Contextual Deliberation of Cognitive Agents in Defeasible Logic

M. Dastani[1], G. Governatori[2]*, A. Rotolo[3], I. Song[2], L. van der Torre[4]

[1] University of Utrecht, Utrecht, The Netherlands, mehdi@cs.uu.nl

[2] School of ITEE, The University of Queensland, Brisbane, Australia, {guido,insu}@itee.uq.edu.au

[3] CIRSFID, University of Bologna, Bologna, Italy, rotolo@cirsfid.unibo.it

[4] University of Luxembourg, Luxembourg, leendert@vandertorre.com

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. CONTEXTUAL DELIBERATION

Cognitive agents often deliberate about preferences among rules. Consider, for example, an agent with the obligation to travel to Paris next week leading to a desire to travel by train, with the preferences that if the desire to travel by train cannot be met, then there is a desire to travel by plane. Such an agent may reason about preferences among rules as follows:

- The rule leading to a desire to travel by train is preferred to the rule leading to a desire to travel by plane, maybe as a second alternative.

The train rule may even be replaced by the plane rule, an example of rule revision, maybe due to experienced train delays.

Since such preferences among rules and rule revisions only hold within some contexts, we refer to this kind of reasoning as *contextual deliberation*. It is a powerful level of abstraction, which can be used to describe a new class of patterns of the coordination of interaction. Such interactions patterns are represented by rule priorities (e.g., obligations override desires or intentions – for social agents) [1, 6] rule conversions (e.g., obligations behave as desires – for norm internalizing agents) [4], and so on. Consider, for example, an agent which is called social when it prefers its obligations over its desires, and selfish when it prefers its desires over its obligations (e.g., in the BOID architecture [1]). The following is an example of contextual deliberation:

- In some circumstances, social agents turn into selfish agents, maybe when the agent does not have sufficient resources.

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

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

To support contextual deliberation, in this paper we introduce a defeasible logic expressing ordered preferences over different options for contextualising non-nested rules. For example, in the logic proposed in this paper we may have meta-rules such as the following: $r : a \Rightarrow_C (r' : b \Rightarrow_{\mathrm{OBL}} c) \otimes \neg(r'' : d \Rightarrow_{\mathrm{INT}} f \otimes g)$. 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.

Our logic builds on extensions of defeasible logic for programming deliberation and meta-deliberation tasks with, amongst others, rule types, preferences [4], actions [3, 4] and nested rules [8]. It is based on the following assumptions:

Modalities: the system develops a constructive account of modalities corresponding to mental states and obligations, in the sense that rules devise the logical conditions for introducing them. Modalities may have a different logical behaviour, as illustrated by 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 occur. For example, the applicability of rule leading to derive OBLp (p is obligatory) may permit, under appropriate conditions, to obtain INTp (p is intended) [4, 6].

Preferences: preferences can be expressed in two ways: using standard priority 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 α is preferred, but if α is violated, then β is preferred; if β is violated, then the third choice is γ.

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.

The *definitions* of the deliberation logics developed in this paper are 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, deliberation logics have to be efficient – with at most linear complexity (in the number of rules), whereas traditional temporal logics have a relatively high computational complexity. Moreover, interaction patterns in such temporal logics have focussed on a relatively small class of agent types such as, for example, kinds of realism and commitment strategies in BDI-CTL [2, 7], whereas a much broader class is studied in the deliberation logics.

Due to space limitations, we focus only on the formal aspects of the new logic.

2. DEFEASIBLE LOGIC

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.

2.1 Contextual agent theory

We divide the rules into atomic rules and meta-rules, where atomic rules are again divided into rules for beliefs, desires, intentions, and obligations. For rule index $X \in \{C, \mathrm{BEL}, \mathrm{DES}, \mathrm{INT}, \mathrm{OBL}\}$, where $\{\mathrm{BEL}, \mathrm{DES}, \mathrm{INT}, \mathrm{OBL}\}$ 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 ψ. $\phi_1, \dots, \phi_n \Rightarrow_X \psi$ is a *defeasible rule* that can be defeated by contrary evidence. $\phi_1, \dots, \phi_n \leadsto_X \psi$ 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 ψ, 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) 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. Finally we will admit negation of rules, thus if r is an atomic rule, then $\neg r$ is a rule.

A contextual agent theory consists of a set of *facts* or indisputable statements, a set of rules for beliefs, a set of meta-rules, a *superiority* or priority 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.

DEFINITION 1. *A* contextual agent theory*D is a structure $(F, R^{\mathrm{BEL}}, R^C, >, c)$ where $F \subseteq \mathrm{L} \cup \mathrm{MLit}$ is a finite set of facts, $R^{\mathrm{BEL}} \subseteq \mathrm{Rule}^{\mathrm{BEL}}$, $R^C \subseteq \mathrm{Rule}^C$, $> \subseteq \mathrm{Rule} \times \mathrm{Rule}$, 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 $r^{\Rightarrow_C}$ with the empty body. That is, a defeasible nested rule $\Rightarrow_C (p \rightarrow_{\mathrm{INT}} q)$ will be just represented as $p \rightarrow_{\mathrm{INT}} q$.

2.2 Incompatible rules

In this section we define when two rules are incompatible. We use some abbreviations, such as superscript for mental attitude or meta-rule, subscript for type of rule, and $\mathrm{Rule}[\phi]$ for rules whose consequent is ϕ, thus, for example $\mathrm{Rule}^{\mathrm{BEL}}_{sd}$ is the set of strict and defeasible rules of type BEL, $\mathrm{Rule}_s[\psi]$ is the set of strict rules whose consequent is ψ, and $\mathrm{Rule}^C_d[r]$ 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 $r^X_d[c_i = q]$ 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 2. *Let $r \in$* Rule *be a non-nested rule and $\rhd \in \{\rightarrow, \Rightarrow, \leadsto\}$. The set $Sub(r)$ of* sub-rules *is defined as follows:*
- $Sub(r) = \{A(r) \rhd_X \otimes_{i=1}^j a_i | C(r) = \otimes_{i=1}^n a_i, j \leq n\}$, if r is atomic
- $Sub(r) = \{\neg(A(r) \rhd_X \otimes_{i=1}^j a_i) | C(r) = \otimes_{i=1}^n a_i, j \leq n\}$, otherwise

E.g., given $r : (a \rightarrow_{\mathrm{INT}} b \otimes c)$, $Sub(r) = \{a \rightarrow_{\mathrm{INT}} b, a \rightarrow_{\mathrm{INT}} b \otimes c\}$.

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 $r : \mathrm{INT}a \rightarrow_{\mathrm{INT}} b$, $L(r)$ is $r : a \rightarrow_{\mathrm{INT}} b$.

DEFINITION 4. *Let D be a contextual agent theory and $r^{\rhd_X} \in$* Rule *a non-nested rule. The set $R^C\langle r^{\rhd_X}\rangle$ of* supporting rules *for $r^{\rhd_X}$ in R^C is:*
- if $r^{\rhd_X} \in \mathrm{Rule}_{atom}$ and $\forall a \in A(r) : a = Xb \in \mathrm{MLit}$,

$$R^C\langle r^{\rhd_X}\rangle = \bigcup_{s^{\rhd_X} \in Sub(r^{\rhd_X})} \left(R^C[c_i = s^{\rhd_X}] \cup \bigcup_{Y : c(Y,X)} R^C[c_i = L(s^{\rhd_Y})] \right)$$

- otherwise $R^C\langle r^{\rhd_X}\rangle = \bigcup_{\forall s^{\rhd_X} \in Sub(r^{\rhd_X})} R^C[c_i = s^{\rhd_X}]$

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

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 $X \in \{\mathrm{DES}, \mathrm{INT}, \mathrm{OBL}\}$,*
- $RP^X = \{Sub(c_i) | C(r) = \bigotimes_{i=1}^n c_i, r \in R^C\} \cup$
$\{Sub(L(c_i^{\rhd_Y})) | \forall Y$ such that $c(Y,X), C(r) = \otimes_{i=1}^n c_i^{\rhd_X}$,
$r \in R^C$, and $\forall a \in A(r) : a = Xb \in \mathrm{MLit}\}$
- $RP^{BEL} = \{Sub(r) | r \in R^{\mathrm{BEL}}\}$.

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 $A(r) = A(r')$, $C(r) = \bigotimes_{i=1}^n a_i$ and $C(r') = \bigotimes_{i=1}^m b_i$, 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 $A(r) = A(r')$, $C(r) = \bigotimes_{i=1}^n a_i$ and $C(r') = \bigotimes_{i=1}^m b_i$, 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 $r^{\rhd_X}$ is:*

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

2.3 Proofs and proof rules

Let $X \in \{C, \mathrm{BEL}, \mathrm{DES}, \mathrm{INT}, \mathrm{OBL}\}$. 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.

DEFINITION 8. *Let $\# \in \{\Delta, \partial\}$, $P = (P(1), \dots, P(n))$ be a proof in a contextual agent theory D, and $X \in \{\mathrm{DES}, \mathrm{INT}, \mathrm{OBL}\}$. A literal $q \in L$ or a rule $r \in$* 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 $P(m) = +\#_{\mathrm{BEL}} q$ or
2. q is a modal literal Xp and $P(m) = +\#_X p$ or
3. q is a modal literal $\neg Xp$ and $P(m) = -\#_X p$ or
4. $r^{\rhd_X}$ is a rule in RP^X and $P(m) = +\#_C r^{\rhd_X}$;

A literal $q \in L$ or a rule $r \in$ 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 $P(m) = -\#_{\mathrm{BEL}} q$ or
2. q is a modal literal Xp and $P(m) = -\#_X p$ or
3. q is a modal literal $\neg Xp$ and $P(m) = +\#_X p$ or
4. $r^{\rhd_X}$ is a rule in RP^X and $P(m) = -\#_C r^{\rhd_X}$.

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

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

Before providing proof procedures to derive rules, let us introduce specific proof tags for this purpose. For $\rhd \in \{\rightarrow, \Rightarrow, \leadsto\}$, $\pm\Delta_C r^{\rhd_X}$ means that rule $r \in R^X$ is (is not) definitely provable using meta-rules; $\pm\partial_C r^{\rhd_X}$ means that rule $r \in R^X$ is (is not) defeasibly provable using meta-rules. In general, $\pm\Delta_C^{\rhd_X}$ and $\pm\partial_C^{\rhd_X}$ mean, respectively, definitive (non-)provability of rules for X, and defeasible (non-)provability of rules for X.

Finally we give 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 $a \rightarrow_{\mathrm{INT}} b$ is defeasibly derived, it acts as a defeasible rule $a \Rightarrow_{\mathrm{INT}} b$. Proof procedures for the strict derivation of atomic rules in a contextual cognitive agent theory $D = (F, R^{\mathrm{BEL}}, R^C, >, c)$ are as follows.[1]

$+\Delta_C^{\rhd_X}$: If $P(i+1) = +\Delta_C r^{\rhd_X}$ then
1) $X = BEL$ and $r^{\rhd_X} \in R^{BEL}$ or
2) $\exists s \in R^C_s\langle r^{\rhd_X}\rangle$ $\forall a \in A(s)$ a is Δ-provable.

For defeasible derivations of rules the conditions are as follows.

$+\partial_C^{\rhd_X}$: If $P(n+1) = +\partial_C r^{\rhd_X}$, then
1) $+\Delta_C r^{\rhd_X} \in P(1..n)$, or
2.1) $\forall r'' \in IC(r^{\rhd_X})$, $\forall r' \in R^C_s\langle r''\rangle$, r' is discarded and
.2) $\exists t \in R^C\langle c_i = r^{\rhd_X}\rangle$ such that
.1) $\forall i' < i$, $c_{i'}$ is applicable,
.2) $\forall i' < i$, $C(c_{i'}) = \bigotimes_{k=1}^n b_k$, s.t. $\forall k : +\partial_{\mathrm{BEL}} {\sim} b_k \in P(1..n)$,
.3) t is applicable, and
.3) $\forall r'' \in IC(r^{\rhd_X}), \forall s \in R^C\langle d_i = r''\rangle$
.1) if $\forall i' < i$, $d_{i'}$ is applicable,
$C(d_{i'}) = \bigotimes_{k=1}^n a_k$ s.t. $\forall k : +\partial_{\mathrm{BEL}} {\sim} a_k \in P(1..n)$, then
.1) s is discarded, or
.2) $\exists z \in R^C\langle p_i = r'''\rangle$ such that $r''' \in IC(C(s))$ such that
$\forall i' < i$, $p_{i'}$ is applicable, and
$C(p_{i'}) = \bigotimes_{k=1}^n d_k$ s.t. $\forall k : +\partial_{\mathrm{BEL}} {\sim} d_k \in P(1..n)$ 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.

[1] Due to space limitations we omit the proof conditions for $-\Delta$ and $-\partial$. They 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.

$+\Delta_X$: If $P(i+1) = +\Delta_X q$ then
1) $Xq \in F$, or $q \in F$ if $X =$ BEL, or
2) $\exists r \in \mathrm{Rule}^X_s[q] : +\Delta_C r$ and $\forall a \in A(r)$ a is Δ-provable or
3) $\exists r \in \mathrm{Rule}^Y_s[q] : +\Delta_C r$, $\forall a \in A(r)$ a is Δ-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) $\exists r \in \mathrm{Rule}_{sd}[c_i = q]$ such that $+\partial_C r$, r is applicable, and
$\forall i' < i$, $+\partial_{\mathrm{BEL}} {\sim} c_{i'} \in P(1..n)$; and
.3) $\forall s \in \mathrm{Rule}[c_j = {\sim} q]$, either $-\partial_C s$, or s is discarded, or
$\exists j' < j$ such that $-\partial_{\mathrm{BEL}} {\sim} c_{j'} \in P(1..n)$, or
.1) $\exists t \in \mathrm{Rule}[c_k = q]$ such that $+\partial_C t$, t is applicable and
$\forall k' < k$, $+\partial_{\mathrm{BEL}} {\sim} c_{k'} \in P(1..n)$ and $t > s$.

3. SUMMARY

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. We therefore extend 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. The main challenge in the formal definition is to introduce a notion of compatibility between rules, for which we have given one possible solution in this paper, based on a subtle introduction of negated rules. 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. ∆eon'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.