Nested Argumentation and Its Application to Decision Making over Actions

S. Modgil

Advanced Computation Lab, Cancer Research UK, London WC2A 3PX
sm@acl.icnet.uk

Abstract. In this paper we describe a framework in which the grounds for one argument's defeat of another is itself subject to argumentation. Hence, given two conflicting arguments, each of which defeat the other, one can then determine the preferred defeat and hence the preferred argument. We then apply this nested argumentation to selection of an agent's preferred 'instrumental' arguments, where each such argument represents a plan of actions for realising an agent's goals.

1 Introduction

There is a growing body of work addressing the uses of argumentation in agent applications. Many of these works define an argumentation system for construction of arguments, and then instantiate Dung's framework [6] to determine which arguments are 'justified' or 'preferred' on the basis of the ways in which they interact. The interactions considered include the binary relations of *attack* and *defeat*. The former represents that two arguments conflict with each other. The latter additionally accounts for some relative valuation of the strength of two attacking arguments. However, given two mutually attacking arguments *A1* and *A2*, it may well be that there are grounds for *defeat(A1,A2)* **and** *defeat(A2,A1)*. For example, strengths of arguments may be evaluated on the basis of different criteria, so that *A1* defeats *A2* based on criterion c, and *A2* defeats *A1* based on criterion c'. Also, for any given criterion, evaluation of an argument's strength may vary according to the context in, or the perspective from, which it is evaluated. For example, reference to one information source for determining argument strength may indicate that *A1* defeats *A2*, whereas from the perspective of another information source, *A2* may defeat *A1*. Given two 'conflicting defeats' *defeat(A1,A2)* and *defeat(A2,A1)*, then one cannot establish which of *A1* or *A2* is preferred. However, such a preference can be established if one can determine which *defeat is preferred*.

We therefore propose that the reasoning underlying relative evaluation of the strength of two attacking arguments should itself be subject to argumentation. Hence, one constructs two 'level 2' arguments *B1* and *B2*, respectively providing grounds for *defeat(A1,A2)* and *defeat(A2,A1)*. To determine which of these conflicting defeats is preferred, we need to determine a preference between the mutually attacking arguments *B1* and *B2*. This in turn requires construction of 'level 3' arguments: *C1* providing grounds for *defeat(B1,B2)* or *C2* providing

S. Parsons et al. (Eds.): ArgMAS 2005, LNAI 4049, pp. 57–73, 2006.

grounds for *defeat(B2,B1)*. Of course, one might be able to construct both *C1* and *C2*, in which case one ascends to another level to determine which of these are preferred. In principle, this nested argumentation can continue indefinitely. Reasoning about the relative strength of arguments is also explored in [9, 11]. They do so by extending the object level language for argument construction with rules that allow context dependent inference of possibly conflicting relative prioritisations of rules. Thus, argument strength is exclusively based on rule priorities. The framework proposed here allows for argument strength to be based on any number of criteria. Furthermore, our framework formalises reasoning about the strength and defeats amongst arguments at the meta rather than object level. These requirements are of particular relevance to the use of argumentation in agent applications.

The issue of conflicting defeats is particularly relevant for agent applications, given the general requirement for a context dependent account of agents' cognitive processes. Specifically, a number of recent works [1, 2, 4, 8, 9] extend theories of argumentation over beliefs, to argumentation over agents' desires and intentions. For example, Amgoud [1, 2], and subsequently Hulstijn [8], define construction of *instrumental* arguments composed of actions and sub-goals for realising some top level goal (these arguments can be thought of as unscheduled plans). The idea is to then choose the preferred instrumental arguments so as to determine which plans the agent should adopt. However, the argumentation systems proposed do not straightforwardly instantiate Dung's framework. Furthermore, 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. However, in practical settings, strengths of arguments need to be established on the basis of multiple additional criteria such as the efficacy and temporal and financial costs of a plan's actions with respect to their goals. This implies a need to handle conflicting defeats in order to determine the preferred instrumental arguments. This need may also be a requirement for argumentation-based multi-agent dialogues [12], where the agents represent different perspectives from which communicated arguments are evaluated.

The main contributions of this paper are as follows. In section 2 we formalise nested argumentation over nested Dung argumentation frameworks. In section 3 we modify and build on Amgoud's system [1, 2] for constructing instrumental arguments. In particular our system is able to instantiate a Dung framework without adapting Dung's central definitions. In section 4 we apply nested argumentation to decide the preferred instrumental arguments on the basis of multiple information sources and criteria. In section 5 we conclude with a discussion of related and future work.

2 Nested Argumentation

Arguments can be said to *rebut* attack or *undercut* attack. In the former case the attack is symmetric; *attack(A1,A2)* and *attack(A2,A1)*. An example of a rebut attack is when the claim of *A1* conflicts with the claim of *A2*. Defeat additionally accounts for some relative valuation of the strength of attacking arguments: *defeat(A1,A2)* if *attack(A1,A2)* and it is not the case that *A2* is stronger than *A1*. Hence, in the case of a rebut attack, *defeat(A1,A2)* **and** *defeat(A2,A1)* if: **i)** their are no grounds for determining the relative strengths of *A1* and *A2*, or **ii)** there are grounds for *A1* being stronger than *A2*, **and** grounds for *A2* being stronger than *A1*.

Unlike rebut attacks, undercut attacks are asymmetric; *attack(A1,A2)* but not *attack(A2,A1)*. We support the view ([3, 11]) that one should not distinguish between undercut attacks and defeats; i.e., undercut defeats should not depend on the relative strength of arguments. To illustrate, consider a Pollock undercut defeat [10] whereby the claim of argument *A1* denies that the premises of *A2* support its claim (an attack on the link between premises and claim of *A2*). Pollock requires that *A2* is not stronger than *A1*. This leads to unintuitive results: if *A2* is stronger than *A1*, or information regarding their relative strength is missing, then neither argument defeats or attacks each other, and hence both arguments can be coherently held to be acceptable.

As discussed in section 1, we aim at a framework in which argumentation over the grounds for one argument being stronger than another can be used to resolve conflicting defeats of type **ii)** above. In this way one can determine a preference amongst mutually defeating arguments. We begin with two notions of a Dung argumentation framework, and then give Dung's standard definition of the preferred extensions of an argumentation framework.

Definition 1. *Let Args be a finite set of arguments. An argumentation framework AF is a pair (Args, Attack), where Attack* $\subseteq$ *(Args* $\times$ *Args). A justified argumentation framework JAF is a pair (Args, Defeat), where Defeat* $\subseteq$ *(Args* $\times$ *Args).*

Definition 2. *For any set of arguments S:*

- *S is **conflict free** iff no argument in S is defeated(attacked) by an argument in S.*
- *An argument A is **acceptable** w.r.t. S iff each argument defeating (attacking) A is defeated (attacked) by an argument in S.*
- *A conflict free set of arguments S is **admissible** iff each argument in S is acceptable with respect to S.*
- *A conflict free set of arguments S is a **preferred extension** iff it is a maximal (w.r.t. set inclusion) admissible set.*

Definition 3. *Let* $\{S_1, \ldots, S_n\}$ *be the preferred extensions of JAF = (Args,Defeat)*[1]*. Then* $\bigcap_{i=1}^{n} S_i$ *is the set of preferred arguments of JAF (denoted **Pf**(JAF))*

[1] Note that there will be a finite number of preferred extensions given the restriction in definition 1 to argumentation frameworks with a finite number of arguments.

We now define nested argumentation frameworks of the form $(AF_1, \ldots, AF_n)$. We make some minimal assumptions about the argumentation system instantiating each AF. In particular, each argument A in a system has a claim *claim(A)* (we write *claims(S)* to denote $\{claim(A) \mid A \in S\}$), and for AF_i, $i > 1$, the language for argument construction is a first order language whose signature contains the binary predicate symbol *defeat* and a set of constants $\mathsf{f\_name}_{i-1}(Args_{i-1})$ $= \{\mathcal{A}_1, \ldots, \mathcal{A}_n\}$ naming arguments in $Args_{i-1}$.

Definition 4. *A nested argumentation framework (NAF) is an ordered finite set of argumentation frameworks* $((Args_1, Attack_1),\ldots,(Args_n, Attack_n))$ *such that for* $i = 1 \ldots n\text{-}1$, $Attack_i \supseteq \{(A,A') \mid defeat(\mathcal{A},\mathcal{A}') \in claims(Arg_{i+1})\}$.

Given a NAF $(AF_1,\ldots,AF_n)$, we now define a justified NAF, mapping each AF_i to a JAF_i. Intuitively, an AF_{i+1} argument B with claim $defeat(\mathcal{A}',\mathcal{A})$ provides the grounds for an AF_i argument A' being stronger than A. The basic idea is that an attack (A,A') in some AF_i is not a defeat in JAF_i iff an argument B with claim $defeat(\mathcal{A}',\mathcal{A})$ is a preferred argument of JAF_{i+1}.

Definition 5. *Let* $\triangle = (AF_1, \ldots, AF_n)$ *be a NAF. Then the justified NAF* $(JAF_1, \ldots, JAF_n)$ *is defined as follows:*
1) *For* $i = 1\ldots n$, $Args_i$ *in* JAF_i $=$ $Args_i$ *in* AF_i
2) $Defeat_n = Attack_n$
3) *For* $i = 1\ldots n\text{-}1$, $Defeat_i = Attack_i - \{(A,A') \mid defeat(\mathcal{A}',\mathcal{A}) \in claims(\boldsymbol{Pf}(JAF_{i+1}))\}$
We say that $\boldsymbol{Pf}(JAF_1)$ *is the set of preferred arguments of* $\triangle$.

Note that the restriction in definition 4 ensures that any undercut attack in AF_i will, as required, be an undercut defeat in JAF_i:

Proposition 1. *Let* $(JAF_1, \ldots, JAF_n)$ *be defined on the basis of* $(AF_1, \ldots, AF_n)$. *Then, for* $i = 1\ldots n$: $(A, A') \in Attack_i$ *and* $(A', A) \notin Attack_i$ *implies* $(A, A') \in Defeat_i$ *and* $(A', A) \notin Defeat_i$.

Proof. Suppose otherwise: i.e., $(A, A') \notin Defeat_i$ *or* $(A', A) \in Defeat_i$. *If* $(A, A') \notin Defeat_i$, *then by def.5(3),* $defeat(\mathcal{A}',\mathcal{A}) \in claims(\boldsymbol{Pf}(JAF_{i+1}))$. *By def.5(1) the arguments in* AF_{i+1} *are the same as those in* JAF_{i+1}. *Hence,* $defeat(\mathcal{A}',\mathcal{A})$ *is the claim of an argument in* AF_{i+1}. *Hence,* $(A', A) \in Attack_i$ *by the restriction* $Attack_i \supseteq \{(A',A) \mid defeat(\mathcal{A}',\mathcal{A}) \in claims(Arg_{i+1})\}$ *in def.4. This contradicts the assumption that* $(A', A) \notin Attack_i$. *If* $(A', A) \in Defeat_i$, *then by def.5(3)* $(A', A) \in Attack_i$, *again contradicting the assumption that* $(A', A) \notin Attack_i$.

Proposition 2. *Let* $(JAF_1, \ldots, JAF_n)$ *be defined on the basis of* $(AF_1, \ldots, AF_n)$. *Assuming* $defeat(\mathcal{A}',\mathcal{A}) \in claims(\boldsymbol{Pf}(JAF))$ *implies* $defeat(\mathcal{A},\mathcal{A}') \notin claims\ (\boldsymbol{Pf}(JAF))$ *(since arguments for these claims conflict and so cannot both be in the preferred set), then for* $i = 1\ldots n$:

E is a conflict free maximal subset of Args in AF_i *iff E is a conflict free maximal subset of* $Args'$ *in* JAF_i.

Proof. *By def.5 $Args = Args'$. It remains to show that: A* ***attacks, or is attacked by, an argument in*** *AF_i* ***iff*** *A* ***defeats, or is defeated by, an argument in*** *JAF_i.*

For $i = n$ this follows from def.5(2). For $i \neq n$, the right to left half follows from def.5(3) which implies that $Defeat_i \subseteq Attack_i$. For the left to right half, consider two cases: i) $(A, A') \in Attack_i$, $(A', A) \notin Attack_i$; ii) $(A, A') \in Attack_i$, $(A', A) \in Attack_i$. Case i) is given by proposition 1. For case ii), we show that (A, A') or $(A', A) \in Defeat_i$. Suppose otherwise. Then by def.5(3), $defeat(\mathcal{A}', \mathcal{A})$ and $defeat(\mathcal{A}', \mathcal{A}) \in claims(\boldsymbol{Pf}(JAF_{i+1}))$, contradicting the assumption.

Given proposition 2, the preferred extensions of JAF_i will be a subset of those of AF_i. It is nested argumentation's substitution of rebut attacks in AF_i by asymmetric defeats in JAF_i that enables choice of a single preferred extension. In the following examples we write $A1 \rightleftharpoons A2$ to denote rebut attacks $attack(A1,A2)$ and $attack(A2,A1)$, and $A1 \rightharpoonup A2$ for the asymmetric undercut $attack(A1,A2)$.

Example 1. Let $\triangle = (AF1, AF2, AF3)$ where:
$AF1 = (\{A1, A2, A3, A4, A5\}, \{A1 \rightleftharpoons A2, A2 \rightharpoonup A3, A4 \rightleftharpoons A5\})$,
$AF2 = (\{B1, B2, B3, B4\}, \{B1 \rightleftharpoons B2, B4 \rightharpoonup B3\})$, where *claim(B1)* = *defeat*($\mathcal{A}$1,$\mathcal{A}$2), *claim(B2)* = *defeat*($\mathcal{A}$2,$\mathcal{A}$1), *claim(B3)* = *defeat*($\mathcal{A}$4,$\mathcal{A}$5)
$AF3 = (\{C1\}, \emptyset)$ where *claim(C1)* = *defeat*($\mathcal{B}$1,$\mathcal{B}$2).

Then: $\boldsymbol{Pf}(JAF_3) = \{C1\}$, $\boldsymbol{Pf}(JAF_2) = \{B1, B4\}$, $\boldsymbol{Pf}(JAF_1) = \{A1, A3\}$ - the set of preferred arguments of $\triangle$. Notice that $B4$'s undercut of $B3$ means that $A4$ is not preferred, despite the fact that there exists no $AF2$ argument for *defeat*($\mathcal{A}$5,$\mathcal{A}$4). If $B3$ were not undercut then $A4$ would also be preferred.

We consider the above to be a general framework for modelling nested argumentation, whereby given a particular argumentation system instantiating $AF1$, one can define suitable mappings from AF_i to AF_{i+1}, and logics for construction of arguments instantiating AF_i, $i > 1$. In what follows we show how this is possible, applying nested argumentation to decision making over plans of action.

3 A System for Constructing Instrumental Arguments

In [1, 2], Amgoud describes how realisation trees for an agent's initial goals can be built from an agent's planning rules. These rules are of a single type, relating goals to their sub-goals, and (sub)goals to the actions they are realised by. These realisation trees are modelled as 'instrumental' arguments for a claim - the initial goal - where the supporting argumentation can be thought of as a plan of actions and subgoals for realising the initial goal. Argument theoretic notions are then used to select the preferred arguments from a set of arguments that may conflict given constraints precluding joint execution of plans. Here we define a modified system for construction of instrumental arguments.

In what follows we define an agent description consisting of formulae in some propositional language $\mathcal{L}1$, where, unlike [1, 2], we distinguish three types of planning rule, and distinguish between literals denoting beliefs, atomic actions (that need no further plan to be achieved) and goals that require further plans to be achieved:

Definition 6. *Let $\mathcal{L}1$ be a propositional language consisting of three sets Ac, G and B of propositional literals denoting actions, goals and beliefs respectively. Let $\bigwedge \overline{Ac}$ ($\bigwedge \overline{G}$) ($\bigwedge \overline{B}$) denote the conjunction of a (possibly empty) subset of literals in Ac (G) (B). A planning rule is of the form $r : (l_1 \wedge \ldots \wedge l_{n-1}) \Rightarrow l_n$, where r is a unique propositional name for the rule, and for $i = 1 \ldots n$, l_i is a propositional literal or its negation. We write head(r) to denote $\{l_n\}$ and body(r) to denote $\{l_1,\ldots,l_{n-1}\}$. There are three types of planning rule:*

1. *precondition-action rules - $\bigwedge \overline{B} \Rightarrow l_n$ where $l_n \in Ac$*
2. *action-effect rules - $(\bigwedge \overline{B}) \wedge (\bigwedge \overline{Ac}) \Rightarrow l_n$ where $l_n \in B$ and $\overline{Ac}$ is non-empty*
3. *goal-realisation rules - $(\bigwedge \overline{B}) \wedge (\bigwedge \overline{Ac}) \wedge (\bigwedge \overline{G}) \Rightarrow l_n$ where $l_n \in G$*

Definition 7. *Let $\langle IG, \mathcal{B}, \mathcal{B}_p \rangle$ denote an agent description, where IG is the agent's set of initial goals ($IG \subseteq G$), the belief base $\mathcal{B}$ is a set of wff of $\mathcal{L}1$, and $\mathcal{B}_p$ is a set of planning rules.*

Note that planning rules are not material implications but behave as production rules. Intuitively, the antecedent $\bigwedge \overline{B}$ of a precondition-action rule represents what must be believed true about the current state of the world for an action to be applicable (i.e., the actions's preconditions). For action-effect rules, $\bigwedge \overline{B}$ represents what must be believed true about the world for actions $\bigwedge \overline{Ac}$ to result in some belief b to be true (i.e., b represents a postcondition or immediate effect of an action or actions). Finally, a goal-realisation rule represents that the goal in the head of the rule is realisable if the beliefs (effects of actions) in the antecedent are true and/or actions in the antecedent are executed and/or subgoals in the antecedent are realised.

Example 2. Let $\triangle$ be a medical agent description consisting of an initial treatment goal g and the planning rules: ***r1*** : *b1* $\Rightarrow$ *a1*; ***r2*** : *a1* $\Rightarrow$ *e1*; ***r3*** : *b2* $\Rightarrow$ *a2*; ***r4*** : *a2* $\Rightarrow$ *e1*; ***r5*** : *e1* $\Rightarrow$ g, where *b1* (*b2*) represents a precondition for a medical action *a1* (*a2*), and *a1* (*a2*) results in an effect *e1* that realises g. For example, *a1* = 'administer aspirin', *a2* = 'administer chlopidogrel', *e1* is the effect 'reduced platelet adhesion' and g = 'prevent blood clotting'.

We now define a realisation tree R for an initial goal ($\vdash$ denotes classical consequence in this and subsequent definitions), where $root(R)$ denotes the root node of R, $child_1(n), \ldots, child_k(n)$ denote the child nodes $n1, \ldots, nk$ of node n, and n is a leaf node if it has no child nodes. Also, a node n in R is the parent of a subtree T of R iff $child(n) = root(T)$.

Definition 8. *A realisation tree based on $\langle IG, \mathcal{B}, \mathcal{B}_p \rangle$ is a finite AND tree R defined as follows:*

- *root(R) is a goal-realisation rule r where head(r)= g, $g \in IG$*
- *If node n of R is a planning rule $r : l_1 \wedge \ldots \wedge l_k \Rightarrow l$, then for $i = 1 \ldots k$:*

1. *if* $l_i \in G$ *or* $l_i \in Ac$ *then* $child_i(n)$ *is a planning rule* r_i *with head* l_i
2. *if* $l_i \in B$, *then if* r *is a precondition-action or action-effect rule, then* $\mathcal{B} \vdash l_i$, *else if* r *is a goal-realisation rule then* $child_i(n)$ *is an action-effect rule* r_i *with head* l_i

From hereon, *nodes(R)* returns the set of rules in R, *ig(R)* denotes the initial goal of R, and we refer to each node (rule) in R as a partial plan. Realisation trees as defined by Amgoud [1] and Hulstijn [8] are instrumental arguments. Two such arguments conflict, and so attack each other, if they contain partial plans that conflict.

Definition 9. *Two partial plans* r_1 *and* r_2 *conflict iff* $head(r_1) \cup head(r_2) \cup body(r_1) \cup body(r_2) \cup \mathcal{B} \cup \mathcal{B}_p \vdash \perp$.

Hence, the defined arguments and their attacks can be used to instantiate a Dung framework. However, employing Dung's attack based definition of a conflict free set of arguments (def.2) may yield a preferred set of arguments that cannot be jointly adopted as plans. For example, suppose $\langle\ IG = \{a,b,c\}$, $\mathcal{B} = \{a' \wedge b' \rightarrow \neg c'\}$, $\mathcal{B}_p = \{a' \Rightarrow a,\ b' \Rightarrow b,\ c' \Rightarrow c\}\rangle$. Then the instrumental arguments as defined in $[1,8]$ are $R1 = (\Rightarrow a', a' \Rightarrow a,)$, $R2 = (\Rightarrow b', b' \Rightarrow b,)$, $R3 = (\Rightarrow c', c' \Rightarrow c,)$ (note that actions a',b',c' are not required to be the heads of planning rules in $[1,8]$). No two arguments attack each other, and so the single preferred extension and hence set of preferred arguments is $\{R1, R2, R3\}$. However, the constraint in $\mathcal{B}$ precludes joint adoption of $R1$, $R2$ and $R3$.

This is rectified in Amgoud [2] by dropping the attack relation and attack based definition of conflict free sets. A conflict free set of instrumental arguments is simply defined on the basis that all the contained partial plans are mutually consistent. Thus, one obtains $\{R1, R2\}$, $\{R1, R3\}$, $\{R2, R3\}$. However, this represents a departure from Dung, so that in [2], the preferred extensions are selected solely on the basis of those sets that maximise the number of initial goals realised by the contained arguments (this is also the only criterion used in [1] and [8]). By this criterion, all the above sets are preferred extensions. Hence, none of the arguments are preferred.

The solution is to recognise that two or more realisation trees can be combined into a single instrumental argument provided that the trees do not conflict. We thus obtain instrumental arguments for more than one initial goal (conceptually, the conjunction of multiple initial goals can be considered as the head of a goal realisation rule whose body includes the individual initial goals). Thus, we will have three instrumental arguments $(R1 + R2)$, $(R1 + R3)$ and $(R2 + R3)$, each of which conflict with, and so attack, each other. We now define our notion of conflict free sets of realisation trees. Note that as in Hulstijn [8] (but unlike Amgoud), we additionally regard two realisation trees as conflicting if they realise the same goal. This is because an agent will at some stage have to decide and commit to a particular plan for realisation of any given goal.

Definition 10. *Let* S *be a set of realisation trees based on* $\langle IG, \mathcal{B}, \mathcal{B}_p\rangle$. *Then* S *is conflict free iff:*

- $\forall$ *R,R′* $\in$ *S, R* $\neq$ *R′* $\rightarrow$ *ig(R)* $\neq$ *ig(R′)*
- $\bigcup_{R \in S} [\bigcup_{r \in nodes(R)}(head(r) \cup body(r))] \cup \mathcal{B} \cup \mathcal{B}_p \nvdash \perp$

An instrumental argument is defined as follows:

Definition 11. *Let $S_1,\ldots,S_m$ be the maximal (w.r.t set inclusion) conflict free sets of realisation trees based on $\langle IG, \mathcal{B}, \mathcal{B}_p\rangle$. Then $\{A_1,\ldots,A_m\}$ is the set of instrumental arguments based on $\langle IG, \mathcal{B}, \mathcal{B}_p\rangle$, where for $i{=}1\ldots m$, A_i is a finite AND tree with root node $n{=}\{ig(R)|R \in S_i\}$ and n is the parent of each tree in $\{R|R{\in}S_i\}$.*

Note that given definition of the planning rules (def.6) and realisation trees (def.8) one can readily show that:

Proposition 3. *Any path from the root to the leaf of an instrumental argument starts with the root node set of initial goals, followed by one or more goal-realisation rules, followed by at most one action-effect rule, and terminating in exactly one precondition-action rule.*

Each instrumental argument conflicts with and attacks all other instrumental arguments. We can now instantiate a Dung argumentation framework *AF1*:

Definition 12. *AF1 = (Args1, Attack1) where Args1 is the set of all instrumental arguments built from an agent description $\langle IG, \mathcal{B}, \mathcal{B}_p\rangle$, and Attack1 = $\{(A,A')|A, A' \in Args1$ and $A \neq A'\}$.*

Example 3. In the following variation of an example in [2], an agent decides over plans of action to realise its initial goals to prepare for a journey to Africa (*pja*) and finish a paper (*fp*). Let the agent description be:

$\langle$ IG = $\{pja, fp\}$, $\mathcal{B} = \{w \rightarrow \neg\, pc\}$,
$\mathcal{B}_p = \{r1{:}w \Rightarrow fp,\ r2{:}t \wedge vac \Rightarrow pja,\ r3{:}int \Rightarrow t,\ r4{:}hop \Rightarrow vac,\ r5{:}pc \Rightarrow vac,\ r6{:}dr \Rightarrow vac,\ r7{:} \Rightarrow int,\ r8{:} \Rightarrow dr,\ r9{:} \Rightarrow pc,\ r10{:} \Rightarrow hop,\ r11{:} \Rightarrow w\}\rangle$

where $G = \{fp, pja, t, vac\}$, $Ac = \{int,\ dr,\ pc, hop, w\}$, and w = 'work', pc = 'go to private clinic', t = 'get a ticket', vac = 'get vaccinated', dr = 'go to the doctor', hop = 'go to the hospital', int = 'log on to internet'. Note that

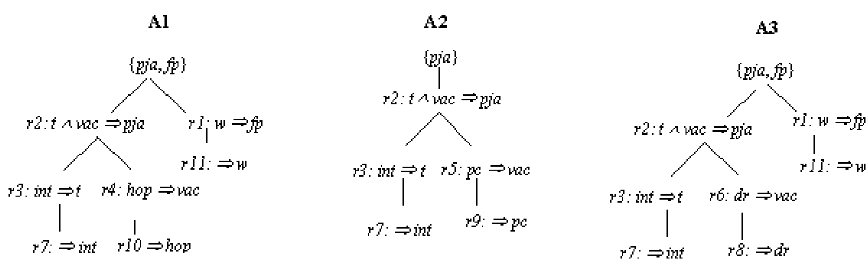

Fig. 1

$w \rightarrow \neg pc$ represents that working to finish the paper would take up to the end of the working day and so exclude going to a private clinic which (unlike the hospital and doctor's surgery) is closed outside of working hours.

Fig. 1 shows the arguments *Args1* based on $\langle IG, \mathcal{B}, \mathcal{B}_p \rangle$. $Attack1 = \{A1 \rightleftharpoons A2, A1 \rightleftharpoons A3, A2 \rightleftharpoons A3\}$. The preferred extensions of *AF1* = *(Args1,Attack1)* are: $\{A1\}$, $\{A2\}$, $\{A3\}$.

To summarise, an instrumental argument is a maximal conflict free set of realisation trees constructed from planning rules. Any two such arguments attack each other on the basis that they contain partial plans that conflict with each other and/or share an initial goal. This means that each maximal conflict free set of instrumental arguments (as defined by def.2) will always be a singleton set. We will have non-singleton sets when we consider other types of argument interacting with instrumental arguments. For example, arguments built from the agent's belief base may attack instrumental arguments by conflicting with beliefs in the antecedent of a precondition-action rule or action-effect rule.

Example 4. To illustrate, in our medical example 2, $AF1 = (\{A1,A2\},\{A1 \rightleftharpoons A2\})$ where *A1* is built from rules *r1*, *r2* *r5*, and *A2* built from rules *r3*, *r4*, *r5*. An argument *A3* with claim $\neg b1$ would be a non-instrumental argument built from the agent's beliefs, which attacks *A1*. One might also account for the desirability of goals and effects realised or effected by an action. Assume the agent description is extended to include a set *U* of *undesirable* effects. Suppose an undesirable side-effect $e2 \in U$, and an action-effect rule *r6*: $b_1,\ldots,b_n,\ a1 \Rightarrow e2$, which represents that action *a1* has effect *e2* if $b_1, \ldots, b_n$ are believed true (e.g., aspirin has the effect gastric ulceration if it is believed that the patient has a history of gastritis). If $\mathcal{B} \vdash b_1, \ldots, b_n$ then *r6* will be used to construct a non-instrumental argument attacking *A1*.

However, the focus of this paper is on determining preferences amongst instrumental arguments that mutually attack and defeat each other, given that the strength of such arguments can be valuated on the basis of different criteria, or for any given criterion, on the basis of different sources. In the following section we show how nested argumentation can be used to resolve these conflicting defeats and thus determine a single preferred instrumental argument.

4 Applying Nested Argumentation to Decide the Preferred Instrumental Arguments

In what follows we define a *NAF* (AF_1,AF_2,AF_3) where AF_1 is defined as in the previous section. Arguments instantiating AF_2 will be for valuations of the strengths of AF_1 arguments and defeats between AF_1 arguments. Arguments instantiating AF_3 will make use of orderings on sources and criteria to construct arguments for defeats between AF_2 arguments. We then apply nested argumentation to determine a single preferred instrumental argument.

4.1 Defining the Argumentation Framework AF_2

Firstly, we define an argumentation system instantiating AF_2. We define the language $\mathcal{L}2$, a logic for argument construction, and a definition of conflict (attack).

Definition 13. *Let AF1 = (Args1, Attack1) be defined by an agent description $\langle IG, \mathcal{B}, \mathcal{B}_p \rangle$. Then $\mathcal{L}2$ is any first order logic language whose signature contains the set of real numbers $\Re$, the binary predicate symbols "attack" and "defeat", the arithmetic less than relation "<", and the following sets of constant symbols:*
- a set of argument names $\mathsf{f\_name}_1(Args1)$
- the set of planning rule names $\{r \mid r : l_1 \wedge \ldots \wedge l_k \Rightarrow l \in \mathcal{B}_p\}$
- a set Π denoting criteria and a set Ψ denoting sources

In what follows, variables $X, Y, \ldots$ range over $\Re$, $\mathcal{A}, \mathcal{A}_1, \mathcal{A}_2 \ldots$ range over $\mathsf{f\_name}_1(Args1)$, $P, P_1, P_2 \ldots$ range over criteria, $S, S_1, S_2 \ldots$ range over sources, and lower case roman letters range over all other constants in $\mathcal{L}2$. Lower case greek letters range over predicate formulae in $\mathcal{L}2$. Also, $\vdash_{FOL}$ denotes first order classical inference, and for any first order theory we assume the usual axiomatisation of $<$. We now define a mapping from *AF1* to a set $\triangle_{map}$ of first order implications and ground predicates in $\mathcal{L}2$. In this way an instrumental argument A is decomposed into its 'sub-arguments', e.g., the initial goals of A, or actions and action goal pairs in A.

Definition 14. *Let AF1 = (Args1, Attack1). Then $\triangle_{map}$ is defined as follows:*

- *attack($\mathcal{A}, \mathcal{A}'$) $\in \triangle_{map}$ iff (A, A') $\in$ Attack1*
- *initial_goal($\mathcal{A}$,g) $\in \triangle_{map}$ iff $A \in$ Args1, $g \in$ root(A)*
- *goal($\mathcal{A}$,g) $\in \triangle_{map}$ iff $A \in$ Args1, r: $l_1 \wedge \ldots \wedge l_k \Rightarrow g$ is a node in A and $g \in G$*
- *action($\mathcal{A}, a$) $\in \triangle_{map}$ iff $A \in$ Args1 and r: $l_1 \wedge \ldots \wedge l_k \Rightarrow a$ is a leaf node in A*
- *rule($\mathcal{A}$,r) $\in \triangle_{map}$ iff $A \in$ Args1 and r: $l_1 \wedge \ldots \wedge l_k \Rightarrow l$ is a node in A*
- *rule_head($\mathcal{A}$,r,h) $\in \triangle_{map}$ iff rule($\mathcal{A}$,r) $\in \triangle_{map}$, head(r) = $\{h\}$*
- *rule_body($\mathcal{A}$,r,b) $\in \triangle_{map}$ iff rule($\mathcal{A}$,r) $\in \triangle_{map}$, $b \in$ body(r)*
- *(action($\mathcal{A}$,a) $\wedge$ goal($\mathcal{A}$,g) $\wedge$ rule_body($\mathcal{A}$,r,a) $\wedge$ rule_head($\mathcal{A}$,r,g) $\rightarrow$ action_goal($\mathcal{A}$,a,g)) $\in \triangle_{map}$*
- *(action($\mathcal{A}$,a) $\wedge$ goal($\mathcal{A}$,g) $\wedge$ rule_body($\mathcal{A}$,r,a) $\wedge$ rule_head($\mathcal{A}$,r,h) $\wedge$ ($h \neq g$) $\wedge$ rule_body($\mathcal{A}$,r',h) $\wedge$ rule_head($\mathcal{A}$,r',g) $\rightarrow$ action_goal($\mathcal{A}$, a, g)) $\in \triangle_{map}$*

Note that the last two rules allow inference of action goal pairs so that one can valuate the temporal or financial cost or efficacy of an action w.r.t. the immediate (sub)goal realised by the action. In the first case, the action is in the antecedent of a goal realisation rule. In the second case, the action is in the body of an action-effect rule whose head (effect) must be (given proposition 3) in the body of a goal realisation rule.

Construction of AF_2 arguments for evaluation of an AF_1 instrumental argument A, proceeds in two steps. Firstly, numerical valuations of sub-arguments of A are inferred from data of the type *temporal_cost(S,a,g,X)*, where S is the source of the valuation of the temporal cost of action a w.r.t goal g. Then second order rules are used to infer a valuation of A from its sub-argument valuations (each of which may be obtained from a different source). In the following, *tc*, *fc*, *eff* and *gp* respectively denote the criteria temporal cost, financial cost, efficacy and goal priority (the importance of a goal to an agent).

Definition 15. $\triangle_{s\_eval}$ *denotes the set of sub-argument evaluation rules :*

- $action\_goal(\mathcal{A},a,g) \wedge \rho(S,a,g,X) \rightarrow eval(S,\rho,\mathcal{A},a,X)$, *where* $\rho \in \{tc,fc,eff\}$
- $initial\_goal(\mathcal{A},g) \wedge gp(S,g,X) \rightarrow eval(S,gp,\mathcal{A},g,X)$

Definition 16. *Let* ρ *denote a constant in* $\{tc, fc, eff, gp\}$ *and* Γ *a first order theory. Then* $\mathcal{D}$ *is the following set of* Γ *specific full-argument evaluation rules.*
$d_\rho(\Gamma)$: $eval(S_1, \rho, \mathcal{A}, l_1, X_1), \ldots, eval(S_n,\rho,\mathcal{A},l_n, X_n) \hookrightarrow eval(\rho, \mathcal{A}, Y)$ *where:*

1. $\{eval(S_1, \rho, \mathcal{A}, l_1, X_1) \ldots eval(S_n, \rho, \mathcal{A}, l_n, X_n)\}$ *is the set of all inferences of the form* $\Gamma \vdash_{FOL} eval(S, \rho, \mathcal{A}, l, X)$
2. $\forall jk, j \neq k \rightarrow l_j \neq l_k$
3. *If* $\rho \in \{tc, fc, eff\}$ *then* $Y = \Sigma_{i=1}^{n} X_i$, *else if* $\rho = gp$ *then* $Y = max_{i=1}^{n} X_i$

Notice that the goal priority of an argument is the maximum of the goal priorities of the argument's initial goals. The financial/temporal cost and efficacy valuation of an argument is the sum of the valuations of the action goal pairs in the argument. The above does not represents an exhaustive list of criteria for evaluating the strength of instrumental arguments. Examples of other criteria include the depth of an argument (preferring arguments of lesser depth favours arguments with fewer intermediate subgoals relating actions to an initial goal), the *certainty level* of an argument (the minimum of the weights associated with rules in an argument), and the number of initial goals in an argument (the criterion used in [1, 2, 8]).

In the following definition we define construction of *AF2* arguments from a first order theory Γ, such that:

- $\Gamma \nvdash_{FOL}$
- $\triangle_{map} \subset \Gamma$, i.e., Γ contains a mapping of instrumental arguments to their sub-arguments in $\mathcal{L}2$
- $\triangle_{s\_eval} \subset \Gamma$, i.e., Γ contains the sub-argument evaluation rules defined in def.15
- $\triangle_{dom} \subset \Gamma$ where $\triangle_{dom}$is a set of domain specific facts of the form $gp(S, g, X)$, $fc(S, a, g, X) \ldots$ used together with rules in $\triangle_{s\_eval}$ to infer valuations of the above sub-arguments
- $\mathcal{ACK} \in \Gamma$ where $\mathcal{ACK}$ is the rule:

  $attack(\mathcal{A}_1, \mathcal{A}_2) \wedge eval(P, \mathcal{A}_1, X) \wedge eval(P, \mathcal{A}_2, Y) \wedge (Y < X) \rightarrow defeat$ $(\mathcal{A}_1, \mathcal{A}_2)$

  for inferring arguments with $defeat$ claims from full-argument valuations
- apart from $\mathcal{ACK}$ there exists no other formula ϕ in Γ such that $defeat(X, Y)$ is a predicate in ϕ. This restriction fulfills the requirement on NAFs in definition 4, viz. a. vie. that $defeat(\mathcal{A}_1, \mathcal{A}_2)$ is a claim of an AF_2 argument built from Γ only if (A_1, A_2) is an attack in $AF_1 =$ (*Args1, Attack1*)

Definition 17. *An argument B based on Γ is a pair (Γ', ϕ), where either:*

1. $\Gamma' = \{\phi_1, \ldots, \phi_n\}$ *where* $d_P(\Gamma) \in \mathcal{D}$ *and* $d_P(\Gamma) = \phi_1, \ldots, \phi_n \hookrightarrow \phi$, *or*
2. $\Gamma' = \Gamma_1 \cup \Gamma_2$, *such that:*
   - $\Gamma_1 = \{\phi_1, \ldots, \phi_n\}$ *where for* $i = 1 \ldots n$, ϕ_i *is the claim of an argument of type 1*
   - $\Gamma_2 \subseteq \Gamma$
   - $\Gamma' \vdash_{FOL} \phi$, *and* Γ' *is consistent and set-inclusion minimal*

Example 5. Continuing with example 3 we list in the left hand column of the table below, the claims of *AF2* sub-argument valuations *J0* - *J5′* (writing 'e' as shorthand for 'eval') obtained by def.17-2. We assume that the temporal cost of logging on to the internet is negligible, the agent *ag1*'s initial goal of finishing a paper has higher priority than preparing for a journey to Africa, and getting a vaccination at the hospital takes more time than at the doctor which takes more time than at the private clinic. These are inferred from valuation data in $\triangle_{p\_dom}$[2]. In the middle column we list the claims of *AF2* full argument valuations *K0* - *K5* that are supported by *J0* - *J5′*. Arguments *K0* - *K5* are obtained by def.17-1. In the right hand column we list *AF2* arguments *L0* - *L4* for *defeat* claims (we write '*d*' instead of *defeat* and show only the K arguments providing support) obtained by def.17-2. Examples of constructed arguments include:

$J0 = (\{\ initial\_goal(\mathcal{A}1,fp)\ ,\ gp(ag1,fp,0.8),\ initial\_goal(\mathcal{A}1,fp) \wedge gp(ag1,fp,0.8) \rightarrow e(ag1,gp,\mathcal{A}1,fp,0.8)\},\ e(ag1,gp,\mathcal{A}1,fp,0.8)\ )$
$K0 = (\{\ e(ag1,gp,\mathcal{A}1,fp,0.8),\ e(ag1,gp,\mathcal{A}1,pja,0.2)\},\ e(gp,\mathcal{A}1,0.8))$
$L0 = (\{\ attack(\mathcal{A}1,\mathcal{A}2),\ e(gp,\mathcal{A}1,0.8)),\ e(gp,\mathcal{A}2,0.2))\} \cup \{\mathcal{ACK}\},\ d(\mathcal{A}1,\mathcal{A}2))$

$J0 = e(ag1,gp,\mathcal{A}1,fp,0.8)$ $J0' = e(ag1,gp,\mathcal{A}1,pja,0.2)$	$K0 = e(gp,\mathcal{A}1,0.8)$	
$J1 = e(ag1,\ gp,\ \mathcal{A}2,\ pja,\ 0.2)$	$K1 = e(gp,\ \mathcal{A}2,\ 0.2)$	$L0 = (K1 \cup K0, d(\mathcal{A}1,\mathcal{A}2))$
$J2 = e(ag1,\ gp,\ \mathcal{A}3,\ fp,\ 0.8)$ $J2' = e(ag1,\ gp,\ \mathcal{A}3,\ pja,\ 0.2)$	$K2 = e(gp,\ \mathcal{A}3,\ 0.8)$	$L1 = (K1 \cup K2, d(\mathcal{A}3,\mathcal{A}2))$
$J3 = e(ag1,\ tc,\ \mathcal{A}1,\ hop,\ 1)$ $J3' = e(ag1,\ tc,\ \mathcal{A}1,\ w,\ 0.5)$	$K3 = e(tc,\ \mathcal{A}1,\ 1.5)$	$L2 = (K3 \cup K4, d(\mathcal{A}2,\mathcal{A}1))$
$J4 = e(ag1,\ tc,\ \mathcal{A}2,\ pc,\ 2)$	$K4 = e(tc,\ \mathcal{A}2,\ 2)$	$L3 = (K3 \cup K5, d(\mathcal{A}3,\mathcal{A}1))$
$J5 = e(ag1,\ tc,\ \mathcal{A}3,\ dr,\ 1.3)$ $J5' = e(ag1,\ tc,\ \mathcal{A}3,\ w,\ 0.5)$	$K5 = e(tc,\ \mathcal{A}3,\ 1.8)$	$L4 = (K4 \cup K5, d(\mathcal{A}2,\mathcal{A}3))$

We now define the binary relation 'conflict' over *wff* of $\mathcal{L}2$. In the first case, two *wff* conflict if they represent two different valuations of the same sub-argument l of an instrumental argument $\mathcal{A}$ (by the same or different sources) w.r.t. the same criterion P. In the second case, two *wff* conflict if they represent two different valuations of the same instrumental argument $\mathcal{A}$ w.r.t the same criterion P. The third case represents two conflicting defeat claims.

[2] Note that temporal valuations are normalised, e.g., if getting a vacination at the hospital takes 120 minutes and at the private clinic 60 minutes, then *tc(S,pc,vac,2)* and *tc(S,hop,vac,1)*.

Definition 18. *Let* ϕ_1 *and* ϕ_2 *be wff of* $\mathcal{L}2$*. Then, conflict*(ϕ_1, ϕ_2) *iff:*

- ϕ_1 = *eval(S, P,* $\mathcal{A}$*, l, X)*, ϕ_2 = *eval(S′, P,* $\mathcal{A}$*, l, Y)*, $X \neq Y$
- ϕ_1 = *eval(P,*$\mathcal{A}$*,X)*, ϕ_2 = *eval(P,*$\mathcal{A}$*,Y)*, $X \neq Y$
- ϕ_1 = *defeat(*$\mathcal{A}$*,*$\mathcal{A}'$*)*, ϕ_2 = *defeat(*$\mathcal{A}'$*,*$\mathcal{A}$*)*

We define the conflict based rebut and undercut attacks on the set *Args2* of arguments given by def.17, and then define *AF2*.

Definition 19. *For all* (Γ, ϕ), $(\Gamma', \phi') \in$ *Args2,*

- (Γ, ϕ) *rebuts* (Γ', ϕ') *iff conflict*(ϕ, ϕ')
- (Γ, ϕ) *undercuts* (Γ', ϕ') *iff* $\exists \phi'' \in \Gamma'$ *such that conflict*(ϕ, ϕ'')

Definition 20. *AF2 = (Args2, Attack2), where for all B,B′* $\in$ *Args2, (B,B′)* $\in$ *Attack2 iff B rebuts B′ or B undercuts B′.*

Example 6. Continuing with example 5, no two sub-argument or full argument valuations conflict. Hence, *AF2* = (*Args2,Attack2)* where *Args2* includes *J0* - *J5′*, *K0* - *K5*, *L0* - *L4* and *Attack2* = $\{L0 \rightleftharpoons L2, L1 \rightleftharpoons L4\}$. The preferred arguments of *AF2* are *J0* - *J5′*, *K0* - *K5* and *L3*.

Example 7. Recall that in e.g.4 two *AF1* arguments *A1* and *A2*, respectively relate medical actions *a1* and *a2* to treatment goal *g*. Suppose sources clinical trial 1 (*ct1*) reporting that *a1* is more efficacious than *a2* w.r.t. *g*, and clinical trial 2 (*ct2*) reporting that *a2* is more efficacious than *a1* w.r.t. *g*. Therefore *AF2* = (*Args2,Attack2)* where:

- *Args2* includes:
  - *J1*, *J2* and *J3* with claims *e(ct1,eff,*$\mathcal{A}$*1,a1,5)*, *e(ct1,eff,*$\mathcal{A}$*2,a2,4)* and *e(ct2, eff,* $\mathcal{A}$*2,a2,6)* respectively
  - The claims of *J1*, *J2* and *J3* respectively support arguments *K1* with claim *e(eff,* $\mathcal{A}$*1, 5)*, *K2* with claim *e(eff,* $\mathcal{A}$*2, 4)*, and *K3* with claim *e(eff,* $\mathcal{A}$*2, 6)*
  - *K1* and *K2*'s claims support argument *L1* with claim *defeat*($\mathcal{A}$1,$\mathcal{A}$2), and *K1* and *K3*'s claims support *L2* with claim *defeat*($\mathcal{A}$2,$\mathcal{A}$1)

- *Attack2* = $\{J2 \rightleftharpoons J3, K2 \rightleftharpoons K3, J3 \rightharpoonup K2, J2 \rightharpoonup K3, L1 \rightleftharpoons L2, K2 \rightharpoonup L2, K3 \rightharpoonup L1\}$

4.2 Defining the Argumentation Framework AF_3

We now define an argumentation system instantiating $AF3$. Priority orderings on sources are used to construct arguments for defeats between $AF2$ sub-argument valuations (e.g., *J2* and *J3* in e.g.7). Priority orderings on criteria are used to construct arguments for defeats between $AF2$ arguments with claims of the form $defeat(\mathcal{A}, \mathcal{A}')$ (e.g., *L0* and *L2* in e.g.6). We will consider a set Π of named partial orderings, where if $\wp$ is the name of an ordering in Π, then this is represented by the usual first order reflexivity and transitivity axioms, and formulae of the form $>(\wp, J, K)$ interpreted as source (criterion) J is prioritised above source (criterion) K. We now define the language $\mathcal{L}3$, a mapping from $AF2$ arguments to first order formulae in $\mathcal{L}3$, and rules for construction of $AF3$ arguments:

Definition 21. *Let $AF2 = (Args2 = \{B_1, \ldots, B_n\}, Attack2)$. Then:*

- *$\mathcal{L}3$ is any first order logic language whose signature contains the signature of $\mathcal{L}2$, and the set of constants $\mathsf{f\_name}_2(Args2) = \mathcal{B}_1, \ldots, \mathcal{B}_n$.*
- *$\triangle_{e\_arg} = \{attack(\mathcal{B}_1, \mathcal{B}_2)\ |(B_1, B_2) \in Attack2\} \cup \bigcup_{i=1}^{n} m(B_i)$, where:*
  - *If $claim(B) = eval(S, P, \mathcal{A}, l, X)$ then $m(B) = \{eval(\mathcal{B}, S, P, \mathcal{A}, l, X)\}$*
  - *Else if $B = (\{attack(\mathcal{A}_1, \mathcal{A}_2), eval(P, \mathcal{A}_1, X), eval(P, \mathcal{A}_2, Y) \cup \{\mathcal{ACK}\}\}, defeat(\mathcal{A}_1, \mathcal{A}_2))$ then $m(B) = \{defeat(\mathcal{B}, P, \mathcal{A}_1, \mathcal{A}_2)\}$*
  - *Else $m(B) = \emptyset$*
- *Let $\triangle_{po\_arg}$ be the set of rules:*
  $(attack(\mathcal{B}, \mathcal{B}') \wedge eval(\mathcal{B}, S_1, P, \mathcal{A}, l, X) \wedge eval(\mathcal{B}', S_2, P, \mathcal{A}, l, Y) \wedge (X \neq Y) \wedge >(\wp, S_1, S_2)) \rightarrow defeat(\mathcal{B}, \mathcal{B}')$
  $(attack(\mathcal{B}, \mathcal{B}') \wedge defeat(\mathcal{B}, P, \mathcal{A}_1, \mathcal{A}_2) \wedge defeat(\mathcal{B}', P', \mathcal{A}_2, \mathcal{A}_1) \wedge > (\wp, P, P')) \rightarrow defeat(\mathcal{B}, \mathcal{B}')$

An input theory for constructing *AF3* arguments contains the above mapping $\triangle_{e\_arg}$ of *AF2* arguments, a set Π of named orderings on criteria and sources, and the rules $\triangle_{po\_arg}$ for construction of *AF3* arguments. We also assume the restriction (for the same reason as outlined in section 4.1 for an input theory for constructing *AF32* arguments) that the predicate $defeat(X, Y)$ is only in formulae in $\triangle_{e\_arg} \cup \triangle_{po\_arg}$.

Definition 22. *Let Γ be a first order theory such that $\Gamma \nvdash_{FOL} \perp$ and $(\triangle_{e\_arg} \cup \Pi \cup \triangle_{po\_arg}) \subseteq \Gamma$. An argument C based on Γ is a pair (Γ', ϕ), where $\Gamma' \subseteq \Gamma$, $\Gamma' \vdash_{FOL} \phi$ and Γ' is set inclusion minimal.*

Definition 23. *Let AF3 = (Args3, Attack3) where Args3 is the set of all arguments given by def.22, and $\forall C, C' \in$ Args3, $(C, C') \in$ Attack3 iff $claim(C) = defeat(\mathcal{B}, \mathcal{B}')$ and $claim(C') = defeat(\mathcal{B}', \mathcal{B})$.*

Note that no $AF3$ argument attacks another under the conditions that there is only a single criterion ordering and a single source ordering, and no source provides more than one valuation of a sub-argument. Suppose the latter was not satisfied. Then we would have $eval(\mathcal{B}, S_1, P, \mathcal{A}, l, X)$ and $eval(\mathcal{B}', S_1, P, \mathcal{A}, l, Y)$, $X \neq Y$ and $> (\wp, S_1, S_1)$ (by reflexivity of $>$) supporting claims $defeat(\mathcal{B}, \mathcal{B}')$ and $defeat(\mathcal{B}', \mathcal{B})$. If the above conditions are not satisfied, then one might need to determine preferences amongst mutually attacking C arguments, which would require construction of *AF4* arguments for preferences amongst criterion/source orderings and sub-argument valuations from a single source.

Definition 24. *Let AF1, AF2 and AF3 = (Args3, Attack3) be defined as in definitions 12, 20 and 23. Let Args3 be defined on the basis of some Γ such that Π contains a single source and a single criterion ordering. Then a nested argumentation framework for agent decision making over instrumental arguments is the triple (AF1,AF2,AF3).*

Example 8. Continuing with example 6, assume a single criterion ordering prioritising goal priority over temporal cost. Then, simply writing this prioritisation in each arguments support, we obtain:
AF3 = (*Args3* = { ({*gp*>*tc*},*defeat(L0,L2)*), ({*gp*>*tc*},*defeat(L1,L4)*) }, ∅).
By def.5:
- *JAF3* = *AF3* and so ***Pf****(JAF3)* = *Args3*
- *JAF2* = *Args2* and *defeat(L0,L2)*, *defeat(L1,L4)*. Hence ***Pf****(JAF2)* now includes *L0*, *L1* and *L3* for claims *defeat*($\mathcal{A}$1,$\mathcal{A}$2), *defeat*($\mathcal{A}$3,$\mathcal{A}$2) and *defeat* ($\mathcal{A}$3,$\mathcal{A}$1).
- *JAF1* is *Args1* and *defeat(A1,A2)*, *defeat(A3,A2)*, *defeat(A3,A1)*. Hence, ***Pf****(JAF1)* = {*A3*}. That is, *A3* is the single preferred instrumental argument given that *A2* is stronger than *A3* is stronger than *A1* on the grounds of temporal cost, but *A3* and *A1* are stronger than *A2* on the grounds of goal priority, where the latter is the preferred criterion.

Example 9. Continuing with example 7, assume a single source ordering *ct1* > *ct2*. Then *AF3* = ({ ({*ct1* > *ct2*},*defeat(J2,J3)*) },∅).
By def.5:
- *JAF3* = *AF3* and so ***Pf****(JAF3)* = ({*ct1* > *ct2*},*defeat(J2,J3)*)
- *JAF2* = (*Args2*,*Defeat2*), where *Defeat2* = *Attack2* - {*(J3,J2)*}. We obtain ***Pf****(JAF2)* = {*J1*, *J2*, *K1*, *K2*, *L1*} where *claim(L1)* = *defeat*($\mathcal{A}$1,$\mathcal{A}$2)
- *JAF1* = {*A1*, *A2*} and *defeat(A1,A2)*. Hence ***Pf****(JAF1)* = *A1*, since although the efficacy of *A2*'s action w.r.t. treatment goal *g* is rated above *A1*'s action by clinical trial 2, the preferred source clinical trial 1 rates *A1*'s action higher than *A2*'s action.

5 Future and Related Work

In this paper we have formalised a framework for nested argumentation, and applied this framework to selection of an agent's preferred instrumental arguments. Future work will more thoroughly investigate properties of nested argumentation frameworks. For example, one might establish the conditions under which arguments are 'objectively' preferred. To illustrate, if *defeat(A2,A1)* and *defeat(A1,A3)* are both based on some criterion *c*, and *defeat(A3,A1)* and *defeat(A1,A2)* are both based on c', then *A2* and *A3* will be preferred irrespective of the ordering of these criteria. One might also consider extending the kinds of 'meta-argumentation' described in frameworks AF_i, $i > 1$. For example, while data concerning the relative strengths of *A1* and *A2* may not be available, a 'transitive' argument for *defeat*($\mathcal{A}1, \mathcal{A}2$) could be constructed from *AF2* arguments for *defeat*($\mathcal{A}1, \mathcal{A}3$) and *defeat*($\mathcal{A}3, \mathcal{A}2$), where the latter two arguments are based on the same criterion. Argumentation over criterion/source orderings will also be investigated. This will require extending *NAF*s to include *AF4* frameworks. For example, a preference for one clinical trial source over another is based on factors including statistical validity, measures taken to eliminate biases e.t.c. This suggests there may be arguments for different orderings on these sources. Finally, application of our work to argumentation-based dialogues [12] would enable agents to engage in the kinds of meta-argumentation described in this paper. For example, an agent justifying to another agent as to why it prefers one argument to another, and this justification itself being challenged. In *deliberation* dialogues, multiple agents cooperate to determine a preferred course of action. A recently proposed model for deliberation [7] describes requirements for communication of arguments for plans of action, and perspectives by which competing arguments are judged. We believe our work has the potential to provide such requirements.

As mentioned in section 1, reasoning about the relative strength of arguments is also explored in [9, 11] in which argument strength is based on rule priorities alone. In value-based argumentation frameworks (VAF) [5] a successful attack (defeat) of one argument by another depends on the comparative strength of the values (analogous to criteria) advanced by the arguments concerned. However, for two arguments that both promote some value v, one cannot defeat the other on the grounds that it promotes v more than the other. Furthermore, VAF is restricted to evaluation of defeats on the basis of value orderings, so that other justifications for defeat are not possible. Also, argumentation over value orderings is not possible.

Section 3 describes how our work on instrumental arguments compares with [1, 2, 8]. To summarise, in our approach arguments more readily instantiate a Dung framework, and preferred arguments are selected on the basis of multiple criteria and sources for valuating the strengths of arguments. Furthermore, as described in example 2, we have defined planning rules so as to 'expose' an instrumental argument's 'potential points of attack'. Future work will further investigate agent argumentation over beliefs and goals and the ways in which these arguments interact with instrumental arguments. Indeed, instrumental arguments can be seen as instantiating a variation on Atkinson et.al's presumptive schema justifying a course of action [4]: *In circumstances R, we should perform action A, whose effects will result in state S which will realise goal G, which promotes some value V.* Arguments attacking an instrumental argument can be seen as instantiating critical questions associated with this schema, e.g.: *does the action have a side effect which demotes some other value?*; *are there alternative ways of realising the same goal?*

Acknowledgements. This work was funded by the European Commission's Information Society Technologies programme, under the IST-FP6-002307 ASPIC project.

References

1. L. Amgoud. A formal framework for handling conflicting desires. In *Proc. 7th European Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty (ECSQARU'2003)*, pages 552–563, 2003.
2. L. Amgoud and C. Cayrol. On the use of an ATMS for handling conflicting desires. In *Proc. Ninth International Conference on Principles of Knowledge Representation and Reasoning (KR'04)*, pages 175–182, 2004.

3. ASPIC. Deliverable D2.2 - Draft formal semantics for inference and decision-making.
4. K. M. Atkinson, T. J. M. Bench-Capon, and P. McBurney. A dialogue game protocol for multi-agent argument for proposals over action. In I. Rahwan, P. Moraitis, and C. Reed, editors, *Proc. First International Workshop on Argumentation in Multi-Agent Systems (ArgMAS 2004)*. Springer, 2004.
5. T. J. M. Bench-Capon. Persuasion in practical argument using value-based argumentation frameworks. *Journal of Logic and Computation*, 13(3):429–448, 2003.
6. P. M. Dung. On the acceptability of arguments and its fundamental role in non-monotonic reasoning, logic programming and n-person games. *Artificial Intelligence*, 77:321–357, 1995.
7. D. Hitchcock, P. McBurney, and S. Parsons. A framework for deliberation dialogues. In H. V. Hansen et.al, editor, *Proc. Fourth Biennial Conference of the Ontario Society for the Study of Argumentation (OSSA 2001)*, Canada, 2001.
8. J. Hulstijn and L. van der Torre. Combining goal generation and planning in an argumentation framework. In *Proc. 15th Belgium-Netherlands Conference on Artificial Intelligence (BNAIC'03)*, 2003.
9. Antonis Kakas and Pavlos Moraitis. Argumentation based decision making for autonomous agents. In *Proc. Second international joint conference on Autonomous agents and multiagent systems*, pages 883–890. ACM Press, 2003.
10. J. L. Pollock. Defeasible reasoning. *Cognitive Science*, 11:481–518, 1987.
11. H. Prakken and G. Sartor. Argument-based extended logic programming with defeasible priorities. *Journal of Applied Non-Classical Logics*, 7:25–75, 1997.
12. C. Reed and T. J. Norman, editors. *Argument and multi-agent systems - Chapter2. In: Argumentation machines: New frontiers in argument and computation.* Kluwer Academic Publishers, 2004.