

# Contextual Deliberation of Cognitive Agents in Defeasible Logic

M. Dastani<sup>1</sup>, G. Governatori<sup>2\*</sup>, A. Rotolo<sup>3</sup>, I. Song<sup>2</sup>, L. van der Torre<sup>4</sup>

<sup>1</sup> University of Utrecht, Utrecht, The Netherlands, mehdi@cs.uu.nl

<sup>2</sup> School of ITEE, The University of Queensland, Brisbane, Australia, {guido,insu}@itee.uq.edu.au

<sup>3</sup> CIRSFID, University of Bologna, Bologna, Italy, rotolo@cirsfid.unibo.it

<sup>4</sup> University of Luxembourg, Luxembourg, leon.vandertorre@uni.lu

## Categories and Subject Descriptors

D.2.8 [Artificial Intelligence]: Knowledge Representation Formalisms and Methods—Representation languages, Representations (procedural and rule-based)

## General Terms

Cognitive Agents, Defeasible Logic, Rule-Based Systems

## 1. INTRODUCTION

Logic is used in agent oriented software engineering not only for specification and verification, but also for programming deliberation and meta-deliberation tasks. For this reason, Defeasible Logic (DL) has been extended with, amongst others, rule types, preferences [4], and actions [3, 4]. In rule based cognitive agents, for example in defeasible logic, detailed interactions among cognitive attitudes like beliefs, desires, intentions and obligations are represented by rules, like the obligation to travel to Paris next week leading to a desire to travel by train ( $r_1$ ), or by preferences, such that if the desire to travel by train cannot be met, than there is a desire to travel by plane ( $p_1$ ). Patterns of such interactions are represented by rule priorities (obligations override desires or intentions – for social agents) [1, 6] rule conversions (obligations behave as desires – for norm internalizing agents) [4], and so on.

As interaction among mental attitudes becomes more complicated, the new challenge in agent deliberation languages is the coordination of such interactions. For example, at one moment an obligation to travel may lead to the desire to travel by train ( $r_1$ ), whereas at another moment it may lead to a desire to travel by plane ( $r_2$ ). Such coordination may be expressed by making the context explicit in rules  $r_1$  and  $r_2$ , and defining when rule  $r_1$  has higher priority than rule  $r_2$ , or it can be defined as a combination of rule  $r_1$  and preference  $p_1$  [3].

In this paper we raise the question how, as a further sophistication to coordinate the interaction among mental attitudes, to define the proof theory of nested rules (see [8] for a general theory of

\*Supported by the Australian Research Council under the Discovery Project DP0558854.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

AAMAS'07 May 14–18 2007, Honolulu, Hawai'i, USA.

Copyright 2007 IFAAMAS.

nested rules) and preferences among rules. Surprisingly, this complex language gives us just the right expressive power to describe a wide class of interaction phenomena: the rule leading to desire to travel by train may be preferred to the rule leading to a desire to travel by plane ( $r_1$  is preferred to  $r_2$ ), maybe as a second alternative, or the train rule may even be replaced by the plane rule ( $r_1$  into  $r_2$ ), maybe due to experienced train delays. The new language can be used to describe a new class of patterns of the coordination of interaction, e.g., when social agents turn into selfish agents, maybe when the agent does not have sufficient resources. Due to space limitations, we focus only on the formal aspects of the new logic.

Finally, the definitions of the deliberation logics developed here are much more complex than the definitions of temporal logics traditionally used in agent based software engineering for specification and verification, since they contain rules, preferences, non-monotonic proof system, and so on. However, whereas these temporal logics have a relatively high computational complexity, deliberation logics have to be efficient – with at most linear complexity (in the number of rules). Moreover, interaction patterns in such temporal logics have focussed on a relatively small class of agent types such as, for example, realisms and commitment strategies in BDI-CTL [2, 7], whereas a much broader class has been studied in the more expressive deliberation logics.

## 2. CONTEXTUAL DELIBERATION

The basic deliberative process uses rules to derive goals (desires, intentions, obligations) based on existing beliefs, desires, intentions and obligations (beliefs concern the knowledge an agent has about the world: they are not in themselves motivations for action). Contextualising the deliberation requires to provide the agent with a mechanism for reasoning with rules, which are conditioned to some additional factors. In the simplest case, this can be done by adding such factors as new antecedents of the rules to be contextualised. But transformations may be problematic when complex reasoning patterns are considered. The framework of this paper is based on the following assumptions:

**Modalities:** the system develops a constructive account of the modalities corresponding to mental states and obligations: rules are meant to devise the logical conditions for introducing them. Modalities may have a different logical behaviour. (Consider the special role played by belief rules, which permit to derive only unmodalised literals, whereas the other rule types allow for deriving modalised conclusions.) [4, 6, 3].

**Conversions:** possible conversions of a modality into another can be accepted, as when the applicability of rule leading to derive, for example,  $OBL p$  ( $p$  is obligatory) may permit, under appropriate conditions, to obtain  $INT p$  ( $p$  is intended) [4, 6].

**Preferences:** preferences can be expressed in two ways: using

standard DL superiority relation over rules and the operator  $\otimes$ . Operator  $\otimes$  [5] applies to literals [3] as well as to rules, and captures the idea of violation. A  $\otimes$ -sequence such as  $\alpha \otimes \beta \otimes \gamma$  means that  $\alpha$  is preferred, but if  $\alpha$  is violated, then  $\beta$  is preferred; if  $\beta$  is violated, then the third choice is  $\gamma$ .

**Meta-rules:** meta-rules permit to reason about rules for deriving goals. This is the main device for contextualising the provability of goals and requires to introduce nested rules.

We extend the language of Defeasible Logic with the modal operators INT, DES and OBL, and the non-classical connective  $\otimes$ . Accordingly, if  $l$  is a literal and  $X$  a modal operator, then  $Xl$  and  $\neg Xl$  are modal literals (we will use L and MLit to denote the sets of literals and modal literals). If  $l_1, \dots, l_n$ ,  $n \geq 1$ , are literals, then we will say that  $l_1 \otimes \dots \otimes l_n$  is an  $\otimes$ -expression.

We divide the rules into meta-rules, and atomic rules. Atomic rules are in addition divided into rules for beliefs, desires, intentions, and obligations. For  $X \in \{C, BEL, INT, DES, OBL\}$ , where  $\{BEL, INT, DES, OBL\}$  is the set of modalities and  $C$  stands for contextual or meta-rules, we have that  $\phi_1, \dots, \phi_n \rightarrow_X \psi$  is a *strict rule* such that whenever the premises  $\phi_1, \dots, \phi_n$  are indisputable so is the conclusion  $\psi$ .  $\phi_1, \dots, \phi_n \Rightarrow_X \psi$  is a *defeasible rule* that can be defeated by contrary evidence.  $\phi_1, \dots, \phi_n \rightsquigarrow_X \psi$  is a *defeater* that is used to defeat some defeasible rules by producing evidence to the contrary. The premises of rules, i.e.,  $\phi_1, \dots, \phi_n$ , are literals or modal literals, while  $\psi$ , the consequent, is (i) a literal for strict atomic rules; (ii) an atomic rule for a strict meta-rule; (iii) an  $\otimes$ -expression for defeasible atomic rules and atomic defeaters and (iv) and a  $\otimes$ -rule for defeasible meta-rules and meta-defeaters, where, if  $r_1, \dots, r_n$ ,  $r \geq 1$ , are atomic rules, then,  $r_1 \otimes \dots \otimes r_n$  is a  $\otimes$ -rule.

A defeasible agent theory consists of a set of *facts* or indisputable statements, a set of rules for beliefs, a set of meta-rules, a *superiority relation*  $>$  among rules saying when one rule may override the conclusion of another rule, and a conversion function  $c$  saying when a rule of one type can be used also as another type. Belief rules are the reasoning core of the agent. Rules for goals (desires, intentions, and obligations) are viewed in any theory as meta-rules with an empty antecedent and a consequent consisting of a  $\otimes$ -sequence of rules for goals.

This extension of DL makes it possible to express ordered preferences over different options for contextualising non-nested rules. In fact, we may have meta-rules such as the following:

$$r : a \Rightarrow_C (r' : b \Rightarrow_{OBL} c) \otimes \neg(r'' : d \Rightarrow_{INT} f \otimes g)$$

Intuitively, meta-rule  $r$  states that, under the condition  $a$ , we should infer rule  $r'$  stating that  $c$  is obligatory if  $b$  is the case; however, if this rule is violated (i.e., if, given  $b$  we obtain  $\neg c$ ) then the second choice is to derive the negation of rule  $r''$ , which would imply to intend  $f$ , as a first choice, or  $g$  as a second choice, if  $d$  is the case.

We use some abbreviations, such as superscript for mental attitude or meta-rule, subscript for type of rule, and  $\text{Rule}[\phi]$  for rules whose consequent is  $\phi$ , thus, for example  $\text{Rule}_{sd}^{BEL}$  is the set of strict and defeasible rules of type BEL,  $\text{Rule}_s[\psi]$  is the set of strict rules whose consequent is  $\psi$ , and  $\text{Rule}_d^C[r]$  denotes the set of defeasible meta-rules whose conclusion is the atomic rule  $r$ . We use  $r_1, \dots, r_n$  to label (or name) rules,  $A(r)$  to denote the set  $\{\phi_1, \dots, \phi_n\}$  of *antecedents* of the rule  $r$ , and  $C(r)$  to denote the *consequent* of the rule  $r$ . For some  $i$ ,  $1 \leq i \leq n$ , such that  $c_i = q$ ,  $R[c_i = q]$  and  $r_d^X[c_i = q]$  denote, respectively, the set of rules and a defeasible rule of type  $X$  with the head  $\otimes_{i=1}^n c_i$  such that  $c_i = q$ .

**DEFINITION 1.** A contextual agent theory  $D$  is a structure  $(F, R^{BEL}, R^C, >, c)$  where  $F \subseteq L \cup \text{MLit}$  is a finite set of facts,  $R^{BEL} \subseteq \text{Rule}^{BEL}$ ,  $R^C \subseteq \text{Rule}^C$ ,  $> \subseteq \text{Rule} \times \text{Rule}$ , the superiority relation is a binary relation over the set of rules,  $c$ , the conversion, is a binary relation over the set of modalities.

For readability reasons, we omit defeasible arrows for defeasible nested-rules  $r \rightsquigarrow^c$  with the empty body. That is, a defeasible nested rule  $\Rightarrow_C (p \rightarrow_{INT} q)$  will be just represented as  $p \rightarrow_{INT} q$ .

Let  $X \in \{C, BEL, DES, INT, OBL\}$ . Proofs are sequences of literals and modal literals together with so-called proof tags  $+\Delta$ ,  $-\Delta$ ,  $+\partial$  and  $-\partial$ . Given a defeasible agent theory  $D$ ,  $+\Delta_X q$  means that literal  $q$  is provable in  $D$  using only facts and strict rules for  $X$ ,  $-\Delta_X q$  means that it has been proved in  $D$  that  $q$  is not definitely provable in  $D$ ,  $+\partial_X q$  means that  $q$  is defeasibly provable in  $D$ , and  $-\partial_X q$  means that it has been proved in  $D$  that  $q$  is not defeasibly provable in  $D$ .

Before introducing proof procedures to derive specific tagged literals in a contextual agent theory, we need to define some auxiliary notions.

**DEFINITION 2.** Let  $r \in \text{Rule}$  be an atomic rule and  $\triangleright \in \{\rightarrow, \Rightarrow, \rightsquigarrow\}$ . The set  $\text{Sub}(r)$  of sub-rules is defined as follows:

- $\text{Sub}(r) = \{A(r) \triangleright_X \otimes_{i=1}^j a_i | C(r) = \otimes_{i=1}^n a_i, j \leq n\}$ , if  $r$  is atomic
  - $\text{Sub}(r) = \{\neg(A(r) \triangleright_X \otimes_{i=1}^j a_i) | C(r) = \otimes_{i=1}^n a_i, j \leq n\}$ , otherwise
- E.g., given  $r : (a \rightarrow_{INT} b \otimes c)$ ,  $\text{Sub}(r) = \{a \rightarrow_{INT} b, a \rightarrow_{INT} b \otimes c\}$ .

**DEFINITION 3.** Given an atomic rule  $r$ , the modal free rule  $L(r)$  of  $r$  is obtained by removing all modal operators in  $A(r)$ .

For example, given  $r : \text{INT}a \rightarrow_{INT} b$ ,  $L(r)$  is  $r : a \rightarrow_{INT} b$ .

**DEFINITION 4.** Let  $D$  be a contextual agent theory. The set  $R^C \langle r^{\triangleright_X} \rangle$  of supporting rules in  $R^C$  for a non-nested rule  $r^{\triangleright_X} \in \text{Rule}$  is:

- if  $r^{\triangleright_X} \in \text{Rule}_{atom}$  and  $\forall a \in A(r) : a = Xb \in \text{MLit}$ ,
- $$R^C \langle r^{\triangleright_X} \rangle = \bigcup_{s^{\triangleright_X} \in \text{Sub}(r^{\triangleright_X})} \left( R^C [c_i = s^{\triangleright_X}] \cup \bigcup_{Y:c(Y,X)} R^C [c_i = L(s^{\triangleright_Y})] \right)$$
- otherwise  $R^C \langle r^{\triangleright_X} \rangle = \bigcup_{\forall s^{\triangleright_X} \in \text{Sub}(r^{\triangleright_X})} R^C [c_i = s^{\triangleright_X}]$

For example, a meta-rule  $\Rightarrow_C (a \Rightarrow_{INT} b \otimes c) \otimes (a \Rightarrow_{INT} d)$  supports the following rules:  $(a \Rightarrow_{INT} b)$ ,  $(a \Rightarrow_{INT} b \otimes c)$ , and  $(a \Rightarrow_{INT} d)$ .

**DEFINITION 5.** Let  $D$  be a contextual agent theory. The maximal provable-rule-sets of non-nested rules that are possibly provable in  $D$  is, for  $X \in \{DES, INT, OBL\}$ ,

- $RP^X = \{ \text{Sub}(c_i) | C(r) = \otimes_{i=1}^n c_i, r \in R^C \} \cup \{ \text{Sub}(L(c_i^{\triangleright_Y})) | \forall Y \text{ such that } c(Y, X), C(r) = \otimes_{i=1}^n c_i^{\triangleright_Y}, r \in R^C, \text{ and } \forall a \in A(r) : a = Xb \in \text{MLit} \}$
- $RP^{BEL} = \{ \text{Sub}(r) | r \in R^{BEL} \}$ .

**DEFINITION 6.** Two non-nested rules  $r$  and  $r'$  are incompatible iff  $r'$  is an incompatible atomic rule of  $r$  or  $r'$  is an incompatible negative rule of  $r$ .

- 1)  $r'$  is an incompatible atomic rule of  $r$  iff  $r$  and  $r'$  are atomic rules and  $A(r) = A(r')$ ,  $C(r) = \otimes_{i=1}^n a_i$  and  $C(r') = \otimes_{i=1}^m b_i$ , such that  $\exists j, 1 \leq j \leq n, m, a_j = \sim b_j$  and,  $\forall j' \leq j, a_{j'} = b_{j'}$ .
- 2)  $r'$  is an incompatible negative rule of  $r$  iff either  $r$  or  $r'$  is not an atomic rule and  $A(r) = A(r')$ ,  $C(r) = \otimes_{i=1}^n a_i$  and  $C(r') = \otimes_{i=1}^m b_i$ , such that  $N = \min\{n, m\}, \forall j \leq N, a_j = b_j$ .

**DEFINITION 7.** Let  $D$  be a contextual agent theory and  $r$  a non-nested rule. The set of all possible incompatible rules for  $r^{\triangleright_X}$  is:

$$IC(r^{\triangleright_X}) = \{r' | r' \in RP^X, r' \text{ is incompatible with } r^{\triangleright_X}\}$$

**DEFINITION 8.** Let  $\# \in \{\Delta, \partial\}$ ,  $P = (P(1), \dots, P(n))$  be a proof in a contextual agent theory  $D$ , and  $X \in \{DES, INT, OBL\}$ . A literal  $q \in L$  or a rule  $r \in \text{Rule}$  are  $\#$ -provable in  $P$  if there is an initial sequence  $P(1), \dots, P(m)$  of  $P$  such that either

1.  $q$  is a literal and  $P(m) = +\#_{\text{BEL}}q$  or
2.  $q$  is a modal literal  $Xp$  and  $P(m) = +\#_Xp$  or
3.  $q$  is a modal literal  $\neg Xp$  and  $P(m) = -\#_Xp$  or
4.  $r^{\triangleright x}$  is a rule in  $RP^X$  and  $P(m) = +\#_C r^{\triangleright x}$ ;

A literal  $q \in L$  or a rule  $r \in \text{Rule}$  are  $\#$ -rejected in  $P$  if there is an initial sequence  $P(1), \dots, P(m)$  of  $P$  such that either

1.  $q$  is a literal and  $P(m) = -\#_{\text{BEL}}q$  or
2.  $q$  is a modal literal  $Xp$  and  $P(m) = -\#_Xp$  or
3.  $q$  is a modal literal  $\neg Xp$  and  $P(m) = +\#_Xp$  or
4.  $r^{\triangleright x}$  is a rule in  $RP^X$  and  $P(m) = -\#_C r^{\triangleright x}$ .

DEFINITION 9. Let  $D$  be a contextual agent theory. Applicable rules and discarded rules are defined as follows:

1. A rule  $r \in R^{\text{BEL}} \cup R^C$  is applicable iff  $\forall a \in A(r)$ :  
if  $a \in L$  then  $+\partial_{\text{BEL}}a \in P(1..n)$ , and  
if  $a = Xb \in \text{MLit}$  then  $+\partial_Xa \in P(1..n)$ .
2. A rule  $r \in R[c_i = q]$  is applicable in the condition for  $\pm\partial_X$  iff  
 $r \in R_{\text{atom}}^X$  and  $\forall a \in A(r)$ : if  $a \in L$  then  $+\partial_{\text{BEL}}a \in P(1..n)$ , and  
if  $a = Zb \in \text{MLit}$  then  $+\partial_Za \in P(1..n)$ , or  
 $r \in R_{\text{atom}}^Y$  and  $c(Y, X) \in c$  and  $\forall a \in A(r)$ :  $+\partial_Xa \in P(1..n)$ .
3. A rule  $r$  is discarded in the condition for  $\pm\partial_X$  iff either:  
if  $r \in R^{\text{BEL}} \cup R^C \cup R^X$  then either  $\exists a \in A(r)$ :  $-\partial_{\text{BEL}}a \in P(1..n)$   
or  $\exists Xb \in A(R)$ ,  $Xb \in \text{MLit}$  and  $-\partial_Xb \in P(1..n)$ ;  
if  $r \in R^Y$ , then  $\exists a \in A(r)$ :  $-\partial_Xa \in P(1..n)$ .

Before providing proof procedures to derive rules, let us introduce specific proof tags for this purpose. Remember that  $\triangleright$  denotes either  $\rightarrow$ ,  $\Rightarrow$  or  $\sim$  to simplify our presentation.  $\pm\Delta_C r^{\triangleright x}$  means that rule  $r \in R^X$  is (is not) definitely provable using meta-rules;  $\pm\partial_C r^{\triangleright x}$  means that rule  $r \in R^X$  is (is not) defeasibly provable using meta-rules. In general,  $\pm\Delta_C^{\triangleright x}$  and  $\pm\partial_C^{\triangleright x}$  mean, respectively, definitive (non-)provability of rules for  $X$ , and defeasible (non-)provability of rules for  $X$ .

Let us see proof procedures to derive rules. In this perspective, however, we have to be careful, as we can distinguish between strict and defeasible derivations of non-nested strict and defeasible rules. Given a contextual agent theory  $D$ , a non-nested rule  $r$  is strictly provable in  $D$  when it is strictly derived using a meta-rule such as  $a \rightarrow_C r$ . A rule  $r$  is defeasibly provable in  $D$  when it is defeasibly derived using a meta-rule such as  $a \rightarrow_C r$  and  $a \Rightarrow_C r$ . When a strict atomic rule  $a \rightarrow_{\text{INT}} b$  is defeasibly derived, it acts as a defeasible rule  $a \Rightarrow_{\text{INT}} b$ . Proof procedures for the strict derivation of atomic rules in a contextual defeasible agent theory  $D = (F, R^C, >, c)$  are as follows<sup>1</sup>.

- $+\Delta_C^{\triangleright x}$ : If  $P(i+1) = +\Delta_C r^{\triangleright x}$  then
- 1)  $X = \text{BEL}$  and  $r^{\triangleright x} \in R^{\text{BEL}}$  or
  - 2)  $\exists s \in R_s^C \langle r^{\triangleright x} \rangle \forall a \in A(s)$   $a$  is  $\Delta$ -provable.

For defeasible derivations of rules the conditions are

- $+\partial_C^{\triangleright x}$ : If  $P(n+1) = +\partial_C r^{\triangleright x}$ , then
- 1)  $+\Delta_C r^{\triangleright x} \in P(1..n)$ , or
  - 2.1)  $\forall r' \in IC(r^{\triangleright x})$ ,  $\forall r' \in R_s^C \langle r' \rangle$ ,  $r'$  is discarded and  
.2)  $\exists t \in R^C \langle c_i = r^{\triangleright x} \rangle$  such that  
.1)  $\forall i' < i$ ,  $c_{i'}$  is applicable,  
.2)  $\forall i' < i$ ,  $C(c_{i'}) = \otimes_{k=1}^n b_k$ , such that  $\forall k : +\partial_{\text{BEL}} \sim b_k \in P(1..n)$ ,  
.3)  $t$  is applicable, and  
.3)  $\forall r'' \in IC(r^{\triangleright x})$ ,  $\forall s \in R^C \langle d_i = r'' \rangle$   
.1) if  $\forall i' < i$ ,  $d_{i'}$  is applicable,

<sup>1</sup>Due to space limitations we omit the proof conditions for  $-\Delta$  and  $-\partial$  since these are the constructive negation of the corresponding positive conditions; i.e., the negative condition is obtained from the positive one swapping  $\forall$  and  $\exists$ , conjunctions and disjunctions and changing the signs of the proof tags

- $C(d_{i'}) = \otimes_{k=1}^n a_k$  such that  $\forall k : +\partial_{\text{BEL}} \sim a_k \in P(1..n)$ , then
- .1)  $s$  is discarded, or
  - .2)  $\exists z \in R^C \langle p_i = r''' \rangle$  such that  $r''' \in IC(C(s))$  s.t.  
 $\forall i' < i$ ,  $p_{i'}$  is applicable, and  
 $C(p_{i'}) = \otimes_{k=1}^n d_k$  such that  $\forall k : +\partial_{\text{BEL}} \sim d_k \in P(1..n)$  and  
 $z$  is applicable and  $z > s$ .

Given the above proof conditions for deriving rules, the following are the procedures for proving literals. Notice that each time a rule  $r$  is used and applied, we are required to check that  $r$  is provable.

- $+\Delta_X$ : If  $P(i+1) = +\Delta_X q$  then
- 1)  $Xq \in F$ , or  $q \in F$  if  $X = \text{BEL}$ , or
  - 2)  $\exists r \in \text{Rule}_s^X [q] : +\Delta_C r$  and  $\forall a \in A(r)$   $a$  is  $\Delta$ -provable or
  - 3)  $\exists r \in \text{Rule}_s^Y [q] : +\Delta_C r$ ,  $\forall a \in A(r)$   $a$  is  $\Delta$ -provable and  $c(Y, X)$ .

- $+\partial_X$ : If  $P(n+1) = +\partial_X q$  then
- 1)  $+\Delta_X q \in P(1..n)$  or
  - 2.1)  $-\Delta_X \sim q \in P(1..n)$  and  
.2)  $\exists r \in \text{Rule}_{sd} [c_i = q]$  such that  $+\partial_C r$ ,  $r$  is applicable, and  
 $\forall i' < i$ ,  $+\partial_{\text{BEL}} \sim c_{i'} \in P(1..n)$ ; and  
.3)  $\forall s \in \text{Rule} [c_j = \sim q]$ , either  $-\partial_C s$ , or  $s$  is discarded, or  
 $\exists j' < j$  such that  $-\partial_{\text{BEL}} \sim c_{j'} \in P(1..n)$ , or  
.1)  $\exists t \in \text{Rule} [c_k = q]$  such that  $+\partial_C t$ ,  $t$  is applicable and  
 $\forall k' < k$ ,  $+\partial_{\text{BEL}} \sim c_{k'} \in P(1..n)$  and  $t > s$ .

### 3. SUMMARY

We extended Defeasible Logic to deal with the contextual deliberation process of cognitive agents. First, we introduce meta-rules to reason with rules. Meta-rules are rules that have, as a consequent, rules to derive goals (obligations, intentions and desires): in other words, meta-rules include nested rules. Second, we introduce explicit preferences among rules to capture complex structures where nested rules can be involved in scenarios where rules are violated. Further research are the development of a methodology to use the language, and a formal analysis of the logic.

### 4. REFERENCES

- [1] J. Broersen, M. Dastani, J. Hulstijn, and L. van der Torre. Goal generation in the BOID architecture. *Cog. Sc. Quart.*, 2(3-4):428–447, 2002.
- [2] P. Cohen and H. Levesque. Intention is choice with commitment. *Artif. Intell.*, 42(2-3):213–261, 1990.
- [3] M. Dastani, G. Governatori, A. Rotolo, and L. van der Torre. Programming cognitive agents in defeasible logic. In *Proc. LPAR 2005*, LNAI 3835, pages 621–636. Springer, 2005.
- [4] G. Governatori and A. Rotolo. Defeasible logic: Agency, intention and obligation. In *Proc. Deon'04*, LNCS 3065, pages 114–128. Springer, 2004.
- [5] G. Governatori and A. Rotolo. Logic of violations: A Gentzen system for reasoning with contrary-to-duty obligations. *Australasian Journal of Logic*, 4:193–215, 2006.
- [6] G. Governatori, A. Rotolo, and V. Padmanabhan. The cost of social agents. In *Proc. AAMAS 2006*, pages 513–520, 2006.
- [7] A. Rao and M. Georgeff. Decision procedures for bdi logics. *J. Log. Comput.*, 8(3):293–342, 1998.
- [8] I. Song and G. Governatori. Nested rules in defeasible logic. In *Proc. RuleML 2005*, volume LNCS 3791, pages 204–208. Springer, 2005.