# Argumentative Deliberation for Autonomous Agents

**Antonis Kakas and Pavlos Moraïtis**
**Dept. of Computer Science**
**University of Cyprus**
**P.O.Box 20537, CY-1678 Nicosia, Cyprus**
**antonis@ucy.ac.cy, moraitis@ucy.ac.cy**

**Abstract.** This paper presents an argumentation based framework, developed as an extension of an existing framework for nonmonotonic reasoning, in order to support an agent's self deliberation process. The framework allows the agent to draw conclusions taking into account in a natural way a given preference policy. After developing the argumentation framework we examine two general cases of such argumentative deliberation: (a) under a preference policy that takes into account the roles agents can have within a context pertaining to an environment of interaction and (b) under a preference policy for the current needs of the agent emerging from his profile. In the first case we apply the argumentative deliberation model within a simple agent interaction scenario where each agent's self-deliberation determines, according to his own policy, his position at each step of the interaction process. In the second case we apply the framework to model motivational factors that apparently drive human behaviors and therefore can define agent personalities. Agents will thus similarly, as it is claimed in psychological literature for human beings, chose at any moment to pursue, those goals that are most compatible with their current motivations.

The proposed argumentation framework allows us to define policy preferences at different levels of deliberation resulting in modular representations of the agent's knowledge or personality profile. This high degree of modularity gives a simple computational model in which the agent's deliberation can be naturally implemented.

## 1 Introduction

Argumentation has had a renewed interest in Artificial Intelligence with several recent works studying its links to various problems such as the formalization of law, non-monotonic and common-sense reasoning, agent deliberation and dialogue and others. Abstract frameworks of argumentation are very powerful as they can encode many different problems but they face the challenge of doing so in a direct and natural way that at the same time is amenable to a simple computational model.

In this paper, we study an argumentation framework developed over the last decade as a result of a series of studies [12, 8, 7, 11, 10, 6] on the links of argumentation to non-monotonic reasoning. This framework, called Logic Programming without Negation as Failure ($LPwNF$), was proposed originally in [10] and can be seen as a realization of the more abstract frameworks of [7, 4]. The abstract attacking relation, i.e. its notion of argument and counter-argument, is realized through monotonic proofs of contrary conclusions and a priority relation on the sentences of the theory that make up these proofs. We extend the framework, following the more recent approach of other works [23, 5] to allow this priority relation and thus the attacking relation to be dynamic, making the framework more suitable for applications.

We claim that this extended argumentation framework is a natural argumentation framework. But how should we define the naturality of an argumentation framework? To do so we can set the following desiderata for naturality:

- the framework must be *simple* employing a small number of basic notions e.g. a uniform single notion of attack between arguments
- the encoding of a problem within the framework must be *directly* related to the high-level specification of the problem
- the representations of problems must be *modular*, with changes in the problem accommodated locally within the argumentation theory
- the argumentative reasoning and its computation must be *modular and local* to the problem task or query at hand

These properties are motivated from the perspective of a viable computational model of argumentation. This list of desiterata is not meant to be a complete list but rather that these are good properties that one would expect from a natural argumentation framework. Ultimately, the best criterion of the naturality of a framework is the test whether it can be applied, exhibiting the above properties, to capture different forms of natural human argumentative reasoning thus formalizing natural behaviour.

For this reason after developing our argumentation framework we test this by studying in detail how it can be used to capture agent deliberation in a dynamic external environment. In particular, we examine two problems: (a) argumentative deliberation of an agent according to a given decision policy on a domain of interest that takes into account the roles filled by the agents and the context of the external environment, and (b) argumentative deliberation of an agent about his needs according to a meta theory of ”personality” related preferences.

In this work, we adopt the idea that an agent is composed of a set of modules each of them being responsible for a particular functionality, and all together implementing the agent's overall behavior (e.g. problem solving, cooperation, communication, etc.). Therefore we consider that the proposed argumentative deliberation model can be used in order to implement the various decision making processes needed by different modules of an agent. For example, the decision for the choice and achievement of a goal (within the problem solving module) or the decision for the choice of the appropriate partners according to a specific cooperation protocol (within the cooperation module), etc.

Over the last few years argumentation is becoming increasingly important in agent theory. Several works have proposed argumentation models in the multi-agent field [28, 27, 21, 16, 3, 1, 2]. Our work can be seen as bringing together work from [27, 2] who have suggested that roles can affect an agent's argumentation, especially within the context of a dialogue, and work from [23, 5] who have shown the need for dynamic priorities within an argumentation framework when we want to apply this to formalize law and other related problems. In this paper, we put together these ideas proposing a new argumentation framework for agent deliberation obtained by extending the argumentation framework of $(LPwNF)$ [10, 6] to include dynamic priorities. We also employ a simple form of abduction to deal with the incompleteness and evolving nature of the agent's knowledge of the external environment.

We show how our framework can encompass the influence that the different relative roles of interacting agents and the context of the particular interaction can have on the deliberation process of the agents. Roles and context define in a natural way dynamic priorities on the argumentative decision rules of the agent at two different levels in the deliberation process. These priorities are represented within the overall argumentation theory of the agent in two corresponding modular parts. The use of this argumentative deliberation framework is demonstrated within an interaction protocol where the agent's deliberation helps him to decide his position.

Our use of the same argumentation framework to model agent motivations and through that, agent personalities, is inspired by the classical work of Maslow [17] in which he sets up a theory of hierarchy of human needs (physiological, safety, affiliation, achievement, learning) corresponding to motivational factors that drive human behavior. According to this theory human beings consider their unsatisfied needs in an order and decide to satisfy first those that are lower (and hence more important) in the hierarchy before considering higher needs. In the agent literature, Maslow's theory is already used by [18, 19] for guiding the behavior of deliberative and reactive agents in various unpredictable environments. To our knowledge our work is the first time where argumentation is used to model Maslow's hierarchy and other similar agent personalities where the mechanism for choosing which need to address next is carried out via a process of argumentative deliberation.

Section 2 presents the extension of the basic argumentation framework of $LPwNF$ with dynamic priorities. It also gives the basic concepts of roles and context and how these are captured through dynamic priorities in argumentation. Section 3 studies a simple interaction protocol based on argumentative deliberation. Section 4 presents how we model within our argumentation framework a hierarchy of needs of an agent and how these are chosen via argumentative deliberation. Section 5 discusses related and future work.

## 2 Argumentative Deliberation

An agent has his own theory expressing the knowledge under which he will take decisions. This decision process needs to compare alternatives and arrive at a conclusion that reflects a certain policy of the agent. In this paper we formalize this type of agent reasoning via argumentation where the deliberation of an agent is captured through an argumentative evaluation of arguments and counter-arguments.

There are several frameworks of argumentation proposed recently (e.g. [22, 4]) that could be adopted for formalizing an agent's deliberation. We will use the framework presented in [10, 6], called *Logic Programming without Negation as Failure* $(LPwNF)$ (The historical reasons for this name are not directly relevant to this paper). We briefly review this framework and then study its extension needed to accommodate roles and context in argumentative deliberation.

In $LPwNF$ a non-monotonic argumentation theory is viewed as a pool of sentences (or rules) from which we must select a suitable subset, i.e. an argument, to reason with, e.g. to support a conclusion. Sentences in a $LPwNF$ theory are written in the usual extended logic programming language with an explicit negation, but without the Negation as Failure (NAF) operator. We will often refer to the sentences of a theory as argument rules. In addition, these rules may be assigned locally a "relative strength" through a partial ordering relation. For example, we may have

$fly(X) \leftarrow bird(X)$ $\qquad \neg fly(X) \leftarrow penguin(X)$
$bird(X) \leftarrow penguin(X)$ $\qquad bird(tweety)$

with an ordering relation between the rules that assigns the second rule higher than the first. This theory captures the usual example of "flying birds" with its exceptions, without the use of explicit qualifications of the default rules with abnormality conditions. We can conclude that $tweety$ flies since we can derive this from the first rule and there is no way to derive $\neg fly(tweety)$. We have an argument (i.e. a proof) for $fly(tweety)$ but no argument for $\neg fly(tweety)$. If we add to the theory $penguin(tweety)$ then we can derive both $fly(tweety)$ and $\neg fly(tweety)$ - we have an argument for either conclusion. But in the non-monotonic argumentation semantics of the theory we can only conclude $\neg fly(tweety)$. This overrides $fly(tweety)$ since the argument that derives $\neg fly(tweety)$ contains the second rule which is designated higher than the first rule which belongs to the argument that derives $fly(tweety)$. We say that the argument for $\neg fly(tweety)$ **attacks** the argument for $fly(tweety)$ but not vice-versa. In general, the argumentation-based framework of $LPwNF$ is defined as follows.

**Definition 1** *Formulae in the* **background logic**[1] $(\mathcal{L}, \vdash)$ *of the framework are defined as* $L \leftarrow L_1, \ldots, L_n$, *where* $L, L_1, \ldots, L_n$ *are positive or explicit negative literals. The derivability relation,* $\vdash$, *of the logic is given by the single inference rule of modus ponens.*

Together with the set of sentences of a theory $\mathcal{T}$, we are given an ordering relation $<$ on these sentences (where $\phi < \psi$ or $<(\phi, \psi)$ means that $\phi$ has lower priority than $\psi$). The role of the priority relation is to encode locally the relative strength of argument rules in the theory. The relation $<$ is required to be irreflexive.

**Definition 2** *An* **argumentation theory** $(\mathcal{T}, <)$ *is a set of sentences* $\mathcal{T}$ *in* $\mathcal{L}$ *together with a priority relation* $<$ *on the sentences of* $\mathcal{T}$. *An* **argument** *for a literal* $L$ *in a theory* $(\mathcal{T}, <)$ *is any subset of* $\mathcal{T}$ *that derives* $L$, $T \vdash L$, *under the background logic.*

In general, we can separate out a part of the theory $\mathcal{T}_0 \subset \mathcal{T}$ (e.g. the last two rules of the example above) and consider this as a non-defeasible part from which any argument rule can draw information that it might need. The notion of attack between arguments in a theory $\mathcal{T}$ is based on the possible conflicts between a literal $L$ and its explicit negation $\neg L$ and on the priority relation $<$ on $\mathcal{T}$.

**Definition 3** *Let* $(\mathcal{T}, <)$ *be a theory and* $T, T' \subseteq \mathcal{T}$. *Then* $T'$ **attacks** $T$ *(or* $T'$ *is a* **counter argument** *of* $T$*) iff there exists* $L$, $T_1 \subseteq T'$ *and* $T_2 \subseteq T$ *s.t.:*

*(i)* $T_1 \vdash_{min} L$ *and* $T_2 \vdash_{min} \neg L$
*(ii)* $(\exists r' \in T_1, r \in T_2$ *s.t.* $r' < r) \Rightarrow (\exists r' \in T_1, r \in T_2$ *s.t.* $r < r')$.

[1] The background logic of this argumentation framework can be replaced with any monotonic first order logic.

Here $T \vdash_{min} L$ means that $T \vdash L$ under the background logic and that $L$ can not be derived from any proper subset of $T$. The second condition in this definition states that an argument $T'$ for $L$ attacks an argument $T$ for the contrary conclusion only if the set of rules that it uses to prove $L$ are at least of the same strength (under the priority relation $<$) as the set of rules in $T$ used to prove the contrary. Note that the attacking relation is not symmetric.

Using this notion of attack we then define the central notions of an *admissible argument* of a given theory and the non-monotonic argumentation consequence relation of a given theory as follows.

**Definition 4** *Let $(\mathcal{T}, <)$ be a theory and $T$ a subset of $\mathcal{T}$. Then $T$ is* **admissible** *iff $T$ is consistent and for any $T' \subseteq \mathcal{T}$ if $T'$ attacks $T$ then $T$ attacks $T'$.*

**Definition 5** *Let $T = (\mathcal{T}, <)$ be a theory and $L$ a ground literal. Then $L$ is a* **credulous (resp. skeptical) consequence** *of $T$ iff $L$ holds in a (resp. every) maximal (wrt set inclusion) admissible subset of $\mathcal{T}$.*

## 2.1 Roles and Context

Agents are always integrated within a (social) environment of interaction. We call this the *context* of interaction. This determines relationships between the possible roles the different agents can have within the environment. We consider, in line with much of the agent literature, (e.g. [20, 30]), a *role* as a set of behaviour obligations, rights and privileges determining its interaction with other roles.

Generally, the substance of roles is associated to a *default context* that defines shared social relations of different forms (e.g. authority, friendship, relationship, etc.) and specifies the behaviour of roles between each others. Consequently, it implicitly installs a partial order between roles that can expresses preferences of behaviour. For instance in the army context an officer gives orders that are obeyed by a soldier, or in a everyday context we respond in favour more easily to a friend than to a stranger. However, a default context that determines the basic roles filled by the agents is not the only environment where they could interact. For example, two friends can also be colleagues or an officer and a soldier can be family friends in civil life. Therefore we consider a second level of context, called *specific context*, which can overturn the pre-imposed, by the default context, ordering between roles and establish a different social relation between them. For instance, the authority relationship between an officer and a soldier would change under the specific context of a social meeting at home or the specific context of treason by the officer.

## 2.2 Argumentation with Roles and Context

In order to accommodate in an agent's argumentative reasoning the roles and context as described above we can extend the framework of $LPwNF$ so that the priority relation of a theory is not simply a static relation but a dynamic relation that captures the non-static preferences associated to roles and context. There is a natural way to do this. Following the same philosophy of approach as in [23], the priority relation can be defined as part of the agent's theory $\mathcal{T}$ and then be given the same argumentation semantics along with the rest of the theory.

We distinguish the part of the theory that defines the priority relation by $\mathcal{P}$. Rules in $\mathcal{P}$ have the same form as any other rule, namely $L \leftarrow L_1, \ldots, L_n$ where the head $L$ refers to the higher-priority relation, i.e. $L$ has the general form $L = h\_p(rule1, rule2)$. Also for any ground atom $h\_p(rule1, rule2)$ its negation is denoted by $h\_p(rule2, rule1)$ and vice-versa. For simplicity of presentation we will assume that the conditions of any rule in the theory do not refer to the predicate $h\_p$ thus avoiding self-reference problems. We now need to extend the semantic definitions of attack and admissibility.

**Definition 6** *Let $(\mathcal{T}, \mathcal{P})$ be a theory, $T, T' \subseteq \mathcal{T}$ and $P, P' \subseteq \mathcal{P}$. Then $(T', P')$* **attacks** *$(T, P)$ iff there exists a literal $L$, $T_1 \subseteq T'$, $T_2 \subseteq T$, $P_1 \subseteq P'$ and $P_2 \subseteq P$ s.t.:*

*(i)* $T_1 \cup P_1 \vdash_{min} L$ *and* $T_2 \cup P_2 \vdash_{min} \neg L$
*(ii)* $(\exists r' \in T_1 \cup P_1, r \in T_2 \cup P_2$ *s.t.* $T \cup P \vdash h\_p(r, r')) \Rightarrow (\exists r' \in T_1 \cup P_1, r \in T_2 \cup P_2$ *s.t.* $T' \cup P' \vdash h\_p(r', r))$.

Here, when $L$ does not refer to $h\_p$, $T \cup P \vdash_{min} L$ means that $T \vdash_{min} L$. This extended definition means that a composite argument $(T', P')$ is a counter-argument to another such argument when they derive a contrary conclusion, $L$, and $(T' \cup P')$ makes the rules of its counter proof at least ”as strong” as the rules for the proof by the argument that is under attack. Note that now the attack can occur on a contrary conclusion $L$ that refers to the priority between rules.

**Definition 7** *Let $(\mathcal{T}, \mathcal{P})$ be a theory, $T \subseteq \mathcal{T}$ and $P \subseteq \mathcal{P}$. Then $(T, P)$ is* **admissible** *iff $(T \cup P)$ is consistent and for any $(T', P')$ if $(T', P')$ attacks $(T, P)$ then $(T, P)$ attacks $(T', P')$.*

Hence when we have dynamic priorities, for an object-level argument (from $\mathcal{T}$) to be admissible it needs to take along with it priority arguments (from $\mathcal{P}$) to make itself at least ”as strong” as the opposing counter-arguments. This need for priority rules can repeat itself when the initially chosen ones can themselves be attacked by opposing priority rules and again we would need to make now the priority rules themselves at least as strong as their opposing ones.

Let us illustrate this extended form of argumentative reasoning with an example adapted from [23]. In this example, we are trying to formalise a piece of legislation that refers to whether or not we should modify an old building. In the first part,$\mathcal{T}$, of the theory we have the object-level law that refers directly to this particular topic:
$r_1(X) : \neg modify(X) \leftarrow protected(X)$
$r_2(X) : modify(X) \leftarrow needs\_repair(X)$

In addition, we have a theory $\mathcal{P}$ that represents the priorities between these (and other) laws as captured by another (more general) part of the law that deals with the relative strength of different types of regulations:
$rr_1(L_a, L_b) : h\_p(L_a(X), L_b(X)) \leftarrow art\_protect\_law(L_a(X)),$
$planning\_law(L_b(X))$
$rr_2(L_a, L_b) : h\_p(L_a(X), L_b(X)) \leftarrow art\_protect\_law(L_b(X)),$
$preservation\_law(L_a(X))$
$rr_3(rr_2, rr_1) : h\_p(rr_2(L_a(X), L_b(X)), rr_1(L_a(X), L_b(X))) \leftarrow$
$dangerous(X).$

The first of these states that a law for artistic protection is generally stronger than a planning law whereas the second says that a law for the preservation of an old building is generally stronger than an artistic protection law. The third statement stipulates that in the particular case of a building that is dangerous to the public then the law that gives higher priority to preservation laws over artistic protection laws is stronger than the law that gives higher strength to artistic protection laws over planning laws.

We also have in the non-defeasible part $\mathcal{T}_0$ of the theory some general information on the type of these laws together with information on a particular case for a $villa_0$:
$preservation\_law(r_2(X)) \leftarrow serious\_damage(X)$
$art\_protect\_law(r_1(X)) \qquad planning\_law(r_2(X))$

$protected(villa_0) \qquad needs\_repair(villa_0)$
$serious\_damage(villa_0) \qquad dangerous(villa_0).$

Should we modify $villa_0$ or not and how do we argue the case for our conclusion? We have two conflicting object-level arguments relating to the modification of $villa_0$. These are $\Delta_1 = (\{r_1(villa_0)\}\{\})$ for $\neg modify(villa_0)$ and $\Delta_2 = (\{r_2(villa_0)\}\{\})$ for $modify(villa_0)$. We can strengthen these arguments by adding priority rules in them. If we extend $\Delta_1$ to $\Delta_1^{'} = (\Delta_1, \{rr_1(r_1(villa_0), r_2(villa_0))\})$ then for $\Delta_2$ to attack back $\Delta_1^{'}$ it needs to extend itself to $\Delta_2^{'} = (\Delta_2, \{rr_2(r_2(villa_0), r_1(villa_0))\})$. Now these extended arguments have another conflict on the priority between the object level rules $r_1, r_2$, i.e. on $h\_p(r_1(villa_0), r_2(villa_0))$. $\Delta_1^{'}$ and $\Delta_2^{'}$ attack each other on this. But $\Delta_2^{'}$ can strengthen its argument for $h\_p(r_2(villa_0), r_1(villa_0))$ by adding in its priority rules the rule $\{rr_3(rr_2, rr_1)\}$. In fact, if we consider the attack on $\Delta_1^{'}$ given by $(\{\}, \{rr_2(r_2(villa_0), r_1(villa_0)), rr_3(rr_2, rr_1)\})$ there is no way to extend $\Delta_1^{'}$ so that it attacks this back. Hence $\Delta_1^{'}$ (and $\Delta_1$) is not admissible. We only have admissible sets that derive $modify(villa_0)$ and hence this is a skeptical conclusion.

This example illustrates in particular how we can take into account the relative strength that different types of law have on the reasoning. The types of law act as roles with relative importance which depends on the particular context under which we are examining the case.

We can now define an agent's argumentation theory for describing his policy in an environment with roles and context as follows.

**Definition 8** *An agent's* **argumentative policy theory or theory,** $T$**,** *is a triple $T = (\mathcal{T}, \mathcal{P}_R, \mathcal{P}_C)$ where the rules in $\mathcal{T}$ do not refer to $h\_p$, all the rules in $\mathcal{P}_R$ are priority rules with head $h\_p(r_1, r_2)$ s.t. $r_1, r_2 \in \mathcal{T}$ and all rules in $\mathcal{P}_C$ are priority rules with head $h\_p(R_1, R_2)$ s.t. $R_1, R_2 \in \mathcal{P}_R \cup \mathcal{P}_C$.*

We therefore have three levels in an agent's theory. In the first level we have the rules $\mathcal{T}$ that refer directly to the subject domain of the agent. We call these the **Object-level Decision Rules** of the agent. In the other two levels we have rules that relate to the policy under which the agent uses his object-level decision rules according to roles and context. We call the rules in $\mathcal{P}_R$ and $\mathcal{P}_C$, **Role (or Default Context) Priorities** and **(Specific) Context Priorities** respectively.

As an example, consider the following theory $\mathcal{T}$ representing (part of) the object-level decision rules of an employee in a company.
$r_1 : give(A, Obj, A_1) \leftarrow requests(A_1, Obj, A)$
$r_2 : \neg give(A, Obj, A_1) \leftarrow needs(A, Obj)$
$r_3 : \neg give(A, Obj, A_2) \leftarrow give(A, Obj, A_1), A_2 \neq A_1.$
In addition, we have a theory $\mathcal{P}_R$ representing the general default behaviour of the code of contact in the company relating to the roles of its employees: a request from a superior is in general stronger than an employee's own need; a request from another employee from a competitor department is in general weaker than its own need. (Here and below we will use capitals to name the priority rules but these are not to be read as variables).
$R_1 : h\_p(r_1(A, Obj, A_1), r_2(A, Obj, A_1)) \leftarrow higher\_rank(A_1, A)$
$R_2 : h\_p(r_2(A, Obj, A_1), r_1(A, Obj, A_1)) \leftarrow competitor(A, A_1)$
$R_3 : h\_p(r_1(A, Obj, A_1), r_1(A, Obj, A_2)) \leftarrow higher\_rank(A_1, A_2)$
Between the two alternatives to satisfy a request from a superior from a competing department or not, the first is stronger when these two departments are in the specific context of working together on a common project. On the other hand, if we are in a case where the employee who has an object and needs it, needs this urgently then s/he would prefer to keep it. Such policy is represented at the third level in $\mathcal{P}_C$:
$C_1 : h\_p(R_1(A, Obj, A_1), R_2(A, Obj, A_1)) \leftarrow common(A, Obj, A_1)$
$C_2 : h\_p(R_2(A, Obj, A_1), R_1(A, Obj, A_1)) \leftarrow urgent(A, Obj).$
Note the *modularity* of this representation. For example, if the company decides to change its policy "that employees should generally satisfy the requests of their superiors" to apply only to the direct manager of an employee we would simply replace $R_1$ by the new rule $R_1'$ without altering any other part of the theory:
$R_1' : h\_p(r_1(A, Obj, A_1), r_2(A, Obj, A_1)) \leftarrow manager(A_1, A).$

Consider now a scenario where we have two agents $ag_1$ and $ag_2$ working in competing departments and that $ag_2$ requests an object from $ag_1$. This is represented by extra statements in the non-defeasible part, $\mathcal{T}_0$, of the theory, e.g. $competitor(ag_2, ag_1)$, $requests(ag_2, obj, ag_1)$. Should $ag_1$ give the object to $ag_2$ or not?

If $ag_1$ does not need the object then, there are only admissible arguments for giving the object, e.g. $\Delta_1 = (\{r_1(ag_1, obj, ag_2)\}, \{\})$ and supersets of this. This is because this does not have any counter-argument as there are no arguments for not giving the object since $needs(ag_1, obj)$ does not hold. Suppose now that $needs(ag_1, obj)$ does hold. In this case we do have an argument for not giving the object, namely $\Delta_2 = (\{r_2(ag_1, obj, ag_2)\}, \{\})$. This is of the same strength as $\Delta_1$ but the argument $\Delta_2^{'}$, formed by replacing in $\Delta_2$ its empty set of rules of priority with $\{R_2(r_2(ag_1, obj, ag_2), r_1(ag_1, obj, ag_2))\}$, attacks $\Delta_1$ and any of its supersets but not vice-versa: $R_2$ gives higher priority to the rules of $\Delta_2$ and there is no set of priority rules with which we can extend $\Delta_1$ to give its object-level rules equal priority as those of $\Delta_2$. Hence we conclude skeptically that $ag_1$ will not give the object. This skeptical conclusion was based on the fact that the theory of $ag_1$ cannot prove that $ag_2$ is of higher rank than himself. If the agent learns that $higher\_rank(ag_2, ag_1)$ does hold then $\Delta_2^{'}$ and $\Delta_1^{'}$, obtained by adding to the priority rules of $\Delta_1$ the set $\{R_1(r_1(ag_1, obj, ag_2), r_2(ag_1, obj, ag_2))\}$, attack each other. Each one of these is an admissible argument for not giving or giving the object respectively and so we can draw both conclusions credulously.

Suppose that we also know that the requested object is for a common project of $ag_1$ and $ag_2$. The argument $\Delta_2^{'}$ is now not admissible since now it has another attack obtained by adding to the priority rule of $\Delta_1^{'}$ the extra priority rule $C_1(R_1(ag_1, obj, ag_2), R_2(ag_1, obj, ag_2))$ thus strengthening its derivation of $h\_p(r_1, r_2)$. The attack now is on the contrary conclusion $h\_p(r_1, r_2)$. In other words, the argumentative deliberation of the agent has moved one level up to examine what priority would the different roles have, within the specific context of a common project. $\Delta_2^{'}$ cannot attack back this attack and no extension of it exists that would strengthen its rules to do so. Hence there are no admissible arguments for not giving and $ag_1$ draws the skeptical conclusion to give the object.

We have seen in the above example that in several cases the admissibility of an argument depends on whether we have or not some background information about the specific case in which we are reasoning. For example, $ag_1$ may not have information on whether their two departments are in competition or not. This means that $ag_1$ cannot build an admissible argument for not giving the object as he cannot use the priority rule $R_2$ that it might like to do. But this information maybe just unknown and if $ag_1$ wants to find a way to refuse the request he can reason further to find *assumptions* related to the unknown information under which he can build an admissible argument. Hence in this example he would build an argument for not giving the object to $ag_2$ that is *conditional* on the fact that they belong to competing departments. Furthermore, this type of information may itself be dynamic and change while the rest of the theory of the agent remains fixed, e.g. $ag_1$ may have in his theory that $ag_2$ belongs to a competing department but he has not yet learned that $ag_2$ has changed department or that his department is no longer a competing one.

We can formalize this conditional form of argumentative reasoning by defining the notion of *supporting information* and extending argumentation with *abduction* on this missing information.

**Definition 9** *Let $T = (\mathcal{T}, \mathcal{P})$ be a theory, and $\mathcal{A}$ a distinguished set of predicates in the language of the theory, called* **abducible** *predicates. Given a goal $G$, a set $S$ of abducible literals consistent with the non-defeasible part $\mathcal{T}_0$ of $T$, is called a* **strong (resp. weak) supporting evidence** *for $G$ iff $G$ is a skeptical (resp. credulous) consequence of $(\mathcal{T} \cup S, \mathcal{P})$.*

The structure of an argument can also be generalized as follows.

**Definition 10** *Let $T = (\mathcal{T}, \mathcal{P})$ be a theory and $\mathcal{A}$ its abducible predicates. A* **supported argument** *in $T$ is a tuple $(\Delta, S)$, where $S$ is a set of abducible literals consistent with $\mathcal{T}_0$ and $\Delta$ is a set of argument rules in $T$, which is not admissible in $T$, but is admissible in the theory $(\mathcal{T} \cup S, \mathcal{P})$. We say that $S$ supports the argument $\Delta$.*

The supporting information expressed through the abducibles predicates refers to the incomplete and evolving information of the external environment of interaction. Typically, this information pertains to the context of the environment, the roles between agents or any other aspect of the environment that is dynamic. We will see in section 3 how agents can acquire and/or validate such information through an interaction protocol where they exchange missing information.

Given the above framework the **argumentative deliberation** of an agent can be formalized via the following basic reasoning functions.

**Definition 11** *Let $Ag$ be an agent, $T$ his argumentation theory, $G$ a goal and $S$ a set of supporting information consistent with $\mathcal{T}_0$. Then we say that $Ag$* **deliberates** *on $G$ to produce $s^{ag}$, denoted by $deliberate(Ag, G, S; s^{ag})$, iff $s^{ag} \neq \{\}$ is a strong supporting evidence for $G$ in the theory $T \cup S$. If $s^{ag} = \{\}$ then we say that $Ag$* accepts *$G$ under $T \cup S$ and is denoted by* **accept(Ag,G,S)**. *Furthermore, given an opposing goal $\overline{G}$ (e.g $\neg G$) to $G$ and $s'$ produced by deliberation on $\overline{G}$, i.e. that $deliberate(Ag, \overline{G}, S; s')$ holds, we say that $s'$ is supporting evidence for agent $Ag$ to* **refuse** *$G$ in $T \cup S$.*

### 2.3 Modularity and Computation

As mentioned above, the proposed framework allows modular representations of problems where a change in the policy of an agent can be effected locally in his theory. The following results formalize some of the properties of modularity of the framework.

**Proposition 12** *Let $\Delta$ be a set of arguments that is admissible separately with respect to the theory $T_1 = (\mathcal{T}, \mathcal{P}_{R1}, \{\})$ and the theory $T_2 = (\mathcal{T}, \mathcal{P}_{R2}, \{\})$. Then $\Delta$ is admissible with respect to the theory $T = (\mathcal{T}, \mathcal{P}_{R1} \cup \mathcal{P}_{R2}, \{\})$. Similarly, we can decompose $\mathcal{P}_C$ into $\mathcal{P}_{C1}$ and $\mathcal{P}_{C2}$.*

**Proposition 13** *Let $\Delta$ be a set of arguments that is admissible with respect to the theory $T_1 = (\mathcal{T}, \mathcal{P}_R, \{\})$. Suppose also that $\Delta$ is admissible with respect to $T_2 = (\mathcal{T} \cup \mathcal{P}_R, \{\}, \mathcal{P}_C)$. Then $\Delta$ is admissible with respect to $T = (\mathcal{T}, \mathcal{P}_R, \mathcal{P}_C)$.*

The later proposition shows that we can build an admissible argument $\Delta = (O, R)$ by joining together an object-level argument $O$ together with a set of priority rules $R$ that makes $O$ admissible and is itself admissible with respect to the higher level of context priorities. These results provide the basis for a modular computational model in terms of interleaving levels of admissibility processes one for each level of arguments in the theory.

In general, the basic $LPwNF$ has a simple and well understood computational model [6] that can be seen as a realization of a more abstract computational model for argumentation [14]. It has been successfully used [13] to provide a computational basis for reasoning about actions and change. The simple argumentation semantics of $LPwNF$, where the attacking relation between arguments depends only on the priority of the rules of a theory, gives us a natural "dialectical" proof theory for the framework. In this we have two types of interleaving derivations one for considering the attacks and one for counter attacking these attacks. The proof theory then builds an admissible argument for a given goal by incrementally considering all its attacks and, whenever an attack is not counter-attacked by the argument that we have build so far, we extend this with other arguments (rules) so that it does so. This in turn may introduce new attacks against it and the process is repeated.

The priorities amongst the rules help us move from one type of derivation to the other type e.g. we need only consider attacks that come from rules with strictly higher priority than the rules in the argument that we are building (as otherwise the argument that we have so far will suffice to counter attack these attacks.) For the more general framework with dynamic priorities we apply the same proof theory extended so that a derivation can be split into levels. Now a potential attack can be avoided by ensuring that its rules are not of higher priority than the argument rules we are building and hence we move the computation one level up to attacks and counter attacks on the priorities of rules. This move one level can then be repeated to bring us to a third level of computation.

This extended proof theory has been implemented and used to build agents that deliberate in the face of complete (relevant) information of their environments. We are currently investigating how to extend this implementation further with (simple forms of ground) abduction, required for the computation of supporting evidence in the face of incomplete information about the environment, using standard methods from abductive logic programming.

## 3 Argumentation based Agent Interaction

In this section, we study the use of the argumentative deliberation of an agent, defined above, within a simple interaction protocol where two agents are trying to agree on some goal, as an example of how this argumentation framework can be used within the different decision making processes of an agent. In our study of this we will be mainly interested how agents can use their argumentative deliberation in order to decide their position at each step of the interaction process. We will not be concerned with the conversation protocol supporting the agent interaction.

Each agent builds his reaction according to his internal argumentative policy theory, his current goal and other supporting information about the external environment that he has accumulated from the other agent. This extra supporting information is build gradually during the interaction and it allows an incremental deliberation of the agents as they acquire more information.

In the specific interaction protocol that we will consider, each agent insists in proposing his own goal as long as his deliberation with his theory and the accumulated supporting information (agreed by the two agents so far) produces new supporting evidence for this goal, suitable to convince the other agent. The first of the two interacting agents, who is unable to produce a new such supporting evidence, abandons his own goal and searches for supporting information, if any, under which he can accept the goal of the other agent (e.g. a seller agent unable to find another way to support his high price considers selling at a cheap price, provided that the buyer has a regular account and pays cash). In such a case, if the receiver agent can endorse the proposed supporting information the interaction ends with an agreement on this goal and the supporting information accumulated so far. Otherwise, if the receiver refuses some of the proposed supporting information the sender takes this into account and tries again to find another way to support the goal of the other agent. If this is not possible then the interaction ends in failure to agree on a common goal.

The following algorithm describes the steps of the interaction process presented above. Let us denote by X and Y the two agents, by $G^X$, $G^Y$ their respective goals, by S the knowledge accumulated during the interaction exchanges and by $s_i^X$, $s_i^Y$ the various supports that the agents generate in their deliberation. Note that when $G^X$, $G^Y$ are opposing goals any supporting evidence for one of these goals also forms a reason for refusing the other goal.

Besides the argumentative functions $deliberate$ and $accept$ given in the previous section, we need three more auxiliary functions, which are external to the argumentative reasoning of an agent and relate to other functions of the agent in the present interaction protocol. The function $propose(Goal, e_j, s_i)$ is used by a sender agent to determine what information to send to the other agent: $Goal$ is a goal proposed, $e_j$ is the evaluation by the sender of the supporting information $s_j$ sent to him in the previous step by the other agent, and $s_i$ is a new supporting evidence produced by the deliberation function of the sender. The function $evaluate(Ag, s_i)$ produces $e_i$ where each (abducible) literal in the supporting information $s_i$ may remain as it is or negated according to some external process of evaluation of this by an agent $Ag$. The function $update(S, e)$ updates, through an external mechanism, the accumulated supporting information $S$ with the new information $e$ consisting of the agent's evaluation of the supporting evidence sent by the other agent and the evaluation of his own supporting information by the other agent.

As described above, the interaction protocol has two phases. Phase 1 where each agent insists on its own goal and Phase 2 where they are trying to agree on the goal of one of the two agents. In the definition below Phase 2 supposes that agent X initiates this phase.

```
Phase 1
  Step 1 (Agent X starts the Interaction)
      Agent X propose($G^X$, $\varnothing$, $s_n^X$) to Y
  Step 2 (Agent Y)
      $e_n^X \leftarrow evaluate(Y, s_n^X)$; $S \leftarrow update(S, e_{n-1}^Y \cup e_n^X)$
      If Y accept($Y, G^X, S$) then End(agreement, $G^X$)
      Else  $n \leftarrow n+1$; agent Y deliberate ($Y, G^Y, S; s_n^Y$)
        If $s_n^Y$ exists then propose($G^Y, e_{n-1}^X, s_n^Y$) to X
        Else Start Phase 2
  Step 3 (Agent Y)
      $e_n^Y \leftarrow evaluate(X, s_n^Y)$; $S \leftarrow update(S, e_{n-1}^X \cup e_n^Y)$
      If X accept($X, G^Y, S$) then End(agreement, $G^Y$)
      Else $n \leftarrow n+1$; agent X deliberate ($X, G^X, S; s_n^X$)
         If $s_n^X$ exists then propose($G^X, e_{n-1}^Y, s_n^X$) to Y
         Goto step2
         Else Start Phase 2
```

```
Phase 2
  Step 1 (Agent X)
      $S \leftarrow update(S, e_n^X)$; $n \leftarrow n+1$
      agent X deliberate($X, G^Y, S; s_n^X$)
       If $s_n^X$ exists then propose($G^Y, \varnothing, s_n^X$) to Y
       Else End(Failure)
  Step 2 (Agent Y)
      $e_n^X \leftarrow evaluate(Y, s_n^X)$
      If $e_n^X = s_n^X$ then End(agreement, $G^Y$)
      Else propose ($G^Y, e^X, \varnothing$); Goto  step 1
```

We illustrate this algorithm with a buying-selling scenario between two agents, a seller called X who has the goal, $G^X$, to sell a product at a high price to another agent, the buyer, called Y, who has the (opposing) goal, $G^Y$, to buy this product at a low price. They are trying to find an agreement on the price by agreeing either on $G^X$ or on $G^Y$. We assume that the seller has the following argumentation policy for selling products. We present only a part of this theory.
The object-level theory $\mathcal{T}^X$ of the seller contains the rules:
$r_1 : sell(Prd, A, high\_price) \leftarrow pay\_normal(A, Prd)$
$r_2 : sell(Prd, A, high\_price) \leftarrow pay\_install(A, Prd)$
$r_3 : sell(Prd, A, low\_price) \leftarrow pay\_cash(A, Prd)$
$r_4 : \neg sell(Prd, A, P_2) \leftarrow sell(Prd, A, P_1), P_2 \neq P_1.$
His role and context priority theories, $\mathcal{P}_R^X$ and $\mathcal{P}_C^X$, are given below. They contain the policy of the seller on how to treat the various types of customers. For example, to prefer to sell with normal paying conditions over payment by installments when the buyer is a normal customer (see $R_1$). Also that there is always a preference to sell at high price (see $R_2, R_3$) but for regular customers there are conditions under which the seller would sell at low price (see $R_4, R_5$). This low price offer to a regular customer applies only when we are not in high season (see $C_1, C_2$).

$R_1 : h\_p(r_1(Prd, A), r_2(Prd, A)) \leftarrow normal(A)$
$R_2 : h\_p(r_1(Prd, A), r_3(Prd, A))$
$R_3 : h\_p(r_2(Prd, A), r_3(Prd, A))$
$R_4 : h\_p(r_3(Prd, A), r_1(Prd, A)) \leftarrow regular(A), buy\_2(A, Prd)$
$R_5 : h\_p(r_3(Prd, A), r_1(Prd, A)) \leftarrow regular(A), late\_del(A, Prd)$
$C_1 : h\_p(R_2(Prd, A), R_4(Prd, A)) \leftarrow high\_season$
$C_2 : h\_p(R_2(Prd, A), R_5(Prd, A)) \leftarrow high\_season$
$C_3 : h\_p(R_4(Prd, A), R_5(Prd, A)).$

Lets us consider the particular interaction scenario given below and study how the seller uses his own argumentative deliberation in this scenario.

**Phase1**
**Seller X (step 1):** propose($G^X$, $\varnothing$, $s_0$={pay normal})
**Buyer Y (step 2):** NO; $e_0=s_0$; S=($e_0$); deliberate (Y, $G^Y$, S; $s_1$={expensive price}); propose ($G^Y$, $e_0$, $s_1$)
**Seller X (step 3):** NO; $e_1=s_1$, S=($e_0 \cup e_1$); deliberate (X, $G^X$, S; $s_2$={pay install}); propose ($G^X$, $e_1$, $s_2$)
**Buyer Y (step 2):** NO; $e_2=\neg s_2$; S=($e_0 \cup e_1 \cup e_2$); deliberate(Y, $G^Y$, S; $s_3$={pay cash}); propose($G^Y$, $e_2$, $s_3$)
**Seller X (step 3):** NO; $e_3=s_3$; S=($e_0 \cup e_1 \cup e_2 \cup e_3$), deliberate (X, $G^X$, S, s); fails

**Phase 2**
**Seller X (step 1):** S=($e_0 \cup e_1 \cup e_2 \cup e_3$); deliberate (X, $G^Y$, S; $s_4$={regular cust, buy 2}); propose ($G^Y$, $\varnothing$ $s_4$)
**Buyer Y (step 2):** NO; $e_4$={regular cust, $\neg$buy 2}; propose($G^Y$, $e_4$, $\varnothing$)
**Seller X (step 1):** S=($e_0 \cup e_1 \cup e_2 \cup e_3 \cup e_4$); deliberate (X, $G^Y$, S; $s_5$={later delivery}); propose ($G^Y$, $\varnothing$, $s_5$)
**Buyer Y (step 2):** $e_5=s_5$;YES; End(agreement, $G^Y$)

At the third step of Phase1 the seller needs to see if he can find an argument to support his goal (of selling high) given the fact that the buyer considers the price expensive. Deliberating on his goal, he now finds another argument for selling high, using the object-level rule $r_2$ since he can no longer consider the buyer a normal customer and $R_1$ does not apply (the seller derives this from some general background knowledge that he has in $\mathcal{T}_0$ e.g. from a rule $\neg normal(A) \leftarrow expensive(A, high\_price)$). This new argument needs the support $pay\_install(buyer, prd)$ and the seller offers this information to the buyer.

At the last step of Phase1 the seller deliberates again on his own goal (to sell high) but cannot find a new solution anymore. He therefore initiates phase2 where he considers the goal of the buyer, i.e. to sell at $low\_price$ and finds that it is possible to do so if the customer is a regular one and he accepts some other conditions. He finds an admissible argument for low price using the object-level rule $r_3$ and the role priority rule $R_4$. This is conditional on the information that the buyer is indeed a regular customer, will pay cash and that he will buy two of the products. Note that for this argument to be admissible the context rule $C_1$ must not apply, i.e. the seller knows that currently they are not in a $high\_season$. The buyer confirms the first two conditions but refuses the third. The seller then has another solution to sell low to a regular customer conditional on late delivery.

It is easy to show the following result of termination and correctness of the above interaction protocol.

**Proposition 14** *Let $X, Y$ be two agents with $T_X$, $T_Y$ their respective argumentation policy theories such that for each goal, $G$, there exists only a finite number of different supporting evidence for $G$ in $T_X$ or $T_Y$. Then any interaction process between $X$ and $Y$ will terminate. Furthermore, if an interaction process terminates with agreement on a goal $G$ and $S$ is the final set of supporting information accumulated during the interaction then $G$ is a skeptical conclusion of both $T_X \cup S$ and $T_Y \cup S$.*

We also remark that the evaluation function, $evaluate(Ag, s_i)$, used by an agent within the interaction process in order to decide if he can accept a proposed supporting information $s_i$, can vary in complexity from a simple check in the agent's database on the one hand to a new (subsidiary) argumentative deliberation on $s_i$ according to a related argumentative policy theory that the agent may have.

## 4 Agent Deliberation on Needs and Motivations

In this section, we will study how the argumentation framework proposed in this paper can help us model the needs and motivations of an agent. In particular, we will examine the argumentative deliberation that an agent has to carry out in order to decide which needs to address at any current situation that he finds himself. This will then allows us to use the argumentation framework to specify different personalities of agents in a modular way independently from the other architectural elements of an agent.

We will apply the same approach as when we model a preference policy of an agent in a certain knowledge or problem domain, described in the previous sections. We now simply consider the domain of an agent's needs and motivations where, according to the type or personality of an agent, the agent has a default (partial) preference amongst the different types of needs. Hence now the type of need, or the motivation that this need addresses, plays an analogous role to that of Roles in the previous section. The motivations will then determine the basic behaviour of the agent in choosing amongst his different needs and whenever we have some specific context this may overturn the default decision of the agent for a particular need.

We will follow the work of Maslow [17] from Cognitive Psychology (see also [18, 19]) where needs are categorized in five broad classes according to the motivation that they address. These are **Physiological, Safety, Affiliation or Social, Achievement or Ego and Self-actualization or Learning**. As the world changes a person is faced with a set of potential goals from which it selects to pursue those that are "most compatible with her/his (current) motivations". We choose to eat if we are hungry, we protect ourselves if we are in danger, we work hard to achieve a promotion etc. The theory states that in general there is an ordering amongst these five motivations that we follow in selecting the corresponding goals. But this ordering is only followed in general under the assumption of "other things being equal" and when special circumstances arise it does not apply.

Our task here is then to model and encode such motivating factors and their ordering in a natural way thus giving a computational model for agent behaviour and personality.

Let us assume that an agent has a theory $\mathcal{T}$ describing the knowledge of the agent. Through this, together with his perception inputs, he generates a set of needs that he could possibly address at any particular situation that he finds himself. We will consider that these needs are associated to goals, G, e.g. to fill with petrol, to rest, to help someone, to promote himself, to help the community etc. For simplicity of presentation and without loss of generality we will assume that the agent can only carry out one goal at a time and thus any two goals activated by $\mathcal{T}$ oppose each other and a decision is needed to choose one. Again for simplicity we will assume that any one goal $G$ is linked only to one of the five motivations above, $m_j$, and we will thus write $G_j$, $j = 1, ..., 5$ to indicate this, with $m_1 = Physiological$, $m_2 = Safety$, $m_3 = Affiliation$, $m_4 = Achievement$, $m_5 = Self - actualization$.

Given this theory, $\mathcal{T}$, that generates potential goals an agent has a second level theory, $\mathcal{P_M}$, of priority rules on these goals according to their associated motivation. This theory helps the agent to choose amongst the potential goals that it has and forms part of his decision policy for this. It can be defined as follows.

**Definition 15** *Let $Ag$ be an agent with knowledge theory $\mathcal{T}$. For each motivation, $m_j$, we denote by $S_j$ the set of conditions, evaluated in $\mathcal{T}$, under which the agent considers that his needs pertaining to motivation $m_j$ are* **satisfied**. *Let us also denote by $N_j$ the set of conditions, evaluated in $\mathcal{T}$, under which the agent considers that his needs pertaining to motivation $m_j$ are* **critical**. *We assume that $S_j$ and $N_j$ are disjoint and hence $N_j$ corresponds to a subset of situations where $\neg S_j$ holds. Then the* **default motivation preference theory of** *$Ag$, denoted by $\mathcal{P_M}$, is a set of rules of the following form:*

- $R^1_{ij} : h\_p(G_i, G_j) \leftarrow N_i$
- $R^2_{ij} : h\_p(G_i, G_j) \leftarrow \neg S_i, \neg N_j$

*where $G_i$ and $G_j$ are any two potential goals, $(i \neq j)$, of the agent associated to motivations $m_i$ and $m_j$ respectively.*

The first rule refers to situations where we have a critical need to satisfy a goal $G_i$ whereas the second rule refers to situations where the need $G_j$ is not critical and so $G_i$ can be preferred.

Hence when the conditions $S_i$ hold an agent would not pursue goals of needs pertaining to this motivation $m_i$. In fact, we can assume that whenever a goal $G_i$ is activated and is under consideration that $\neg S_i$ holds. On the other side of the spectrum when $N_i$ holds the agent has an urgency to satisfy his needs under $m_i$ and his behaviour may change in order to do so. Situations where $\neg S_j$ and $\neg N_j$ both hold are in between cases where the decision of an agent to pursue a goal $G_j$ will depend more strongly on the other simultaneous needs that he may have. These conditions $S_i$ and $N_i$ vary from agent to agent and their truth is evaluated by the agent using his knowledge theory.

For example, when a robotic agent has $low\_energy$, that would make it non-functional, the condition $N_1$ is satisfied and a goal like $G_1 = fill\_up$ has, through the rules $R^1_{1j}$ for $j \neq 1$, higher priority than any other goal. Similarly, when the energy level of the robotic agent is at some middle value, i.e. $\neg S_1$ and $\neg N_1$ hold, then the robot will again consider, through the rules $R^2_{1j}$ for $j \neq 1$, the goal $G_1$ to fill up higher than other goals provided also that in such a situation there is no other goal whose need is critical. Hence if in addition the robotic agent is in great danger and hence $N_2$ holds then rule $R^2_{12}$ does not apply and the robot will choose goal $G_2 = self\_protect$ which gets a higher priority through $R^1_{21}$.

In situations as in this example, the agent has a clear choice of which goal to select. Indeed, we can show that under some suitable conditions the agent can decide deterministically in any situation.

**Proposition 16** *Let $\mathcal{P}_\mathcal{M}$ be a preference theory for an agent and suppose that $N_i \cap N_j = \emptyset$ $(i \neq j)$ and that $\neg S_j = N_j$ for each $j$. Then given any two goals $G_i, G_j$ only one of these goals belongs to an admissible extension of the agents theory and thus the agent at any situation has a deterministic choice of which need to address.*

Similarly, if we have $N_i \cap N_j = \emptyset$ and $\neg S_i \cap \neg S_j = \emptyset$ $(i \neq j)$ then the agent can always make a deterministic choice of which goal to choose to address in any current situation. But these conditions are too strong. There could arise situations where, according to the knowledge of the agent, two needs are not satisfied and/or where they are both urgent/critical. How will the agent decide which one to perform? The agent is in a *dilemma* as its theory will give an admissible argument for each need. For example, a robotic agent may at the same time be low in energy and in danger. Similarly, the robotic agent may be in danger but also need to carry out an urgent task of helping someone.

According to Maslow's theory decisions are then taken following a basic hierarchy amongst needs. For humans this basic hierarchy puts the Physiological needs above all other needs, Safety as the second most important with Affiliation, Achievement and Self-Actualization following in this order. Under this hierarchy a robotic agent would choose to fill its battery despite the danger or avoid a danger rather than give help. One way to model in $\mathcal{P}_\mathcal{M}$ such a hierarchy of needs that helps resolve the dilemmas is as follows. For each pair $k, l$ $s.t.$ $k \neq l$ the theory $\mathcal{P}_\mathcal{M}$ contains only one of the rules $R^1_{kl}$ or $R^1_{lk}$. Deciding in this way which priority rules, $R^1$, to include in the theory gives a basic profile to the agent.

But this would only give us a partial solution to the problem not resolving dilemmas that are not related to urgent needs and a similar decision needs to be taken with respect to the second category of rules, $R^2$, in $\mathcal{P}_\mathcal{M}$. More importantly this approach is too rigid in the sense that the chosen hierarchy in this way can never be overturned under any circumstance. Often we may want a higher degree of flexibility in modeling an agent and indeed Maslow's hierarchy of needs applies under the assumption of "other things being equal". In other words, there maybe special circumstances where the basic hierarchy in the profile of an agent should not be followed. For example, an agent may decide, despite his basic preference to avoid danger rather than help someone, to help when this is a close friend or a child.

We can solve these problems by extending the agent theory with a third level analogous to the specific context level presented in the previous sections.

**Definition 17** *An agent theory expressing his decision policy on needs is a theory $T = (\mathcal{T}, \mathcal{P}_\mathcal{M}, \mathcal{P}_\mathcal{C})$ where $\mathcal{T}$ and $\mathcal{P}_\mathcal{M}$ are defined as above and $\mathcal{P}_\mathcal{C}$ contains the following types of rules. For each pair of rules $R^k_{ij}, R^k_{ji}$ in $\mathcal{P}_\mathcal{M}$ we have the following rules in $\mathcal{P}_\mathcal{C}$:*

- $H^k_{ij} : h\_p(R^k_{ij}, R^k_{ji}) \leftarrow true$
- $E^k_{ji} : h\_p(R^k_{ji}, R^k_{ij}) \leftarrow sc^k_{ji}$
- $C^k_{ji} : h\_p(E^k_{ji}, H^k_{ij}) \leftarrow true$

*where $sc^k_{ji}$ are (special) conditions whose truth can be evaluated in $\mathcal{T}$. The rules $H^k_{ij}$ are called the* **basic hierarchy** *of the theory $T$ and the rules $E^k_{ji}$ the* **exception policy** *of the theory $T$. The theory $\mathcal{P}_\mathcal{C}$ contains exactly one of the basic hierarchy rules $H^k_{ij}$ and $H^k_{ji}$ for each $k = 1, 2$ $and$ $i \neq j$.*

Choosing which one of the basic hierarchy rules $H^k_{ij}$ or $H^k_{ji}$ to have determines the default preference of needs $G_i$ over $G_j$ or $G_j$ over $G_i$ respectively (for $k = 1$ in critical situations and for $k = 2$ in non-critical situations). The special conditions $sc_{ij}$ define the specific contexts under which this preference is overturned. They are evaluated by the agent in his knowledge theory $\mathcal{T}$. They could have different cases of definition that depend on the particular nature of the goals and needs that we are considering in the dilemma.

Each choice of the rules $H^k_{ij}$ to include in the agent theory, determining a basic hierarchy of needs, in effect gives a different agent with a different basic profile of behaviour. For example, if we have $H^k_{34}$ in $\mathcal{P}_\mathcal{C}$ (remember that $m_3 = Affiliation$ and $m_4 = Achievement$) we could say that this is an *altruistic* type of agent, since under normal circumstances (i.e. not exceptional cases defined by $sc^k_{43}$) he would give priority to the affiliation needs over the self-achievement needs. Whereas if we have $H^k_{43}$ we could consider this as a *selfish* type of agent.

To illustrate this let us consider the specific theory $\mathcal{P}_\mathcal{C}$ corresponding to Maslow's profile for humans. This will contain the following rules to capture the basic hierarchy of Physiological ( $m_1$) over Safety ($m_2$) and Safety over Affiliation ($m_3$):

- $H^k_{12} : h\_p(R^k_{12}, R^k_{21}) \leftarrow true, \ for \ k = 1, 2$
- $H^k_{13} : h\_p(R^k_{13}, R^k_{31}) \leftarrow true, \ for \ k = 1, 2$
- $H^k_{23} : h\_p(R^k_{23}, R^k_{32}) \leftarrow true, \ for \ k = 1, 2$
- $E^2_{21} : h\_p(R^2_{21}, R^2_{12}) \leftarrow sc^2_{21}$
- $C^2_{21} : h\_p(E^2_{21}, H^2_{12}) \leftarrow true$
- $E^2_{31} : h\_p(R^2_{31}, R^2_{13}) \leftarrow sc^2_{31}$
- $C^2_{31} : h\_p(E^2_{31}, H^2_{13}) \leftarrow true$
- $E^2_{32} : h\_p(R^2_{32}, R^2_{23}) \leftarrow sc^2_{32}$
- $C^2_{32} : h\_p(E^2_{32}, H^2_{23}) \leftarrow true.$

The conditions $sc^2_{21}$ are exceptional circumstances under which we prefer a safety need over a physiological need, e.g. $sc^2_{21}$ could be true if an alternative supply of energy exists. Similarly for $sc^2_{31}$ and $sc^2_{32}$. Note that if we are in a situation of critical physiological need (i.e $N_1$ holds and hence $R^1_{12}$ applies) then this theory has no exceptional circumstances (there is no $E^1_{21}$ rule) where we would not prefer to satisfy this physiological need over a critical safety need. Similarly, this profile theory does not allow any affiliation need to be preferred over a critical safety need; it does not allow a "heroic" behaviour of helping. If we want to be more flexible on this we would add the following rules in the profile:

- $E^1_{32} : h\_p(R^1_{32}, R^1_{23}) \leftarrow sc^1_{32}$
- $C^1_{32} : h\_p(E^1_{32}, H^1_{23}) \leftarrow true$

where the conditions $sc^1_{32}$ determine the circumstances under which the agent prefers to help despite the risk of becoming non-functional, e.g. when the help is for a child or a close friend in great danger.

Given any such profile theory $\mathcal{P}_\mathcal{C}$ we can show that an agent can always decide which goal to pursue once he can evaluate the $sc^k_{ij}$ special conditions independently in $\mathcal{T}$ alone.

**Proposition 18** *Let $T = (\mathcal{T}, \mathcal{P}_\mathcal{M}, \mathcal{P}_\mathcal{C})$ be an agent theory according to definition 17 and $G_i, G_j$ $(i \neq j)$ be any two potential goals addressing different needs. Then given any situation there exists an admissible argument for only one of the two goals.*

In practice, the agent when in a dilemma will need to deliberate on each of the two goals and produce supporting information for each goal. This information is the incomplete information from $N_i, \neg S_i$ and $sc^k_{ij}$ that the agent may be missing at the current situation. He would then be able to test (or evaluate) in the real world which one of these supporting information holds and thus enable him to make the decision which need to pursue.

Our argumentation based approach allows a high degree of flexibility in profiling deliberative agents. An agent's profile, defined via his $\mathcal{P}_\mathcal{M}$ and $\mathcal{P}_\mathcal{C}$ theories, is parametric on the particular rules we choose to adopt in both of these theories. In this paper we have adopted one possibility but this is certainly not the only one. For example, we could adopt a different underlying theory $\mathcal{P}_\mathcal{M}$ containing the basic priority rules amongst needs, rather than the fixed theory we have used in this paper, and use this as a new basis for profiling the agents. This issue needs to be studied further to examine the spectrum of different agents that can be build in this way.

## 5 Related Work and Conclusions

In this paper we have proposed an argumentative deliberation framework for autonomous agents and presented how this could be applied in different ways. We have argued that this framework has various desired properties of simplicity and modularity and in particular we have shown how it can capture some natural aspects of the behaviour of an autonomous agent. The framework can embody in a direct and modular way any preference policy of the agent and hence can be used to support the various decision making processes of an agent. It can be incorporated within different models of agent architecture. For example, it could be used within the BDI model to implement (with the necessary adaptations) the filter function [29] which represents the agent's deliberation process, for determining the agent's new intentions based on its current beliefs, desires and intentions. The proposed argumentation framework also has a simple and modular computational model that facilitates the implementation of deliberative agents.

The main characteristic of our argumentation framework is its modularity of representation and associated computation. Our work rests on the premise that for a computational framework of argumentation to be able to encapsulate *natural* forms of argumentation it is necessary for this framework to have a high degree of modularity. The argumentation theory of the agent should be able to capture locally and in a direct way the decision policy and accompanied knowledge of the agent. This modularity is needed for the agent to be able to carry out his argumentative deliberation efficiently, where at each particular instance of deliberation the computational argumentative process for this can be localized to the relevant (for this instance) part of the agent's argumentation theory. In a complex problem domain where an agent needs to address different types of problems and take into account different factors this ability to ”home in” on the relevant part of the theory is very important. Furthermore, the dynamic environment of an agent where new information is acquired and changes to his existing theory (or policy) can be made, requires that the representation framework is able to encode the agent's theory in a highly modular way so that these changes can be easily localized and accommodated effectively.

The argumentation framework developed and used in this paper is based on the more general and abstract notions that have emerged from a series of previous studies on argumentation [12, 8, 11, 7, 10]. The basic notion that is used is that of admissibility [7] which is itself a special case of acceptability [10]. It also follows the more recent approach of [23, 5] who have shown the need for dynamic priorities within argumentation when we want to apply this to formalize law and other related problems. Our framework is close to that of [23] in that it uses a similar background language of logic programming. They also both have a computational model that follows a dialectical pattern in terms of interleaving processes one for each level of arguments in the theory. In comparison our framework is simpler using only a single notion of attack and avoids the separate use of negation as failure that is subsumed by the use of rule priorities. In [5] dynamic priorities are related to the argumentation protocols, also called rules of order, describing which speech acts are legal in a particular state of the argumentation. Although the interests for application of our framework are different its formal relation to these frameworks is an interesting problem for further study.

In the development of agent deliberation we have introduced, in the same spirit as [27, 2], roles and context as a means to define non-static priorities between arguments of an agent. This helps to capture the social dimension of agents, as it incorporates in a natural way the influence of the environment of interaction (which includes other agents) on the agents ”way of thinking and acting”. We have shown how we can encompass, within this framework, the relative roles of agents and how these can vary dynamically depending on the external environment. The representation of this role and context information is expressed directly in terms of priority rules which themselves form arguments and are reasoned about in the same way as the object level arguments. This gives a high-level encapsulation of these notions where changes are easily accommodated in a modular way.

The use of roles and dynamic context is a basic difference with most of other works [28, 27, 21, 16, 3, 1] on agent argumentation. Our work complements and extends the approaches of [27, 2] with emphasis on enriching the self argumentative deliberation of an agent. It complements these works by linking directly the preferences between different contexts, which these works propose, to a first level of roles that agents can have in a social context, called default context, showing how roles can be used to define in a natural way priorities between arguments of the agents filling these roles. It extends this previous work by incorporating reasoning on these preferences within the process of argumentative deliberation of an agent. This is done by introducing another dimension of context, called specific context, corresponding to a second level of deliberation for the agent. This allows a higher degree of flexibility in the adaptation of the agents argumentative reasoning to a dynamically changing environment. In [2] the context preferences can also be dynamic but the account of this change is envisaged to occur outside the argumentative deliberation of the agent. An agent decides a-priori to change the context in which he is going to deliberate. In our case the change is integrated within the deliberation process of the agent.

This extra level of deliberation allows us to capture the fact that recognized roles in a context have their impact and substance only within this default context where they are defined, although these roles always ”follow” agents filling them, as a second identity in any other context they find themselves. Therefore agents who have some relationships imposed by their respective roles can be found in a specific context where the predefined (according to their relationships) order of importance between them has changed.

In comparison with other works on agent argumentation our work also integrates abduction with argumentation to handle situations where the information about the environment, currently available to the agent, is incomplete. This use of abduction is only of a simple form and more work is needed to study more advanced uses of abduction drawing from recent work on abduction in agents [26]. Another direction of future work concerns dialogue modeling. Our aim is to use our argumentative deliberation model for determining dialogue acts and protocols thus extending the framework of [15].

We have also studied, following the work of Maslow’s hierarchy of needs [17], the use of our argumentative deliberation framework to model an agent’s needs corresponding to motivational factors. This allows the expression of different personality profiles of an agent in a modular and flexible way. In the agent literature [18, 19] have already used Maslow’s theory for guiding the behaviour of deliberative and reactive agents in various unpredictable environments. However, to our knowledge, this is first time that an argumentative deliberation framework is used to model these motivation factors, in a way that, we believe, allows a more natural expression of several behaviours. Also in comparison with the various behavior-based approaches for agent personalities (e.g. [25, 24]), our work gives an alternative model for specifying different personalities in a modular way independently from the other architectural elements of the agent. In addition, our approach uses a uniform representation framework for encoding an agent’s personality and other policies or protocols associated with some of his different functionalities, e.g. with his problem solving capability.

More work is needed in this direction. On the technical side we need to extend the framework to allow an agent to decide amongst goals which address more than one need simultaneously. Also a deeper study is needed to explore the flexibility of the framework in modeling different agent personalities with respect to the way that they address their needs. Here we can draw further from work in cognitive science (see e.g. [9]) on the characteristics of human personalities. It is also important to study how these different personalities play a role in the interaction among agents especially in relation to the problem of forming heterogeneous communities of different types of agents, where the deliberation process of an agent may need to take into account the personality profile of the other agents.

In our work so far we have considered as separate the different processes of (i) generating an agent’s needs and associated goals and (ii) deciding which one of these is prevalent under the current circumstances. The potential goals that an agent generates at any situation can be influenced by the personality of the agent and his previous decisions of which goal and need to address. According to Maslow when a more important need is satisfied then new goals for other less important needs are generated. We are currently studying how to integrate together these processes into a unified model for the overall deliberation of an argumentative agent, where these two processes are interleaved into each other, taking also into account the deliberative decision making of the agent on how to satisfy his chosen goals.

## REFERENCES

[1] L. Amgoud, N. Maudet, and S. Parsons, ‘Modelling dialogues using argumentation’, in *ICMAS-00, pp. 31-38,* (2000).

[2] L. Amgoud and S. Parsons, ‘Agent dialogues with conflicting preferences’, in *ATAL01*, (2001).

[3] L. Amgoud, S. Parsons, and N. Maudet, ‘Arguments, dialogue and negotiation’, in *ECAI-00, pp. 338-342*, (2000).

[4] A. Bondarenko, P. M. Dung, R. A. Kowalski, and F. Toni, ‘An abstract, argumentation-theoretic framework for default reasoning’, *Artificial Inelligence*, **93(1-2)**, 63–101, (1997).

[5] G. Brewka, ‘Dynamic argument systems: a formal model of argumentation process based on situation calculus’, in *Journal of Logic and Computation, 11(2), pp. 257-282,* (2001).

[6] Y. Dimopoulos and A. C. Kakas, ‘Logic programming without negation as failure’, in *Proc. ILPS’95, pp. 369-384,* (1995).

[7] P.M. Dung, ‘On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games’, in *Artificial Intelligence, 77, pp. 321-357 (also in IJCAI’93),* (1995).

[8] P.M. Dung, A. C. Kakas, and P. Mancarella, ‘Negation as failure revisited’, in *University of Pisa Technical Report,* (1992).

[9] Great Ideas in Personality, ‘Five-factor model’, in *www.personalityresearch.org/bigfive.html,* (2002).

[10] A. C. Kakas, P. Mancarella, and P.M. Dung, ‘The acceptability semantics for logic programs’, in *Proc. ICLP’94, pp. 504-519,* (1994).

[11] A.C. Kakas, ‘Default reasoning via negation as failure’, in *LNAI, Vol. 810, pp. 160-179,* (1994).

[12] A.C. Kakas, R.A. Kowalski, and F. Toni, ‘Abductive logic programming.’, in *Journal of Logic and Computation, 2(6), pp. 719-770,* (1992).

[13] A.C. Kakas, R.S. Miller, and F. Toni, ‘E-res: Reasoning about actions, events and observations’, in *LPNMR’01, pp. 254-266,* (2001).

[14] A.C. Kakas and F. Toni, ‘Computing argumentation in logic programming’, in *JLC 9(4), 515–562, O.U.P,* (1999).

[15] N. Karacapilidis and P. Moraitis, ‘Engineering issues in inter-agent dialogues’, in *Proc. of 15th European Conference on Artificial Intelligence (ECAI02), Lyon, France,,* (2002).

[16] S. Kraus, K. Sycara, and A. Evenchik, ‘Reaching agreements through argumentation: a logical model and implementation’, in *Artificial Intelliegence, 104 pp. 1-69,* (1998).

[17] A. Maslow, ‘Motivation and personality’, in *Harper and Row, New York,* (1954).

[18] P. Morignot and B. Hayes-Roth, ‘Adaptable motivational profiles for autonomous agents’, in *Knowledge Systems Laboratory, Report No. KSL 95-01, Dept of Computer Science, Stanford University,* (1995).

[19] P. Morignot and B. Hayes-Roth, ‘Motivated agents’, in *Knowledge Systems Laboratory, Report No. KSL 96-22, Dept of Computer Science, Stanford University,* (1996).

[20] P. Panzarasa, N.R. Jennings, and T. Norman, ‘Formalising collaborative decision-making and practical reasoning in multi-agent systems’, in *Journal of Logic and Computation 12 (1), to appear,* (2002).

[21] S. Parsons, C. Sierra, and N.R. Jennings, ‘Agents that reason and negotiate by arguying’, in *Logic and Computation 8 (3), 261-292,* (1998).

[22] J.L. Pollock, ‘Justification and defeat’, in *Artficial Intelligence Vol 67, pp. 377-407,* (1994).

[23] H. Prakken and G. Sartor, ‘A dialectical model of assessing conflicting arguments in legal reasoning’, in *Artficial Intelligence and Law Vol 4, pp. 331-368,* (1996).

[24] P. Rizzo, M. Veloso, M. Miceli, and A. Cesta, ‘Goal-based personalities and social behaviors in believable agents’, *Applied Artificial Inelligence*, **13**, 239–272, (1999).

[25] D. Rousseau and B. Hayes-Roth, ‘Improvisational synthetic actors with flexible personalities’, in *Technical Report, KSL 97-10, Stanford University,* (1997).

[26] F. Sadri, F. Toni, and P. Torroni, ‘Dialogues for negotiation: agent varieties and dialogue sequences’, in *ATAL01,* (2001).

[27] C. Sierra, N.R. Jennings, P. Noriega, and S. Parsons, ‘A framework for argumentation-based negotiation’, in *ATAL-97, pp. 167-182,* (1997).

[28] K. Sycara, ‘Argumentation: Planning other agents’ plans’, in *IJCAI-89, pp. 517-523,* (1989).

[29] M. Wooldridge, *Introduction to Multi-Agent Systems*, John Wiley and Sons, 2002.

[30] M. Wooldridge, N.R. Jennings, and D. Kinny, ‘The gaia methodology for agent-oriented analysis and design’, in *JAAMAS 3 (3), pp. 285-312,* (2000).