Norm Compliance of Rule-Based Cognitive Agents

Antonino Rotolo
CIRSFID, University of Bologna
Italy
antonino.rotolo@unibo.it

Abstract

This paper shows how belief revision techniques can be used in Defeasible Logic to change rule-based theories characterizing the deliberation process of cognitive agents. We discuss intention reconsideration as a strategy to make agents compliant with the norms regulating their behavior.

1 Introduction and Background

A significant research effort in MAS combine two perspectives [Broersen *et al.*, 2002; Dastani *et al.*, 2005; Dignum, 1999; van der Torre *et al.*, 2008]: (a) a classical (BDI-like) cognitive model of agents; (b) a model of agents' behavior based on normative concepts. This combination leads to an account of agents' deliberation in terms of the interplay between mental attitudes and normative factors such as obligations.

In this approach agents' reasoning is typically embedded in rule-based non-monotonic systems, as one significant problem concerns the cases where the agent's intentions are in conflict with obligations.

If intentions prevail over obligations, this poses the question of agents' norm compliance. There are two strategies to get compliance in MAS. First, norm compliance is achieved by design, i.e., by stating that rules supporting the derivation of obligations always prevail over rules supporting conflicting intentions [Broersen *et al.*, 2002; Governatori and Rotolo, 2008]. Second, compliance is ensured by stating that violations should result in sanctions or other normative effects [Governatori and Rotolo, 2009], as norms cannot limit in advance agents' behavior, but provide soft constraints which can be violated [Boella and van der Torre, 2004].

There are pros and cons in both approaches. But, independently of this, a research issue is still overlooked in the literature: what's the relation between norm compliance and intention reconsideration? Most of the existing models of intentional systems view the reconsideration of intentions as either a costly computational process or mainly dependent on the dynamics of beliefs [Singh, 1991; Rao and Georgeff, 1991; Shoham, 1993; Meyer *et al.*, 1999; Wooldridge, 2000; Lorini, 2007; Lorini and Herzig, 2008]. Decision making in most agent systems is composed of two main activities like *deliberation* (deciding *what intentions* to achieve) and *means/ends reasoning* (deciding *how* to achieve these intentions). Deliberation itself can be a computationally costly process and requires an appropriate intention reconsideration policy which helps the agent to deliberate only when necessary. In this picture, it is still overlooked the problem of changing intentions not because of the change of beliefs, but because the normative constraints require to do so.

This paper explores how different types of intention reconsiderations can be modeled by applying techniques from revision theory to an extension of Defeasible Logic (DL) which embeds modalities for obligations and intentions [Governatori and Rotolo, 2008; Governatori *et al.*, 2009] and also—this is another novelty—makes use of path labels to keep track of the reasoning chains leading to "illegal" intentions.

The layout of the paper is as follows. Section 2 presents the extension of DL with path labels to reason about intentions and obligations; the purpose is to develop a formalism able to handle intention reconsideration when intentions conflict with obligations. Section 3 recalls a classification between different types of intentions and then shows how different techniques from revision theory can used in the proposed logical framework.

2 The Logical Framework

In line with [Governatori and Rotolo, 2008; Governatori *et al.*, 2009] we develop a constructive account of the modalities **O** and **I** corresponding to obligations and intentions: rules for these concepts are thus meant to devise suitable logical conditions for introducing modalities. For example, rules such as $a_1,\ldots,a_n \Rightarrow_{\mathbf{O}} b$ and $d_1,\ldots,d_n \Rightarrow_{\mathbf{I}} e$ if applicable, will allow for deriving $\mathbf{O}b$ and $\mathbf{I}e$, meaning the former that b is obligatory, the latter that e is an intention of an agent.

In our language, for $X \in \{\mathbf{O},\mathbf{I}\}$, *strict rules* have the form $\phi_1,\ldots,\phi_n \rightarrow_X \psi$. *Defeasible rules* have the form $\phi_1,\ldots,\phi_n \Rightarrow_X \psi$. A rule of the form $\phi_1,\ldots,\phi_n \rightsquigarrow_X \psi$ is a *defeater*. Strict rules support indisputable conclusions whenever their antecedents, too, are indisputable; defeasible rules can be defeated by contrary evidence; defeaters cannot lead to any conclusion but are used to defeat some defeasible rules by producing evidence to the contrary.

Definition 1 (Language). *Let* Prop *be a set of propositional atoms,* Mod $= \{\mathbf{O},\mathbf{I}\}$*, and* Lbl *be a set of labels. The sets defined below are the smallest ones closed under the given*

construction conditions:

Literals

$$\text{Lit} = \text{Prop} \cup \{\neg p | p \in \text{Prop}\}$$

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);

Modal literals

$$\text{ModLit} = \{X\,l, \neg X\,l | X \in \{\mathbf{O}, \mathbf{I}\}, l \in \text{Lit}\}$$

Rules $\text{Rul} = \text{Rul}^X_{\text{s}} \cup \text{Rul}^X_{\text{d}} \cup \text{Rul}^X_{\text{dft}}$, *where* $X \in \{\mathbf{I}, \mathbf{O}\}$, *s.t.*

$$\begin{aligned}
\text{Rul}^X_{\text{s}} &= \{r : \phi_1, \ldots, \phi_n \rightarrow_X \psi | \\
&\quad r \in \text{Lbl}, A(r) \subseteq \text{Lit} \cup \text{ModLit}, \psi \in \text{Lit}\} \\
\text{Rul}^X_{\text{d}} &= \{r : \phi_1, \ldots, \phi_n \Rightarrow_X \psi | \\
&\quad r \in \text{Lbl}, A(r) \subseteq \text{Lit} \cup \text{ModLit}, \psi \in \text{Lit}\} \\
\text{Rul}^X_{\text{dft}} &= \{r : \phi_1, \ldots, \phi_n \rightsquigarrow_X \psi | \\
&\quad r \in \text{Lbl}, A(r) \subseteq \text{Lit} \cup \text{ModLit}, \psi \in \text{Lit}\}
\end{aligned}$$

We use some obvious abbreviations, such as superscripts for the rule mode ($\mathbf{I}, \mathbf{O}$), subscripts for the type of rule, and $\text{Rul}[\phi]$ *for rules whose consequent is ϕ, for example:*

$$\begin{aligned}
&\text{Rul}^{\mathbf{I}} = \{r : \phi_1, \ldots, \phi_n \hookrightarrow_{\mathbf{I}} \psi | \hookrightarrow \in \{\rightarrow, \Rightarrow, \rightsquigarrow\}\} \\
&\text{Rul}^X_{\text{sd}} = \{r : \phi_1, \ldots, \phi_n \hookrightarrow_X \psi | X \in \text{Mod}, \hookrightarrow \in \{\rightarrow, \Rightarrow\}\} \\
&\text{Rul}_{\text{s}}[\psi] = \{\phi_1, \ldots, \phi_n \rightarrow_X \psi | \forall X \in \text{Mod}\}
\end{aligned}$$

We use $A(r)$ 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.*

An agent theory is the knowledge base which is used to reason about the agent's intentions and their interplay with a set of normative rules regulating the agent's deliberation.

Definition 2 (Agent Theory). *An agent theory D is a structure $(F, R^{\mathbf{O}}, R^{\mathbf{I}}, \succ)$ where (i) $F \subseteq \text{Lit} \cup \text{ModLit}$ is a finite set of facts; (ii) $R^{\mathbf{O}} \subseteq \text{Rul}^{\mathbf{O}}$ is a finite set of obligation rules; (iii) $R^{\mathbf{I}} \subseteq \text{Rul}^{\mathbf{I}}$ is a finite set of intention rules; (iv) $\succ$ is an acyclic (superiority) relation over $(R^{\mathbf{I}} \times R^{\mathbf{I}}) \cup (R^{\mathbf{O}} \times R^{\mathbf{O}})$.*

Definition 3. *A* path *based on an agent theory D is a structure*

$$[\alpha_{1_1}, \ldots, \alpha_{n_1}]_1 [\alpha_{1_2}, \ldots, \alpha_{n_2}]_2 \ldots [\omega]_j$$

where α_{l_w}, $1 \leq w \leq j-1$ and $1_w \leq l_w \leq n_w$, is either a literal, modal literal, or a rule, such that $j \geq 0$. If $j \geq 1$, then

- *either*
  - *$\omega = -r$ where $r \in R^{\mathbf{O}} \cup R^{\mathbf{I}}$; or*
  - *$\omega \in R^{\mathbf{O}}_{\text{sd}} \cup R^{\mathbf{I}}_{\text{sd}}$ such that, if $j = 1$ then $A(\omega) = \emptyset$; or*
  - *if $\omega \in R^{\mathbf{O}}_{\text{sd}} \cup R^{\mathbf{I}}_{\text{sd}}$ and $j > 1$, then $\forall b \in A(\omega)$ $\exists \alpha_{k_{j-1}}$ such that either*
    - *if $b \in \text{Lit}$, then $\alpha_{k_{j-1}} = b \in F$, or*
    - *if $b = Xl \in \text{ModLit}$, then either $\alpha_{k_{j-1}} = b \in F$ or $l = C(\alpha_{k_{j-1}}) : \alpha_{k_{j-1}} \in R^X_{\text{sd}}$;*
- *$\forall \alpha_{x_t}$, $1 < t \leq j-1$, $\alpha_{x_t} \in R^{\mathbf{O}}_{\text{sd}} \cup R^{\mathbf{I}}_{\text{sd}}$ such that $\forall b \in A(\alpha_{x_t})$ $\exists \alpha_{y_{t-1}}$ such that either*
  - *if $b \in \text{Lit}$, then $\alpha_{y_{t-1}} = b \in F$, or*
  - *if $b = Xl \in \text{ModLit}$, then either $\alpha_{y_{t-1}} = b \in F$ or $l = C(\alpha_{y_{t-1}}) : \alpha_{y_{t-1}} \in R^X_{\text{sd}}$.*

An empty path *is a path where $j = 0$. A* broken path *is a path where $\omega = -r$. A rule r occurs in a path iff $r = \alpha_{l_w}$, or $r = \omega$.*

Example: if $d \in F$ and we have the following rules

$$\begin{array}{lll}
r : Oe \Rightarrow_{\mathbf{I}} f & s : \mathbf{I}f \Rightarrow_{\mathbf{O}} a & t : d \Rightarrow_{\mathbf{I}} g \\
u : Ig \Rightarrow_{\mathbf{I}} b & w : Oa, Ib \Rightarrow_{\mathbf{O}} c & z : \Rightarrow_{\mathbf{O}} e
\end{array}$$

then we can obtain, e.g., the path $[z,d]_1[r,t]_2[s,u]_3[w]_4$.

Proofs are sequences of literals and modal literals together with the so-called proof tags $+\Delta$, $-\Delta$, $+\partial$ and $-\partial$. These tags can be labeled by modalities and paths: the modality indicates the mode of the conclusion (if it is an intention or an obligation), the path keeps track of the facts and rules used to obtain it. Hence, if $X \in \{\mathbf{O}, \mathbf{I}\}$, given an agent theory D, $+\Delta^X \mathscr{L} q$ means that literal q is provable as modalized with X (e.g., $\mathbf{O}q$, if $X = \mathbf{O}$) in D using the facts and strict rules in the path $\mathscr{L}$, $-\Delta^X \mathscr{L} q$ means that it has been proved in D that q is not definitely provable in D, $+\partial^X \mathscr{L} q$ means that q is defeasibly provable as modalized with X in D using the facts and rules in $\mathscr{L}$, and $-\partial^X \mathscr{L} q$ means that it has been proved in D that q is not defeasibly provable in D. We will clarify later the structure of paths in the case of the negative proof tags.

Definition 4. *Given an agent theory D, a proof in D is a linear derivation, i.e, a sequence of labelled formulas of the type $+\Delta^X \mathscr{L} q$, $-\Delta^X \mathscr{L} q$, $+\partial^X \mathscr{L} q$ and $-\partial^X \mathscr{L} q$, where the proof conditions defined in the rest of this section hold.*

Definition 5. *Let D be an agent theory. Let $\# \in \{\Delta, \partial\}$ and $X \in \{\mathbf{O}, \mathbf{I}\}$, $\mathscr{L}$ be any path based on D, and $P = (P(1), \ldots, P(n))$ be a proof in D. A literal q is $\#\mathscr{L}$-*provable *in P if there is a line $P(m)$, $1 \leq m \leq n$, of P s.t. either*

1. *q is a modal literal Xp and $P(m) = +\#^X \mathscr{L} p$ or*
2. *q is a modal literal $\neg Xp$ and $P(m) = -\#^X \mathscr{L} p$.*

*A literal q is $\#\mathscr{L}$-*rejected *in P if there is a line $P(m)$ of P s.t.*

1. *q is a modal literal Xp and $P(m) = -\#^X \mathscr{L} p$, or*
2. *q is a modal literal $\neg Xp$ and $P(m) = +\#^X \mathscr{L} p$.*

The definition of Δ^X describes just forward (monotonic) chaining of strict rules: given $1 \leq j \leq n$

If $P(n+1) = +\Delta^X \mathscr{L}[\alpha_1, \ldots, \alpha_n][r]q$ then
(1) $\exists x \in R^X_s[q]$:
(1.1) $x = r$ and
(1.2) $\forall a \in A(r)$ either
(1.2.1) $a \in F$, or
(1.2.2) a is $\Delta\mathscr{L}[\alpha_j]$-provable.

If $P(n+1) = -\Delta^X \mathscr{L}[\alpha_1, \ldots, \alpha_n][r]q$ then
(1) $\forall x \in R^X_s[q]$ either
(1.1) $x \neq r$ or
(1.2) $\exists a \in A(r)$:
(1.2.1) $a \notin F$, and
(1.2.2) a is $\Delta\mathscr{L}[\alpha_j]$-rejected.

The path supporting q is built step by step by including the rules and facts used to obtain it. In the case of negative proof tags, any path involved is in fact empty, since there is no reasoning chain supporting q. See also Proposition 1 below.

Example 1. *Consider the following agent theory:*

$$F = \{\mathbf{O}a, \mathbf{I}b\}$$
$$R = \{r_1 : \mathbf{I}b \rightarrow_{\mathbf{I}} c, r_2 : \mathbf{I}c \rightarrow_{\mathbf{I}} a\}$$
$$\succ = \emptyset$$

Let us work on the proof conditions for Δ. The obligation $\mathbf{O}a$ does not trigger any rule. The fact $\mathbf{I}b$ triggers r_1 (condition (1.1)): hence we obtain $+\Delta^{\mathbf{I}}[\mathbf{I}b][r_1]c$. Now, using proof condition (1.2), we trigger r_2 to get $+\Delta^{\mathbf{I}}[\mathbf{I}b][r_1][r_2]a$.

Consider now proof conditions for ∂^X: given $1 \leq j \leq n$,

If $P(n+1) = +\partial^X \mathscr{L}[\alpha_1, \ldots, \alpha_n][r]q$ then
(1) $+\Delta^X \mathscr{L}[\alpha_1, \ldots, \alpha_n][r]q$ or
(2) (2.1) $-\Delta^X \mathscr{X} \sim q \in P[1..n]$ and
(2.2) $\exists x \in R^X_{\mathrm{sd}}[q]$:
(2.2.1) $x = r$ and
(2.2.2) $\forall a \in A(r)$
(2.2.2.1) $a \in F$, or
(2.2.2.2) a is $\partial\mathscr{L}[\alpha_j]$-provable, and
(2.3) $\forall s \in R^X[\sim q]$ either $\exists a \in A(s)$:
a is $\partial\mathscr{Y}$-rejected, or
(2.3.1) $\exists t \in R^X[q]$: $\forall a \in A(r)$
a is $\partial\mathscr{Z}$-provable and $t \succ s$.

If $P(n+1) = -\partial^X \mathscr{L}[\alpha_1, \ldots, \alpha_n][r]q$ then
(1) $-\Delta^X \mathscr{L}[\alpha_1, \ldots, \alpha_n][r]q$ and
(2) (2.1) $+\Delta^X \mathscr{X} \sim q \in P[1..n]$ or
(2.2) $\forall x \in R^X_{\mathrm{sd}}[q]$ either
(2.2.1) $x \neq r$, or
(2.2.2) $\exists a \in A(r)$:
(2.2.2.1) $a \notin F$, and
(2.2.2.2) a is $\partial\mathscr{L}[\alpha_j]$-rejected, or
(2.3) $\exists s \in R^X[\sim q]$: $\forall a \in A(s)$
a is $\partial\mathscr{Y}[s]$-provable, and
(2.3.1) $\forall t \in R^X[q]$, $\exists a \in A(r)$:
a is $\partial\mathscr{Z}$-rejected or $t \not\succ s$,
where $[r] = [-r]$ if $\forall a \in A(r)$
$a \in F$, or
a is $\partial\mathscr{L}[\alpha_j]$-provable.

To show that a literal q is defeasibly provable with the mode X we have two choices: (a) We show that q is already definitely provable; or (b) We need to argue using the defeasible part of an agent theory D. For this second case, some (sub)conditions must be satisfied: First, we need to consider possible reasoning chains in support of $\sim q$ with the mode X, and show that $\sim q$ is not definitely provable with that mode (2.1 above). Second, we require that there must be a strict or defeasible rule with mode X for q which can be applied (2.2 above). Third, we must consider the set of all rules which are not known to be inapplicable and which permit to get $\sim q$ with the mode X (2.3 above). Essentially, each such a rule s attacks the conclusion q. For q to be provable, s must be counterattacked by a rule t for q with the following properties: t must be applicable and must prevail over s. Thus each attack on the conclusion q must be counterattacked by a stronger rule. In other words, r and the rules t form a team (for q) that defeats the rules s. The mechanism for handling paths is basically the one for definite conclusions. The only difference is that here we can have broken paths when a rule is made applicable but is defeated by a stronger rule: in this case, the path keeps track of the defeated rule r, which is marked as $-r$.

Proposition 1. *(a) For $-\Delta$: if condition (1) holds, then $\mathscr{L}[\alpha_1, \ldots \alpha_n][r]$ is an empty path.*

(b) For $-\partial$: if condition (2.2) holds, then $\mathscr{L}[\alpha_1, \ldots \alpha_n][r]$ is an empty path;

(c) For $-\partial$: if condition (2.3) holds and $\mathscr{L}[\alpha_1, \ldots \alpha_n][r]$ is broken, then rule r is applicable.

Sketch. Consider the case (a): if condition (1) holds, this means that there is no path and proof supporting q, and so $\mathscr{L}[\alpha_1, \ldots \alpha_n][r]$ must be empty. The same argument applies to the case (b). Consider case (c): here, by construction there is a path and a proof supporting the antecedents of r, even though any r is defeated. Hence r is applicable. □

Example 2. *Let us expand the theory in Example 1:*

$$F = \{\mathbf{O}a, \mathbf{I}b, d\}$$
$$R = \{r_1 : \mathbf{I}b \rightarrow_{\mathbf{I}} c, r_2 : \mathbf{I}c \rightarrow_{\mathbf{I}} a, r_3 : \rightsquigarrow_{\mathbf{O}} \neg e, r_4 : \mathbf{I}b \Rightarrow_{\mathbf{O}} e,$$
$$r_5 : d \Rightarrow_{\mathbf{I}} \neg e, r_6 : \Rightarrow_{\mathbf{I}} \neg e, r_7 : \mathbf{I}\neg e, \mathbf{I}b \Rightarrow_{\mathbf{O}} f\}$$
$$\succ = \{r_4 \succ r_3\}$$

Since the defeasible part of the theory cannot affect the derivation obtained using the monotonic part, which is the same of Example 1, the definite conclusions do not change. The fact $\mathbf{I}b$ triggers r_4, which conflicts with r_3; if r_3 could prevail, we would have $-\partial^{\mathbf{O}}[-r_3]e$, but this is not the case since r_4 is stronger than r_3, thus leading to $+\partial^{\mathbf{O}}[\mathbf{I}b][r_4]e$. The fact d makes r_5 applicable, and so $+\partial^{\mathbf{I}}[d][r_5]\neg e$. The rule r_6 is always applicable, thus supporting $+\partial^{\mathbf{I}}[r_6]\neg e$. Finally, from the $\mathbf{I}b$ and the last two conclusions we obtain $+\partial^{\mathbf{O}}[d][r_5, \mathbf{I}b][r_7]f$ and $+\partial^{\mathbf{O}}[r_6, \mathbf{I}b]f$.

Definition 6. *Given an agent theory D, $D \vdash \pm\#^X l$ (i.e., $\pm\#^X l$ is a conclusion of D), where $\# \in \{\Delta, \partial\}$ and $X \in \{c, \mathbf{O}, \mathbf{I}\}$, iff there is a proof $P = (P(1), \ldots, P(n))$ in D such that $P(n) = \pm\#^X l$.*

Definition 7. *Given a theory D, the* universe *of D (U^D) is the set of all the atoms occurring in D; the* extension *of D (E^D), is defined as follows:*

$$E^D = (\Delta^+(D), \Delta^-(D), \partial^+(D), \partial^-(D))$$

where for $X \in \{\mathbf{I}, \mathbf{O}\}$

$$\Delta^+(D) = \{Xl | D \vdash +\Delta^X \mathscr{L} l\} \quad \Delta^-(D) = \{Xl | D \vdash -\Delta^X \mathscr{L} l\}$$
$$\partial^+(D) = \{Xl | D \vdash +\partial^X \mathscr{L} l\} \quad \partial^-(D) = \{Xl | D \vdash -\partial^X \mathscr{L} l\}.$$

3 Compliance and Revising Intentions

3.1 Conceptual Background

Suppose the agent's intentions conflict with some obligations. If we assume that obligations are unchangeable, the possibility to avoid violations relies on the possibility to handle rules for intentions.

As we argued elsewhere [Governatori and Rotolo, 2008; Governatori *et al.*, 2009], we can conceptually distinguish between different types of intentions: unchangeable intentions, strong intentions and weak intentions. The first type corresponds in our logic to intentional facts of an agent theory (elements of F), the second type to definite conclusions ($+\Delta^{\mathbf{I}}$),

the third to defeasible conclusions ($+\partial^{\mathbf{I}}$). Unchangeable intentions cannot be reconsidered in any case (see [Governatori *et al.*, 2009] for a discussion on this issue). To give up a strong intention we have necessarily to change (revise) the theory (i.e., we have to modify the strict rules), while we can abandon a weak intention if we have an exception to it without having to change the theory. To illustrate this point let us consider the following rules:

$$r_1 : a \rightarrow_{\mathbf{I}} b \qquad r_2 : c \rightarrow_{\mathbf{I}} \neg b \tag{1}$$

Suppose the same connections are expressed as defeasible rules:

$$r'_1 : a \Rightarrow_{\mathbf{I}} b \qquad r'_2 : c \Rightarrow_{\mathbf{I}} \neg b \qquad r'_2 \succ r'_1 \tag{2}$$

In both cases we obtain $\mathbf{I}b$ given a as a fact. However if both a and c are given then from (1) we get an inconsistency, since definite conclusions cannot be blocked and we have to revise the theory. If we use belief revision to change the theory then we have to remove r_1 from the theory. A consequence of this operation is that we are no longer able to derive $\mathbf{I}b$ from a. An alternative would be to use base revision instead of belief revision. If this strategy is taken then r_1 is changed into

$$r''_1 : a, \neg c \rightarrow_{\mathbf{I}} b \tag{3}$$

Again, it is not possible to obtain $\mathbf{I}b$ from a. To derive it we have to supplement the theory with the information whether c or $\neg c$ is definitely the case, increasing then the cognitive burden on the agent.

If the same information were encoded as weak intentions, as in (2), then we would not suffer from the above drawback, since (2) prevents the conclusion of an inconsistency (in case we do not specify that r'_2 is stronger than r_1, we are not able to conclude $\mathbf{I}b$ nor $\mathbf{I}\neg b$). Indeed, the defeasibility of weak intentions makes it possible to block the application of the intention to the particular case without reconsidering it. This is in agreement with [Bratman, 1987]. This way the amount of deliberation required for intention re-consideration can be minimized to some extent.

Unfortunately, if the first and compelling purpose is to make agents compliant, not in all cases the defeasibility of intentions is *the* solution. Indeed, if a theory containing (2) allows for deriving $\mathbf{O}b$, there is no way to recover, unless we change the theory.

3.2 A Simple Model

Let us first formally characterize the notion of compliance to a norm:

Definition 8 (Rule Fulfilment and Violation). *An agent theory $D = (F, R^{\mathbf{O}}, R^{\mathbf{I}}, \succ)$* fulfil *a rule $r \in R^{\mathbf{O}}_{\mathrm{sd}}$ iff, if $D \vdash +\partial^{\mathbf{O}} \mathscr{L} C(r)$, then, either*

- *if $C(r)$ is a positive literal l (r is a conditional obligation), then there is an $\mathscr{Y}$ such that $D \vdash +\partial^{\mathbf{I}} \mathscr{Y} l$, or*
- *if $C(r)$ is a negative literal $\neg l$ (r is a conditional prohibition) for any $\mathscr{L}$, $D \vdash -\partial^{\mathbf{I}} \mathscr{L} l$ or $D \vdash +\partial^{\mathbf{I}} \mathscr{L} \neg l$.*

D violates *the rule r whenever D does not fulfil r. D is compliant iff D does not violate any rule in it.*

As we briefly discussed in Section 3.1, an option to recover from violations and reinstate compliance is to revise intentions by using AGM techniques. This idea looks natural (see e.g. [Cawsey *et al.*, 1993; Lorini, 2007]). However, it is far from obvious how to do it in DL. Fortunately, AGM fundamental operations have been defined for propositional DL in [Billington *et al.*, 1999].

The first step is thus to extend [Billington *et al.*, 1999]'s notions of expansion and contraction to cover DL with modalities, which is trivial. Consider an agent theory D and suppose we want to expand the extension of D with $c = \mathbf{I}p_1, \ldots, \mathbf{I}p_n$:

$$D^+_c = \begin{cases} D & \text{if } \exists i \in \{1,\ldots n\}\text{: } \mathbf{I}{\sim}p_i \in \partial^+(D) \\ D & \text{if } \exists i,j \in \{1,\ldots n\}\text{: } {\sim}p_i = p_j \\ (F, R^{\mathbf{O}}, R^{\mathbf{I}'}, \succ') & \text{otherwise} \end{cases}$$

where

$$\begin{aligned} R^{\mathbf{I}'} =& R^{\mathbf{I}} \cup \{w_1 :\Rightarrow_{\mathbf{I}} p_1, \ldots, w_n :\Rightarrow_{\mathbf{I}} p_n\} \\ \succ' =& (\succ \cup \{w_i \succ r \mid 1 \leq i \leq n, r \in R^{\mathbf{I}}[{\sim}p]\}) - \\ & \{r \succ w_i \mid 1 \leq i \leq n, r \in R^{\mathbf{I}}[{\sim}p]\}. \end{aligned} \tag{4}$$

Thus, we add rules that prove $p_1, \ldots p_n$ as intentions; these rules are always applicable and are strictly stronger than any possibly contradicting rules. This solution looks useful to deal with many cases of violation.

Example 3. *Consider the following theory D.*

$$\begin{aligned} & F = \{a\} \\ & R = \{r_1 : a \rightarrow_{\mathbf{I}} b, r_2 : \mathbf{I}b \Rightarrow_{\mathbf{O}} c, r_3 : a \rightsquigarrow_{\mathbf{I}} \neg c, r_4 : \mathbf{I}b \Rightarrow_{\mathbf{I}} c\} \\ & \succ = \emptyset \end{aligned}$$

Here we obtain, among other conclusions, $+\partial^{\mathbf{O}}[a][r_1][r_2]c$. To be compliant, we should be able to derive that c is intended, but this is not possible. We have here that $-\partial^{\mathbf{I}}[a][r_1][-r_4]c$. What we can do is to expand D with $\mathbf{I}c$ by simply adding a rule w and applying (4). Since this operation satisfies AGM postulates for expansion [Billington et al.*, 1999], this guarantees that $\mathbf{I}c$ is added to the positive extension of D. Hence, we obtain $+\partial^{\mathbf{I}}[w]c$ and make D compliant.*

Let us define the procedure explained in Example 3.

Definition 9 (Positive Revision). *Let $D = (F, R^{\mathbf{O}}, R^{\mathbf{I}}, \succ)$ be an agent theory. If D* violates *the rules $r_1, \ldots, r_n \in R^{\mathbf{O}}$, then D^+_c where $c = \mathbf{I}p_1, \ldots, \mathbf{I}p_n$ such that $C(r_1) = p_1, \ldots, C(r_n) = p_n$.*

Let us adjust [Billington *et al.*, 1999]'s definition of contraction. Here, too, we trivially extends [Billington *et al.*, 1999]'s. If we want to contract $c = \mathbf{I}p_1, \ldots, \mathbf{I}p_n$ in D, then:

$$D^-_c = \begin{cases} D & \text{if } \mathbf{I}p_1, \ldots, \mathbf{I}p_n \notin \partial^+(D) \\ (F, R^{\mathbf{O}} R^{\mathbf{I}'}, \succ') & \text{otherwise} \end{cases}$$

where

$$\begin{aligned} R^{\mathbf{I}'} =& R^{\mathbf{I}} \cup \{s : \mathbf{I}p_1, \ldots, \mathbf{I}p_{i-1}, \mathbf{I}p_{i+1}, \ldots, \mathbf{I}p_n \rightsquigarrow_{\mathbf{I}} {\sim}p_i \mid \\ & 1 \leq i \leq n\} \\ \succ' =& \succ - \{r \succ s \mid r \in R^{\mathbf{I}'} - R^{\mathbf{I}}\}. \end{aligned} \tag{5}$$

(5) blocks the proof of $\mathbf{I}p_1,\ldots,\mathbf{I}p_n$. It is ensured that at least one of the $\mathbf{I}p_i$s will not be derived. The new rules in $R^{\mathbf{I}'}$ are such that, if all but one $\mathbf{I}p_i$ have been obtained, a defeater with head $\sim p_j$ is triggered. The defeaters are not weaker than any other rules, so the defeater cannot be "counterattacked" by another rule, and p_j will not be proven as an intention.

Example 4. *Consider the following theory D.*

$$F = \{a, \mathbf{I}d\}$$
$$R = \{r_1 : a \rightarrow_{\mathbf{I}} b, r_2 : \mathbf{I}b \Rightarrow_{\mathbf{O}} c, r_3 : \mathbf{I}d \Rightarrow_{\mathbf{I}} \neg c\}$$
$$\succ = \emptyset$$

We obtain, among other conclusions, $+\partial^{\mathbf{O}}[a][r_1][r_2]c$. To be compliant, we should derive that $\mathbf{I}c$, but we obtain the opposite through r_3. More precisely, we get $+\partial^{\mathbf{I}}[\mathbf{I}d][r_3]\neg c$. What we can do is to contract $\mathbf{I}d$ by simply adding a defeater $s :\rightsquigarrow_{\mathbf{I}} \neg c$ and thus applying (5). Since this operation satisfies AGM postulates for contraction [Billington et al.*, 1999], this guarantees that $\mathbf{I}d$ is removed from the positive extension of D and added to the negative extension.*

Definition 10 (Negative Revision). *Let $D = (F, R^{\mathbf{O}}, R^{\mathbf{I}}, \succ)$ be an agent theory. If D* violates *the rules $r_1,\ldots,r_n \in R^{\mathbf{O}}$, then D_c^- where $c = \mathbf{I}p_1,\ldots,\mathbf{I}p_n$ such that $C(r_1) = \neg p_1,\ldots,C(r_n) = \neg p_n$.*

Definitions 9 and 10 guarantee to recover from violations, are very simple, and directly exploit techniques and results from [Billington *et al.*, 1999]. Also, they do not make any essential use of paths, which sometimes may look cumbersome. However, they have two serious drawbacks: (a) They work only on the defeasible part of agent theories, and so cannot be used to recover from violations when these are caused by strong intentions (see the discussion in Section 3.1); (b) They apply only to the last rule of the reasoning chains supporting "illegal" intentions.

To overcome the above difficulties, DL with paths is useful.

3.3 Refinements: Using Paths

The advantage of using paths is that we can easily identify (i) which rules have been violated, and (ii) which rules for intentions have determined the violation of an obligation.

Let us see when Definitions 9 and 10 clearly fail while DL with paths succeeds.

Example 5 (Strong Intentions). *Consider this theory:*

$$F = \{a, \mathbf{I}b\}$$
$$R = \{r_1 : a \rightarrow_{\mathbf{I}} \neg c, r_2 : \mathbf{I}b \Rightarrow_{\mathbf{O}} c, r_3 : \mathbf{I}b \rightarrow_{\mathbf{I}} d,$$
$$r_4 : \mathbf{I}d, a \rightarrow_{\mathbf{I}} \neg c\}$$
$$\succ = \emptyset$$

We have two reasons for the violation of r_2 (indeed, we obtain $+\partial^{\mathbf{O}}[\mathbf{I}b][r_2]c$). In fact, we can derive $+\Delta^{\mathbf{I}}[a][r_1]\neg c$ and $+\Delta^{\mathbf{I}}[\mathbf{I}b][r_3,a][r_4]\neg c$. Since strict rules cannot be defeated, the only solution is rule removal. Hence, we have to operate over r_1 but we are free to remove either r_3 or r_4. For example, if we prefer not to remove r_4, we will successfully get compliance by removing r_1 and only r_3.

Definition 11 (Rule Removal). *Let $D = (F, R^{\mathbf{O}}, R^{\mathbf{I}}, \succ)$ be an agent theory. For each $r \in R^{\mathbf{O}}_{\mathrm{sd}}$ such that the paths $\mathscr{L}_1,\ldots\mathscr{L}_n$ are the ones based on D such that $D \vdash +\Delta^{\mathbf{I}}\mathscr{L}_1 p,\ldots,D \vdash +\Delta^{\mathbf{I}}\mathscr{L}_n p$ and $D \vdash +\partial^{\mathbf{O}}\mathscr{Y}C(r)$, where $C(r) = \neg p$, the theory D_{-X} is such that*

- *$X = \{w_1,\ldots,w_m\}$ is the smallest set of rules in $R^{\mathbf{I}}$ such that, for each $k \in \{1,\ldots,n\}$, there is at least a $w_j \in X$ that occurs in $\mathscr{L}_k$,*
- *$R^{\mathbf{I}}_{-X} = R^{\mathbf{I}} - X$, and*
- *$F_{-X} = F$, $R^{\mathbf{O}}_{-X} = R^{\mathbf{O}}$, and $\succ_{-X} = \succ$.*

Let us work on weak intentions only. The following definition proposes intention retraction for DL with paths by exploiting the contraction of intentions as framed in (5).

Definition 12 (Contraction with Paths). *Let $D = (F, R^{\mathbf{O}}, R^{\mathbf{I}}, \succ)$ be an agent theory. For each $r \in R^{\mathbf{O}}_{\mathrm{sd}}$ such that the paths $\mathscr{L}_1,\ldots\mathscr{L}_n$ are the ones based on D such that $D \vdash +\partial^{\mathbf{I}}\mathscr{L}_1 p,\ldots,D \vdash +\partial^{\mathbf{I}}\mathscr{L}_n p$ and $D \vdash +\partial^{\mathbf{O}}\mathscr{Y}C(r)$, where $C(r) = \neg p$, the theory $D_{\flat p} = (F, R^{\mathbf{O}}, R^{\mathbf{I}'}, \succ')$ is such that*

(i) $R^{\mathbf{I}'} = R^{\mathbf{I}} \cup \{s :\rightsquigarrow_{\mathbf{I}} \sim q\} \cup \{t :\rightsquigarrow_{\mathbf{I}} \sim x\}$,

(ii) $\succ' = \succ - [\{r_k \succ s | r_k \in R^{\mathbf{I}}[\sim C(s)], r_k$ *occurs in* $\mathscr{L}_k\ \forall k \in \{1,\ldots,n\}\} \cup \{w \succ t \mid$ *for each path* $\mathscr{M}[-w]$ *based on D such that* $C(w) = x$, *either* $x = p$ *or* w *occurs in* $\mathscr{L}_k\ \forall k \in \{1,\ldots,n\}\}]$.

Example 6 (Paths). *Consider the following agent theory:*

$$F = \{a, \mathbf{I}b\}$$
$$R = \{r_1 : a \Rightarrow_{\mathbf{I}} \neg c, r_2 : \mathbf{I}b \Rightarrow_{\mathbf{O}} c, r_3 : \mathbf{I}b \Rightarrow_{\mathbf{I}} d, r_4 : \mathbf{I}d, a \Rightarrow_{\mathbf{I}} \neg c$$
$$r_5 : g \Rightarrow_{\mathbf{I}} \neg c\}$$
$$\succ = \emptyset$$

Like in Example 5, we obtain $+\partial^{\mathbf{O}}[\mathbf{I}b][r_2]c$. We also derive $+\partial^{\mathbf{I}}[a][r_1]\neg c$ and $+\partial^{\mathbf{I}}[\mathbf{I}b][r_3,a][r_4]\neg c$, which violate rule r_2. Definition 12 allow us to add, for example, a defeater for c which is stronger than r_1 and another defeater for $\neg d$ which is stronger than r_3. Hence, as we have already seen in Example 5, Definition 12 does not only provide tools to affect the rules r_1 and r_4 that directly prove illegal intentions, but also rules preceding them in the involved path (e.g., r_3).

Proposition 2 (Success). *Let $D = (F, R^{\mathbf{O}}, R^{\mathbf{I}}, \succ)$ be an agent theory. If, for each $r \in R^{\mathbf{O}}_{\mathrm{sd}}$ we have $D \vdash +\partial^{\mathbf{O}}\mathscr{Y}C(r)$, where $C(r) = \neg p$, and for the paths $\mathscr{L}_1,\ldots\mathscr{L}_n$ based on D*

(a) $D \vdash +\Delta^{\mathbf{I}}\mathscr{L}_1 p,\ldots,D \vdash +\Delta^{\mathbf{I}}\mathscr{L}_n p$, *then* $\mathbf{I}p \notin \Delta^+(D_{-X})$;

(b) $D \vdash +\partial^{\mathbf{I}}\mathscr{L}_1 p,\ldots,D \vdash +\partial^{\mathbf{I}}\mathscr{L}_n p$, *then* $\mathbf{I}p \notin \partial^+(D_{\flat p})$ *unless* $\mathbf{I}p \in \Delta^+(D)$.

Sketch. Case (a): By construction, Definition 11 guarantees that at least one strict intention rule is removed in every path based on D supporting $\mathbf{I}p$.

Case (b): An inspection of the proof conditions for ∂ shows that Definition 12 successfully blocks the derivation of $\mathbf{I}p$, unless it is derived using only strict rules. Notice that condition (ii) in Definition 12 ensures that, in case the attacks made by the defeaters s activate other (previously defeated) rules supporting $\mathbf{I}p$, these last potential derivations are made unsuccessful. □

4 Summary and Future Work

In this paper we presented an extension of DL with path labels to reason about intentions and obligations The formalism was able to handle intention reconsideration when the agent's intentions conflict with obligations. In particular, we showed that the reconsideration of different types of intentions can be modeled using different techniques from revision theory.

This is a preliminary step towards modeling intention reconsideration in DL. A number of open issues should be addressed. First: according to [Governatori and Rotolo, 2010] if I violate a norm r but I comply with an obligation which is meant to compensate the violation of r, I am still compliant. In [Dastani *et al.*, 2005] we introduced the operator $\otimes$ to handle compensations in a version of DL with modalities but without paths. What happens if we combine $\otimes$ with DL with paths? Second: in [Governatori and Rotolo, 2008] we showed that the extensions of agent theories, in some modal versions of DL, can be computed in linear time. We will have to check whether this is preserved in the new logic. Third: we have to investigate the properties of the new operations over agent theories. In particular, we have to better study how to minimize changes. Finally: another possibility is not to revise the set of rules for intention, but to change rule priorities [Governatori *et al.*, 2010]. Also this question is left to a future research.

Acknowledgments

This paper benefited from conversations with Guido Governatori and Leon van der Torre. Another version of it was presented at the RuleML 2011 symposium [Rotolo, 2011].

References

[Billington *et al.*, 1999] D. Billington, G. Antoniou, G. Governatori, and M.J. Maher. Revising nonmonotonic belief sets: The case of defeasible logic. In *Proc. KI-99*. Springer, 1999.

[Boella and van der Torre, 2004] G. Boella and L. van der Torre. Fulfilling or violating obligations in multiagent systems. In *Proc. IAT04*, 2004.

[Bratman, 1987] M.E. Bratman. *Intentions, Plans and Practical Reason*. Harvard University Press, Cambridge, MA, 1987.

[Broersen *et al.*, 2002] J. Broersen, M. Dastani, J. Hulstijn, and L. van der Torre. Goal generation in the BOID architecture. *Cognitive Science Quarterly*, 2(3-4):428–447, 2002.

[Cawsey *et al.*, 1993] A. Cawsey, J. Galliers, B. Logan, S. Reece, and K. Sparck Jones. Revising beliefs and intentions: A unified framework for agent interaction. In *Proc. 9th Biennial Conference of the Society for the Study of Artificial Intelligence and Simulation of Behaviour*. IOS Press, 1993.

[Dastani *et al.*, 2005] M. Dastani, G. Governatori, A. Rotolo, and L. van der Torre. Programming cognitive agents in defeasible logic. In *Proc. LPAR 2005*. Springer, 2005.

[Dignum, 1999] F. Dignum. Autonomous agents with norms. *Artificial Intelligence and Law*, 7(1):69–79, 1999.

[Governatori and Rotolo, 2008] G. Governatori and A. Rotolo. Bio logical agents: Norms, beliefs, intentions in defeasible logic. *Autonomous Agents and Multi-Agent Systems*, 17(1):36–69, 2008.

[Governatori and Rotolo, 2009] G. Governatori and A. Rotolo. How do agents comply with norms? In *Web Intelligence/IAT Workshops*, pages 488–491, 2009.

[Governatori and Rotolo, 2010] Guido Governatori and Antonino Rotolo. A conceptually rich model of business process compliance. In *APCCM*, pages 3–12, 2010.

[Governatori *et al.*, 2009] Guido Governatori, Vineet Padmanabhan, Antonino Rotolo, and Abdul Sattar. A defeasible logic for modelling policy-based intentions and motivational attitudes. *Logic Journal of the IGPL*, 17(3):227–265, 2009.

[Governatori *et al.*, 2010] G. Governatori, F. Olivieri, S. Scannapieco, and M. Cristani. Superiority based revision of defeasible theories. In *Proc. RuleML 2010*. Springer, 2010.

[Lorini and Herzig, 2008] E. Lorini and A. Herzig. A logic of intention and attempt. *Synthese*, 163(1):45–77, 2008.

[Lorini, 2007] E. Lorini. *Variations on intentional themes: From the generation of an intention to the execution of an intentional action*. PhD thesis, University of Siena, 2007.

[Meyer *et al.*, 1999] J.-J. Ch. Meyer, W. van der Hoek, and B. van Linder. A logical approach to the dynamics of commitments. *Artif. Intell.*, 113:1–40, September 1999.

[Rao and Georgeff, 1991] A.S. Rao and M.P. Georgeff. Modelling rational agents within a BDI-architecture. In *Proc. KR'91*. Morgan Kaufmann, 1991.

[Rotolo, 2011] A. Rotolo. Rule-based agents, compliance, and intention reconsideration in defeasible logic. In *Proc. RuleML 2011*, Berlin, 2011. Springer.

[Shoham, 1993] Yoav Shoham. Agent-oriented programming. *Artif. Intell.*, 60:51–92, March 1993.

[Singh, 1991] M.P. Singh. On the commitments and precommitments of limited agents. In *Proc. IJCAI-91 Workshop on Theoretical and Practical Design of Rational Agents*, 1991.

[van der Torre *et al.*, 2008] L. van der Torre, G. Boella, and H. Verhagen, editors. *Normative Multi-agent Systems*, Special Issue of *JAAMAS*, vol. 17(1), 2008.

[Wooldridge, 2000] M. Wooldridge. *Reasoning about rational agents*. MIT Press, 2000.