Contextual Agent Deliberation in Defeasible Logic

Mehdi Dastani[1], Guido Governatori[2], Antonino Rotolo[3],
Insu Song[2], Leendert van der Torre[4]

[1] Department of Information and Computing Sciences, University of Utrecht
[2] School of Information Technology and Electrical Engineering, The University of Queensland
[3] CIRSFID, University of Bologna
[4] Université du Luxembourg, Computer Science and Communication

Abstract. This article extends 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 for motivational components, such as obligations, intentions and desires. In other words, they include nested rules. Second, we introduce explicit preferences among rules. They deal with complex structures where nested rules can be involved.

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 [**?**], and actions [**?**,**?**]. 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) [**?**,**?**] rule conversions (obligations behave as desires – for norm internalizing agents) [**?**], 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 [**?**].

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 [**?**] for a general theory of nested rules, [**?**] for theory of nested rules with time) 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

A. Ghose and G. Governatori (eds)
10 Pacific Rim International Workshop on Multi-Agents (PRIMA 2008)
LNAI 5044, 2008.

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, for example 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 [**?**,**?**], whereas a much broader class has been studied in the more expressive deliberation logics.

This paper is organized as follows. We first give some general intuitions of how the logical system works and define the formal language we use to contextualise the deliberation of cognitive agents. A running example will help us illustrate the system. The second step consists in describing the logical machinery for reasoning about contextualised agents. This requires to provide proof procedures to derive goals (desires, intentions, and obligations) as well as to derive rules for proving goals. A further example concludes the paper.

2 Contextualising Cognitive Agents

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 about rules for goals, 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 here are not contextualised and permit to derive only unmodalised literals, whereas the other rule types allow for deriving modalised conclusions.) [**?**,**?**,**?**].

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

Preferences: preferences can be expressed in two ways: using standard DL superiority relation over rules and the operator $\otimes$. Operator $\otimes$ [**?**] applies to literals [**?**] 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.

We extend the language of Defeasible Logic with the modal operators BEL, INT, DES and OBL, and the non-classical connective $\otimes$. 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 $X \in \{C, \mathrm{BEL}, \mathrm{INT}, \mathrm{DES}, \mathrm{OBL}\}$, where $\{\mathrm{BEL}, \mathrm{INT}, \mathrm{DES}, \mathrm{OBL}\}$ is the set of modalities and C stands for contextual or meta-rules, we have that $\phi_1, \ldots, \phi_n \rightarrow_X \psi$ is a *strict rule* such that whenever the premises $\phi_1, \ldots, \phi_n$ are indisputable so is the conclusion ψ. $\phi_1, \ldots, \phi_n \Rightarrow_X \psi$ is a *defeasible rule* that can be defeated by contrary evidence. $\phi_1, \ldots, \phi_n \rightsquigarrow_X \psi$ is a *defeater* that is used to defeat some defeasible rules by producing evidence to the contrary.

Definition 1 (Language). *Let* PROP *be a set of propositional atoms and* Lab *be a set of labels.*

- *The set of modal operators is* $\mathrm{MOD} = \{\mathrm{BEL}, \mathrm{OBL}, \mathrm{INT}, \mathrm{DES}\}$*;*
- *The set of literals is* $\mathrm{L} = \mathrm{PROP} \cup \{\neg p | p \in \mathrm{PROP}\}$*. If q is a literal, $\sim q$ denotes the complementary literal (if q is a positive literal p then $\sim q$ is $\neg p$; and if q is $\neg p$, then $\sim q$ is p);*
- *The set of modal literals is*

$$\mathrm{MLit} = \{Xl, \neg Xl | l \in L, X \in \{\mathrm{DES}, \mathrm{INT}, \mathrm{OBL}\}\};$$

- *The set of $\otimes$-expressions is*

$$PREF = \{l_1 \otimes \cdots \otimes l_n : n \geq 1, \{l_1, \ldots, l_n\} \subseteq L\}.$$

  We also write $\bigotimes_{i=1}^{n} l_i$ for $l_1 \otimes \cdots \otimes l_n \in$ Pref*.*
- *The set of labeled atomic rules is* $\mathrm{Rule}_{atom} = \mathrm{Rule}_{atom,s} \cup \mathrm{Rule}_{atom,d} \cup \mathrm{Rule}_{atom,dft}$*, where for* $X \in \mathrm{MOD}$

$$\begin{aligned}
\mathrm{Rule}_{atom,s} &= \{r : \phi_1, \ldots, \phi_n \rightarrow_X \psi | r \in \mathrm{Lab}, A(r) \subseteq \mathrm{L} \cup \mathrm{MLit}, \psi \in \mathrm{L}\} \\
\mathrm{Rule}_{atom,d} &= \{r : \phi_1, \ldots, \phi_n \Rightarrow_X \psi | r \in \mathrm{Lab}, A(r) \subseteq \mathrm{L} \cup \mathrm{MLit}, \psi \in \mathrm{Pref}\} \\
\mathrm{Rule}_{atom,dft} &= \{r : \phi \rightsquigarrow_X \psi | r \in \mathrm{Lab}, A(r) \subseteq \mathrm{L} \cup \mathrm{MLit}, \psi \in \mathrm{Pref}\}
\end{aligned}$$

- *the set of labeled rules is*

$$\begin{aligned}
\mathrm{Rule} = \mathrm{Rule}_{atom} \cup \{\neg(r : \phi_1, \ldots, \phi_n \rhd_Y \psi) | (r : \phi_1, \ldots, \phi_n \rhd_Y \psi) \in \mathrm{Rule}_{atom}, \\
\rhd \in \{\rightarrow, \Rightarrow, \rightsquigarrow\}, Y \in \{\mathrm{DES}, \mathrm{INT}, \mathrm{OBL}\}\}
\end{aligned}$$

  By convention, if r is a rule, $\sim r$ denotes the complementary rule (if $r : \phi_1, \ldots, \phi_n \rhd_X \psi$ then $\sim r$ is $\neg(r : \phi_1, \ldots, \phi_n \rhd_X \psi)$; and if $r : \neg(r : \phi_1, \ldots, \phi_n \rhd_X \psi)$ then $\sim r$ is $r : \phi_1, \ldots, \phi_n \rhd_x \psi$).

– *the set of* $\otimes$*-rules is*

$$Q = \{a_1 \otimes \cdots \otimes a_n | n \geq 1, \{a_1, \ldots, a_n\} \subseteq \text{Rule}\}$$

– *The set of labeled meta-rules is* $\text{Rule}^C = \text{Rule}^C_s \cup \text{Rule}^C_d \cup \text{Rule}^C_{dft}$*, where, for* $X \in \{\text{DES}, \text{INT}, \text{OBL}\}$

$$\begin{array}{l} \text{Rule}^C_s = \{r : \phi \rightarrow_C \psi | r \in \text{Lab}, \phi \subseteq \text{L} \cup \text{MLit}, \psi \in \text{Rule}^X\} \\ \text{Rule}^C_d = \{r : \phi \Rightarrow_C \psi | r \in \text{Lab}, \phi \subseteq \text{L} \cup \text{MLit}, \psi \in Q\} \\ \text{Rule}^C_{dft} = \{r : \phi \rightsquigarrow_C \psi | r \in \text{Lab}, \phi \subseteq \text{L} \cup \text{MLit}, \psi \in Q\} \end{array}$$

– *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* ϕ*, for example:*

$$\begin{array}{l} \text{Rule}^{\text{BEL}} = \{r : \phi_1, \ldots, \phi_n \rhd_{\text{BEL}} \psi | (r : \phi_1, \ldots, \phi_n \rhd_{\text{BEL}} \psi) \in \text{Rule}, \rhd \in \{\rightarrow, \Rightarrow, \rightsquigarrow\}\} \\ \text{Rule}_s[\psi] = \{\phi_1, \ldots, \phi_n \rightarrow_X \psi | \{\phi_1, \ldots, \phi_n\} \subseteq \text{L} \cup \text{MLit}, \psi \in \text{L}, X \in \text{MOD}\} \end{array}$$

Other abbreviations are the following. We use $r_1, \ldots, r_n$ *to label (or name) rules,* $A(r)$ *to denote the set* $\{\phi_1, \ldots, \phi_n\}$ *of* antecedents *of the rule r, and* $C(r)$ *to denote the* consequent ψ *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* $\otimes^n_{i=1} c_i$ *such that* $c_i = q$*.*

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.

Definition 2 (Contextual Defeasible Agent Theory). *A contextual defeasible agent theory is a structure* $D = (F, R^{\text{BEL}}, R^C, >, c)$ *where*

– $F \subseteq \text{L} \cup \text{MLit}$ *is a finite set of facts,*
– $R^{\text{BEL}} \subseteq \text{Rule}^{\text{BEL}}$*,*
– $R^C \subseteq \text{Rule}^C$*,*
– $> \subseteq (\text{Rule} \times \text{Rule}) \cup (R^C \times R^C)$*, the superiority relation, is an acyclic binary relation over the set of rules;*
– $c \subseteq \text{MOD} \times \text{MOD}$ *is a set of conversions.*

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

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

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

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.

The following running example illustrates the contextual defeasible agent theory.

Example 1. (RUNNING EXAMPLE). Frodo, our Tolkienian agent, intends to be entrusted by Elrond to be the bearer of the ring of power, a ring forged by the dark lord Sauron. Frodo has the task to bring the ring to Mordor, the realm of Sauron, and to destroy it by throwing it into the fires of Mount Doom. Given this task, if Frodo does not destroy the ring, he is obliged to leave the Middle Earth, while, if his primary intention is not to destroy it, but he does accomplish the task anyway, he will intend to go back to the Shire. If Frodo is a brave hobbit, rule r_0 should hold, which states the intention to kill Sauron; however, if r_0 is violated, then rule r_3 should not hold. On the other hand, if Frodo is selfish, that he has the intention to be entrusted implies that he has also the intention to kill Elrond. However, if he has this last intention, he is not obliged destroy the ring if he is obliged to be the ring bearer. As facts of the theory, we know that Frodo intends to be entrusted by Elrond, that he is selfish and brave at the same time, and that he does not kill Sauron.

$$
\begin{aligned}
F &= \{\mathrm{INT}\mathit{Entrusted}, \mathit{Selfish}, \mathit{Brave}, \neg \mathit{KillSauron}\} \\
R &= \{r_1 : \mathrm{OBL}\mathit{Mordor} \Rightarrow_{\mathrm{OBL}} \mathit{DestroyRing} \otimes \mathit{LeaveMiddleEarth} \\
&\quad r_2 : \mathrm{INT}\mathit{RingBearer} \Rightarrow_{\mathrm{OBL}} \mathit{Mordor} \\
&\quad r_3 : \mathrm{INT}\mathit{RingBearer} \Rightarrow_{\mathrm{INT}} \neg \mathit{DestroyRing} \otimes \mathit{BackToShire} \\
&\quad r_4 : \mathit{Entrusted} \Rightarrow_{\mathrm{OBL}} \mathit{RingBearer} \\
&\quad r_5 : \mathit{Brave} \Rightarrow_C (r_0 : \emptyset \Rightarrow_{\mathrm{INT}} \mathit{KillSauron}) \otimes \\
&\qquad\qquad \neg(r_3 : \mathrm{INT}\mathit{RingBearer} \Rightarrow_{\mathrm{INT}} \neg \mathit{DestroyRing} \otimes \mathit{BackToShire}) \\
&\quad r_6 : \mathit{Selfish} \Rightarrow_C (r_7 : \mathrm{INT}\mathit{Entrusted} \rightarrow_{\mathrm{INT}} \mathit{KillElrond}) \\
&\quad r_8 : \mathit{Selfish}, \mathrm{INT}\mathit{KillElrond} \Rightarrow_C (r_9 : \mathrm{INT}\mathit{RingBearer} \Rightarrow_{\mathrm{OBL}} \neg \mathit{DestroyRing})\} \\
{>} &= \{r_3 > r_1, r_5 > r_8\} \\
c &= \{c(\mathrm{OBL}, \mathrm{INT})\}
\end{aligned}
$$

3 Reasoning about Contextual Deliberation

Let $X \in \{C, \mathrm{BEL}, \mathrm{DES}, \mathrm{INT}, \mathrm{OBL}\}$. Proofs are sequences of literals, modal literals, and rules together with so-called proof tags $+\Delta$, $-\Delta$, $+\partial$ and $-\partial$. Given a defeasible agent theory D, $+\Delta_X q$ means that a conclusion 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 presenting proof procedures to derive specific tagged literals and rules in a contextual agent theory, we need to introduce some auxiliary notions.

Definitions **??** and **??** are propaedeutic for Definition **??** (which defines the set of meta-rules supporting the derivation of a chosen rule) and Definition **??** (which defines the maximal-provable sets of rules that are provable in a theory).

Definition 3 (Sub Rule). *Let $r \in$ Rule be an atomic rule and $\rhd \in \{\rightarrow, \Rightarrow, \rightsquigarrow\}$. 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 $r : (a \rightarrow_{\text{INT}} b \otimes c)$, $Sub(r) = \{a \rightarrow_{\text{INT}} b, a \rightarrow_{\text{INT}} b \otimes c\}$.

Definition 4 (Modal Free Rule). *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 $r : \text{INT}a \rightarrow_{\text{INT}} b$, $L(r)$ is $r : a \rightarrow_{\text{INT}} b$.

Definition 5 (Rule-Supporting Rules). *Let D be a contextual agent theory. The set $R^C\langle r^{\rhd_X}\rangle$ of supporting rules in R^C for a non-nested rule $r^{\rhd_X} \in$ Rule is:*

- *if $r^{\rhd_X} \in \text{Rule}_{atom}$ and $\forall a \in A(r) : a = Xb \in$ 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 $\Rightarrow_C (a \Rightarrow_{\text{INT}} b \otimes c) \otimes (a \Rightarrow_{\text{INT}} d)$ supports the following rules: $(a \Rightarrow_{\text{INT}} b)$, $(a \Rightarrow_{\text{INT}} b \otimes c)$, and $(a \Rightarrow_{\text{INT}} d)$.

Definition 6 (Maximal Provable-Rule-Sets). *Let D be a contextual agent theory. The maximal provable-rule-sets of non-nested rules that are possibly provable in D is, for $X \in \{\text{DES}, \text{INT}, \text{OBL}\}$,*

$$\begin{aligned} RP^X = & \left\{ Sub(c_i) | C(r) = \bigotimes_{i=1}^{n} c_i, r \in R^C \right\} \cup \\ & \left\{ Sub(L(c_i^{\rhd_Y})) | \forall Y \text{ such that } sc(Y,X), C(r) = \otimes_{i=1}^{n} c_i^{\rhd_X}, r \in R^C, \text{ and} \right. \\ & \qquad \left. \forall a \in A(r) : a = Xb \in \text{MLit} \right\} \\ RP^{\text{BEL}} = & \{ Sub(r) | r \in R^{\text{BEL}} \}. \end{aligned}$$

Since we want to derive rules for goals, this requires defining when two rules are incompatible. In this regard, notice that defeasible rules and defeaters for goals may have $\otimes$-expressions as their consequents, which is something we do not have for strict rules. Accordingly, the notion of incompatibility has to take also into account when two $\otimes$-expressions occur in the the heads of two rules.

Definition 7. *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 8. *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:*

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

Definitions **??** and **??** define, respectively, when a literal or a rule is provable (and non-provable: rejected), and when a rule is applicable (and non-applicable: discarded). Both notions are used in the proof procedures presented in the remainder.

Definition 9 (Provable). *Let $\# \in \{\Delta, \partial\}$, $P = (P(1), \ldots, P(n))$ be a proof in a contextual agent theory D, and $X \in \{\text{DES}, \text{INT}, \text{OBL}\}$. A literal $q \in L$ or a rule $r \in$ Rule are* #-provable *in P if there is an initial sequence $P(1), \ldots, P(m)$ 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 $P(m) = +\#_X p$ or*
3. *q is a modal literal $\neg Xp$ and $P(m) = -\#_X p$ 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 $r \in$ Rule are #-rejected *in P if there is an initial sequence $P(1), \ldots, P(m)$ 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 $P(m) = -\#_X p$ or*
3. *q is a modal literal $\neg Xp$ and $P(m) = +\#_X p$ or*
4. *$r^{\triangleright X}$ is a rule in RP^X and $P(m) = -\#_C r^{\triangleright X}$.*

Definition 10. *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 $a = Xb \in$ 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 $+\partial_{\text{BEL}} a \in P(1..n)$, 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 $r \in R^{\text{BEL}} \cup R^C \cup R^X$ then either*
     $\exists a \in A(r) : -\partial_{\text{BEL}} a \in P(1..n)$ or
     $\exists Xb \in A(R)$, $Xb \in$ MLit and $-\partial_X b \in P(1..n)$; or

– *if* $r \in R^Y$, *then* $\exists a \in A(r) : -\partial_X a \in P(1..n)$.

Notice that the notion of applicability needs to take conversions into account.

Remark 1. Conversions affect applicability of rules. In many cases we want that this possibility can be admitted:

$$+\Delta_{\mathrm{INT}} \mathit{GoToRome},\ \mathit{GoToRome} \rightarrow_{\mathrm{BEL}} \mathit{GoToItaly} \mid\sim +\Delta_{\mathrm{INT}} \mathit{GoToItaly}$$
$$+\partial_{\mathrm{INT}} \mathit{GoToRome},\ \mathit{GoToRome} \Rightarrow_{\mathrm{OBL}} \mathit{VisitVatican} \mid\sim +\partial_{\mathrm{INT}} \mathit{VisitVatican}$$

The first rule states that the agent believes that going to Rome strictly implies going to Italy; if we can derive that the agent has the intention to visit Rome, we have reasons to say that a rational agent may have the intention to visit Italy. The second rule says that visiting Rome defeasibly implies the obligation to visit Vatican City. With norm-complying agents, agent's intention to visit Rome rationally implies to having the intention to go to Vatican City.

Example 2. Suppose we allow a deontic rule to be converted into a rule for intention, i.e. $c(\mathrm{OBL}, \mathrm{INT})$. Consider rule $r : a, b \Rightarrow_{\mathrm{OBL}} p$: if $\partial_{\mathrm{INT}} a$ and $\partial_{\mathrm{INT}} b$, then r is applicable in the proof condition for $+\partial_{\mathrm{INT}}$.

Before providing proof procedures to derive rules, let us introduce specific proof tags for this purpose. Remember that $\triangleright$ denotes either $\rightarrow$, $\Rightarrow$ or $\rightsquigarrow$ 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 $a \rightarrow_{\mathrm{INT}} b$ 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 defeasible agent theory D are as follows:

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

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

Strict derivations of rules are based on the following intuition. If the rule r we want to derive is for belief, r must be in the set of belief rules of the theory. Otherwise (for the other rule types), r must be proved using a strict meta-rule whose antecedents are strictly provable. Defeasible derivations of rules are based on the following procedures.

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

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

Remark 2. The defeasible proof of a rule runs in three phases. We have to find an argument in favour of the rule we want to prove. Second, all counter-arguments are examined (rules for the opposite conclusion). Third, all the counter-arguments have to be rebutted (the counter-argument is weaker than the pro-argument) or undercut (some of the premises of the counter-argument are not provable). Let us exemplify positive proof conditions ($+\partial_C^{\triangleright x}$) step by step. Suppose we want to derive $r : \text{INT}a \rightarrow_{\text{INT}} b$, namely, that $X = \text{INT}$ and $\triangleright = \rightarrow$. We have the following options. Condition (1): r is definitely provable; or, Condition (2): We use a strict or defeasible meta-rule to derive r. This must exclude, as a precondition, that any rule, which is incompatible with r, is definitely supported: (condition 2.1). That is, rules such as

$$r' : \neg(\text{INT}a \rightarrow_{\text{INT}} b) \qquad r'' : \text{INT}a \rightarrow_{\text{INT}} \neg b \qquad r''' : a \rightarrow_{\text{OBL}} \neg b$$

should not be supported, if we have that r (the rule we want to prove) is applicable, $+\partial_{\text{INT}} a$ and $c(\text{OBL}, \text{INT})$, namely, if we may convert a rule for obligation into one for intention. In fact, if we have this conversion, r'' behaves like a rule for intention. With this done, condition (2.2) states that there should exist a meta-rule such as

$$t : d \Rightarrow_C (w : p \Rightarrow_{\text{OBL}} q) \otimes (r : \text{INT}a \rightarrow_{\text{INT}} b)$$

such that t is applicable, $+\partial_{\text{BEL}} d$, and the first choice, rule w, is violated, namely that $+\partial_{\text{BEL}} p$ and $+\partial_{\text{BEL}} \neg q$. But this fact must exclude that any meta-rule s supporting an

incompatible conclusion against r is applicable (see condition 2.3.1.1). Alternatively, if s is applicable, we have to verify that there exists a meta-rule z supporting r such that z is applicable and is stronger that s (see condition 2.3.1.2). Notice that when we say that that a meta-rule supports a rule, we take into account that meta-rules may have $\otimes$-rules in their consequents. For example, if s is as follows

$$s : d \Rightarrow_C (w : b \rightarrow_{\mathrm{DES}} c) \otimes r' : \neg(\mathrm{INT}a \rightarrow_{\mathrm{INT}} b),$$

but we prove $-\partial_{\mathrm{BEL}}\neg c$, this means that w cannot be violated and so s cannot be used to attack the derivation of r: in this case, using s, we could only prove rule w.

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 $X = \mathrm{BEL}$, or
(2) $\exists r \in \mathrm{Rule}_s^X[q] : +\Delta_C r$ and $\forall a \in A(r)$ a is Δ-provable or
(3) $\exists r \in \mathrm{Rule}_s^Y[q] : +\Delta_C r$, $\forall a \in A(r)$ a is Δ-provable and $c(Y,X)$.

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

$+\partial_X$: 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 $t > s$.

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

Example 3. (RUNNING EXAMPLE, CONTINUED). The fact that Frodo has the intention to be entrusted by Elrond makes it possible to derive $+\Delta_{\mathrm{INT}} \mathit{Entrusted}$. Since we have

the conversion $c(\mathrm{OBL},\mathrm{INT})$, this would make both r_2 and r_3 applicable. However, r_5 and r_6, too, are applicable. Rule r_6 permits to derive r_7, which is applicable, as Frodo has the intention to be entrusted. This allows in turn for the derivation of the intention to kill Elrond, which makes r_8 applicable. Here we have a conflict between r_5 and r_8, but the former is stronger. On the other hand, r_5 states, as a first choice, that Frodo has the intention to kill Sauron. But this intention is violated, as $\neg$*KillSauron* is a fact, which makes it possible to derive $+\Delta_{\mathrm{BEL}}\neg$*KillSauron* and so $+\partial_{\mathrm{BEL}}\neg$*KillSauron*. Thus we have to derive the second choice, namely, the negation of r_3. In this way, even if r_3 is stronger than r_1, the applicability of meta-rule r_5 makes r_3 inapplicable.

4 Context-Detection: A Further Example

Context can play as a disambiguating function in agent communication. For example, when an agent A receives a command "On" from another agent B, the meaning of the message usually depends on the common context between the two agents. Without context information, the command "On" is too ambiguous because "On" could mean many things, for example, turn on a water tap, turn on a light, turn on a projector, or run a weekly meeting presentation.

To show how contexts can be detected and decisions can be made using contextual information within our reasoning model, we consider a simple theory representing an office assistant that can help office workers to control their modernized office environment for non-interrupted work flow:

$$
\begin{aligned}
F &= \{MRoom, Monday, Morning, onProjector\}\\
R^{\mathrm{BEL}} &= \{r_1 : MRoom, Monday, Morning \Rightarrow_{\mathrm{BEL}} CWMeeting\\
&\quad\ r_2 : MRoom, Monday, Morning \Rightarrow_{\mathrm{BEL}} \neg CDMeeting\\
&\quad\ r_3 : MRoom, Morning \Rightarrow_{\mathrm{BEL}} CDMeeting\\
&\quad\ r_4 : onProjector \rightarrow_{\mathrm{BEL}} \neg trunOnProjector\}\\
R^{C} &= \{r_5 : CWMeeting \Rightarrow_{C} (MessageOn \Rightarrow_{\mathrm{INT}} turnOnProjector \otimes\\
&\qquad\qquad openWMPresentation \otimes openDMPresentation)\\
&\quad\ r_6 : CDMeeting \Rightarrow_{C} (MessageOn \Rightarrow_{\mathrm{INT}} turnOnProjector \otimes\\
&\qquad\qquad openDMPresentation \otimes openWMPresentation)\}
\end{aligned}
$$

Herein, rules r_1 and r_2 say that if the assistant agent is in a meeting room (*MRoom*) and it is Monday morning, then usually the context is that of a weekly meeting (*CWMeeting*) and not of a daily meeting (*CDMeeting*). However, rule r_3 says that if the assistant agent is in a meeting room in the morning, then the context is usually that of a daily meeting. That is, r_1, r_2, and r_3 are used to detect the context. Once the context is determined, the assistant agent can properly process the command "On". For instance, r_5 says that if it is a weekly meeting, enable the following rule:

$$MessageOn \Rightarrow_{\mathrm{INT}} trunOnProjector \otimes openWMPresentation \otimes openDMPresentation$$

meaning that, if the agent receives a command "On", then the agent usually should form the intention to turn on the data projector, but if it cannot be turned on (because

it is already turned on), then it should try to open a weekly meeting presentation file. However, for some reason, if the weekly meeting presentation cannot be run (maybe it is just a daily meeting), then it should try to open a daily meeting presentation file.

On the other hand, r_6 says that if it is a daily meeting, enable the following rule:

$$Message\_On \Rightarrow_{INT} trunOnProjector \otimes openDMPresentation \otimes openWMPresentation$$

This rule sasys that, in this context, opening a daily meeting presentation file ($openDMPresentation$) has higher priority than opening a weekly meeting presentation file ($WMPresentation$). With the given theory, the office agent will conclude $+\partial_{INT}$ *openWMPresentation* if it receives $Message\_On$ because it is a weekly meeting and a projector is already turned on. This example clearly illustrates how contextual information is naturally represented within our reasoning model. That is, contextual information can be used to enable certain rules and to change priority between deliberations.

5 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. They deal with complex structures where nested rules can be involved to capture scenarios where rules are violated. Further research are the development of a methodology to use the language, and a formal analysis of the logic.

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.R. Cohen and H.J. 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, V. Padmanabhan, and A. Rotolo. Rule-based agents in temporalised defeasible logic. In *Proc. PRICAI 2006*, LNAI 4099, pages 31–40. Springer, 2006.
5. G. Governatori and A. Rotolo. Defeasible logic: Agency, intention and obligation. In *Proc. Δeon'04*, LNCS 3065, pages 114–128. Springer, 2004.
6. 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.
7. G. Governatori, A. Rotolo, and V. Padmanabhan. The cost of social agents. In *Proc. AAMAS 2006*, pages 513–520, 2006.
8. A.S. Rao and M.P. Georgeff. Decision procedures for bdi logics. *J. Log. Comput.*, 8(3):293–342, 1998.
9. Insu Song and Guido Governatori. Nested rules in defeasible logic. In *Proc. RuleML 2005*, volume LNCS 3791, pages 204–208. Springer, 2005.