Dung framework in which each attack  $(A, B)$  is replaced by a set of attacks  $\{(A, \bar{A}), (\bar{A}, \overrightarrow{AB}), (\overrightarrow{AB}, B)\}$ , and the additional arguments  $\bar{A}$  and  $\overrightarrow{AB}$  are included. Intuitively,  $\bar{A}$  stands for “ $A$  is not acceptable”, and  $\overrightarrow{AB}$  stands for “ $A$  defeats  $B$ ”.

**Definition 10** Let  $AF = (Args, \mathcal{R})$  be a Dung argumentation framework. Then the *expansion* of  $AF$  is the framework  $AF' = (Args', \mathcal{R}')$ , where:

- $\mathcal{R}' = \bigcup_{(X,Y) \in \mathcal{R}} \{(X, \bar{X}), (\bar{X}, \overrightarrow{XY}), (\overrightarrow{XY}, Y)\}$
- $Args' = Args \cup \bigcup_{(X,Y) \in \mathcal{R}} \{\bar{X}, \overrightarrow{XY}\}$

Proposition 1 below follows immediately from lemma 1 in the Appendix.

**Proposition 1** Let  $AF' = (Args', \mathcal{R}')$  be the expansion of  $AF = (Args, \mathcal{R})$ . Then  $A \in Args$  is a sceptically, respectively credulously, justified argument of  $AF$  iff  $A$  is a sceptically, respectively credulously, justified argument of  $AF'$ .

We now formally define the rewrite of an *EAF* as a Dung argumentation framework:

**Definition 11** Let  $\Delta = (Args, \mathcal{R}, \mathcal{D})$ . Let  $(Args', \mathcal{R}')$  be the expansion of  $(Args, \mathcal{R})$ . Then  $AF_\Delta = (Args_\Delta, \mathcal{R}_\Delta)$  where:

- $Args_\Delta = Args'$
- $\mathcal{R}_\Delta = \mathcal{R}' \cup \{ (C, \overrightarrow{AB}) \mid (C, (A, B)) \in \mathcal{D} \}$ .

Figure 2b) shows the rewrite of the *EAF* in figure 2a). The single preferred extension of the *EAF* is  $\{A, v1 > v2, v1|v2\}$ , and (where  $C = v1 > v2, D = v2 > v1$ ) the single preferred extension of the rewrite is  $\{A, \overrightarrow{AB}, \overrightarrow{B}, C, \overrightarrow{CD}, \overrightarrow{D}, v1|v2\}$ . Theorem 1 follows immediately from lemma 2 in the Appendix.

**Theorem 1** Let  $\Delta = (Args, \mathcal{R}, \mathcal{D})$  be a hierarchical *EAF*.  $A \in Args$  is a sceptically, respectively credulously, justified argument of  $\Delta$ , iff  $A$  is a sceptically, respectively credulously, justified argument of the rewrite  $AF_\Delta$ .

As mentioned earlier, one might additionally include more than one audience argument in level 3 of an *EAF*. Given a *VAF*  $\langle Args, \mathcal{R}, V, val, P \rangle$ , then its *EAF* is obtained as in definition 8, except that now  $\{a|a \in P\} \subseteq Args_3$  (recall that  $P$  is the set of all possible audiences). If  $\{a|a \in P\} = Args_3$ , then we say that the *VAF* and its obtained *EAF* are *equivalent*. Notice that if for any  $a, a' \in P$ ,  $(a, (v > v', v' > v))$ ,  $(a', (v' > v, v > v')) \in \mathcal{D}_2$ , then it follows from the definition of an *EAF* (definition 2) that  $a$  and  $a'$  attack each other, i.e.  $(a, a'), (a', a) \in \mathcal{R}_3$ . Figure 2c) shows the rewrite of such an *EAF* as an *AF*, extending the example in figure 2b) to include the alternative audience choice.

Recall that each possible audience argument corresponds to a different total orderings on the values. Also,  $\forall v, v' \in V, v > v'$  and  $v' > v$  are value preference arguments in  $Args_2$ . Hence, every audience argument will attack every other audience argument. Moreover, each audience argument will give rise to a corresponding preferred extension (under the assumption of no cycles in the same value), so that we no longer have a unique preferred extension.

None the less, there are some nice properties of the *AF* in figure 2c). In [2], the arguments that appear in the preferred extension for every, respectively at least one, audience, are referred to as *objectively*, respectively *subjectively*, acceptable. One can straight-forwardly show that:

*A is an objectively, respectively subjectively, acceptable argument of a VAF iff A is a sceptically, respectively credulously, justified argument of the AF rewrite of the VAF's equivalent EAF.*

Moreover our task is bounded in that all the preferred extensions depend on a single choice of audience argument. However, if all possible audience arguments are included, then there are  $|V|$  factorial many audience arguments, rendering the augmented Dung graph impractical for even moderately large number of values. None the less, for small values of  $|V|$  there may be utility in presenting the complete picture in the manner shown in figure 2c). Moreover, choice of an audience can be effected through submission of other arguments attacking audience arguments, or indeed ascending to level 4 to argue about preferences between audiences (as described in [7]).

#### 4. Emergence of Value Orderings as a Product of Reasoning

We have thus far assumed that value orderings are available at the outset. However, as Searle [9] points out, this is unrealistic; preference orderings are more often the product of practical reasoning rather than an input to it. This has motivated development of dialogue games [3,1] in which value orderings emerge from the attempts of a proponent to defend an argument, in much the same manner as [4] used a dialogue game to establish the preferred extension of an *AF*. We illustrate use of these games using a three cycle in two values as shown in Figure 3a).

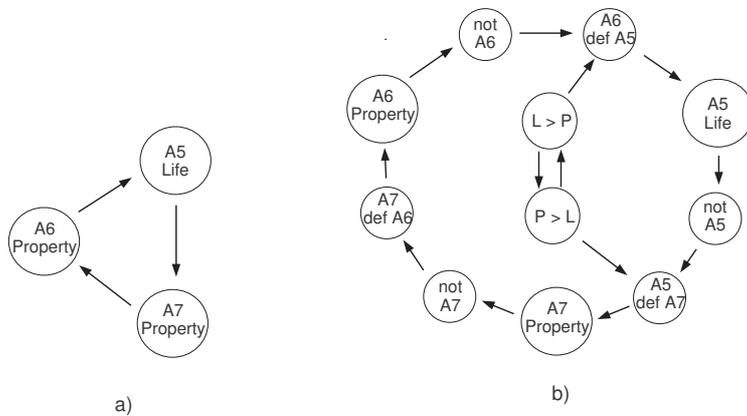


Figure 3.

Consider this first as an *AF*. [4]’s *TPI* game begins with Proponent (Prop) choosing an argument he wishes to defend by showing that it is included in some preferred extension. Suppose Prop chooses *A5*. Since in an *AF* attacks always succeed, this means that any argument attacked by a chosen argument - in this case *A7* - cannot be in the same preferred extension, and so cannot be used by Prop subsequently. The game proceeds by the Opponent (Opp) playing an argument - *A6* - which attacks the argument played by Prop’. Now the game ends with a victory for Opp, because the attacker *A7* of *A6* is not available to Prop. Thus *A5* cannot be incorporated in a preferred extension. This is as it should be: in an *AF* the preferred extension of a three cycle is empty.

In a *VAF*, however, there are two preferred extensions of the three cycle shown in figure 3a), depending on whether Life (*L*) is preferred to Property (*P*) or vice versa. If  $L > P$  the preferred extension is  $\{A5, A6\}$ , and if  $P > L$  it is  $\{A5, A7\}$ . Note that *A5* is objectively accepted. In [1]’s *VTPI* game, *TPI* is augmented to enable Prop to defend an argument by claiming a preference for the value of the attacked argument over its attacker. Thus if Prop chooses *A5*, and Opp plays *A6*, Prop may defend *A5* by claiming  $L > P$ . Under *VTPI*, *A7* is also unavailable to Prop because it is attacked by the originally chosen argument *A5*. There is a major problem with this: since *A5* is objectively acceptable, it should not have been necessary for Prop to commit to a value preference in order to defend it. The objective acceptability of *A5* is lost. This might be remedied by keeping *A7* available. However this requires that Prop, when playing *A5*,

**Table 1.** Game establishing objective acceptance of  $A5$

Ply	Prop Moves	Opp Moves	Prop Commitments	Opp Commitments
1	$A5$	$A6 \text{ def } A5$	$A5$	$A6 \text{ def } A5$
2	$L > P$	$P > L$	$A5, L > P$	$A6 \text{ def } A5, P > L$
3	RETRACT $L > P$ : $\text{not } A6$	$A6$	$A5, \text{not } A6$	$A6 \text{ def } A5, P > L, A6$
4	$A7 \text{ def } A6$	$\text{not } A7$	$A5, \text{not } A6, A7 \text{ def } A6$	$A6 \text{ def } A5, P > L, A6, \text{not } A7$
5	$A7$		$A5, \text{not } A6, A7 \text{ def } A6, A7$	$A6 \text{ def } A5, P > L, A6, \text{not } A7$

declares that  $A7$  is not defeated by  $A5$ , which in turn forces Prop to commit to  $P > L$ . Now Prop can use  $A7$  to defend  $A5$  against  $A6$ , but the damaging commitment has already been made. Moreover this greatly complicates the game, as is shown by the game presented in [3], where a similar move requires a distinction between attacks, successful attacks and definite attacks, in order to keep arguments available for subsequent use. The game in [3] also suffers the problem of being forced to choose a particular value order to defend even an objectively acceptable argument.

We now informally describe a game w.r.t. a *partial* rewrite of a VAF's EAF. The rewrite is partial in the sense that only value preference arguments are included, and attacks between these arguments are not expanded as described in definition 10. Consider now the three cycle partial rewrite in figure 3b) (in which we write  $\text{not } X$  and  $X \text{ def } Y$  instead of  $\overline{X}$  and  $\overline{X} \overline{Y}$ ). Since this is effectively an AF, we can use the TPI game to find a preferred extension. In TPI, a player who cannot defend an argument against an attack can retract the argument if there is another line of defense which does not use it. Now consider Prop's defense of  $A5$ . Choosing  $A5$  now has no effect on  $A7$ : it only excludes the argument  $\text{not } A5$ , which seem intuitively correct. Importantly therefore  $A7$  will remain available for Prop's later use. Prop has a choice of two moves to defend  $A5$ :  $L > P$  and  $\text{not } A6$ . Suppose Prop plays  $L > P$ : although this appears to be a commitment to a value ordering this will only be temporary. Opp has no choice but to attack this with  $P > L$ . Prop cannot defend against this and so must retract and so abandon his commitment to  $L > P$ , and pursue the alternative defence of  $A5$ . Since this retraction is conditional on  $P > L$ , Opp remains committed to  $P > L$ . Now the game continues as shown in Table 1. At Ply 5 Prop will win, since  $A5 \text{ def } A7$  is no longer available to Opp, as Opp is committed to an attacker  $P > L$ . Nor is Opp able to retract the commitment to  $P > L$ , since there is no other way to attack  $L > P$ , i.e., there is no alternative defence available to Opp, so backtracking is not possible. Thus Prop is able to defend  $A5$  without embracing any commitment to value preferences: the only commitment to a preference when the game is complete, is on the part of Opp.

Similar advantages accrue if we consider the defence of a subjectively acceptable argument such as  $A6$ . This is defensible only for audiences with  $L > P$ , so that  $A6$  does not defeat  $A5$ , allowing  $A5$  to defeat  $A7$ . But in VTPI,  $A5$  is unavailable once we have included  $A6$ . However use of TPI for the framework in figure 3b) now gives rise to the game in Table 2. Here, playing  $A6$  does not exclude  $A5$ , but only  $\text{not } A6$ . Prop is forced to commit to  $L > P$ , but this also enables Prop to defend against  $A6 \text{ def } A5$ , whereupon Opp cannot play the already retracted  $P > L$ . Thus  $A6$  is defensible for the audience which prefers  $L$  to  $P$ , although note that Opp is not obliged to agree with this value preference.

Another advantage of the augmented framework is that the explicit representation of value preferences ensures that commitments to value preferences are propagated im-

**Table 2.** Game establishing subjective acceptance of  $A_6$ 

Ply	Prop Moves	Opp Moves	Prop Commitments	Opp Commitments
1	$A_6$	$A_7 \text{ def } A_6$	$A_6$	$A_7 \text{ def } A_6$
2	$\text{not } A_7$	$A_7$	$A_6, \text{not } A_7$	$A_7 \text{ def } A_6, A_7$
3	$A_5 \text{ def } A_7$	$P > L$	$A_6, \text{not } A_7, A_5 \text{ def } A_7$	$A_7 \text{ def } A_6, A_7, P > L$
4	$L > P$	RETRACT $P > L$ : $\text{not } A_5$	$A_6, \text{not } A_7, A_5 \text{ def } A_7$ $L > P$	$A_7 \text{ def } A_6, A_7, \text{not } A_5$
5	$A_5$	$A_6 \text{ def } A_5$	$A_6, \text{not } A_7, A_5 \text{ def } A_7,$ $L > P, A_5$	$A_7 \text{ def } A_6, A_7, \text{not } A_5,$ $A_6 \text{ def } A_5$
6	$L > P$		$A_6, \text{not } A_7, A_5 \text{ def } A_7,$ $L > P, A_5$	$A_7 \text{ def } A_6, A_7, \text{not } A_5,$ $A_6 \text{ def } A_5$

mediately, instead of requiring potentially complicated housekeeping to ensure that the emerging value order is consistent. Because an argument of the form  $L > P$  attacks (and defeats) every argument of the form  $A \text{ def } B$ , making this commitment in respect of one argument anywhere in the framework will protect all other arguments which are rendered admissible by this value preference.

Finally, note that while [3] suggests a defence at the object level should always be attempted before resorting to value preferences, as this has fewer ramifications elsewhere in the framework, this is less clearly the correct strategy to use with *TPI* played over our augmented framework. Thus, in the example in Table 1, it was tactically useful for Prop to first make a defence with a value preference. As this preference is only attacked by the opposite value preference, this forces the opponent to play this preference. Now when Prop retracts the value preference and makes the alternative defence, he is not committed to a particular audience, and can exploit the opponent's commitment to force acceptance of his argument.

## 5. Conclusions and Future Work

We have reviewed an extension of Dung's argumentation framework that enables integration of meta-level reasoning about which arguments should be preferred, and shown how certain useful cases, of which value based argumentation frameworks are an example, can be rewritten as a standard Dung framework. This enables results and techniques that apply to, and have been developed for, standard frameworks to be used directly for frameworks integrating meta-level reasoning about preferences in general, and values in particular. As an illustration of the advantages that accrue, we showed how a dialogue game devised for standard *AFs* can be used to identify the value order under which a particular argument can be defended. Now that arguments committing to a preference are of the same sort as other arguments in the framework, the problems arising from the need to give special treatment to these commitments in existing games can be avoided.

Future work will explore how our extended framework can assist in handling arguments which promote values to different degrees, previously treated in [6] and arguments which promote multiple values [8]. In both cases this will provide an extra source of attacks on arguments of the form 'A defeats B'. For example, in current *VAFs* an attack by an argument promoting the same value always succeeds. However, allowing for different degrees of promotion will offer the possibility of the attacked argument defending itself by promoting the value to a greater degree than its attacker.

## 6. Appendix

**Lemma 1** Let  $AF' = (Args', \mathcal{R}')$  be the expansion of  $AF = (Args, \mathcal{R})$ . Then  $S \subseteq Args$  is an admissible extension of  $AF$  iff  $T \subseteq Args'$  is a admissible extension of  $AF'$ , where  $T = S \cup \{\overrightarrow{X} \mid Y \in S, (X, Y) \in \mathcal{R}\} \cup \{\overrightarrow{Y\overline{Z}} \mid Y \in S, (Y, Z) \in \mathcal{R}\}$ .

**Proof:**

*Left to Right half:* Suppose some  $A \in S, (B, A) \in \mathcal{R}$ . By definition of  $AF'$ :  $\overrightarrow{BA}$   
 $\forall A \in Args, (B, A) \in \mathcal{R}$  iff  $(X, A) \in \mathcal{R}'$ , where  $X = \overrightarrow{BA}$  and  $(\overrightarrow{B}, \overrightarrow{BA}) \in \mathcal{R}'$  (1)

By definition of  $T, \overrightarrow{B} \in S$ . Hence:

$\forall A \in Args, A \in S$  is acceptable w.r.t.  $S$  implies  $A \in T$  is acceptable w.r.t.  $T$  (2)

We show that  $\overrightarrow{B}$  is acceptable w.r.t.  $T$ . By assumption of  $A$  acceptable w.r.t.  $S, \exists C \in S, (C, B) \in \mathcal{R}$ . By definition of  $T: C, \overrightarrow{CB} \in T$ . By definition of  $AF', (X, \overrightarrow{B}) \in \mathcal{R}'$  implies  $X = B$ , and  $(\overrightarrow{CB}, B) \in \mathcal{R}'$ . Hence  $\overrightarrow{B}$  is acceptable w.r.t.  $T$ .

We show that  $\forall A \in Args$  s.t.  $A \in T$ , if  $(A, C) \in \mathcal{R}$ , then  $\overrightarrow{AC}$  is acceptable w.r.t.  $T$ . This follows from the definition of  $AF'$ , where  $(X, \overrightarrow{AC}) \in \mathcal{R}'$  implies  $X = A$ , and  $(A, \overline{A}) \in \mathcal{R}'$ .

*Right to Left half:* For any  $A \in Args, A \in T$  we need to show that  $A$  is acceptable w.r.t.  $S$ . Suppose  $(X, A) \in \mathcal{R}'$ . By (1),  $X$  is some  $\overrightarrow{BA}$  s.t.  $(B, A) \in \mathcal{R}, (\overrightarrow{B}, \overrightarrow{BA}) \in \mathcal{R}'$ , and by definition of  $T, \overrightarrow{B} \in T$ .

By definition of  $AF'$ , if  $(X, \overrightarrow{B}) \in \mathcal{R}'$  then  $X = B$ . Since  $\overrightarrow{B} \in T$  and  $T$  is admissible,  $\exists \overrightarrow{CB} \in T$  s.t.  $(\overrightarrow{CB}, B) \in \mathcal{R}'$ , and so  $(C, B) \in \mathcal{R}$ .

If  $(X, \overrightarrow{CB}) \in \mathcal{R}'$ , then  $X = \overline{C}$ . If  $(Y, \overline{C}) \in \mathcal{R}'$  then  $Y = C$ , where  $C \in Args$ . Hence, since  $\overrightarrow{CB}$  is acceptable w.r.t.  $T$ , it must be that  $C \in T$ . By assumption,  $C \in S$ . Hence  $A$  acceptable w.r.t.  $S$ .

Proof of lemma 2 makes use of the following partition of a hierarchical  $EAF$ 's rewrite:

**Definition 12** Let  $\Delta_H = ((Args_1, \mathcal{R}_1), \mathcal{D}_1), \dots, ((Args_n, \mathcal{R}_n), \mathcal{D}_n)$  be the partition of the hierarchical  $EAF \Delta = (Args, \mathcal{R}, \mathcal{D})$ . Let  $AF_\Delta = (Args_\Delta, \mathcal{R}_\Delta)$ . Then  $AF_\Delta$  can be represented by the partition  $((Args'_1, \mathcal{R}'_1), \mathcal{R}'_{1-\mathcal{D}}), \dots, ((Args'_n, \mathcal{R}'_n), (\mathcal{R}'_{n-\mathcal{D}}))$  where:

- $Args_\Delta = \bigcup_{i=1}^n Args'_i$  and  $\mathcal{R}_\Delta = \bigcup_{i=1}^n (\mathcal{R}'_i \cup \mathcal{R}'_{i-\mathcal{D}})$
- for  $i = 1 \dots n$ ,  $(Args'_i, \mathcal{R}'_i)$  is an expansion of  $(Args_i, \mathcal{R}_i)$  and for  $i = 1 \dots n - 1$ ,  $(C, \overrightarrow{AB}) \in \mathcal{R}'_{i-\mathcal{D}}$  iff  $(C, (A, B)) \in \mathcal{D}_i$ , where  $C \in Args'_{i+1}, \overrightarrow{AB} \in Args'_i$

**Lemma 2** Let  $\Delta = (Args, \mathcal{R}, \mathcal{D})$  be a hierarchical  $EAF$ .  $S$  is an admissible extension of  $\Delta$  iff  $T$  is an admissible extension of  $AF_\Delta = (Args_\Delta, \mathcal{R}_\Delta)$ , where:

$T = S \cup \{\overrightarrow{X} \mid Y \in S, X \rightarrow^S Y\} \cup \{\overrightarrow{Y\overline{Z}} \mid Y \in S, Y \rightarrow^S Z$  and there is a reinstatement set for  $Y \rightarrow^S Z\}$

**Proof** Let  $((Args_1, \mathcal{R}_1), \mathcal{D}_1), \dots, ((Args_n, \mathcal{R}_n), \mathcal{D}_n)$  be the partition of  $\Delta$ , and  $((Args'_1, \mathcal{R}'_1), \mathcal{R}'_{1-\mathcal{D}}), \dots, ((Args'_n, \mathcal{R}'_n), (\mathcal{R}'_{n-\mathcal{D}}))$  the partition of  $AF_\Delta$ . For  $i = 1 \dots n$ :

1)  $(Args_i, \mathcal{R}_i)$  and  $(Args'_i, \mathcal{R}'_i)$  are Dung argumentation frameworks, where  $(Args'_i, \mathcal{R}'_i)$  is an expansion of  $(Args_i, \mathcal{R}_i)$

2)  $\forall (X, Y) \in \mathcal{R}, \forall (Z, (X, Y)) \in \mathcal{D}, (X, Y) \in \mathcal{R}_i$  iff  $(Z, (X, Y)) \in \mathcal{D}_i, Z \in Args_{i+1}$

3) For  $i = 1 \dots n - 1, (Z, W) \in \mathcal{R}'_{i-\mathcal{D}}$  implies  $Z \in Args'_{i+1}, W \in Args'_i$ , and  $W$  is an argument of the form  $\overrightarrow{XY}$ .

4) For  $i = 1 \dots n, (Z, (X, Y)) \in \mathcal{D}_i$  iff  $(Z, \overrightarrow{XY}) \in \mathcal{R}'_{i-\mathcal{D}}$ .

1) - 3) imply that  $S$  can be partitioned into  $S_1 \cup \dots \cup S_n, T$  into  $T_1 \cup \dots \cup T_n$ , and that the theorem is shown by proving by induction on  $i$ , the following result:

$S_i = S_1 \cup \dots \cup S_n$  is admissible iff  $T_i = T_1 \cup \dots \cup T_n$  is admissible, where:

$T_i = S_i \cup \{\overrightarrow{X} \mid Y \in S_i, X \rightarrow^{S_i} Y\} \cup \{\overrightarrow{Y\overline{Z}} \mid Y \in S_i, Y \rightarrow^{S_i} Z$  and there is a reinstatement set for  $Y \rightarrow^{S_i} Z\}$ .

*Base case ( $i = n$ ):* Since  $\mathcal{D}_n = \emptyset, \mathcal{R}'_{n-\mathcal{D}} = \emptyset$ , and  $(Args'_n, \mathcal{R}'_n)$  is the expansion of  $(Args_n, \mathcal{R}_n)$ ,

then the result is given by lemma 1, where trivially:

$\forall Y \in S_n, T_n: \forall X$  s.t.  $(X, Y) \in \mathcal{R}_n, X \rightarrow^{S_n} Y, \forall Z$  s.t.  $(Y, Z) \in \mathcal{R}_n, Y \rightarrow^{S_n} Z$  and there is a reinstatement set  $\{Y \rightarrow^{S_n} Z\}$  for  $Y \rightarrow^{S_n} Z$ .

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

*General case:*

*Left to Right half:* Let  $A \in S_i$ . Suppose some  $B \in S_i$  s.t.  $B \rightarrow^{S_i} A$ , based on the attack  $(B, A) \in \mathcal{R}_i$ . Since  $A$  acceptable w.r.t.  $S_i, \exists C \in S_i, \text{ s.t. } C \rightarrow^{S_i} B$ , based on  $(C, B) \in \mathcal{R}_i$ . By definition of  $T_i$ , and given **1**, **3**) and lemma 1,  $A, \overline{B}$  and  $C$  are all in  $T_i$  and are acceptable w.r.t.  $T_i$  if we can show that  $\overline{CB}$  is acceptable w.r.t.  $T_i$  given some  $D \in T_{i+1}, (D, \overline{CB}) \in \mathcal{R}'_{i-D}$ .

By **4**),  $(D, (C, B)) \in \mathcal{D}_i$ . By assumption of  $A$  acceptable w.r.t.  $S_i$ , then by **2**),  $\exists E \in S_{i+1}$  s.t.  $E \rightarrow^{S_i} D$  based on the attack  $(E, D) \in \mathcal{R}_{i+1}$ , and there is a reinstatement set for  $E \rightarrow^{S_i} D$ . By *IH*,  $E \in T_{i+1}, \overline{ED} \in T_{i+1}$ , and since  $(\overline{ED}, D) \in \mathcal{R}'_{i+1}, \overline{CB}$  is acceptable w.r.t.  $T_i$ .

It remains to show that for  $A \in S_i, A \in \overline{AX}$ , for any  $X$  such that  $A \rightarrow^{S_i} X$ , and there is a reinstatement set for  $A \rightarrow^{S_i} X$ , then  $\overline{AX} \in S_i$ . Since  $A \rightarrow^{S_i} X, (A, X) \in \mathcal{R}_i$ , then  $\{(A, \overline{A}), (\overline{A}, \overline{AX}), (\overline{AX}, X)\} \subseteq \mathcal{R}'_i$ . Since  $A \rightarrow^{S_i} X, \neg \exists Z \in S_{i+1}$  s.t.  $(Z, (A, X)) \in \mathcal{D}(\mathcal{D}_i)$ . By *IH*,  $\neg \exists Z \in T_{i+1}$  s.t.  $(Z, \overline{AX}) \in \mathcal{R}'_{i-D}$ . Hence,  $\overline{AX} \in T_i$  is acceptable w.r.t.  $T_i$  as it is reinstated from  $\overline{A}$ 's attack by  $A \in T_i$ .

*Right to Left half:* Let  $A \in T_i$  for some  $A \in \text{Args}$ . We show that  $A$  is acceptable w.r.t.  $S_i$ . Suppose some  $\overline{BA}$  such that  $(\overline{BA}, A) \in \mathcal{R}'_i$ . In which case, by definition of  $T_i$ , and given **1**), lemma 1 shows that:

1.  $\overline{B} \in T_i$ , and if  $(X, \overline{B}) \in \mathcal{R}'_i$  then  $X = B$ . Hence  $(B, A) \in \mathcal{R}$ . Assume  $\neg \exists X \in T_{i+1}$  s.t.  $(X, \overline{BA}) \in \mathcal{R}'_{i-D}$ . By *IH* and **4**),  $\neg \exists X \in S_{i+1}, (X, (B, A)) \in \mathcal{D}_i$ , and so given **2**),  $B \rightarrow^{S_i} A$ .
2.  $\exists \overline{CB} \in T_i$  s.t.  $(\overline{CB}, B) \in \mathcal{R}'_i, (\overline{C}, \overline{CB}) \in \mathcal{R}'_i, (C, \overline{C}) \in \mathcal{R}'_i$ , where  $C \in \text{Args}, C \in S_i$ . Since  $T_i$  is conflict free,  $\neg \exists X \in T_{i+1}$  s.t.  $(X, \overline{CA}) \in \mathcal{R}'_{i-D}$ . By *IH* and **4**),  $\neg \exists X \in S_{i+1}, (X, (C, B)) \in \mathcal{D}_i$ , and so given **2**),  $C \rightarrow^{S_i} B$ .

Suppose some  $X \notin T_{i+1}, (X, \overline{CA}) \in \mathcal{R}'_{i-D}$ . Given **4**),  $(X, C, A) \in \mathcal{D}_i$ . By assumption of  $\overline{CA}$  acceptable w.r.t.  $T_i, \exists \overline{YX} \in T_{i+1}, (\overline{YX}, X) \in \mathcal{R}'_{i+1}$ . By *IH* and definition of  $T_i, Y \in T_{i+1}, Y \in S_{i+1}, Y \rightarrow^{S_{i+1}} X, Y$  is acceptable w.r.t.  $S_i + 1$ , and there is a reinstatement set  $R_{S_{i+1}}$  for  $Y \rightarrow^{S_{i+1}} X$ . Hence, there is a reinstatement set  $R_{S_i} = R_{S_{i+1}} \cup \{C \rightarrow^{S_i} B\}$  for  $C \rightarrow^{S_i} B$ . Hence,  $A$  is acceptable w.r.t.  $S_i$ .

## References

- [1] T.J.M. Bench-Capon. Agreeing to Differ: Modelling Persuasive Dialogue Between Parties Without a Consensus About Values, *Informal Logic*, 22(3), 231-45, 2002.
- [2] T.J.M. Bench-Capon. Persuasion in Practical Argument Using Value-based Argumentation Frameworks, *Journal of Logic and Computation*, 13(3), 429-448, 2003.
- [3] T.J.M. Bench-Capon, S.Doutre and P.E. Dunne. Audiences in Argumentation Frameworks, *Artificial Intelligence*, 171, 42-71, 2007.
- [4] P.E. Dunne and T.J.M. Bench-Capon. Two party immediate response disputes: Properties and efficiency, *Artificial Intelligence*, 149, 221-250, 2002.
- [5] 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.
- [6] S. Modgil, *Value Based Argumentation in Hierarchical Argumentation Frameworks*. In: Proc. 1st Int. Conference on Computational Models of Argument, 297-308, Liverpool, UK, 2006.
- [7] 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, Tunisia, 2007.
- [8] S. Kaci and L. van der Torre. Preference-based argumentation: Arguments supporting multiple values. To appear in *International Journal of Approximate Reasoning*.
- [9] J.R. Searle. *Rationality in Action*, MIT Press, Cambridge, MA, 2001.