Argumentation based Resolution of Conflicts Between Desires and Normative Goals

Sanjay Modgil and Michael Luck

Department of Computer Science, Kings College London

Abstract. Norms represent what ought to be done, and their fulfillment can be seen as benefiting the overall system, society or organisation. However, individual agent goals (desire) may conflict with system norms. If a decision to comply with a norm is determined exclusively by an agent or, conversely, if norms are rigidly enforced, then system performance may be degraded, and individual agent goals may be inappropriately obstructed. To prevent such deleterious effects we propose a general framework for argumentation-based resolution of conflicts amongst desires and norms. In this framework, arguments for and against compliance are arguments justifying rewards, respectively punishments, exacted by 'enforcing' agents. The arguments are evaluated in a recent extension to Dung's abstract argumentation framework, in order that the agents can engage in *met-alevel* argumentation as to whether the rewards and punishments have the required motivational force. We provide an example instantiation of the framework based on a logic programming formalism.

1 Introduction

Requirements for conflict resolution arise in open multi-agent systems in which goals of individual agents conflict with norms imposed by the system to regulate individual agent behaviour. If the decision to comply with a norm is determined exclusively by an individual, then system performance may be degraded. Hence, institutional or social pressure to comply may be brought about by *system agents* exacting punishments and grants rewards [17, 11]. This may be appropriate for closed static systems, but compromises the flexibility of dynamic open systems in which rigid enforcement of norms may lead to both unwarranted obstruction of agent goals and degraded system performance. For example, an agent's goal may be obstructed by enforcing compliance with a norm that is justified by system-held beliefs about the context. However, these beliefs may be erroneous. In addition, it may not always be able to anticipate at design time, contexts in which compliance with norms does or does not coincide with the best interests of the system, and when enforcement mechanisms have insufficient motivational force. In such cases, an agent might appeal to higher level *motivations* [9], arguing that in pursuing its own goal it is indeed acting in the interests of the system as a whole, or that exacted punishments (or rewards) for non-compliance (or compliance) are outweighed by the benefits of pursing its own goal.

In this paper we propose a general argumentation-based framework that evaluates arguments for and against compliance with norms, in order to prevent unwarranted obstruction of individual goals and degraded system performance. As in [11, 6], norms

are interpreted as *system goals* that individual agents are required to realise, and that may conflict with the *individual goals* or *desires* of an agent. Punishments and rewards are the individual goals of system agents responsible for enforcement. In general, an argument for a goal justifies realisation of that goal based on beliefs that are themselves the outcome of argumentation based reasoning about what is the case. An argument for a system goal may then mutually attack an argument for a conflicting individual goal, and arguments for punishment and reward goals attack the argument for an individual goal. It is the success of these attacks that determines which of the arguments prevail and thus whether or not there is a reasoned case for compliance[1]. In general, an attack succeeds as a defeat if the attacked argument is not stronger than or *preferred* to its attacker [1]. As in [4], preferences may be derived from a relative ordering on the values that the arguments promote. In this paper, preferences among arguments for goals are similarly evaluated. For example, a 'reward argument' will successfully attack (defeat) an argument for an individual goal if an agent is persuaded that the reward is of greater utility to it than the individual goal it is required to abandon in favour of compliance with the system goal. The proposed framework will thus need to account for:

1. **Social mechanisms for enforcing compliance**: An agent Ag's argument for an individual goal g may be attacked by arguments for the (punishment and reward) goals $g', \ldots$ of *other* agents, where the attacks are not based on direct conflicts between g and $g', \ldots$. For example, a reward (punishment) may facilitate (hinder) some other goal that Ag is already committed to realising.
2. **Motivational argumentation**: Flexible and adaptive agents need to engage in motivational argumentation over the respective merits of goals. Hence, argumentation frameworks in which preferences [1] and value orderings [4] on arguments are 'given', and not themselves subject to reasoning, do not suffice. Rather, there is a requirement for argumentation based reasoning over the preferences themselves.

Existing work addresses argumentation-based resolution of conflicts among goals ([2], [10], [16]), and [16] explicitly considers conflicts between individual goals and norms. However, no existing work accounts for social mechanisms, whereby an agent's decision as to which goals to pursue is influenced by other agents' goals. Only [10] accounts for argumentation over preferences, but does so in the object level logic programming language, whereby rules express priorities over other rules. In this paper, we aim at an abstract framework in which preferences are not restricted to rule priorities, but can account for any criteria for valuating argument strength, including those that relate to the argument as a whole (e.g., as in [4]). We therefore make use of a recent extension to Dung's seminal abstract argumentation semantics [8]. In a Dung framework, arguments are related by a binary conflict-based relation, and the winning (justified) arguments under different extensional semantics are evaluated. The underlying logic, and definition of the logic's constructed arguments and conflict relation, is left unspecified, enabling instantiation by various logical formalisms. Dung's semantics thus serves as a general framework capturing various species of non-monotonic reasoning [5], and, more generally for conflict resolution. Hence, approaches to argumentation based agent

---

[1] In philosophical parlance we are adopting an *externalist* rather than *internalist* view, where the latter consider norms to be intrinsically motivating.

reasoning often conform to these semantics, whereby an agent's inferences (e.g. denoting beliefs or goals) can be defined in terms of the claims of the justified arguments constructed from the underlying theory (an argument essentially being a proof of a candidate inference — the argument's claim — in the underlying logic). In [12, 13], Dung's semantics have been extended to accommodate arguments that *express preferences between other arguments*, where no assumption is made as to how these preferences are defined in the instantiating formalism.

In Section 2 we review the extended semantics described in [12, 13]. The main contributions of this paper are then described in Sections 3, 4 and 5. In Section 3 we describe a general framework for argumentation based resolution of conflicts between system norms and agent goals. Specifically, we combine the extended argumentation semantics with the normative model of [11] in which compliance with norms is enforced through punishments and rewards modelled as the goals of enforcement agents. The framework thus provides for social mechanisms for enforcing compliance, and motivational argumentation. Section 4 then describes a logic programming instantiation of the general framework. Section 5 illustrates the instantiation with an extended example. Finally, Section 6 concludes with a discussion of related and future work.

2 Extended Argumentation Frameworks for Agent Reasoning

2.1 Dung's Argumentation Framework

A Dung argumentation framework is a tuple $(Args, \mathcal{R})$ where $\mathcal{R} \subseteq (Args \times Args)$ can denote either an 'attack' or 'defeat' relation, and where the latter can be understood as an attack that succeeds given the available preference information. 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. Dung then defines the acceptable extensions of $(Args, \mathcal{R})$ under different extensional semantics. In this paper we focus on the admissible and preferred semantics. Letting $S \subseteq Args$ be conflict free if no two arguments in S are related by $\mathcal{R}$, then:

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*

An argument is said to be justified if it belongs to all preferred extensions of a framework.

2.2 Motivating Extended Argumentation Frameworks

We now motivate extending Dung's framework with the following example (that will be referred to later in Section 3).

Example 1. Consider two agents $Ag1$ and $Ag2$ exchanging arguments A, B ... about the weather forecast for Hawaii.

$Ag2$: "According to the BBC it will be cool in Hawaii" = A

$Ag1$: "According to CNN it will be hot in Hawaii" = B

$Ag2$: "But the BBC are more trustworthy than CNN" = C

$Ag1$: "However, statistics show CNN are more accurate than the BBC" = D

$Ag1$: "And a statistical comparison 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 the preferred extensions, and so neither argument is justified. 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 A's repulsion of, or defence against, B's attack on A; i.e., C attacks B**'s attack on** A ($\Delta2$ in Figure 1a)) 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. Evaluating the preferred extensions on the basis of $\mathcal{R}$ denoting the defeat relation, we now have the single preferred extension containing A. Now, given D claiming a preference for B over A and so attacking A's attack on B, neither defeat the other and so once again we have two preferred extensions. Since C and D claim contradictory preferences they attack each other ($\Delta3$). These attacks can themselves be subject to attacks in order to determine the defeat relation between C and D and, in so doing A and B. E attacks the attack from C to D ($\Delta4$), so that D defeats C, B defeats A, and $Ag1$'s argument that it will be hot in Hawaii is now justified.
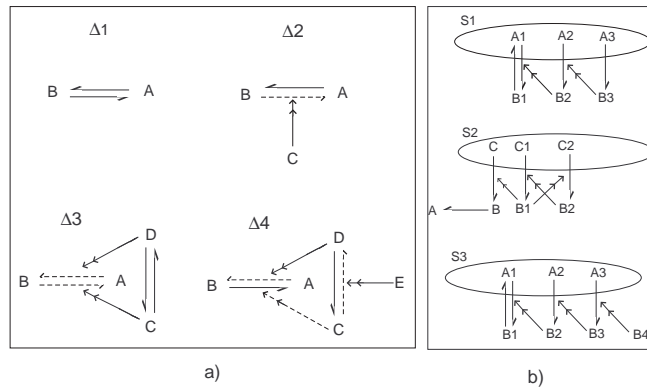

Fig. 1. Motivating *EAF*s

2.3 Defining Extended Argumentation Frameworks

Example 1 illustrates requirements for arguments attacking attacks. Hence, as in [12, 13], an *Extended Argumentation Framework* is defined as follows:

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

Notation 1 We may write $A \rightharpoonup B$ to denote $(A, B) \in \mathcal{R}$. If in addition $(B, A) \in \mathcal{R}$, then $A \rightleftharpoons B$. Also, $D \twoheadrightarrow (A \rightharpoonup B)$ denotes $(D, (A, B)) \in \mathcal{D}$

The defeat relation is now 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 defeats$_S$ B, denoted by $A \rightarrow^S B$, iff $(A, B) \in \mathcal{R}$ and $\neg \exists D \in S$ s.t. $(D,(A, B)) \in \mathcal{D}$.*

Referring to Example 1, A defeats$_\emptyset$ B but A does not defeat$_{\{D\}}$ B. The notion of a conflict free set S of arguments is now defined. Notice that it may be that an argument A asymmetrically attacks an argument B, so that given $D \twoheadrightarrow (A \rightharpoonup B)$, neither A nor B defeat$_S$ each other if $D \in S$. This means that both A and B may be accepted together in the same extension (where any extension is required to be conflict free). For example, if B is an argument for an action, and A claims that (for example) the action is too costly, it may be that an agent decides to execute the action while accepting that it is expensive (in value based argumentation [4], D is an argument claiming that the value promoted by B's action is greater than A's value of 'cost'). In the following section we will show that such *preference dependent asymmetric* attacks are also appropriate when resolving conflicts between norms and desires.

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

The definition of acceptability of an argument A w.r.t. a set S for an *EAF* is motivated in some detail in [12, 13]. It references the notion of a *reinstatement set* for a defeat, in order that an intuitive requirement on what it means for an argument to be acceptable w.r.t. an admissible set S of arguments is satisfied: *if A is acceptable with respect to S, then $S \cup \{A\}$ is admissible* (referred to as the fundamental lemma in Dung [8]).

Definition 5. *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.* $(Y',(X, Y)) \in \mathcal{D}$, $\exists X' \rightarrow^S Y' \in R_S$

Definition 6. *Let $(Args, \mathcal{R}, \mathcal{D})$ be an* EAF. $A \in Args$ *is acceptable w.r.t.* $S \subseteq Args$ *iff* $\forall B \in Args$ *s.t.* $B \to^S A$, $\exists C \in S$ *s.t.* $C \to^S B$ *and there is a* reinstatement set *for* $C \to^S B$.

In Figure 1b), $A1$ is acceptable w.r.t. $S1$. We have that $B1 \to^{S1} A1$, and $A1 \to^{S1} B1$. The latter defeat is itself *challenged* by $B2$. However, $A2 \to^{S1} B2$, which in turn is challenged by $B3$. But then, $A3 \to^{S1} B3$. We have the reinstatement set $\{A1 \to^{S1} B1, A2 \to^{S1} B2, A3 \to^{S1} B3\}$ for $A1 \to^{S1} B1$. Also, A is acceptable w.r.t. $S2$ given the reinstatement set $\{C \to^{S2} B, C1 \to^{S2} B1, C2 \to^{S2} B2\}$ for $C \to^{S2} B$. Finally $A1$ is not acceptable w.r.t $S3$ since no argument in $S3$ defeats$_{S3}$ $B4$.

Admissible and preferred semantics for *EAF*s are now given by Definition 1, where conflict free is defined as in Definition 4. (In [12, 13], the complete, stable and grounded semantics are similarly defined for *EAF*s, i.e., in the same way as for Dung frameworks). Referring to Example 1, $\{B, D, E\}$ is the single preferred extension. In [12, 13] we show that *EAF*s inherit many of the fundamental results holding for extensions of a Dung framework. This suggests that much of the work building on Dung's framework can readily be reformulated for *EAF*s, including work on argument game proof theories and dialogue frameworks. In particular, Dung's fundamental lemma is satisfied, implying that the set of all admissible extensions of an *EAF* form a complete partial order w.r.t. set inclusion, and so for each admissible S there exists a preferred extension S' such that $S \subseteq S'$.

To conclude, the extended semantics accommodates arguments that express preferences between other arguments, while preserving the abstraction of a Dung framework; no commitments are made to how preferences are defined in the instantiating logical formalism. We now make use of the extended semantics in a framework for conflict resolution in normative systems, and show that the ability to engage in argumentation based reasoning *about*, as well as *with*, defeasible and possibly conflicting preference information, provides for agent flexibility and adaptability.

3 A Framework for Conflict Resolution in Normative Systems

This section describes a framework in which agents engage in dialogues to decide which amongst conflicting desire based and normative goals are to be pursued. Agent submit arguments for goals, where these arguments attack each other, and then engage in motivational argumentation over the relative utility of states in which the goals are realised. This equates to arguing over preferences between arguments, and so which attacks succeeds as defeats. The arguments and attacks defined in the course of a dialogue thus instantiate an *EAF*, and the goals to be pursued are those claimed by the justified arguments of the *EAF*. Note that the agents also argue over the beliefs justifying adoption of goals. In this way, the agents are first required to agree that the goal being proposed for adoption is indeed warranted by what is believed to be the case. Section 3.1 first sets out some general assumptions about the kinds of agents modelled by the framework, and the dialogues these agents participate in. Section 3.2 then describes how conflicts between individual agent goals and system norms are resolved through argumentation based dialogues over beliefs, and goals proposed by individual agents and agents acting on behalf of the system.

3.1 Agents and Dialogues

The proposed framework abstracts from the logics for agent reasoning, assuming only *BDI* type agents (e.g. those instantiating the *BOID* architecture [7]) and a declarative interpretation of goals as beliefs holding in some future state. Each agent has a belief base consisting of facts and rules, and a goal base containing rules for deriving goals. From amongst all the goals that are derivable, those that an agent commits to realising are referred to as intentions. An intention persists in an agent's intention base until such a time as it is realised by a plan (the agent's planning component is not modelled here).

As in [11]'s model of normative multi-agent systems, four types of goal are distinguished. Individual agent goals, which we refer to here as *desires*, may conflict with *normative goals*. For example, an agent $Ag1$'s desire to stay on Waikiki beach in Hawaii, may conflict with the normative goal of staying in a cheap hotel. $Ag1$ may decide to comply or not comply with the norm, based on rewards and punishments exacted by system agents (specifically *enforcement* agents). Rewards and punishments are also individual goals of enforcement agents, but are punishment, respectively reward goals, from the perspective of the agent being punished, respectively rewarded. *Punishment goals* hinder the punished agent's intentions if that agent decides not to comply with the norm. For example, a punishment may be to deny the funding that $Ag1$ needs to fulfill its intention to visit Leipzig for a meeting. *Reward goals* benefit the achievement of the rewarded agent's intentions if it decides to comply. For example, a reward for an agent who intends to have a laptop, may be to provide the agent with a laptop.

In general, goals are derived by rules whose antecedents refer to what the agent believes and its current intentions. Extending the scenario described in Example 1, suppose agent $Ag1$ believes it will be hot in Hawaii, and it intends to attend a conference in Hawaii. Then it derives the desire to stay on Waikiki beach. The goals of system agents are derived in the same way, and may additionally refer to the intentions of other agents. For example, if $Ag1$ intends to attend a conference, then the normative goal of staying in a cheap hotel is derived (in either $Ag1$'s goal base or the goal base of a system agent responsible for informing other agents of their obligations). An enforcement agent Ag_P may derive the punishment goal of denying $Ag1$ the funding for a meeting, given $Ag1$'s intention to attend the meeting, and Ag_P's belief that the meeting is not related to an EU project. Rules in the goal base can also capture the sub-goal relationship. For example, if $Ag1$ intends to visit Leipzig for a meeting, then it derives the sub-goal goal of having funding for the visit. Finally, we assume argument construction from agents' bases is defined in some underlying logic.

Definition 7. *Let $\{Ag_1, \ldots, Ag_n\}$ be a set of agents, where for $i = 1 \ldots n$, Ag_i is equipped with a belief base $\mathcal{B}_i$, an intention base $\mathcal{I}_i$, and a goal base $\mathcal{G}_i$. For $i = 1$, let argument A be constructed from $\mathcal{B}i \cup \mathcal{G}i \cup \bigcup_{i=1}^{n} \mathcal{I}i$.*

If A is constructed only from $\mathcal{B}i$, then A is a belief argument *of Ag_i, otherwise A is a* goal argument *of Ag_i.*

In general, we write $bel(A)$ to denote the beliefs in A. We also write claim*(A) to identify an argument A's claim.*

The basic idea is that individual and system agents engage in argumentation-based conflict resolution (*persuasion*) dialogues to determine which amongst the arguments

for beliefs and goals are justified in the *EAF*s of Section 2. The goals that are the claims of justified arguments are then adopted as intentions. In persuasion dialogues (reviewed in [3]) a proponent makes a claim — the *topic* of the dialogue — and (one or more) opponents attempt to persuade the proponent that the claim does not hold. In general such a dialogue d is a sequence of moves $m_1, \ldots, m_i, \ldots$, where the first move m_1 is a locution introducing the topic as an assertion or claim of an argument. Here, we simply assume that the topic of d can be referred to as *topic(d)*. Dialogue *protocols* vary from model to model, and specify the legal moves at each stage of the dialogue, where a move can be an assertion of a proposition or an argument, a challenge to a premise in an argument, a concession of a proposition or argument, and so on. Models also vary on the rules for termination of a dialogue. However, in general, the arguments submitted and constructed (from the propositions asserted) during the course of a dialogue can be organised into an argumentation framework [15]. If an argument for the topic is justified, then the proponent wins the dialogue. Formalising dialogue models is to be addressed in future work. Here, we refer only to an *EAF* constructed on the basis of a dialogue.

Definition 8. *Let $d = m_1, \ldots, m_n$ be a terminated dialogue where* topic(d) = α, *and* $AG = \{Ag_1, \ldots, Ag_m\}$ *the participants in d. We say that the* EAF $\Delta = (Args, \mathcal{R}, \mathcal{D})$ *constructed on the basis of d, is:*

- *a belief EAF iff every argument in $Args$ is a belief argument of some $Ag \in AG$*
- *a goal EAF iff every argument in $Args$ is either a belief or goal argument of some $Ag \in AG$* [2]

3.2 Arguing about Beliefs and Goals

An agent's argument A for a desire may conflict with (and so mutually attack) an argument B for a normative goal. Arguments for punishment and reward goals may in turn attack A and so reinstate the argument B for the normative goal. The success of these attacks as defeats depends on argumentation over preferences between the arguments (corresponding to meta-level motivation-based argumentation over the relative utility of states in which the goals are realised).

Prior to agents submitting goal arguments in a dialogue, the beliefs in the argument justifying the goal may themselves by subject to debate [3]. In our running example, $Ag1$'s desire to stay on Waikiki beach is contingent on its belief that it will be hot in Hawaii. A system agent may successfully persuade $Ag1$ that it will be cool in Hawaii. Hence $Ag1$ will not submit the argument for its desire, precluding the possibility of norm violation (in Example 1 the outcome is in favour of $Ag1$'s argument that it *will* be hot). Furthermore, the beliefs in arguments for system goals may be challenged.

---

[2] Of course, in the limiting case where only arguments can be submitted as locutions, then each m_i in d corresponds to an argument in $Args$, and a protocol for d would require that m_i attack some m_j, $j < i$, or some attack between m_j and m_k, $j < i$, $k < i$

[3] Arguing over beliefs justifying a goal *prior* to arguing over the relative merits of goals precludes 'wishful thinking'; i.e., one wouldn't want that argumentation over which goals to adopt (which future state to realise) influences what is believed about the current state of the world.

Thus, an agent may successfully argue that the beliefs justifying a normative goal may be erroneous; hence the normative goal does not have to be adopted and unwarranted obstruction of the conflicting desire is prevented. Suppose arguments A and B for the conflicting goals of staying on Waikiki and staying at a cheap hotel have been submitted. Ag_P will not submit an argument C for the punishment goal of denying $Ag1$ funding for the Leipzig meeting, if $Ag1$ successfully persuades Ag_P that the meeting *is* related to an EU project. Again, this prevents unwarranted obstruction of $Ag1$'s intention to attend the meeting. Of course, Ag_P may then be motivated to submit an argument for an alternative punishment goal to enforce compliance.

Definition 9. *Let $AG = \{Ag_1, \ldots, Ag_n\}$. Then A is an* **agreed** *goal argument of $Ag \in AG$ if for every $\alpha \in bel(A)$:*
if there is a terminated dialogue d with topic α, participants $AG \subseteq \{Ag_1, \ldots, Ag_n\}$, and Δ is a belief EAF constructed on the basis of d, then α is the conclusion of a justified argument of Δ.

We now describe how argumentation over goals proceeds. Consider the case where a normative goal g' conflicts with a desire g (in the simplest case $g' \equiv \neg g$ in the underlying logic). In general, we say that the goal argument A' for g' *conflicts* with the goal argument A for g. In a goal *EAF*, A and A' attack each other since an agent can either adopt g and not g', or g' and not g.

Suppose such an EAF, where $Ag1$ submits A claiming 'stay on Waikiki beach', and A' claiming 'stay in cheap hotel', mutually attacks A. An enforcement agent can then submit an argument P for a punishment goal p, that, in the terminology of [11], *hinders* some intention of $Ag1$. In our running example, p = 'deny funding for meeting'. Now P does not directly attack on A's goal; it does so in the sense that if the attack succeeds, then $Ag1$ will not pursue its desire, and will comply with the norm. Note also, that the attack is a preference dependent asymmetric attack. $Ag1$ might argue (B) that it is of more value to him to stay on Waikiki beach then attend the meeting. That is, $B \twoheadrightarrow (P \rightharpoonup A)$, and it may now be that A *and* P are justified; $Ag1$ adopts its desire, *and* accepts the punishment. An alternative punishment may then need to be submitted to see if it has the required enforcing effect. Finally, an enforcement agent can submit an argument R for a reward goal r, that, in the terminology of [11], *benefits* some intention of Ag. For example r = 'provide the agent with a laptop', benefiting $Ag1$'s intention to have a laptop. R symmetrically attacks A. Either $Ag1$ accepts the reward and drops the desire, or vice versa.

Definition 10. *Let $AG = \{Ag_1, \ldots, Ag_n\}$ be a set of agents. Let $Args_G = \bigcup_{i=1\ldots n}\{A | A$ is a goal argument for $Ag_i\}$. Let $\mathcal{I}_{AG} = \bigcup_{i=1\ldots n} \mathcal{I}i$. Then:*

- conflicts $\subseteq Args_G \times Args_G$
- hinders $\subseteq Args_G \times \mathcal{I}_{AG}$
- benefits $\subseteq Args_G \times \mathcal{I}_{AG}$[4]

---

[4] Note that an agent's desires may 'internally' conflict. We so not here directly address conflict resolution in such cases. Note also that an agent's goals may benefit/hinder its own intentions.

Definition 11. *Let $AG = \{Ag_1, \ldots, Ag_n\}$, and for some $Ag \in AG$, let A be a goal argument of Ag, $\mathcal{I}$ the intention base of Ag.*
Let A' be the goal argument of some $Ag' \in AG$, $Ag' \neq Ag$. Then:

- *A' goal attacks A and A goal attacks A' if conflicts(A', A) or benefits(A', ι) for some $\iota \in \mathcal{I}$*
- *A' goal attacks A if hinders(A', ι) for some $\iota \in \mathcal{I}$*

We now specify some constraints on a dialogue that begins with a topic that is a goal proposed for adoption as an intention. We do so by expressing constraints on the goal *EAF* constructed on the basis of the dialogue. These are that the goal arguments are agreed, and can only be attacked by goal arguments as defined above, and only belief arguments are used in arguing over the relative merits of the goals.

Definition 12. *Let $AG = \{Ag_1, \ldots, Ag_n\}$ be a set of agents. Let d be a terminated dialogue with topic α, and participants $AG' \subseteq AG$, where:*

- *α is the conclusion of an agreed goal argument A of some agent $Ag \in AG'$.*
- *$\Delta = (Args, \mathcal{R}, \mathcal{D})$ is the goal EAF constructed on the basis of d, where:*
  i) for any goal arguments $B, A \in Args$, $(B, A) \in \mathcal{R}$ iff A and B are agreed goal arguments, and B goal attacks A.
  ii) If $(C, (B, A)) \in \mathcal{D}$ then C is a belief argument for some agent in AG'

If the topic α of the dialogue is an agent's desire, and α is the claim of a justified argument in the dialogue's goal *EAF*, then α is updated to the agent's intention base, and any punishment goal that is the claim of a justified argument is updated to the corresponding enforcement agent's intention base. If α is not the claim of a justified argument, and there is a justified argument for a normative goal β, then β is updated to the agent's intention base, and any reward goal that is the claim of a justified argument is updated to the corresponding enforcement agent's intention base.

4 Instantiating the Framework

In this section we describe an example instantiation of the framework. Agent goals, beliefs and intentions are represented in [14]'s *argument based logic programming with defeasible priorities* (ALP-DP). An ALP-DP theory's arguments are defined as sequences of chained rules. Some rules can express priorities on other rules, so that one can construct *priority arguments* whose claims determine preferences between other mutually attacking arguments. Preferences between priority arguments can also be established on the basis of other priority arguments. [14] then defines the justified arguments of a theory under Dung's grounded semantics. In [12, 13] the arguments and attacks defined by an ALP-DP theory instantiate an *EAF*, and an equivalence result with the *EAF*'s justified arguments (under the grounded semantics) is shown. By giving an *EAF* semantics for ALP-DP one can, unlike [14], also:

1. characterise the justified arguments of an ALP-DP theory under the preferred semantics; and

2. model preference dependent asymmetric attacks.

Both these features are employed when instantiating an *EAF*. Note that ALP-DP models both negation as failure and strict negation. To simplify the presentation, we describe a restricted version of ALP-DP — ALP-DP* — which does not include negation as failure.

Definition 13. *Let (S, D) be a ALP-DP* theory where S is a set of strict rules of the form $s : L_0 \wedge \ldots \wedge L_m \to L_n$, 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. Henceforth,* head(r) *denotes 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 $\neg$.*

Strict rules represent information that is beyond debate (note that neither $\to$ nor $\Rightarrow$ admit contraposition). We also assume that the language contains a two-place predicate symbol $\prec$ for expressing priorities on rule names, and that any S includes the following strict rules expressing the properties of a strict partial order on $\prec$:

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

Definition 14. *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 ($0 \le i \le n$), for every literal L_j in the antecedent of r_i there is a $k < i$ such that* head(r_k) $= L_j$.
     We say that claim(A) $=$ head(r_n), *and 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' and* claim(A) $=$ head(r_{n_1})$\ldots$head(r_{n_m}) *is the ordering claimed by A*

In [14], arguments are exclusively defined by item 1. We additionally define composite priority arguments so that an ordering, and hence a preference, can be claimed by a single argument rather than a set of arguments (as in [14]).

Definition 15. *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*

Note that an argument has only one claim, but may have many conclusions corresponding to the heads of the contained rules. We now instantiate the abstract definition 7 of an agent and its constructed arguments. Note that intentions are represented by the goal arguments that have previously been used to justify their adoption. Hence, an agent's goal arguments will be constructed from its belief and goal base, and the claims (named by the name of the rule whose head is the claim) of intention arguments in all agents' intention bases.

Definition 16. *Let $\{Ag_1, \ldots, Ag_n\}$ be a set of agents, where for $i = 1 \ldots n$:*
- $\mathcal{B}_i$ and $\mathcal{G}_i$ are ALP-DP* theories, and $\mathcal{I}_i$ is a set of arguments.
- A is a belief argument of Ag_i iff it is based on $\mathcal{B}_i$
- A is a goal argument of Ag_i iff it is based on $\mathcal{B}_i \cup \mathcal{G}_i \cup \bigcup_{i=1 \ldots n} \{r : claim(B) | B \in \mathcal{I}_i, head(r) = claim(B)\}$

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

Definition 17. *A1 attacks A2 on the pair $(L, \neg L)$ if 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 $\neg L$.*

In the following example illustrating attacks between belief arguments, we will without loss of generality simply assume that all beliefs are contained in a single theory. Only in the example at the end of this section, in which we illustrate argumentation over goals, will we identify the individual agents involved. Following [14], every rule with terms $t_1, \ldots, t_n$ is named with a function expression $r(t_1, \ldots, t_n)$ where r is the rules's informal name. For example, $r(p(X, Y), q(X, Y))$ names the rule $p(X, Y) \Rightarrow q(X, Y)$. To maintain readability we will only write the function-symbol part of the rule name, and as an abuse of notation, arguments will be represented as sequences of rule names rather than the rules these names identify.

Example 2. Let $tr(X, Y)$, $st(X, Y)$ and $ra(X, Y)$ respectively denote that X is more trustworthy, statistically accurate, and rational than Y.
Let $S = \{o1 \ldots o4\} \cup \{s1 : temp(X, cool) \rightarrow \neg temp(X, hot),$
$\qquad\qquad\qquad s2 : temp(X, hot) \rightarrow \neg temp(X, cool)\}$.
Let $D = \{bbc :\Rightarrow temp(hawaii, cool),$
$\qquad\qquad cnn :\Rightarrow temp(hawaii, hot),$
$\qquad\qquad c1 :\Rightarrow tr(bbc, cnn),$
$\qquad\qquad d1 :\Rightarrow st(cnn, bbc),$
$\qquad\qquad c2 : tr(X, Y) \Rightarrow Y \prec X,$
$\qquad\qquad d2 : st(X, Y) \Rightarrow Y \prec X,$
$\qquad\qquad e1 :\Rightarrow ra(d2, c2),$
$\qquad\qquad e2 : ra(X, Y) \Rightarrow Y \prec X\}$
$A = [bbc]$, $B = [cnn]$, $C = [c1, c2]$, $D = [d1, d2]$.
$E = [e1, e2]$ with conclusions $ra(d2, c2)$ and $c2 \prec d2$, and claim $c2 \prec d2$.

A and B attack each other since $A + s1$ has conclusion $\neg temp(hawaii, hot)$ and B has conclusion $temp(hawaii, hot)$. C and D attack each other since C has conclusion $cnn \prec bbc$ and $D + o4$ has conclusion $\neg(cnn \prec bbc)$

We now define the relations *conflicts*, *hinders* and *benefits*, and goal attacks for ALP-DP* goal arguments. Note that the notion of *benefits* requires that a goal argument of a rewarding agent be extended (as in the definition of attack) with strict rules that link the reward goal to the intention that it benefits (this will be illustrated in the example concluding this section).

Definition 18. *Let A be a goal argument of an agent Ag and $\mathcal{I}$ the intention base of Ag.*
Let B be any goal argument of an agent Ag', where $\mathcal{B}' = (S', D')$ is the belief base of Ag'. We say that:

● conflicts*(B, A) if B attacks A as in definition 17.*

For any $I \in \mathcal{I}$:

● hinders*(B, I) if B attacks I as in definition 17*

● benefits*(B, I) if* claim*($B + S1$) =* claim*(I) for some possibly empty sequence of strict rules $S1$ in S'*

Then:

  – *B and A goal attack each other on the pair $(claim(B), claim(A))$ if* conflicts*(B, A)*
  – *B and A goal attack each other on the pair $(claim(B), claim(A))$ if* benefits*(B, I) for some $I \in \mathcal{I}$*
  – *B goal attacks A on the pair $(claim(B), claim(A))$ if* hinders*(B, I) for some $I \in \mathcal{I}$*

To determine a preference amongst attacking arguments, [14] defines the sets of relevant *defeasible* rules to be compared, and an ordering on these sets. Here, the ordering on such sets is based on the ordering claimed by a given priority argument.

Definition 19. *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 $s : L_1 \wedge \ldots \wedge L_n \to L$*

Definition 20. *Let C be a priority argument claiming 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 $r \prec r'$.*

Intuitively, 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 for belief arguments they attack on more than one conclusion. For goal arguments they goal attack on a single pair of conclusions (the goals claimed by the arguments). Given a priority ordering $\prec$ claimed by 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 21. *Let C be a priority argument claiming $\prec$. Let $(L_1, L'_1), \ldots, (L_n, L'_n)$ be the pairs on which A attacks, or goal attacks B, 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, C and D attack each other on the pair $(cnn \prec bbc, \neg(cnn \prec bbc))$, and $R_{cnn \prec bbc}(C) = \{c2\}$, $R_{\neg(cnn \prec bbc)}(D) = \{d2\}$. E claims $c2 \prec d2$, and so D is preferred$_{c2 \prec d2}$ to C. Note also, that given C, A is preferred$_{cnn \prec bbc}$ to B, and given D, B is preferred$_{bbc \prec cnn}$ to A. We can now instantiate an *EAF* with the arguments, their attacks, and priority arguments claiming preferences and so attacking attacks:

Definition 22. *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 14, and $\forall A, B, C \in Args$:*

1. *$(C,(B, A)) \in \mathcal{D}$ iff C claims $\prec$ and A is preferred$_\prec$ to B*
2. *$(A, B),(B, A) \in \mathcal{R}$ if A and B attack as in definition 17, or A and B goal attack as in definition 18*

The belief *EAF* obtained by the arguments and attacks for our running example is shown in figure 1a). $\{E, D, B\}$ is the single preferred extension of the *EAF*. We can now constrain a goal EAF constructed on the basis of a dialogue between agents, as defined in definition 12.

5   An Extended Example

We now illustrate the previous section's formalism with an extended version of our Hawaiian example, in which we assume that every goal argument is agreed.

In what follows we use the following shorthand:

ha = 'Hawaii', wa = 'Waikiki beach', le = 'Leipzig', att, = 'attend', $conf$ = 'conference', $meet$, = 'meeting', $cheap$ = 'cheap hotel', lap = 'laptop', $fund$ = 'have funding', and $deny\_f$ = 'deny funding'.

Also, predicates may refer to the agents themselves. For example, $att(ag, conf, ha)$ denotes the goal of ag to attend a conference in Hawaii. Also, variables will begin with uppercase letters and constants with lowercase letters. For example, $deny\_f(ag_P, AgX, meet, L)$ denotes the goal of agent ag_P to deny funding for any agent AgX to attend a meeting in some location L.

Let $\{ag_a, ag_N, ag_P, ag_R\}$ be a set of agents. We describe each agent's knowledge bases. Note that we may not show all the goal rules used to construct arguments in the intention base of each agent. Also, as before, we may simply write the rule name rather than the rule the name identifies.

ag_a:

$\mathcal{I} =$

$\{ [ia1 :\Rightarrow att(ag_a, conf, ha)], [ia2 :\Rightarrow att(ag_a, meet, le)],$

$[ia2 :\Rightarrow att(ag_a, meet, le), ia3 : att(ag_a, meet, le) \Rightarrow funds(ag_a, meet, le)],$

$[ia4 :\Rightarrow have(ag_a, lap)] \}$

$\mathcal{G} =$

$\{ ga1 : temp(ha, hot) \wedge att(ag_a, conf, ha) \Rightarrow stay(ag_a, wa)\}$

$\mathcal{B} =$

$\{ ba0 :\Rightarrow temp(ha, hot),$

$ba1 :\rightarrow norm\_des(gn1, ga1),$

$ba\_self : norm\_des(X, Y) \Rightarrow X \prec Y,$

$ba2 :\Rightarrow project\_funds(high),$

$ba3 : project\_funds(high) \Rightarrow except(ba\_self, bn\_social),$

$ba\_excep : except(X, Y) \Rightarrow Y \prec X,$

$ba4 :\Rightarrow gp1 \prec ga1\}$

ag_N:

$\mathcal{I} = \emptyset$

$\mathcal{G} = \{ gn1 : att(AgX, conf, L) \Rightarrow stay(AgX, cheap, L)\}$

$\mathcal{B} = \{ bn1 : stay(AgX, cheap, ha) \rightarrow \neg stay(AgX, wa),$

$bn2 :\rightarrow norm\_des(gn1, ga1),$

$bn\_social : norm\_des(X, Y) \Rightarrow Y \prec X \}$

ag_P:

$\mathcal{I} = \emptyset$

$\mathcal{G} = \{ gp1 : att(AgX, meet, L) \wedge \neg type(meet, eu, L)$

$\Rightarrow deny\_f(ag_P, AgX, meet, L)\}$

$\mathcal{B} = \{ bp1 :\Rightarrow \neg type(meet, eu, le),$

$bp2 : deny\_f(ag_P, AgX, meet, L) \rightarrow \neg funds(AgX, meet, L)\}$

ag_R:

$\mathcal{I} = \emptyset$

$\mathcal{G} = \{ gr1 : have(AgX, lap) \Rightarrow provide(ag_R, AgX, lap)\}$

$\mathcal{B} = \{ br1 : provide(ag_R, AgX, lap) \rightarrow have(AgX, lap),$

$br2 :\rightarrow rew\_des(gr1, ga1),$

$br\_rew\_suffice : rew\_des(X, Y) \Rightarrow Y \prec X \}$

1) ag_a initiates a dialogue with goal argument $A1 = [ba0, ia1, ga1]$ claiming the goal $stay(ag_a, wa)$, having already persuaded a system agent that it will indeed be hot in Hawaii.

2) ag_N submits $A2 = [ia1, gn1]$ ($AgX = ag_a$, $L = ha$), where $A2$ and $A1$ goal attack each other (see figure 2) on the pair

$(stay(ag_a, cheap, ha), stay(ag_a, wa))$.

This symmetric goal attack is based on *conflicts* $(A2, A1)$ which obtains because $A2 +$ $[bn1]$ and $A1$ attack (as in def.17) on the conclusion pair $(\neg stay(ag_a, wa), stay(ag_a, wa))$

ag_N also submits the social ordering argument $B1 = [bn2, bn\_social]$ claiming $ga1 \prec$ $gn1$, and so $B1 \twoheadrightarrow (A1 \rightharpoonup A2)$.
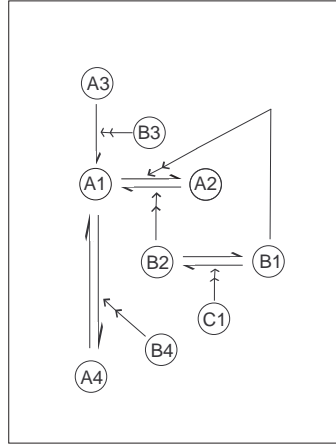
Fig. 2. *EAF* based on argumentation based dialogue over goals

3) ag_a submits:

- the selfish ordering argument $B2 = [ba1, ba\_self]$ claiming $gn1 \prec ga1$, and so $B2 \twoheadrightarrow (A2 \rightharpoonup A1)$

- an argument claiming that the selfish behaviour type is preferred to the social behaviour type given the exceptional circumstances in which the remaining project budget is high:

$C1 = [ba2, ba3, ba\_excep,]$ claiming $bn\_social \prec ba\_self$, and so $C1 \twoheadrightarrow (B1 \rightharpoonup B2)$.

The single preferred extension contains $A1$

4) ag_P attempts to enforce compliance by submitting $A3 = [ia2, bp1, gp1]$ given that it is agreed that the meeting is not an Eu project meeting.

$A3 + [bp2]$ attacks (as in def.17), and so hinders, ag_a's intention $[ia2, ia3]$. Hence, $A3$ goal attacks $A1$ on the pair $(deny\_f(ag_P, ag_a, meet, le), stay(ag_a, wa))$.

5) However, ag_a prefers to stay on the beach and be denied funding by ag_P for the leipzig meeting. It may be that ag_a has another source of funding in mind. We do not encode the rationale for the preference, but simply assume the priority argument $B3 = [ba4]$ claiming $gp1 \prec ga1$. Hence $B3 \twoheadrightarrow (A3 \rightharpoonup A1)$. Since $A3$'s attack on $A1$ is asymmetric:

The single preferred extension contains $A1$ **and** $A3$

6) ag_R attempts to enforce compliance with $A4 = [ia4, gr1]$ offering to provide ag_a with a laptop. This benefits ag_a's intention to have a laptop since $claim([ia4, gr1] + [br1]) = claim[ia4]$. Hence, $A4$ and $A1$ goal attack each other ($A4 \rightleftharpoons A1$) on the pair $(provide(ag_R, ag_a, lap), stay(ag_a, wa))$.

ag_R believes the reward is of sufficient strength that ag_a will prefer the reward to staying on Waikiki beach. ag_R submits $B4 = [br2, br\_rew\_suffice]$ claiming $ga1 \prec gr1$. Hence, $B4 \twoheadrightarrow (A1 \rightharpoonup A4)$. This is accepted by ag_a and the dialogue terminates.

The single preferred extension contains $A2$ **and** $A4$

ag_a's intention set can then be updated with $A2$. ag_R's intention set can then be updated with $A4$. ag_a intends now to book a cheap room in Hawaii, and ag_R intends to provide ag_a with a laptop.

6 Conclusions

In this paper we have proposed a framework for argumentation-based resolution of conflicts in normative multi-agent systems, and have illustrated instantiation of the framework with a logic programming formalism. The framework provides for agents to argue over the beliefs justifying goals, conflicting preferences brought to bear in argumentation over beliefs, and metalevel motivational argumentation over the states represented by desire based goals, and normative, punishment and reward goals argued for by other agents. In this way, unwarranted obstruction of individual agents' desires is precluded, and enforcement of compliance can appropriately account for the motivations of the agents and erroneously held beliefs about the contexts in which the agents find themselves.

As mentioned in Section 1, existing approaches to argumentation-based resolution of conflicts amongst goals ([2],[10],[16]) do not model social mechanisms deployed to enforce compliance with norms. In [16], norms are represented as bridge rules that describe the relationships between mental attitudes. Argumentation based resolution of conflicts amongst goals derived using these rules exploits a preference relation on these rules. In [2], only conflicts amongst desire based goals are addressed. Argumentation over the beliefs that justify desires conforms to the Dung semantics. However selection of desires does not account for their relative importance and does not conform to the Dung semantics. Rather, the maximal (under set inclusion) sets of desires that can be consistently realised are chosen. However, goal selection *does* account for the feasibility of plans for realising goals, and this is a factor that our work needs to account for in future work.

Future work will also investigate instantiation of the framework by formalisms with explicit BDI type modalities. Further work is also required before evaluation of the framework based on prototypical implementations can proceed. In particular, we intend development of argument game proof theories, algorithms and dialogue protocols for *EAF*s. Since *EAF*s inherit the fundamental results shown for Dung frameworks, our approach will adopt the methodologies deployed in specification of game based algorithms [18] and protocols [15] based on the Dung semantics. We also believe that our approach is applicable to resolution of conflicts arising between an individual agent's conflicting desires, and between conflicting norms. Both cases often require reasoning about abstract values and motivations. Furthermore, conflict resolution may lead to refinement and evolution of a system's norms. Finally, one of the key novel features of our framework is that an agent's decision as to which goals to pursue is influenced by other agents' goals. We believe that we can abstract from the normative application of the framework to consider other contexts in which the impact of other agents' goals can be modelled through argumentation based mechanisms.

References

1. L. Amgoud. Using Preferences to Select Acceptable Arguments. In: *Proc. 13th European Conference on Artificial Intelligence*, 43-44, 1998.
2. L. Amgoud and S. Kaci. On the Generation of Bipolar Goals in Argumentation-Based Negotiation. In: *Proc. 1st Int. Workshop on Argumentation in Multi-Agent Systems*, 2004.
3. ASPIC Deliverable D2.1: Theoretical frameworks for argumentation. http://www.argumentation.org/ Public Deliverables.htm, June 2004.
4. T. J. M. Bench-Capon. Persuasion in Practical Argument Using Value-based Argumentation Frameworks, *Journal of Logic and Computation*, 13(3), 429-448, 2003.
5. A. Bondarenko, P. M. Dung, R. A. Kowalski and F. Toni. An abstract, argumentation-theoretic approach to default reasoning, *Artificial Intelligence*, 93:63-101, 1997.
6. J. Broersen, M. Dastani, J. Hulstijn and L. W. N. van der Torre. Goal Generation in the BOID Architecture. In: Cognitive Science Quarterly Journal, 2(3-4), 428-447, 2002.
7. M. Dastani and L. van der Torre. Programming BOID-Plan Agents: Deliberating about Conflicts among Defeasible Mental Attitudes and Plans. In: *Proc 3rd Int. Joint Conference on Autonomous Agents and Multiagent Systems*, 706-713, 2004.
8. 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.
9. M. d'Inverno and M. Luck. Understanding agent systems. *2nd edn Springer-Verlag*.
10. A. Kakas and P. Moraitis. Argumentation based decision making for autonomous agents. In: *Proc. Second international joint conference on autonomous agents and multiagent systems*, 883–890, 2003.
11. F. Lopez Y Lopez, M. Luck and M. D'Inverno. A normative framework for agent-based systems. In: *J. Computational and Mathematical Organization Theory*, 12 (2-3):227–250, 2006.
12. S. Modgil. An Abstract Theory of Argumentation That Accommodates Defeasible Reasoning About Preferences. In: *9th European Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty*, 648–659, 2007.
13. S. Modgil. *Reasoning About Preferences in Argumentation Frameworks*. Technical Report: http://www.dcs.kcl.ac.uk/staff/modgilsa/ArguingAboutPreferences.pdf
14. H. Prakken and G. Sartor. Argument-based extended logic programming with defeasible priorities, *Journal of Applied Non-Classical Logics*,7:25–75, 1997.
15. H. Prakken. Coherence and flexibility in dialogue games for argumentation. In: *Journal of logic and computation* 15 (6):1009–1040 , 2005.
16. D. Gaertner, F. Toni. Conflict-free normative agents using assumption-basedargumentation. In: *Proc. 4th International Workshop on Argumentation in Multi-Agent Systems*, Hawaii, 2007.
17. Y. Moses and M. Tennenholtz. Artificial Social Systems. In: *Computers and AI*, 14(6), 533–562, 1995.
18. G. Vreeswijk. An algorithm to compute minimally grounded and admissible defence sets in argument systems. In: *Proc. 1st International Conference on Computational Models of Argument*,109-120, 2006.