An Argumentation Based Semantics for Agent Reasoning

Sanjay Modgil

Department of Computer Science, Kings College London

Abstract. A key challenge for agent architectures and programming paradigms is to account for defeasible reasoning over mental attitudes and to provide associated conflict resolution mechanisms. A growing body of work is looking to address these challenges by proposing argumentation based approaches to agent defeasible and practical reasoning. This work conforms to Dung's seminal argumentation semantics. In this paper we review our previous work in which we extend Dung's semantics to allow for inclusion of arguments that express preferences between other arguments. In this way we account for the fact that preference information required to resolve conflicts is itself defeasible and may be conflicting. We then propose the extended semantics as a semantics for agent defeasible and practical reasoning, and substantiate this claim by showing how our semantics can characterise, and indeed provide a framework for extending, existing approaches to agent reasoning over beliefs, goals, and actions.

1 Introduction

A key challenge for agent architectures and programming paradigms is the need to formalise defeasible (non-monotonic) and practical reasoning, and associated conflict resolution mechanisms for mental attitudes such as beliefs, desires, intentions and obligations. Conflicts can arise within mental attitudes. For example, two beliefs may logically contradict, or two goals may logically preclude realisation of each other. Plans can be represented in terms of atomic actions related to the adopted goals (intentions) they are intended to realise [9],[20]. Alternative plans for realising a given intention can be viewed as conflicting (in the sense that one must be chosen at the expense of the other), or two plans for realising different intentions can be said to conflict if resource bounds preclude their joint execution. Conflicts can also arise *between* mental attitudes; e.g. a desire derived goal conflicting with an obligation derived goal [8]. Hence, non-monotonic formalisms such as Default Logic [21], have been adopted as a semantics for agent reasoning [23]. For example, the BOID architecture characterises generated candidate goal sets as extensions of a prioritised default logic theory in which rules for inferring goals are modelled as defaults, and a prioritisation of these defaults resolves conflicts between mental attitudes [8]. Consider two default rules respectively inferring the conflicting desire derived goal of 'being on the beach' and obligation derived goal of 'being at work'. Prioritising the former rule over the later will result in the generation of a single extension containing the goal to be on the beach. Indeed, goals are generated through the interaction of beliefs, intentions, desires and obligations, and prioritisations on these attitudes to resolve conflicts, correspond to different agent *types*

(12 primitive types are identified in [8]). For example, a selfish agent will prioritise desire over obligation derived goals, whereas a social agent will adopt the reverse prioritisation. Default Logic semantics have also been proposed for primitives in agent programming languages [22]. For example, the GenGoals and GenPlan primitives in [9] are defined in terms of generation of prioritised default logic extensions of goals, respectively plans.

In recent years, a growing body of work (e.g. [2],[4],[12],[14],[20]) has proposed argumentation based approaches to agent defeasible and practical reasoning. These works propose logical formalisms that conform to Dung's seminal argumentation semantics [11] (and its variants). A Dung argumentation framework consists of a set of arguments $Args$ and a binary conflict based relation $\mathcal{R}$ on $Args$ ($\mathcal{R} \subseteq Args \times Args$). A 'calculus of opposition' is then applied to the framework to evaluate the winning (justified) arguments under different extensional semantics. The underlying logic, and definition of the logic's constructed arguments $Args$ and relation $\mathcal{R}$, is left unspecified, thus enabling instantiation of a framework by various logical formalisms. Dung's semantics have thus become established as a general framework for non-monotonic reasoning, and, more generally, reasoning in the presence of conflict. A theory's inferences can 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 of the major species of logic programming and non-monotonic logics (e.g. default, autoepistemic, non-monotonic modal logics) turn out to be special forms of Dung's theory [6,11]. Hence, Dung's semantics can be seen to generalise and subsume the above Default Logic semantics proposed for agent architectures and programming languages.

To determine a unique set of justified arguments invariably requires preference information to resolve conflicts between pairs of attacking arguments. The role of preferences has been formalised in both the underlying logical formalisms that instantiate a Dung framework, and at the abstract level of the framework itself. Examples of the former (e.g., [19]) define the relation $\mathcal{R}$ in terms of the conflict based interaction between two arguments, and a preference based on their relative strength. Examples of the latter (e.g., [1] [5]) augment Dung's framework to include a preference ordering on arguments. Hence, given a conflict based *attack* relation $\mathcal{R}$ on the arguments, a *defeat* relation $\mathcal{R}'$ is defined, where defeat represents a successful attack by additionally accounting for the relative strengths of (preferences between) attacking arguments. The justified arguments are then evaluated on the basis of the defeat relation $\mathcal{R}'$.

However, the preference information required to determine the success of an attack is often assumed pre-specified as a given ordering, and external to the logical formalism. This does not account for the fact that preferences may vary according to context, and because information sources (be they agents or otherwise) may disagree as to the criteria by which the strengths of arguments should be valuated, or the valuations assigned for a given criterion. Hence, to facilitate agent flexibility and adaptability, requires argumentation based reasoning *about*, as well as *with*, defeasible and possibly conflicting preference information. For example, a 'social' agent uniformly prioritises arguments for obligation derived goals above arguments for desire derived goals. However, certain contexts may warrant selfish behavior. Such behavioural heterogeneity requires

argumentation based reasoning as to which prioritisation (agent type) is appropriate in a given context. In a practical reasoning context, consider two 'instrumental' arguments (that can be understood as denoting unscheduled plans as in [12],[20],[?]) each of which relate alternative drugs for realising a medical treatment goal. Different clinical trials reporting on the relative efficacy of the drugs may lead to contradictory preferences, requiring that the agent justify selecting one clinical trial valuation over another.

Requirements for reasoning about preferences have been addressed in works extending the object level logical languages for argument construction with rules for deriving priorities amongst rules, e.g., in default logic [7] and logic programming [14,19]. One can then construct 'priority arguments' whose claims determine preferences between other mutually attacking arguments to determine the successful attacks (defeats). Arguments claiming conflicting priorities may be constructed and preferences between these can be established on the basis of other priority arguments. However, these works are restricted to basing argument strength on a single criterion; one based on the priorities of the argument's constituent rules. In previous work [16] we extended Dung's abstract argumentation theory so as to allow for argumentation about preferences between arguments. An extended framework can include 'preference arguments' that *claim preferences between other arguments*. This is achieved by defining a new attack relation that originates from a preference argument, and *attacks an attack* between the arguments that are the subject of the preference claim. In section 2 of this paper we present an improved (in the sense that it simplifies) version of the extended semantics described in [16] (more fully described, with associated proofs, in [17]). In the spirit of Dung's abstract approach, no commitment is made to how preferences are defined in the underlying logical formalism instantiating the extended framework. Thus, if C is a preference argument expressing that an argument A is preferred to an argument B, then this preference may be based on any criterion for valuating argument strength, including criteria that relate to the argument as a whole, such as the value promoted by the argument [5]. We therefore claim that the extended semantics can serve as a general semantics for flexible and adaptive agent defeasible and practical reasoning. We substantiate this claim in sections 3 and 4. In section 3 we show how logic programming approaches such as [14,19] can be formalised as instances of our extended framework. We illustrate with examples demonstrating reasoning about preferences between conflicting arguments for beliefs and goals. In section 4 we show how our framework provides for extending an existing argumentation based formalism for agent practical reasoning ([4]) so as to accommodate defeasible reasoning about preference information that is assumed pre-defined in [4]. Finally, we conclude and discuss future work in section 5.

2 Argumentation Semantics That Accommodate Defeasible Reasoning about Preferences

A Dung argumentation framework is of the form $(Args, \mathcal{R})$ where $\mathcal{R} \subseteq (Args \times Args)$ can denote either attack or defeat. A **single** 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. Dung then defines the acceptability of a **set** of arguments under different extensional semantics. The definition is given here, in

which $S \subseteq Args$ is conflict free if no two arguments in S are related by $\mathcal{R}$, and F is a characteristic function of a framework, such that:

- $F : 2^{Args} \mapsto 2^{Args}$
- $F(S) = \{A \in Args | A \text{ is acceptable w.r.t. } S\}$.

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

- *S is admissible iff each argument in S is acceptable w.r.t. S (i.e. $S \subseteq F(S)$)*
- *S is a preferred extension iff S is a set inclusion maximal admissible extension*
- *S is a complete extension iff each argument which is acceptable w.r.t. S is in S (i.e. $S = F(S)$)*
- *S is a stable extension iff $\forall B \notin S, \exists A \in S$ such that $(A, B) \in \mathcal{R}$*
- *S is the grounded extension iff S is the least fixed point of F.*

Consider the following example in which two individuals **P** and **Q** exchange arguments $A, B \ldots$ about the weather forecast:

P : "Today will be dry in London since the BBC forecast sunshine" = A
Q : "Today will be wet in London since CNN forecast rain" = B
P : "But the BBC are more trustworthy than CNN" = C
Q : "However, statistics show that CNN are more accurate than the BBC" = C'
Q : "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 an argument C that claims that A is preferred to B. Hence B does not successfully attack (defeat) A, but A does defeat B. Evaluating admissibility on the basis of this binary defeat relation, $\{A\}$ and not $\{B\}$ is admissible. The impact of argument C could conceivably be modelled by letting C attack B (see $\Delta 1$ in figure 1 in which an attack is visualised as an arrow from the attacker to the attacked). This would yield the required result, but if an argument D then attacked A (e.g. D = "the BBC forecast is for Glasgow and not London") then $\{B\}$ would still not be admissible ($\Delta 2$ in figure 1). This is clearly inappropriate. C expresses a preference for A over B, but if A is attacked (and defeated) by another argument, then we should recover B. Intuitively, C is an argument about the relationship between A and B. Specifically, in expressing a preference for A over B, C is an argument for A's repulsion of, or defence against, B's attack on A, i.e., C *defence* attacks B***'s attack on*** A ($\Delta 3$ in figure 1) so that B's attack on A does not succeed as a defeat. B's attack on A is, as it were, cancelled out, and we are left with A defeating B. Now, if D attacks A we will recover $\{B\}$ as an admissible extension ($\Delta 4$ in figure 1). 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. Intuitively, C and C' claim contradictory preferences and so attack each other ($\Delta 5$ in figure 1). 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. In the example, E d attacks the attack from C to C' ($\Delta 6$ in figure 1), and so determines that C' defeats C. Hence, C's d-attack on B's attack on A is cancelled out, and we are left with B defeating A; the discussion concludes in favour of **Q**'s argument that it will be a wet day in London.
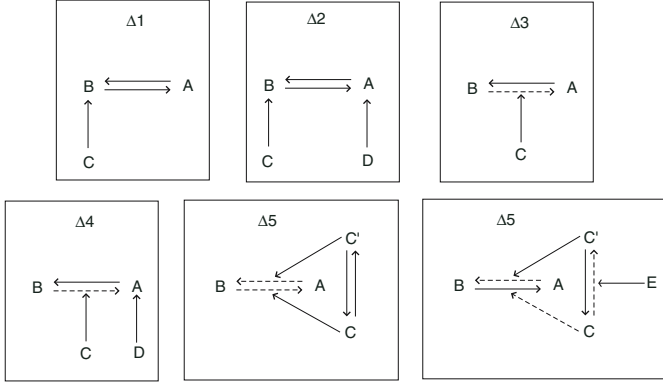
Fig. 1.

We now formally define the elements of an *Extended Argumentation Framework*:

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 $A \rightharpoonup B$ 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 $C \twoheadrightarrow (A \rightharpoonup B)$ to denote $(C, (A, B)) \in \mathcal{D}$, and say that C *defence* (d) attacks A's attack on B.

From hereon, definitions are assumed relative to an *EAF* ($Args$, $\mathcal{R}$, $\mathcal{D}$), where arguments $A, B, \ldots$ are assumed to be in $Args$, and S is a subset of $Args$. We now formally define the defeat relation that is parameterised w.r.t. some set S of arguments. This accounts for an attack's success as a defeat being relative to preference arguments already accepted in some set S, rather than relative to some externally given preference ordering.

Definition 3. *A S-defeats B, denoted by $A \rightarrow^s B$, iff $(A, B) \in \mathcal{R}$ and $\neg \exists C \in S$ s.t. $(C,(A, B)) \in S$. A strictly S-defeats B iff A S-defeats B and B does not S-defeat A.*

Example 1. Let Δ be the *EAF*: $A \rightleftharpoons B$, $C \twoheadrightarrow (A \rightharpoonup B)$

A and B S-defeat each other for $S = \emptyset$, $\{A\}$ and $\{B\}$.
B $\{C\}$-defeats A but A does not $\{C\}$-defeat B (B strictly $\{C\}$-defeats A).

We now define the notion of a conflict free set S of arguments. One might define such a set as one in which no two arguments attack each other. However, when applying argumentation to practical reasoning, it may be that if B asymmetrically attacks A, but A is preferred to B, then neither B or A defeat each other, and so both may end up being justified arguments. This is appropriate only when the arguments are not inherently

contradictory. For example, in [4], arguments relating to deciding a course of action appeal to values [5]. If A is an argument for a medical action a, appealing to the value of *health*, and B is an argument claiming that a is prohibitively expensive, then B asymmetrically attacks A, and B appeals to the value of *cost*. If a given value ordering ranks the value of *health* above *cost*, then B does not defeat A, and both arguments may then be justified; one commits to the action while accepting that it is expensive. Since in what follows, an admissible extension of an *EAF* is defined on the basis of a conflict free set, we need to allow for the above when defining a conflict free set:

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. $(C, (B, A)) \in \mathcal{D}$.*

In [16] we suggest that special attention be given to symmetric *EAF*s in which preference arguments can only *d*-attack attacks between arguments that symmetrically attack:

Definition 5. *Let $\Delta = (Args, \mathcal{R}, \mathcal{D})$ be an* EAF. *We say that Δ is a symmetric EAF iff: if $(C, (B, A)) \in \mathcal{D}$, then $(A, B) \in \mathcal{R}$.*

The restriction on $\mathcal{D}$ is appropriate when the arguments are inherently contradictory, as when arguing about beliefs. This is because no conflict free subset S of $Args$ in a symmetric *EAF* (and so admissible extension) can contain arguments that attack, since it could not satisfy the condition in definition 4. [1]

We now define the acceptability of an argument A w.r.t. a set S for an *EAF*. Consider figure 2-a) in which the acceptability of $A1$ w.r.t. S is under consideration. $B1 \rightarrow^s A1$, and $A1$ reinstates itself via the defeat $A1 \rightarrow^s B1$. However, the latter defeat is based on an attack $A1 \rightharpoonup B1$ that is itself under attack: $B2 \twoheadrightarrow (A1 \rightharpoonup B1)$. We therefore need to check that $A1 \rightarrow^s B1$ is 'defeat reinstated' by an argument in S that S-defeats $B2$. In general, $X \rightarrow^S Y$ is *defeat reinstated* iff for any Z s.t. $(Z, (X, Y)) \in \mathcal{D}$, there is a $Z' \in S$ s.t. $Z' \rightarrow^S Z$.

In figure 2-a) $A1 \rightarrow^s B1$ is 'defeat reinstated' by $A2 \in S$. In general, an argument X can then be said to be 'locally' acceptable w.r.t. a set S if for any Y that S-defeats X, there is a $Z \in S$ such that Z S-defeats Y and this defeat is defeat reinstated. However,
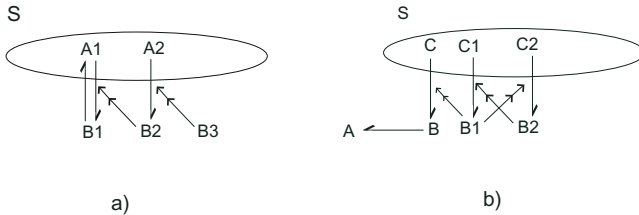

Fig. 2. $A1$ is not acceptable w.r.t. S in a). A is acceptable w.r.t. S in b).

---

[1] Note that other restrictions on $\mathcal{D}$ may be appropriate. E.g in logic programming systems such as [19], if B claims what was assumed non-provable (through negation as failure) by a rule in A, then B asymmetrically attacks A, and defeats A irrespective of their relative preference. One would then preclude such attacks from being *d*-attacked by preference arguments.

local acceptability does not suffice in the sense that the obtained notion of admissibility (as defined in definition 1) does not satisfy the property that if X is acceptable w.r.t. an admissible S, then $S \cup \{X\}$ is admissible. This result is shown by Dung's *fundamental lemma* in [11]. Intuitively, S represents a coherent 'position' defending each of its contained arguments. Having established such a position in the course of an argument, one would want that a proposed argument A that is defended by the position, does not, when included in the position, undermine the position that defends it. Referring to figure 2-a), S is admissible under local acceptability. $B3$ is locally acceptable w.r.t S, but $S' = S \cup \{B3\}$ is not admissible since $A1$ is not locally acceptable w.r.t S' (because the presence of $B3$ now invalidates the defeat from $A2$ to $B2$). What is required when checking the acceptability of $A1$ w.r.t. S is not only that $A1 \rightarrow^s B1$ is defeat reinstated, but that the defeat's reinstater $A2 \rightarrow^s B2$ is itself defeat reinstated. It is not, since no argument in S S-defeats $B3$. In general:

Definition 6. *Let* $S \subseteq Args$ *in* $(Args, \mathcal{R}, \mathcal{D})$. *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* $C \rightarrow^S B$, *iff:*

- $C \rightarrow^S B \in R_S$, *and*
- $\forall X \rightarrow^S Y \in R_S$, $\forall Y'$ *s.t.* $(Y', (X, Y)) \in \mathcal{D}$, *there* $\exists X' \rightarrow^S Y' \in R_S$

Definition 7. *A is acceptable w.r.t. $S \subseteq Args$ in $(Args, \mathcal{R}, \mathcal{D})$, 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 $C \rightarrow^S B$.*

Under this definition of acceptability, Dung's fundamental lemma is satisfied (see [16]). In figure 2-b), A is acceptable w.r.t. S given the reinstatement set $\{C \rightarrow^S B, C1 \rightarrow^S B1, C2 \rightarrow^S B2\}$ for $C \rightarrow^S B$. With the above definition of acceptability, extensional semantics for *EAF*s are now given by definition 1, where conflict free defined as in definition 4, for the stable semantics, 'A S-defeats B' replaces '$(A, B) \in \mathcal{R}$', and (for technical reasons) the domain of an *EAF*'s characteristic function F is restricted to the set of all *conflict free* subsets of $Args$.

For the complete, preferred and stable semantics, an argument is sceptically justified if it belongs to all extensions, and credulously justified if it belongs to at least one extension. The grounded semantics return a single extension, and so are inherently sceptical. In [16] we show that for symmetric *EAF*s, S is an admissible extension obtained on the basis of local acceptability iff S is an admissible extension obtained on the basis of acceptability in definition 7. Hence, the extensions of a symmetric *EAF* can equivalently be defined under local acceptability, and Dung's fundamental lemma holds for such *EAF*s under local acceptability. In [16] we also show the following results that have been shown to hold for Dung argumentation frameworks (the importance of these results are discussed in more detail in [16] and [11]). Let Δ be any *EAF*. Then:

1. The set of all admissible extensions of Δ form a complete partial order w.r.t. set inclusion
2. For each admissible S there exists a preferred extension S' of Δ such that $S \subseteq S'$
3. Δ possesses at least one preferred extension.
4. Every stable extension of Δ is a preferred extension but not vice versa.
5. Defining a sequence $F^1 = F(\emptyset)$, $F^{i+1} = F(F^i)$, then $F^{i+1} \supseteq F^i$ (where each F^j in the sequence is conflict free)

Suppose an *EAF* Δ is defined as finitary iff for any argument A or attack (B, C), the set of arguments attacking A, respectively (B, C), is finite. Referring to the sequence in 5 above, one can also show that the least fixed point (grounded extension) of a symmetric Δ is given by $\bigcup_{i=1}^{\infty}(\mathrm{F}^i)$. For arbitrary *EAF*s we can not guarantee that the fixed point obtained in 5 is the least fixed point (the existence of a least fixed point is guaranteed by the monotonicity of F, which only holds for symmetric EAFs). This means that the grounded extension of an EAF that is not symmetric is defined by the sequence in 5.

To conclude, we have defined an extended semantics that allows for representation of arguments that express preferences between other arguments. No commitments are made to the underlying logics for argument construction, or to the criteria used to valuate the strength of, and so preferences between, arguments. This work is to be contrasted with approaches that extend the underlying **object level** logical languages with rules for deriving priorities amongst rules. In the following section we provide support for our claim that the above extended semantics can serve as a semantics for agent defeasible and practical reasoning, in the sense that given an agent theory T defined in a logic based formalism L, then α is an inferred belief or goal, or chosen action iff α is the claim of a justified argument of the extended argumentation framework instantiated by the arguments and attacks defined by the theory T in L.

3   Argumentation Based Reasoning about Goals

In this section we illustrate how inferences obtained in logic programming formalisms that facilitate defeasible reasoning about rule priorities (e.g.[14],[19]), can be characterised in terms of the claims of the justified arguments of an instantiated *EAF*. We will adopt an approach to reasoning about goals in [14], in which subsets of a set of agent rules for deriving goals are associated with one of five agent's *needs* or *motivations* (based on Maslow's work in cognitive psychology [15]): **Physiological**, **Safety**, **Affiliation** (**Social**), **Achievement** (**Ego**) and **Self-Actualisation**). Simplifying the representation in [14], one can express an agent's default personality by rules of the form $R_{def} : true \Rightarrow hp(r_X^m, r_Y^{m'})$ expressing that a rule with name r_X associated with motivation m has higher priority than a rule with name r_Y associated with motivation m' (where X and Y range over the rule name indicies used). Exceptional circumstances may warrant prioritisation of a specific m' over m goal: $R_{excep} : S \Rightarrow hp(r_Y^{m'}, r_X^m)$ (where S denotes a conjunction of first order literals with explicit negation). Hence, if S holds, then one can construct an argument $A1$ based on R_{def} and $A2$ based on R_{excep}. Given an argument based on $R_{override} : true \Rightarrow hp(R_{excep}, R_{def})$, then $A2$ defeats $A1$, so that an argument for a goal based on $r_Y^{m'}$ will now defeat an argument for a goal based on r_X^m. Note that in agent architectures [8]and programming paradigms [9] that address conflict resolution amongst goals, rules for deriving goals are expressed as conditionals of the form $a \rightarrow^M b$ [8] or modal rules of the form $a \rightarrow Mb$ [9] where a and b are propositional, respectively propositional modal, wff, and $M \in \{B(Belief),$ $O(Obligation), I(Intention), D(Desire)\}$. As mentioned in the introduction, a prioritised default logic semantics resolves conflicts amongst the goals generated by different mental attitudes.The relationship with [14] is clear. Agent personalities represented by orderings on motivations, correspond with agent types represented by orderings on

mental attitudes in [9] and [8]. However, behavioural heterogeneity and adaptability is limited in the latter works, in the sense that an altruistic agent's goals will always be characterised by default extensions that contain obligation (social) derived goals at the expense of conflicting desire (ego) derived goals.

In what follows we formalise reasoning of the above type in [19]'s *argument based logic programming with defeasible priorities* (ALP-DP). To simplify the presentation, we present a restricted version of ALP-DP - ALP-DP* - that does not include negation as failure (as this is not needed for any of the examples). We describe ALP-DP* arguments, their attacks, and how priority arguments define preferences. We refer the reader to [19] for details of the proof theory.

Definition 8. *Let (S, D) be a ALP-DP* theory where S is a set of strict rules of the form $s : L_0 \wedge \ldots \wedge L_m \rightarrow L_n$, D a set of defeasible rules $r : L_0 \wedge \ldots \wedge L_j \Rightarrow L_n$, and:*

- *Each rule name r (s) is a first order term. From hereon we may write* head(r) *to denote the consequent L_n of the rule named r.*
- *Each L_i is an atomic first order formula, or such a formula preceded by strong negation $\neg$.*

We also assume that the language contains a two-place predicate symbol $\prec$ for expressing priorities on rule names. Strict rules are intended to represent information that is beyond debate. We assume that any S includes the following strict rules expressing properties on the relation $\prec$.

- $o1 : (x \prec y) \wedge (y \prec z) \rightarrow (x \prec z)$ • $o2 : (x \prec y) \wedge \neg(x \prec z) \rightarrow \neg(y \prec z)$
- $o3 : (y \prec z) \wedge \neg(x \prec z) \rightarrow \neg(x \prec y)$ • $o4 : (x \prec y) \rightarrow \neg(y \prec x)$

Definition 9. *An argument A based on the theory (S, D) is:*

1. *a finite sequence $[r_0, \ldots, r_n]$ of ground instances of rules such that:*
   - *for every i ($0 \leq i \leq n$), for every literal L_j in the antecedent of r_i there is a $k < i$ such that* head(r_k) $= L_j$. *If* head(r_n) $= x \prec y$ *then A is called a 'singleton priority argument'.*
   - *no distinct rules in the sequence have the same head;*

   or:
2. *a finite sequence $[r_{0_1}, \ldots r_{n_1}, \ldots, r_{0_m}, \ldots r_{n_m}]$, such that for $i=1 \ldots m$, $[r_{0_i}, \ldots r_{n_i}]$ is a singleton priority argument. We say that A is a 'composite priority argument' that concludes the ordering $\bigcup_{i=1}^{m}$ head(r_{n_i})*

In [19], arguments are exclusively defined by 1). Here, we have additionally defined composite priority arguments so that an ordering, and hence a preference, can be claimed (concluded) by a single argument rather than a set of arguments. Note that from hereon we will, for ease of presentation, show only propositional examples. Also, abusing the notation, arguments will be represented as sequences of *rule names* rather that the rules these names identify. The following example adopts the categorisation of goals in the BOID architecture [8] and programming paradigm [9]. Rules named int_i represent goals committed to in previous deliberations. Rules named bel_i allow for derivation of beliefs, and act as a filter on desire and obligation derived goals derived by rules named des_i and

ob_i respectively. The filtering rules ensure that agents do not engage in wishful thinking. This could be expressed as an agent type, by prioritising these filtering rules above desire and obligation rules. Here, we make the filtering rules strict.

Example 2. Let $S = \{o1 \ldots o4\} \cup \{bel_1 : cheap\_room \rightarrow \neg close\_to\_conf,$
$bel_2 : close\_to\_conf \rightarrow \neg cheap\_room\}$, and D be the set of rules:

- $bel_3 :\Rightarrow remaining\_budget\_high$ $soc_1 :\Rightarrow des_1 \prec ob_1$
- $int_1 :\Rightarrow go\_to\_conference$ $self_1 :\Rightarrow ob_1 \prec des_1$
- $des_1 : go\_to\_conference \Rightarrow close\_to\_conf$ $def\_agent\_type :\Rightarrow self_1 \prec soc_1$
- $ob_1 : go\_to\_conference \Rightarrow cheap\_room$
- $exception\_soc : remaining\_budget\_high \Rightarrow soc_1 \prec self_1$
- $overide\_social :\Rightarrow def\_agent\_type \prec exception\_soc$

The following are a subset of the arguments that can be constructed:

$A1 = [int_1, des_1]$ is an argument for the desire derived goal to be close to the conference, given the intention to go to the conference.
$A2 = [int_1, ob_1]$ is an argument for the obligation derived goal to book a cheap room.
$B1 = [soc_1]$ and $B2 = [self_1]$ respectively express social and selfish agent types.
$C1 = [def\_agent\_type]$ expresses that the reasoning agent is by default a social agent, and $C2 = [bel_3, exception\_soc]$ expresses exceptional conditions (when the remaining project funds are high) under which the agent behaves selfishly. Finally, $D1 = [overide\_social]$ expresses that the selfish agent type warranted by exceptional conditions overrides the default social behaviour.

The following definitions assume arguments are relative to a theory (S, D). [19] defines:

Definition 10. *For any arguments A, A' and literal L:*

- *A is strict iff it does not contain any defeasible rule; it is defeasible otherwise.*
- *L is a conclusion of A iff L is the head of some rule in A*
- *If T is a sequence of rules, then A + T is the concatenation of A and T*

[19] motivates definition of attacks between arguments that account for the ways in which arguments can be extended with strict rules:

Definition 11. *A1 attacks A2 iff there are sequences S1 and S2 of strict rules such that A1 + S1 is an argument with conclusion L and A2 + S2 is an argument with a conclusion ¬L.*

By definition, if $A1$ attacks $A2$ then $A2$ attacks $A1$. Note that A1 may attack A2 on a number of conclusions and corresponding concatenations of strict rules. In example 2, $A1$ and $A2$ attack each other since $A1 + [bel_2]$ has the conclusion $\neg cheap\_room$ and $A2$ has the conclusion $cheap\_room$. Also $B1 + [o4]$ has the conclusion $\neg(ob_1 \prec des_1)$ and so $B1$ attacks and is attacked by $B2$. To determine a preference amongst attacking arguments ALP-DP* defines the sets of relevant defeasible rules to be compared:

Definition 12. *If $A+S$ is an argument with conclusion L, the defeasible rules $R_L(A+S)$ **relevant** to L are:*

1. $\{r_d\}$ iff A includes defeasible rule r_d with head L
2. $R_{L_1}(A + S) \cup \ldots \cup R_{L_n}(A + S)$ iff A is defeasible and S includes a strict rule s : $L_1 \wedge \ldots \wedge L_n \to L$

For example, $R_{cheap\_room}(A2) = \{ob_1\}$ and $R_{\neg cheap\_room}(A1 + [bel_2]) = \{des_1\}$. We define ALP-DP*'s ordering on these sets and hence preferences amongst arguments, w.r.t. an ordering concluded (as defined in definition 9-2) by a priority argument:

Definition 13. *Let C be a priority argument concluding the ordering $\prec$. Let R and R' be sets of defeasible rules. Then $R' > R$ iff $\forall r' \in R'$, $\exists r \in R$ such that $r \prec r'$.*

The intuitive idea behind the above definition is that R can be made better by replacing some rule in R with any rule in R', while the reverse is impossible. Now, given two arguments A and B, it may be that they attack on more than one conclusion. Given a priority ordering $\prec$ concluded by an argument C, we say that A is preferred$_\prec$ to B if for every pair (L, L') of conclusions on which they attack, the set of A's defeasible rules relevant to L is stronger ($>$) than the set of B's defeasible rules relevant to L'.

Definition 14. *Let C be a priority argument concluding $\prec$. Let $(L_1, L_1'), \ldots, (L_n, L_n')$ be the pairs on which A and B attack, where for $i = 1 \ldots n$, L_i and L_i' are conclusions in A and B respectively. Then A is preferred$_\prec$ to B if for $i = 1 \ldots n$, $R_{L_i}(A + S_i) > R_{L_i'}(B + S_i')$*

In example 2, $B1$ concludes $des1 \prec ob1$, and so $R_{cheap\_room}(A2) > R_{\neg cheap\_room}$ $(A1)$, and so $A2$ is preferred$_{des1 \prec ob1}$ to $A1$. We can now instantiate a symmetric *EAF* with the arguments, their attacks, and priority arguments claiming preferences and so d attacking attacks:

Definition 15. *The* EAF *$(Args, \mathcal{R}, \mathcal{D})$ for a theory (S, D) is defined as follows. $Args$ is the set of arguments given by definition 9, and $\forall A, B, C \in Args$:*

1. *$(C,(B, A)) \in \mathcal{D}$ iff C concludes $\prec$ and A is preferred$_\prec$ to B*
2. *$(A, B),(B, A) \in \mathcal{R}$ iff A and B attack as in definition 11*

Note that it can be shown that if $(C,(B, A))$ and $(C',(A, B)) \in \mathcal{R}_d$ then C and C' attack each other as in definition 11. The following result is a special case of proposition 8 in [16] which shows an equivalence for full ALP-DP with negation as failure:

Proposition 1. *Let Δ be the* EAF *defined by a theory (S, D) as in definition 15. Then L is the conclusion of a justified argument as defined in [19] iff L is the conclusion of an argument in the grounded extension of Δ.*[2]

For example 2 we obtain the *EAF* $\Delta 1$ in figure 3. $D1$, $C2$, $B2$ and $A1$ are sceptically justified under all the semantics. Intuitively, the normally social agent can behave selfishly if the remaining budget for the project is high. We conclude with another example that illustrates argumentation to resolve conflicts amongst goals derived from the same class of mental attitudes. The example also illustrates how argumentation over beliefs is incorporated into argumentation over goals.

---

[2] Note that all of Dung's extensional semantics can be defined for an ALP-DP theory's *EAF*. In [19] only the grounded, stable and complete semantics can be defined.
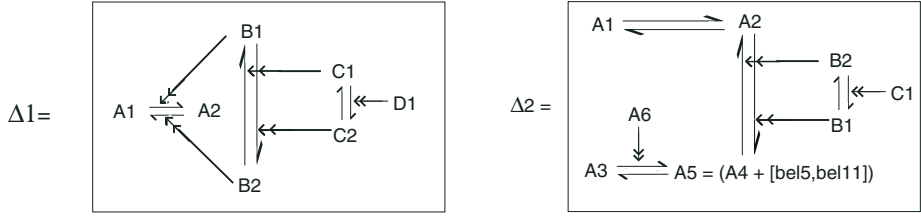
Fig. 3. *EAFs* $\Delta 1$ *and* $\Delta 2$

Example 3. Let (S, D) be a theory where $S = \{o1 \ldots o4\} \cup \{bel_7 : forecast\_storm \rightarrow \neg forecast\_calm, bel_8 : forecast\_calm \rightarrow \neg forecast\_storm, bel_9 : close\_to\_conf \rightarrow \neg close\_to\_beach, bel_{10} : close\_to\_beach \rightarrow \neg close\_to\_conf\}$

D is the set of defeasible rules:

- $int_1 :\Rightarrow go\_to\_conference$
- $des_1 : go\_to\_conference \Rightarrow close\_to\_conf$
- $des_2 : go\_to\_conference \Rightarrow close\_to\_beach$
- $bel_3\_bravo :\Rightarrow forecast\_calm$
- $bel_4\_bbc :\Rightarrow forecast\_storm$
- $bel_5 : forecast\_storm \Rightarrow beach\_hotel\_closed$
- $bel_{11} : beach\_hotel\_closed \rightarrow \neg close\_to\_beach$
- $bel_6 :\Rightarrow bel_3\_bravo \prec bel_4\_bbc$
- $realistic :\Rightarrow des_2 \prec bel_{11}$
- $wishful :\Rightarrow bel_{11} \prec des_2$
- $def\_agent\_type :\Rightarrow wishful \prec realistic$

Amongst the arguments that can be constructed, we obtain:

$A1 = [int_1, des_1]$ and $A2 = [int_1, des_2]$ are mutually attacking arguments for the desire derived goals of being close to the conference and being close to the beach respectively.

$A3 = [bel_3\_bravo]$ and $A4 = [bel_4\_bbc]$ are also mutually attacking arguments for contradictory weather forecasts, and $A3$ also mutually attacks $A5 = [bel_4\_bbc, bel_5, bel_{11}]$ which expresses the belief that if the forecast is for storms then the beach side hotel will be closed and so the agent cannot be close to the beach. Hence, $A5$ also mutually attacks $A2$.

$A6 = [bel_3\_bravo \prec bel_4\_bbc]$ claims that the bbc is acknowledged to be the more reliable than the bravo channel, and so expresses that $A4$ and $A5$ are preferred to $A3$.

$B1 = [realistic]$ and $B2 = [wishful]$ characterise agent types that respectively give priority to belief over desire, and desire over belief derived goals. $C1 = [def\_agent\_type]$ expresses that the default agent behaviour is realistic.

The EAF $\Delta 2$ instantiated by the above arguments is shown in figure 2. Since the agent is realistic and the BBC more reliable than Bravo, $C1$, $B1$, $A6$, $A5$ and $A1$ are sceptically justified under all the semantics. The agent adopts the goal of being close to the conference.

4   Argumentation Based Reasoning over Actions

A number of works apply argumentation to decide a preferred course of action, including logic programming formalisms of the type described in the previous section, and agent programming paradigms such as [9], in which mutually conflicting or 'incoherent' candidate sets of plans are also characterised in terms of prioritised default logic extensions. In works such as [2] and [12], tree structured *instrumental* arguments are composed by chaining propositional rules, and relate the root node top level goal to sub-goals, and primitive actions in the leaf nodes (these arguments can be thought of as unscheduled plans). Given conflict free sets of instrumental arguments, the preferred sets are chosen solely on the basis of those that maximise the number of agent goals realised. These works have been extended in [20] to accommodate decision theoretic notions. An instrumental argument additionally includes the resources required for execution of the leaf node actions. The strength of, and so preferences amongst, instrumental arguments is then based on their utility defined in terms of numerical valuations of the 'worth' of the goals and cost of the resources represented in the arguments. A more general conception of how to valuate the strength of arguments for action is partially realised in [4]. In this work, arguments for alternative courses of action for realising a given goal instantiate a value based argumentation framework (VAF) [5]. A given ordering on values (an 'audience') advanced by arguments, is then used to determine relative preferences and so defeats amongst arguments:

If A attacks B, then A defeats B only if the value advanced by B is not ranked higher than the value advanced by A according to some audience a.

Examples of values include cost, safety, efficacy (of the action w.r.t. goal) e.t.c. In [18] we proposed an extension to VAF that associates for a given value V, the degree to which an argument promotes or demotes a value. In this way, one can prefer $A1$ for action $a1$ to $A2$ for action $a2$, if both promote the same value, but the later does so to a lesser degree. Note that the same can be seen to apply to goals, so that, for example, a preference between two conflicting arguments for obligation derived goals may be based on the relative 'strength' or 'importance' of the obligations. In general then, one can consider arguments for goals or actions as being associated with, or advancing, meta-level criteria (be they motivations, values, etc), where:

1. The degree to which a criterion is advanced may vary according to context and or perspective
2. The ordering on criteria may vary according to context and or perspective

Both cases may warrant argumentation based resolution of conflicting preferences. In what follows, we illustrate some of the above ideas with an example taken from [18] that we now formalise in an *EAF*. We thus demonstrate how the extended semantics provide for integration of object level argumentation about action and metalevel reasoning about values and value orderings. The example builds on a schemes and critical questions approach to value based argumentation over action [4], whereby an argument for an action instantiates the following presumptive scheme **AS1**:

In the current circumstances R, we should perform action A, to achieve new circumstances S, which will realise some goal G, which will promote some value V

An extensive set of critical questions associated with **AS1** are then described. If $A1$ is an argument instantiating **AS1**, then the critical questions serve to identify arguments that attack $A1$. For example, an argument $A2$ stating that the action in $A1$ has an unsafe side-effect, asymmetrically attacks $A1$. $A2$ is identified with the critical question: *does the action have a side effect which demotes some value?* $A2$ may then itself be attacked by arguments identified by $A2$'s critical questions, and so on. Every argument advances a value. The arguments, their attacks, and a value ordering, instantiate a VAF, and based on the derived defeat relation, the justified arguments are evaluated under Dung's preferred semantics. In the following example, we will assume arguments instantiating schemes and critical questions. [4] described formal construction of these arguments in an underlying BDI type logic. We also assume arguments for possibly conflicting valuations of arguments and value orderings. Formal construction of these arguments is described in [18].
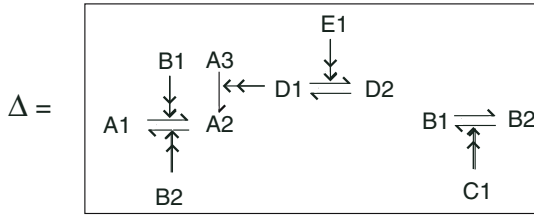

Fig. 4.

Example 4. Consider Δ in fig.3 in which $A1$ and $A2$ are arguments for the medical actions 'give aspirin' and 'give chlopidogrel' respectively. These arguments instantiate **AS1** and refer to beliefs, actions, goals and values, as described in [4]. $A1$ and $A2$ both promote the value of *efficacy*. They symmetrically attack since they claim alternative actions for the goal of preventing blood clotting. Argument $B1$ is based on clinical trial 1's conclusion that $A2$'s chlopidogrel is more efficacious than $A1$'s aspirin at preventing blood clotting. Hence $B1 \twoheadrightarrow (A1 \rightharpoonup A2)$. However, $B2$ is based on clinical trial 2's conclusion that the opposite is the case. Hence $B1 \rightleftharpoons B2$. At this stage neither $A1$ or $A2$ are sceptically justified under any semantics. However, $C1$ is an argument claiming that trial 1 is preferred to trial 2 since the former uses a more statistically robust design. Hence, $C1 \twoheadrightarrow (B2 \rightharpoonup B1)$. Now $A2$ and not $A1$ is sceptically justified. However, $A3$ promoting the value of *cost*, states that chlopidogrel is prohibitively expensive. We now have an example of an asymmetric attack: $A3 \rightharpoonup A2$. However, $D1 \twoheadrightarrow (A3 \rightharpoonup A2)$ where $D1$ is a value ordering ranking *efficacy* over *cost*. Hence, $A3$ does not defeat $A2$ and so $A2$ remains sceptically justified. Here, we see that $A3$ is also sceptically justified. Administering chlopidogrel is the preferred course of action, while acknowledging that it is costly. It's just that efficacy is deemed to be more important than cost. However, $D2$ now ranks *cost* over *efficacy*. Now neither $A2$ or $A1$ are sceptically justified. Finally, $E1$ is a utilitarian argument stating that since financial resources are low, use of chlopidogrel will compromise treatment of other patients, and so one should preferentially rank *cost* over *efficacy* (such a trade of is often made in medical contexts). Hence, $A1$ is now sceptically justified; aspirin is now the preferred course of action.

5   Discussion and Future Work

In this paper we describe an extended Dung semantics that meets requirements for agent flexibility and adaptability, when engaging in defeasible reasoning about beliefs and goals, and practical reasoning about actions. We claim that the extended semantics can serve as a semantics for agent defeasible and practical reasoning. In our view what appropriately accounts for the correctness of an inference for a belief, or goal, or choice of an action, is that the inference or choice can be shown to rationally prevail in the face of opposing inferences or choices. Dung's, and our extended semantics, in abstracting to the level of a 'calculus of opposition', provides logic neutral, rational means for establishing such standards of correctness.

This paper supports the above claim by formalising agent reasoning about goals in an existing logic programming based formalism [19] that facilitates defeasible reasoning about priorities. It remains to show that other non-monotonic formalisms accommodating reasoning about preferences in the object level, can instantiate extended argumentation frameworks. A notable example is the work of [10] in which defeasible logic rules for agent reasoning are extended with graded preferences, so that one can conclude preferences between mental attitudes contingent on mental attitudes in premises of the rules. We also discussed how the BOID architecture [8] and programming paradigm [9], neither of which accommodate defeasible reasoning about preferences, relate to the logic programming formalism described. This suggests that our extended argumentation semantics can serve as a framework in which to further develop existing agent reasoning formalisms of the above kind, in order to facilitate argumentation based reasoning about preferences, and so agent flexibility and adaptability. We also referred to works formalising construction of arguments in a BDI type logic for value based argumentation over action [4], and works formalising construction of arguments for valuations and value orderings [18], and showed how such arguments can instantiate an *EAF*.

The issue of how arguments for goals and actions interact remains to be addressed in future work. For example, a goal g may be selected at the expense of goal g'. However, it may be that no feasible plan exists for realising g (in the sense that resources are not available or that the plan is prohibitively expensive). Of course, arguments about the feasibility of plans can be used to express preferences between arguments for goals. For example, an argument that there is no feasible plan for g expresses a preference for the argument for g' over the argument for g. However, this requires that arguments for plans be constructed for all candidate goals, which may be computationally very expensive. The two stage process whereby goals are selected, and then arguments for plans are constructed and chosen is more efficient.

We also note that the inherently dialectical nature of argumentation has led to development of formal frameworks for argumentation based dialogues (see [3] for a review), where, for example, one agent seeks to persuade another to adopt a belief, or when agents communicate in order to deliberate about what actions to execute, or to negotiate over resources. These dialogues illustrate requirements for communicating and challenging reasons for preferring one argument to another. We aim to address these requirements by developing frameworks for argumentation based dialogues that build on the extended semantics described in this paper.

References

1. Amgoud, L.: Using Preferences to Select Acceptable Arguments. In: Proc. 13th European Conference on Artificial Intelligence, pp. 43–44 (1998)
2. Amgoud, L., Kaci, S.: On the Generation of Bipolar Goals in Argumentation-Based Negotiation. In: Proc. 1st Int. Workshop on Argumentation in Multi-Agent Systems, pp. 192–207 (2004)
3. ASPIC Deliverable D2.1: Theoretical frameworks for argumentation (June 2004), http://www.argumentation.org/PublicDeliverables.htm
4. Atkinson, K.M., Bench-Capon, T.J.M., McBurney, P.: Computational Representation of Practical Argument. Synthese 152(2), 157–206 (2006)
5. Bench-Capon, T.J.M.: Persuasion in Practical Argument Using Value-based Argumentation Frameworks. Journal of Logic and Computation 13(3), 429–448 (2003)
6. Bondarenko, A., Dung, P.M., Kowalski, R.A., Toni, F.: An abstract, argumentation-theoretic approach to default reasoning. Artificial Intelligence 93, 63–101 (1997)
7. Brewka, G.: Reasoning about priorities in default logic. In: Proc. 12th National Conference on Artificial Intelligence (AAAI 1994), pp. 940–945 (1994)
8. Broersen, J., Dastani, M., Hulstijn, J., van der Torre, L.W.N.: Goal Generation in the BOID Architecture. Cognitive Science Quarterly Journal 2(3-4), 428–447 (2002)
9. Dastani, M., van der Torre, L.: Programming BOID-Plan Agents: Deliberating about Conflicts among Defeasible Mental Attitudes and Plans. In: Proc 3rd international Joint Conference on Autonomous Agents and Multiagent Systems, pp. 706–713 (2004)
10. Dastani, M., Governatori, G., Rotolo, A., van der Torre, L.: Preferences of Agents in Defeasible Logic. In: Zhang, S., Jarvis, R. (eds.) AI 2005. LNCS (LNAI), vol. 3809, pp. 695–704. Springer, Heidelberg (2005)
11. Dung, P.M.: On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. Artificial Intelligence 77, 321–357 (1995)
12. Hulstijn, J., van der Torre, L.: Combining Goal Generation and Planning in an Argumentation Framework. In: 15th Belgium-Netherlands Conference on AI, pp. 155–162 (2003)
13. Hindriks, K., de Boer, F., van der Hoek, W., Ch, J.-J.: Agent Programming in 3APL. In: Autonomous Agents and Multi-Agent Systems, vol. 2(4), pp. 357–401 (1999)
14. Kakas, A., Moraitis, P.: Argumentation based decision making for autonomous agents. In: 2nd Int. Joint Conference on Autonomous Agents and Multiagent Systems, pp. 883–890 (2003)
15. Maslow, A.: Motivation and Personality. Harper and Row, New York (1954)
16. Modgil, S.: An Abstract Theory of Argumentation That Accommodates Defeasible Reasoning About Preferences. In: 9th European Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty, pp. 648–659 (2007)
17. Modgil, S.: Reasoning About Preferences in Argumentation Frameworks. Technical Report, http://www.dcs.kcl.ac.uk/staff/modgilsa/ArguingAboutPreferences.pdf
18. Modgil, S.: Value Based Argumentation in Hierarchical Argumentation Frameworks. In: Proc. 1st International Conference on Computational Models of Argument, pp. 297–308 (2006)

19. Prakken, H., Sartor, G.: Argument-based extended logic programming with defeasible priorities. Journal of Applied Non-Classical Logics 7, 25–75 (1997)
20. Rahwan, I., Amgoud, L.: An argumentation based approach for practical reasoning. In: Proc. 5th Int. Joint Conference on Autonomous agents and Multiagent systems, pp. 347–354 (2006)
21. Reiter, R.: A logic for default reasoning. Artificial Intelligence 13, 81–132 (1980)
22. van Riemsdijk, B., Dastani, M., Meyer, J.: Semantics of declarative goals in agent programming. In: 4th Int. Joint Conference on Autonomous agents and Multiagent systems, pp. 133–140 (2005)
23. Thomason, R.: Desires and defaults: a framework for planning with inferred goals. In: 7th International Conference on Knowledge Representation and Reasoning, pp. 702–713 (2002)