Picking Up The Best Goal

An Analytical Study in Defeasible Logic

Guido Governatori[1,4], Francesco Olivieri[1,2,4], Antonino Rotolo[3], Simone Scannapieco[1,2,4], and Matteo Cristani[2]

[1] NICTA, Queensland Research Laboratory, Australia
[2] Department of Computer Science, University of Verona, Italy
[3] CIRSFID and DSG, University of Bologna, Italy
[4] Institute for Integrated and Intelligent Systems, Griffith University, Australia

Abstract. In this paper we analyse different notions of the concept of goal starting from the idea of sequences of "alternative acceptable outcomes". We study the relationships between goals and concepts like agent's beliefs, norms, and desires, and we propose a formalisation using Defeasible Logic that will be able to provide a computationally feasible approach. The resulting system captures various nuances of the notion of goal against different normative domains, for which the right decision is not only context-dependent, but it must be chosen from a pool of alternatives as wide as possible.

1 Motivation and Basic Intuitions

The BDI architecture is a prominent approach to model rational agents [1,2,3]. BDI agents are means-ends reasoners equipped with: (i.) Desires, Goals, Intentions (or Tasks); (ii.) a description of the current state of the environment (Beliefs); (iii.) Actions. The key tenet of this architecture is that the agent's behaviour is the outcome of a rational balance among different mental states. Previous seminal works on the BDI paradigm [2,3,4], or implementing the BDI architecture in Defeasible Logic [5,6,7] have assumed that mental states are primitive and independent from each other, even though some mutual influences are considered (e.g., intentions are seen as desires satisfied up to commitment).

We work here on a different perspective to provide a fresh and efficient rule-based framework that considers goals, desires, and intentions as facets of the same phenomenon (all of them being *goal-like attitudes*): the notion of *outcome*, which is simply something an agent would like or is expected to achieve. An advantage of the proposed framework is that it allows agents to compute different degrees of motivational attitudes, and degrees of commitment that take into account other factors, such as beliefs and norms.

While different schemas for generating and filtering agents' outcomes are possible, we will restrict ourself to schemas where we adopt the following principles:

– When an agent faces alternative outcomes in a given context, it is natural to rank them in a preference order;
– Beliefs prevail over conflicting motivational attitudes, thus avoiding various cases of wishful thinking [8,9];

RuleML 2013,

– Norms and obligations are used to filter social motivational states (*social intentions*) and compliant agents [9,6];
– Goal-like attitudes can be derived via *conversion* using other mental states, such as beliefs (e.g., believing that Madrid is in Spain may imply that the goal to go to Madrid implies the goal to go to Spain) [6].

Consider the following example, Alice, during her holidays, plans to pay a visit to her friend John, who lives close to her parents. The plan can be described by the sentence *I shall come over to John's place to visit him on Monday, but if he is not home or the visit is not possible, I am going to visit my parents. If this is not possible as well, I shall take some rest at home.* This idea can be easily implemented by building for each alternative a sequence of other alternatives $A_1, \dots, A_n$ that are preferred when the first choice is no longer feasible. Normally, each set of alternatives is the result of a specific context C. This scenario can be represented as $C \Rightarrow A_1, \dots, A_n$ which is closely related of contrary-to-duty obligations [10], where a norm is represented by an Obligation rule of the type:

$$r_1 : drive\_car \Rightarrow_{\mathsf{O}} \neg damage \odot compensate \odot foreclosure.$$

Rule r_1 states that, if an agent drives a car, she has the obligation not to cause any damage to others; if this happens, she is obliged to compensate; if she fails to compensate, there is an obligation of foreclosure. The previous setting can be rewritten as:

$$r_2 : holiday \Rightarrow_{\mathsf{U}} visit\_friend \odot visit\_parents \odot stay\_home.$$

where r_2 is a rule introducing the oUtcome mode. In both examples, the sequences express a preference ordering among alternatives, which means that also *stay_home* and *foreclosure*, though not the best options, still correspond to acceptable situations.

Besides rules for outcomes and obligations, we also have rules for beliefs such as

$$r_3 : friend\_away \Rightarrow_{\mathsf{B}} \neg visit\_friend$$

for which we assume there is no preference ordering, since they do not express expected outcomes but simply describe how the world is.

These building blocks allow us to introduce different types of goal-like attitudes and degrees of commitment to outcomes: desires, goals, intentions, and social intentions.

Desires as acceptable outcomes Suppose an agent is equipped with the following outcome rules expressing two preference orderings:

$$r : a_1, \dots, a_n \Rightarrow_{\mathsf{U}} b_1 \odot \cdots \odot b_m \qquad s : a'_1, \dots, a'_n \Rightarrow_{\mathsf{U}} b'_1 \odot \cdots \odot b'_k$$

and that the situation described by $a_1, \dots, a_n$ and $a'_1, \dots, a'_n$ are mutually compatible but b_1 and b'_1 are not, namely $b_1 = \neg b'_1$. In this case $b_1, \dots, b_m, b'_1, \dots, b'_k$ are anyway all *acceptable outcomes*, including the incompatible outcomes b_1 and b'_1. *Desires* are expected or acceptable outcomes, independently of whether they are compatible with other expected or acceptable outcomes.

Goals as preferred outcomes For rule r alone the preferred outcome is b_1, and for rule s alone it is b'_1. But if both rules are applicable, then a state where both b_1 and b'_1 hold is not possible: the agent would not be rational if she considers both b_1 and $\neg b_1$ as her preferred outcomes. Hence, the agent has to decide if she prefers a state where b_1 holds to one where b'_1 (i.e., $\neg b_1$) holds, or *vice versa*. If the agent cannot make up her mind, i.e., she has no way to decide which is the most suitable option for her, then neither the chain of r nor that of s can produce preferred outcomes.

Suppose that the agent opts for the latter option; this can be done if the agent establishes that the second rule overrides the first one, i.e., $s > r$. Accordingly, the preferred outcome is b'_1 for the chain of outcomes defined by s, and b_2 is the preferred outcome of r. b_2 is the second best alternative according to rule r: in fact b_1 has been discarded as an acceptable outcome given that s prevails over r.

Two degrees of commitment: intentions and social intentions The next issue is to clarify which are the acceptable outcomes for an agent to commit to. Naturally, if the agent values some outcomes more than others, she should strive for the best, i.e., for the most preferred outcomes.

Let us start by considering the case where only rule r applies. Here, the agent should commit to the outcome she values the most, i.e., b_1. But what if the agent *believes* that b_1 cannot be achieved in the environment where she is currently situated in, or she knows that $\neg b_1$ holds? Committing to b_1 would result in a waste of agent's resources; rationally, she should target the next best outcome, in this case b_2. Suppose, now, that b_2 is *forbidden*, and the agent is social (an agent is social if the agent would not knowingly commit to anything that is forbidden [6]). Once again, in this situation the agent has to lower her expectation and settle for b_3, which is the next acceptable outcome.

To complete the analysis, consider the situation where both rules r and s apply and the agent prefers s to r. As we have seen before, $\neg b_1$ (b'_1) and b_2 are the preferred outcomes based on the preference of the agent over the two rules. Assume that, this time, the agent knows she cannot achieve $\neg b_1$ (or equivalently, b_1 holds). If the agent is rational, she cannot commit to $\neg b_1$. Thus, the best option for her is to commit to b'_2 and b_1, where she is guaranteed to be successful. In this scenario, the best course of action for the agent is where she commits herself to some outcomes that are not her preferred ones, or even that she would consider not acceptable based only on her preferences, but such that they influence her decision process given that they represent relevant external factors (either her beliefs or the norms that apply to her).

The layout of the paper is as follows. Section 2 presents the new logical framework. Section 3 describes the algorithms to prove that the logic has linear complexity. Section 4 ends the paper with some conclusions and a discussion of related work.

2 Logic

Defeasible Logic (DL) [11] is a simple but flexible and efficient rule based non-monotonic formalism. The strength of DL lays in the constructive proof theory, which has an argumentation-like structure and allows us to draw meaningful conclusions from (potentially) conflicting and incomplete knowledge bases. The framework provided by the proof theory accounts for the possibility of extensions of the logic, in particular extensions with modal operators. Several extensions have been proposed, which resulted in applications in the area of normative reasoning [12], modelling agents [6,13,7], and business process compliance [14], as well as efficient implementations of the logic (including the modal variants), able to handle very large knowledge bases [15,16,17].

2.1 Language

The main aim of this subsection is to establish an inference process to compute factual knowledge, desires, intentions, goals and obligations from existing facts, primitive desires, intentions, goals and unconditional obligations. As a first step, we introduce the language adopted. Let PROP be a set of propositional atoms, $\mathsf{MOD} = \{\mathsf{B}, \mathsf{O}, \mathsf{D}, \mathsf{G}, \mathsf{I}, \mathsf{SI}\}$ the set of modal operators[1] and Lbl be a set of arbitrary labels. The set $\text{Lit} = \text{PROP} \cup \{\neg p | p \in \text{PROP}\}$ denotes the set of *literals*. The *complementary* of a literal q is denoted by $\sim q$; if q is a positive literal p, then $\sim q$ is $\neg p$, and if q is a negative literal $\neg p$ then $\sim q$ is p. The set of *modal literals* is $\text{ModLit} = \{\Box l, \neg \Box l | l \in \text{Lit}, \Box \in \{\mathsf{O}, \mathsf{D}, \mathsf{G}, \mathsf{I}, \mathsf{SI}\}\}$. We assume that the "$\Box$" modal operator for belief B is the empty modal operator, thus a modal literal $\mathsf{B}l$ is equivalent to literal l. Accordingly, we state that the complementary of $\mathsf{B}{\sim}l$ as well as $\neg\mathsf{B}l$ is $\sim l$.

We define a *defeasible theory* D as a structure $(F, R, >)$, where (i.) F is a set of *facts* or indisputable statements, (ii.) R contains three sets of *rules*: for beliefs, obligations, and outcomes and (iii.) $> \subseteq R \times R$ is a *superiority relation* to determine the relative strength of conflicting rules. *Belief rules* are used to relate the factual knowledge of an agent (her vision of the environment), and defines the relationships between states of the world. As such, provability for beliefs does not generate modal literals. *Obligation rules* determine when and which obligations are in force. The conclusions generated by obligation rules are modalised with obligation. Finally, *outcome rules* establish the possible outcomes of an agent depending on the particular context. Apart from obligation rules, outcome rules are used to derive conclusions for all modes representing possible types of outcomes: desires, goals, intentions, and social intentions.

Following ideas given in [10], rules can gain more expressiveness when a *preference operator* $\odot$ is used: an expression like $a \odot b$ means that if a is possible, then a is the first choice and b is the second one; if $\neg a$ holds, then the first choice is not attainable and b is the actual choice. This operator is used to build chains of preferences, called $\odot$*-expressions*. The formation rules for $\odot$-expressions are: (i.) every literal is an $\odot$-expression, (ii.) if A is an $\odot$-expression and b is a literal then $A \odot b$ is an $\odot$-expression. In addition we stipulate that $\odot$ obeys to the following properties: (i.) $a \odot (b \odot c) = (a \odot b) \odot c$ (associativity); (ii.) $\bigodot_{i=1}^{n} a_i = (\bigodot_{i=1}^{k-1} a_i) \odot (\bigodot_{i=k+1}^{n} a_i)$ where exists j such that $a_j = a_k$ and $j < k$ (duplication and contraction on the right). $\odot$-expressions are given by the agent designer, or obtained through *construction rules* based on the particular logic [10].

In this paper we exploit the classical definition of *defeasible rule* in DL [11]. A defeasible rule is an expression $r : A(r) \Rightarrow_{\Box} C(r)$, where

[1] The reading of the modal operators is B for *belief*, O for *obligation*, D for *desire*, I for *intention* and SI for *social intention*.

1. $r \in \text{Lbl}$ is the name of the rule;
2. $A(r) = \{a_1, \dots, a_n\}$ with $a_i \in \text{Lit} \cup \text{ModLit}$ is the set of the premises (or the *antecedent*) of the rule;
3. $\Box \in \{\mathsf{B}, \mathsf{O}, \mathsf{U}\}$ represents the *mode* of the rule (from now on, we omit the subscript B in rules for beliefs, i.e., $\Rightarrow$ is used as a shortcut for $\Rightarrow_{\mathsf{B}}$);
4. $C(r)$ is the *consequent* (or *head*) of the rule, which is a single literal if $\Box = \mathsf{B}$, or an $\odot$-expression otherwise[2].

We use the following abbreviations on sets of rules: $R^{\Box}$ ($R^{\Box}[q]$) denotes all rules of mode $\Box$ (with consequent q), and $R[q] = \bigcup_{\Box \in \{\mathsf{B},\mathsf{O},\mathsf{U}\}} R^{\Box}[q]$. $R[q,i]$ denotes the set of rules whose head is $\odot_{j=1}^{n} c_j$ and $c_i = q$, with $1 \leq i \leq n$.

Most of the terminology defined so far appears in [6], where an extension of DL with modal operators is introduced to differentiate modal and factual rules. However, labelling the rules of DL produces nothing more but a simple treatment of the modalities, thus two interaction strategies between modal operators are analysed.

Rule conversions It is sometimes meaningful to use rules for a modality X as they were for another modality Y, i.e., to convert one type of conclusions into a different one. For example, if 'a car industry has the purpose of assembling perfectly working cars' and 'it is known that in every working car there is a working engine', then 'a car industry has also the purpose of assembling working engines in every car produced'. Formally, we define an asymmetric binary relation $\text{Convert} \subseteq \text{MOD} \times \text{MOD}$ such that $\text{Convert}(X,Y)$ means 'a rule of mode X can be used also to produce conclusions of mode Y'. This intuitively corresponds to the following logical schema:

$$\frac{Ya_1, \dots, Ya_n \quad a_1, \dots, a_n \Rightarrow_X b}{Yb} \ \text{Convert}(X,Y).$$

In our framework obligations and goal-like attitudes cannot change what the agent believes or how she perceives the world, thus we only consider conversion with mode for belief as the first element of the relation (i.e., $\text{Convert}(\mathsf{B}, X)$ with $X \in \{\mathsf{O}, \mathsf{D}, \mathsf{G}, \mathsf{I}, \mathsf{SI}\}$).

Conflict-detection/resolution It is crucial to identify criteria for detecting and solving conflicts between different modalities. Formally, we define an asymmetric binary relation $\text{Conflict} \subseteq \text{MOD} \times \text{MOD}$ such that $\text{Conflict}(X,Y)$ means 'modes X and Y are in conflict and mode X prevails over Y'. Consider the following theory:

$$F = \{\, sunny\_day,\ school\_day \},$$
$$R = \{\, r_1 : Sunny\_day \Rightarrow_{\mathsf{U}} go\_outside,\ r_2 : school\_day \Rightarrow_{\mathsf{O}} \neg go\_outside \}.$$

Even if there is a sunny day, a responsible parent would not go outside and play with her kid but will bring him to school; this behaviour is captured by $\text{Conflict}(\mathsf{O}, \mathsf{SI})$, which means that the rule that forbids to go outside prevents the agent from obtaining the (social) intention of going outside, and the parent will not derive the (social) intention.

In our framework, we consider conflicts between beliefs and intentions, beliefs and social intentions, and obligations and social intentions. In other words, we have:

[2] It is worth noting that modal literals can occur only in the antecedent of rules: the reason is that the rules are used to derive modal conclusions and we do not conceptually need to iterate modalities. The motivation of a single literal as a consequent for belief rules is dictated by the intended reading of the belief rules, where these rules are used to describe the environment.

– Conflict(B,I), Conflict(B,SI) meaning that the agents are realistic (cf. [9]), and
– Conflict(O,SI) meaning that the agents are social (cf. [6]).

Observation 1 Convert *and* Conflict *relations behave differently in our framework than the usual deployed in the literature [6]. Typically, there is a bijective correspondence between a* mode *and the* type of rule *"representing" it. For example, there are rules with mode* O *to derive obligations, or rules with mode* I *to derive intentions. This is not the case in our logic where outcome rules are used to derive conclusions for all goal-like attitudes. Thus, we can have* Conflict(O,SI) *exhibiting the sociality of the agent but not* Conflict(O,U) *since desires and obligation do not attack each other.*

There are two applications of the *superiority relation*: the first considers rules of the same mode; the latter compares rule of different mode. Given $r \in R^X$ and $s \in R^Y$, notice that $r > s$ iff r converts X into Y, or s converts Y into X, i.e., the superiority relation is used when rules, each with a different mode, are used to produce complementary conclusions of the same mode. Consider the following theory with Convert(B,G):

$$
\begin{aligned}
F = \{ & \, go\_to\_Rome, parent\_anniversary, August\},\\
R = \{ & \, r_1 : go\_to\_Rome \Rightarrow_{\mathsf{B}} go\_to\_Italy\\
& \, r_2 : parent\_anniversary \Rightarrow_{\mathsf{U}} go\_to\_Rome\\
& \, r_3 : August \Rightarrow_{\mathsf{U}} \neg go\_to\_Italy\},\\
>= \{ & \, (r_1, r_3)\}.
\end{aligned}
$$

Typically, I have the goal not to go to Italy in August since the weather is too hot and it is too crowded. However, it is my parents' anniversary and they are going to celebrate it this August in Rome, which is the capital of Italy. Nonetheless, I have the goal to go to Italy for my parents' wedding anniversary, since I am a good son. Here, the superiority applies because we use r_1 through a conversion from belief to goal.

2.2 Inferential Mechanism

A *proof* P of *length* n is a finite sequence $P(1),\dots,P(n)$ of *tagged literals* of the type $+\partial_X q$ and $-\partial_X q$, where $X \in$ MOD. The proof conditions below define the logical meaning of such tagged literals. As a conventional notation, $P(1..i)$ denotes the initial part of the sequence P of length i. Given a defeasible theory D, $+\partial_X q$ means that q is defeasibly provable in D with the mode X, and $-\partial_X q$ that it has been proved in D that q is not defeasibly provable in D with the mode X. As usual, we use $D \vdash \pm\partial_\Box l$ iff there is a proof P in D such that $P(n) = \pm\partial_\Box l$ for an index n.

In order to characterise the notions of provability for beliefs ($\pm\partial_{\mathsf{B}}$), obligations ($\pm\partial_{\mathsf{O}}$), desires ($\pm\partial_{\mathsf{D}}$), goals ($\pm\partial_{\mathsf{G}}$), intentions ($\pm\partial_{\mathsf{I}}$) and social intentions ($\pm\partial_{\mathsf{SI}}$), it is essential to define when a rule is *applicable* or *discarded*. To this end, the preliminary notion of when a rule is *body-applicable/discarded* must be introduced, stating that each literal in the body of the rule must be proved/rejected with the suitable mode.

Definition 1. *Let P be a proof and* $\Box \in \{\mathsf{O},\mathsf{D},\mathsf{G},\mathsf{I},\mathsf{SI}\}$. *A rule* $r \in R$ *is* body-applicable *(at step* $n+1$*) iff for all* $a_i \in A(r)$*:*

1. *if* $a_i = \Box l$ *then* $+\partial_\Box l \in P(1..n)$,

2. *if* $a_i = \neg\Box l$ *then* $-\partial_\Box l \in P(1..n)$*,*
3. *if* $a_i = l \in \mathrm{Lit}$ *then* $+\partial l \in P(1..n)$*.*

A rule $r \in R$ *is* body-discarded *(at step* $n+1$*) iff there is* $a_i \in A(r)$ *such that*

1. $a_i = \Box l$ *and* $-\partial_\Box l \in P(1..n)$*, or*
2. $a_i = \neg\Box l$ *and* $+\partial_\Box l \in P(1..n)$*, or*
3. $a_i = l \in \mathrm{Lit}$ *and* $-\partial l \in P(1..n)$*.*

As already stated, belief rules allow us to derive literals with different modes. The applicability mechanism must take into account this constraint.

Definition 2. *Let P be a proof. A rule* $r \in R$ *is 1.* Conv-applicable*, 2.* Conv-discarded *(at step* $n+1$*) for X iff*

1. $r \in R^{\mathsf{B}}$, $A(r) \neq \emptyset$ *and for all* $a \in A(r)$, $+\partial_X a \in P(1..n)$*;*
2. $r \notin R^{\mathsf{B}}$ *or* $A(r) = \emptyset$ *or* $\exists a \in A(r)$, $-\partial_X a \in P(1..n)$*.*

Let us consider the following theory

$$F = \{a, b, \mathsf{O}c\}, \quad R = \{r_1 : a \Rightarrow_{\mathsf{O}} b, r_2 : b, c \Rightarrow d\},$$

r_1 is applicable, while r_2 is not since c is not proved as a belief. Instead, r_2 is *Conv-applicable* in the condition for $\pm\partial_{\mathsf{O}}$, since $\mathsf{O}c$ is a fact and r_1 proves $\mathsf{O}b$.

The notion of applicability gives guidelines on how to consider the next element in a given chain. Since a rule for belief cannot generate reparative chains but only single literals, we can conclude that the applicability condition for belief collapses into body-applicability. The same happens to desires, where we also consider the Convert relation. For obligations, each element before the current one must be a violated obligation. A literal is a candidate to be a goal only if none of the previous elements in the chain have been proved as a goal. For intentions, the elements of the chain must pass the wishful thinking filter, while social intentions are also constrained not to violate any norm.

Definition 3. *Given a proof P,* $r \in R[q, i]$ *is* applicable *(at index i and step* $n+1$*) for*

1. B *iff* $r \in R^{\mathsf{B}}$ *and is body-applicable.*
2. O *iff either: (2.1.1)* $r \in R^{\mathsf{O}}$ *and is body-applicable, (2.1.2)* $\forall c_k \in C(r), k < i$, $+\partial_{\mathsf{O}} c_k \in P(1..n)$ *and* $-\partial c_k \in P(1..n)$*, or (2.2) r is Conv-applicable.*
3. D *iff either: (3.1)* $r \in R^{\mathsf{U}}$ *and is body-applicable, or (3.2) Conv-applicable.*
4. X*,* $X \in \{\mathsf{G}, \mathsf{I}, \mathsf{SI}\}$ *iff either: (4.1.1)* $r \in R^{\mathsf{U}}$ *and is body-applicable, (4.1.2)* $\forall c_k \in C(r), k < i$*,* $+\partial_Y {\sim} c_k \in P(1..n)$ *for some Y such that* $\mathrm{Conflict}(Y, X)$ *and* $-\partial_X c_k \in P(1..n)$ *or (4.2) r is Conv-applicable.*
   For G *there are no conflicts; for* I *we have* $\mathrm{Conflict}(\mathsf{B}, \mathsf{I})$*, and for* SI *we have* $\mathrm{Conflict}(\mathsf{B}, \mathsf{SI})$ *and* $\mathrm{Conflict}(\mathsf{O}, \mathsf{SI})$*.*

Conditions to establish that a rule is discarded correspond to the constructive failure to prove that the same rule is applicable, and follow the principle of *strong negation*.[3]

[3] The strong negation principle is closely related to the function that simplifies a formula by moving all negations to an inner most position in the resulting formula, and replaces the positive tags with the respective negative tags, and the other way around [18,7].

We can now describe the proof conditions for the various modal operators; we start with those for desires:

$+\partial_{\mathsf{D}}$: If $P(n+1) = +\partial_{\mathsf{D}} q$ then
(1) $\mathsf{D}q \in \mathrm{F}$ or
(2) (2.1) $\neg\mathsf{D}q \notin \mathrm{F}$ and
    (2.2) $\exists r \in R[q,i]$: r is applicable for D and
    (2.3) $\forall s \in R[\sim q, j]$ either
        (2.3.1) s is discarded for D, or
        (2.3.2) $s \not> r$.

We say that a *desire* is each element in a chain of an outcome rule for which there is no stronger argument for the opposite desire. The proof conditions for $+\partial_X$, with $X \in \{\mathsf{B},\mathsf{O},\mathsf{G},\mathsf{I},\mathsf{SI}\}$ are as follows:

$+\partial_X$: If $P(n+1) = +\partial_X q$ then
(1) $Xq \in \mathrm{F}$ or
(2) (2.1) $\neg Yq \notin \mathrm{F}$ for $Y = X$ or Convert(Y,X) and
    (2.2) $\exists r \in R[q,i]$: r is applicable for X and
    (2.3) $\forall s \in R^Y[\sim q, j]$ either
        (2.3.1) s is discarded for Y, or
        (2.3.2) $\exists t \in R^T[q,k]$: t is applicable for T and either
            (2.3.2.1) $t > s$ if $Y = T$, Convert(Y,T), or Convert(T,Y); or
            (2.3.2.2) Conflict(T,Y).

To show that a literal q is defeasibly provable with modality X we have two choices: (1) modal literal Xq is a fact; or (2) we need to argue using the defeasible part of D. In this case, we require that a complementary literal (of the same modality, or of a conflictual modality) does not appear in the set of facts (2.1), and that there must be an applicable rule for q for mode X (2.2). Moreover, each possible attack brought by a rule s for $\sim q$ has to be either discarded (3.1), or successfully counterattacked by another stronger rule t for q (2.3.2). We recall that the superiority relation combines rules of the same mode, rules with different modes that produce complementary conclusion of the same mode through conversion (both considered in clause (2.3.2.1)), and conflictual modalities (clause 2.3.2.2). Obviously, if $\Box = \mathsf{B}$, then the proof conditions reduce to those of classical defeasible logic [11].

Again, the negative counterparts ($-\partial_{\mathsf{D}}$ and $-\partial_X$) are derived by strong negation applied to conditions for $+\partial_{\mathsf{D}}$ and $+\partial_X$, respectively. As an example, consider the theory:

$$F = \{\neg b_1, \mathsf{O}\neg b_2, \mathsf{SI} b_4\} \quad R = \{r : \Rightarrow_{\mathsf{U}} b_1 \odot b_2 \odot b_3 \odot b_4\}.$$

Then r is trivially applicable for D and $+\partial_{\mathsf{D}} b_i$ holds, for $1 \leq i \leq 4$. Moreover, we have $+\partial_{\mathsf{G}} b_1$ and r is discarded for G after b_1. Since $+\partial\neg b_1$, $-\partial_{\mathsf{I}} b_1$ holds (as well as $-\partial_{\mathsf{SI}} b_1$); the rule is applicable for I and b_2, and we are able to prove $+\partial_{\mathsf{I}} b_2$, thus the rule becomes discarded for I after b_2. Given that $\mathsf{O}\neg b_2$ is a fact, r is discarded for SI and b_2 and $-\partial_{\mathsf{SI}} b_2$ is proved, which in turn makes the rule applicable for SI at b_3, proving $+\partial_{\mathsf{SI}} b_3$. As we have argued before, this would make the rule discarded for b_4. Nevertheless, b_4 is still provable with mode SI (in this case because it is a fact, but in other theories there could be more rules with b_4 in their head).

The logic resulting enjoys properties describing the appropriate behaviour of the modal operators.

Definition 4. *A defeasible theory $D = (\mathrm{F}, R, >)$ is* consistent *iff $>$ is acyclic and* F *does not contain pairs of complementary (modal) literals, that is pairs like (i.) l and $\sim l$, (ii.) $\Box l$ and $\neg \Box l$, $\Box \in$ MOD, and (iii.) $\Box l$ and $\Box \sim l$, $\Box \in \mathrm{MOD} \setminus \{\mathsf{D}\}$.*

Proposition 1 *Let D be a consistent modal defeasible theory. For any literal l, it is not possible to have both*

1. *$D \vdash +\partial_{\Box} l$ and $D \vdash -\partial_{\Box} l$ with $\Box \in$ MOD;*
2. *$D \vdash +\partial_{\Box} l$ and $D \vdash +\partial_{\Box} \sim l$ with $\Box \in \mathrm{MOD} \setminus \{\mathsf{D}\}$.*

Moreover, given $\Box \in \mathrm{MOD} \setminus \{\mathsf{D}\}$, then:

3. *if $D \vdash +\partial_{\Box} l$, then $D \vdash -\partial_{\Box} \sim l$.*

Proof. *Omissis.*

3 Algorithmic Results

We now present the algorithms apt to compute the *extension* of a *finite* defeasible theory, i.e., with finite set of facts and rules, in order to bind the complexity of the logic introduced in the previous sections. The algorithms are inspired by ideas of [19,20].

For the sake of clarity, from now on $\blacksquare$ denotes a generic mode in MOD, $\Diamond$ a generic mode in $\mathrm{MOD} \setminus \{\mathsf{B}\}$, and $\Box$ a fixed mode chosen in $\blacksquare$. Moreover, we will treat literals $\Box l$ and l as synonyms whenever $\Box = \mathsf{B}$. To accommodate the Convert relation to the algorithms, we denote with $R^{\mathsf{B},\Diamond}$ the set of belief rules with non-empty body that can be used for a conversion to mode $\Diamond$. Furthermore, for each literal l, $l_{\blacksquare}$ is the set (initially empty) such that $\pm\Box \in l_{\blacksquare}$ iff $D \vdash \pm\partial_{\Box} l$. Given a modal defeasible theory D, a set of rules R, and a rule $r \in R^{\Box}[l]$, we expand $>$ by incorporating the Conflict relation to ease the computation. Then, we define: (i.) $r_{sup} = \{s \in R : (s,r) \in >\}$ and $r_{inf} = \{s \in R : (r,s) \in >\}$ for any $r \in R$; (ii.) HB_D as the set of literals such that the literal or its complement appears in D, where 'appears' means that it is a sub-formula of a modal literal occurring in D; (iii.) the modal Herbrand Base of D as $HB = \{\Box l | \ \Box \in \mathrm{MOD}, l \in HB_D\}$. Accordingly, the extension of a defeasible theory is defined as follows.

Definition 5. *Given a modal defeasible theory D, the* defeasible extension *of D is defined as $E(D) = (+\partial_{\Box}, -\partial_{\Box})$ where $\pm\partial_{\Box} = \{l \in HB_D : D \vdash \pm\partial_{\Box} l\}$ with $\Box \in$ MOD. Two defeasible theories D and D' are* equivalent *whenever $E(D) = E(D')$.*

The next definition extends the concept of complement presented in Section 2 for modal literals and establishes the logical connection among proved and refuted literals.

Definition 6. *The* complement *of a given modal literal l, denoted by $\tilde{l}$, is:*

1. *if $l = \mathsf{D}m$, then $\tilde{l} = \{\neg \mathsf{D}m\}$;*
2. *if $l = \Box m$, then $\tilde{l} = \{\neg \Box m, \Box \sim m\}$, with $\Box \in \{\mathsf{O}, \mathsf{G}, \mathsf{I}, \mathsf{SI}\}$;*
3. *if $l = \neg \Box m$, then $\tilde{l} = \{\Box m\}$.*

Truncation and *removal* are two syntactical operations on the consequent of rules.

Definition 7. *Let* $c_1 = a_1 \odot \cdots \odot a_{i-1}$ *and* $c_2 = a_{i+1} \odot \cdots \odot a_n$ *be two (possibly empty)* $\odot$*-expressions such that* a_i *does not occur in them, and* $c = c_1 \odot a_i \odot c_2$ *is an* $\odot$*-expression. Let* r *be a rule with form* $A(r) \Rightarrow_X c$. *We define the*

- truncation *of the consequent* c *at* a_i *as* $A(r) \Rightarrow_X c!a_i = A(r) \Rightarrow_X c_1 \odot a_i$;
- removal *of* a_i *from the consequent* c *as* $A(r) \Rightarrow_X c \ominus a_i = A(r) \Rightarrow_X c_1 \odot c_2$.

Given $\Box \in$ MOD, the sets $\pm\partial_\Box$ denote the global sets of defeasible conclusions (i.e., the set of literals for which condition $\pm\partial_\Box$ holds), while $\partial_\Box^\pm$ are the corresponding temporary sets. Moreover, to simplify the calculus we do not operate on outcome rules: for each rule $r \in R^{\mathsf{U}}$ we create instead a new rule for all the other goal-like modes (resp. r^{D}, r^{G}, r^{I}, and r^{SI}). Consequently, we will use expressions like “the intention rule” as a shortcut for “the clone of outcome rule used to derive intentions”.

The idea of all algorithms is to use the operations of truncation and elimination in order to obtain, step after step, a simpler but equivalent theory. Indeed, proving a literal does not give just local information about the element itself, but reveals which rules will be applicable, discarded, or reduced in their head or tail.

Observation 2 *Assume that, at a given step, the algorithm proves* l. *At the next step,*

1. *the applicability of any rule* r *with* l *in its antecedent* $A(r)$ *does not depend on* l *any longer. Accordingly, we can safely remove* l *from* $A(r)$.
2. *Any rule* s *where* $\widetilde{l}$ *is in its antecedent* $A(s)$ *is discarded. Consequently, any superiority tuple involving this rule is now meaningless and can be removed from the superiority relation as well.*
3. *We can shorten chains by exploiting conditions of Definition 3. For example, if* $l = \mathsf{O}m$, *we can truncate chains for obligations at* $\sim m$ *and eliminate* $\sim m$.

Algorithm 1 DEFEASIBLEEXTENSION is the core algorithm to compute the extension of a defeasible theory. The first part (lines 1–4) sets up the data structure needed for the computation. Lines 5–8 handle facts as immediately provable literals. The main idea of the algorithm is to check whether there are rules whose body is empty. Since defeasible rules can have $\odot$-expressions as their head, the literal we are interested in is the first element of the $\odot$-expression (loop **for** at lines 16–33 and **if** condition at line 17). Such rules are clearly applicable and they can produce conclusions with the right modality. However, before asserting that the first element of the conclusion is provable, we have to check whether there are no rules for the complement (again with the appropriate mode), otherwise such rules for the complement must be weaker than the applicable rules. This information is stored in R_{infd} inspired by the technique of [20]. If no rule stronger than the current one exists, the complementary conclusion must be refuted by condition (2.3) of $-\partial_\Box$ (line 25). A straightforward consequence of $D \vdash -\partial_\Box l$ is that literal l is also refutable in D with any modality conflicting with $\Box$ (line 26). Notice that this reasoning does not hold for desires: since we can have $\mathsf{D}l$ and $\mathsf{D}{\sim}l$ at the same time, when $\Box = \mathsf{D}$ the algorithm invokes procedure 2 PROVED (line 23).

The next step is to check whether there exist rules for the complement of the literal with the same (or conflicting) mode. The rules for the complement should not be defeated by an applicable rule, i.e., they should not be in R_{infd}. If all these rules are defeated by r (line 27), then conditions for deriving $+\partial_\Box$ are satisfied. If a literal is

Algorithm 1 DEFEASIBLEEXTENSION

$+\partial_\blacksquare, \partial^+_\blacksquare \leftarrow \emptyset$; $-\partial_\blacksquare, \partial^-_\blacksquare \leftarrow \emptyset$
$R \leftarrow R \cup \{r^\Box : A(r) \Rightarrow_\Box C(r) | r \in R^{\mathsf{U}}\} \setminus R^{\mathsf{U}}$, with $\Box \in \{\mathsf{D}, \mathsf{G}, \mathsf{I}, \mathsf{SI}\}$
$R^{\mathsf{B},\Diamond} \leftarrow \{r^\Diamond : \Diamond a_1, \dots, \Diamond a_n \Rightarrow_\Diamond C(r) | r \in R^{\mathsf{B}}, A(r) \neq \emptyset, A(r) \subseteq \text{Lit}, a_i \in A(r)\}$
$> \leftarrow > \cup \{(r^\Diamond, s^\Diamond) | r^\Diamond, s^\Diamond \in R^{\mathsf{B},\Diamond}, r > s\} \cup \{(r,s) | r \in R^\blacksquare, s \in R^\Diamond \cup R^{\mathsf{B},\Diamond}, \text{Conflict}(\blacksquare, \Diamond)\}$
for $l \in F$ **do**
    if $l = \Box m$ **then** PROVED(m, $\Box$)
    if $l = \neg\Box m \land \Box \neq \mathsf{D}$ **then** REFUTED(m, $\Box$)
end for
$+\partial_\blacksquare \leftarrow +\partial_\blacksquare \cup \partial^+_\blacksquare$; $-\partial_\blacksquare \leftarrow -\partial_\blacksquare \cup \partial^-_\blacksquare$
$R_{infd} \leftarrow \emptyset$
repeat
    $\partial^+_\blacksquare \leftarrow \emptyset$; $\partial^-_\blacksquare \leftarrow \emptyset$
    for $\Box l \in HB$ **do**
        if $R^\Box[l] \cup R^{\mathsf{B},\Box}[l] = \emptyset$ **then** REFUTED(l, $\Box$)
    end for
    for $r \in R^\Box \cup R^{\mathsf{B},\Box}$ **do**
        if $A(r) = \emptyset$ **then**
            $r_{inf} \leftarrow \{r \in R : (r,s) \in >, s \in R\}$; $r_{sup} \leftarrow \{s \in R : (s,r) \in >\}$
            $R_{infd} \leftarrow R_{infd} \cup r_{inf}$
            Let l be the first literal of $C(r)$ in HB
            if $r_{sup} = \emptyset$ **then**
                if $\Box = \mathsf{D}$ **then**
                    PROVED(m, D)
                else
                    REFUTED($\sim l$, $\Box$)
                    REFUTED($\sim l$, $\Diamond$) for $\Diamond$ s.t. Conflict($\Box$, $\Diamond$)
                    if $R^\Box[\sim l] \cup R^{\mathsf{B},\Box}[\sim l] \cup R^\blacksquare[\sim l] \setminus R_{infd} \subseteq r_{inf}$, for $\blacksquare$ s.t. Conflict($\blacksquare$, $\Box$) **then**
                        PROVED(m, $\Box$)
                    end if
                end if
            end if
        end if
    end for
    $\partial^+_\blacksquare \leftarrow \partial^+_\blacksquare \setminus +\partial_\blacksquare$; $\partial^-_\blacksquare \leftarrow \partial^-_\blacksquare \setminus -\partial_\blacksquare$
    $+\partial_\blacksquare \leftarrow +\partial_\blacksquare \cup \partial^+_\blacksquare$; $-\partial_\blacksquare \leftarrow -\partial_\blacksquare \cup \partial^-_\blacksquare$
until $\partial^+_\blacksquare = \emptyset$ and $\partial^-_\blacksquare = \emptyset$
return $(+\partial_\blacksquare, -\partial_\blacksquare)$

assessed to be provable (with the appropriate modality) the algorithm calls procedure 2 PROVED, otherwise the procedure 3 REFUTED is invoked. The algorithm finally returns the extension of the input theory when no modifications are done on sets $\partial^\pm_\blacksquare$.

Algorithm 2 PROVED is invoked when literal l is proved with modality $\Box$. The computation starts by updating the relative positive extension set for modality $\Box$ and the local information on literal l (line 2); $\Box l$ is then removed from HB at line 3. Proposition 1 Part *3.* defines the modalities literal $\sim l$ can be refuted with (**if** condition at line 4). Lines 5 to 7 modifies the sets of rules R and $R^{\mathsf{B},\Box}$, and the superiority relation accordingly to ideas of Observation 2.

Depending on the modality $\Box$ of l, we have to perform some specific operations on chains (condition **switch** at lines 8–27). Entering into the detail of each **case** would be redundant without giving more information than conditions of a rule being applicable or

Algorithm 2 PROVED

1: **procedure** PROVED($l \in \text{Lit}$, $\Box \in \text{MOD}$)
2: $\quad \partial^+_\Box \leftarrow \partial^+_\Box \cup \{l\}$; $l_\blacksquare \leftarrow l_\blacksquare \cup \{+\Box\}$
3: $\quad HB \leftarrow HB \setminus \{\Box l\}$
4: $\quad$ **if** $\Box \neq \mathsf{D}$ **then** REFUTED($\sim l$, $\Box$)
5: $\quad R \leftarrow \{r : A(r) \setminus \{\Box l, \neg\Box{\sim}l\} \hookrightarrow C(r) \mid r \in R,\ A(r) \cap \widetilde{\Box l} = \emptyset\}$
6: $\quad R^{\mathsf{B},\Box} \leftarrow \{r : A(r) \setminus \{\Box l\} \hookrightarrow C(r) \mid r \in R^{\mathsf{B},\Box},\ A(r) \cap \widetilde{\Box l} = \emptyset\}$
7: $\quad > \leftarrow > \setminus \{(r,s),(s,r) \in > \mid A(r) \cap \widetilde{\Box l} \neq \emptyset\}$
8: $\quad$ **switch** ($\Box$)
9: $\quad\quad$ **case** B**:**
10: $\quad\quad\quad R^X \leftarrow \{A(r) \Rightarrow_X C(r)!l \mid r \in R^X[l,n]\}$ with $X \in \{\mathsf{O},\mathsf{I}\}$
11: $\quad\quad\quad R^X \leftarrow \{A(r) \Rightarrow_X C(r) \ominus {\sim}l \mid r \in R^X[{\sim}l,n]\}$ with $X \in \{\mathsf{I},\mathsf{SI}\}$
12: $\quad\quad\quad$ **if** $+\mathsf{O} \in {\sim}l_\blacksquare$ **then** $R^{\mathsf{O}} \leftarrow \{A(r) \Rightarrow_{\mathsf{O}} C(r) \ominus {\sim}l \mid r \in R^{\mathsf{O}}[{\sim}l,n]\}$
13: $\quad\quad\quad$ **if** $-\mathsf{O} \in {\sim}l_\blacksquare$ **then** $R^{\mathsf{SI}} \leftarrow \{A(r) \Rightarrow_{\mathsf{SI}} C(r)!l \mid r \in R^{\mathsf{SI}}[l,n]\}$
14: $\quad\quad$ **case** O**:**
15: $\quad\quad\quad R^{\mathsf{O}} \leftarrow \{A(r) \Rightarrow_{\mathsf{O}} C(r)!{\sim}l \ominus {\sim}l \mid r \in R^{\mathsf{O}}[{\sim}l,n]\}$
16: $\quad\quad\quad R^{\mathsf{SI}} \leftarrow \{A(r) \Rightarrow_{\mathsf{SI}} C(r) \ominus {\sim}l \mid r \in R^{\mathsf{SI}}[{\sim}l,n]\}$
17: $\quad\quad\quad$ **if** $-\mathsf{B} \in l_\blacksquare$ **then** $R^{\mathsf{O}} \leftarrow \{A(r) \Rightarrow_{\mathsf{O}} C(r) \ominus l \mid r \in R^{\mathsf{O}}[l,n]\}$
18: $\quad\quad\quad$ **if** $-\mathsf{B} \in {\sim}l_\blacksquare$ **then** $R^{\mathsf{SI}} \leftarrow \{A(r) \Rightarrow_{\mathsf{SI}} C(r)!l \mid r \in R^{\mathsf{SI}}[l,n]\}$
19: $\quad\quad$ **case** D**:**
20: $\quad\quad\quad$ **if** $+\mathsf{D} \in {\sim}l_\blacksquare$ **then**
21: $\quad\quad\quad\quad R^{\mathsf{G}} \leftarrow \{A(r) \Rightarrow_{\mathsf{G}} C(r)!l \ominus l \mid r \in R^{\mathsf{G}}[l,n]\}$
22: $\quad\quad\quad\quad R^{\mathsf{G}} \leftarrow \{A(r) \Rightarrow_{\mathsf{G}} C(r)!{\sim}l \ominus {\sim}l \mid r \in R^{\mathsf{G}}[{\sim}l,n]\}$
23: $\quad\quad\quad$ **end if**
24: $\quad\quad$ **otherwise:**
25: $\quad\quad\quad R^{\Box} \leftarrow \{A(r) \Rightarrow_{\Box} C(r)!l \mid r \in R^{\Box}[l,n]\}$
26: $\quad\quad\quad R^{\Box} \leftarrow \{A(r) \Rightarrow_{\Box} C(r) \ominus {\sim}l \mid r \in R^{\Box}[{\sim}l,n]\}$
27: $\quad$ **end switch**
28: **end procedure**

discarded in Section 2. Therefore, we propose one significative example by considering the scenario where l has been proved as a belief (**case** at lines 9–13). Here, chains of obligation (resp. intention) rules can be truncated after l since are discarded for all following elements (line 10). Analogously, condition (4.1.2) of Definition 3 allows us to eliminate $\sim l$ from intention and social intention rules (line 11). If $+\partial_{\mathsf{O}}{\sim}l$ has been already proved, then we eliminate $\sim l$ since it represents a violated obligation. *Vice versa*, if $-\partial_{\mathsf{O}}{\sim}l$ is the case, then each element after l cannot be a social intention (resp. **if** conditions at lines 12 and 13).

Algorithm 3 REFUTED performs all necessary operations in case literal l is refuted with modality $\Box$. The initialisation steps at lines 2–6 follow the same schema exploited at lines 2–7 of Algorithm 2 PROVED. Again, the operations to be performed on chains vary according to the current mode $\Box$ (**switch** at lines 7–19). For example, if $\Box = \mathsf{B}$ (lines 8–11), then applicability condition (4.1.2) for $\pm\partial_{\mathsf{I}}$ cannot be satisfied for any literal after $\sim l$ in chains for intentions, and such chains can be truncated at $\sim l$. Furthermore, if the algorithm has already proven $+\partial_{\mathsf{O}} l$, then l represents a violated obligation. Thus, l can be removed from all chains for obligations. If instead $-\partial_{\mathsf{O}} l$ holds, then the elements after $\sim l$ in chains for social intentions do not satisfy applicability condition (4.1.2) of $\pm\partial_{\mathsf{SI}}$ for $\sim l$, and the algorithm removes them.

We conclude by showing the computational properties of the algorithms proposed.

Algorithm 3 REFUTED

procedure REFUTED($l \in \text{Lit}, \Box \in \text{MOD}$)
  $\partial_{\Box}^{-} \leftarrow \partial_{\Box}^{-} \cup \{l\}$; $l_{\blacksquare} \leftarrow l_{\blacksquare} \cup \{-\Box\}$
  $HB \leftarrow HB \setminus \{\Box l\}$
  $R \leftarrow \{r : A(r) \setminus \{\neg\Box l\} \hookrightarrow C(r) | \; r \in R, \; \Box l \notin A(r)\}$
  $R^{\mathsf{B},\Box} \leftarrow R^{\mathsf{B},\Box} \setminus \{r \in R^{\mathsf{B},\Box} : \Box l \in A(r)\}$
  $> \leftarrow > \setminus \{(r,s),(s,r) \in > | \Box l \in A(r)\}$
  switch ($\Box$)
    case B**:**
      $R^{\mathsf{I}} \leftarrow \{A(r) \Rightarrow_{\mathsf{I}} C(r)!{\sim}l | r \in R^{\mathsf{I}}[{\sim}l, n]\}$
      if $+\mathsf{O} \in l_{\blacksquare}$ **then** $R^{\mathsf{O}} \leftarrow \{A(r) \Rightarrow_{\mathsf{O}} C(r) \ominus l | r \in R^{\mathsf{O}}[l, n]\}$
      if $-\mathsf{O} \in l_{\blacksquare}$ **then** $R^{\mathsf{SI}} \leftarrow \{A(r) \Rightarrow_{\mathsf{SI}} C(r)!{\sim}l | r \in R^{\mathsf{SI}}[{\sim}l, n]\}$
    case O**:**
      $R^{\mathsf{O}} \leftarrow \{A(r) \Rightarrow_{\mathsf{O}} C(r)!l \ominus l | r \in R^{\mathsf{O}}[l, n]\}$
      if $-\mathsf{B} \in l_{\blacksquare}$ **then** $R^{\mathsf{SI}} \leftarrow \{A(r) \Rightarrow_{\mathsf{SI}} C(r)!{\sim}l | r \in R^{\mathsf{SI}}[{\sim}l, n]\}$
    case D**:**
      $R^{X} \leftarrow \{A(r) \Rightarrow_{X} C(r) \ominus l | r \in R^{X}[l, n]\}$ with $X \in \{\mathsf{D}, \mathsf{G}\}$
    otherwise:
      $R^{\Box} \leftarrow \{A(r) \Rightarrow_{\Box} C(r) \ominus l | r \in R^{\Box}[l, n]\}$
  end switch
end procedure

Theorem 1. *Algorithm 1* DEFEASIBLEEXTENSION *terminates and its computational complexity is* $O(|R| * |HB|)$.

Proof Sketch. Termination is ensured since at every iteration either no modification occurs and line 36 ends the computation, or a literal is removed from HB. This set is finite, since the sets of facts and rules are finite, thus the process eventually empties HB. This bounds also the complexity to the number of rules and literals in HB, since each modal literal is processed once, and every time we scan the set of rules.

Theorem 2. *Algorithm 1* DEFEASIBLEEXTENSION *is sound and complete.*

Proof *Omissis.* A similar result can be found in [21].

4 Conclusions and Related Work

This article provides a fresh characterisation for motivational states as the concepts of desire, goal, intention, and social intention obtained through a deliberative process based on various types of preferences among desired outcomes. In this sense, this contribution has strong connections with [5,6,7] but presents significant improvements in at least two respects. First, while in those works the agent deliberation is simply the result of the derivation of mental states from *precisely* the corresponding rules of the logic, here the proof theory is much more aligned with the BDI intuition, according to which intentions and goals are the results of desire manipulation. This allowed us to encode this idea within a logical language and a proof theory, by exploiting the different interaction patterns between the basic mental states, as well as the derived ones. In this perspective, our framework is significantly more expressive than the one in BOID [9], which uses different rules to derive the corresponding mental states and proposes simple criteria to solve conflicts between rule types.

Second, the framework proposes a rich language expressing two orthogonal concepts of preference among motivational attitudes. One is encoded within $\odot$ sequences, which state reparative orders among homogeneous mental states or motivations, and which are contextual. The second type of preference is encoded via the superiority relation between rules which can work locally, as well as via the Conflict relation. The interplay between these two preference mechanisms can help us to isolate different and complex ways for deriving mental states, but the resulting logical machinery is still computationally tractable.

Since the preferences allow us to determine what preferred outcomes can be chosen by an agent (in a specific scenario) when previous goals in $\odot$-sequences are not (or no longer) feasible, our logic in fact provides an abstract semantics for several types of goal and intention reconsideration. Intention reconsideration was expected to play a crucial role in the BDI paradigm [22,2] since intentions obey the law of inertia and resist retraction or revision, but they can be reconsidered when new relevant information comes in [22]. Despite that, the problem of revising intentions in BDI frameworks has received little attention. A very sophisticated exception is [23], where revisiting intentions mainly depends on the dynamics of beliefs but the process is incorporated in a very complex framework for reasoning about mental states. Recently, [24] discussed how to revise the commitments to planned activities because of mutually conflicting intentions, which interestingly has connections with our work. How to employ our logic to give a semantics for intention reconsideration is not the main goal of the paper and is left to future work.

Finally, we recall that our focus was on the deliberation aspects of an agent determining what are her mental states in a given moment. That is the case, our investigation is orthogonal with respect to the works of [25,26,27].

Acknowledgements

NICTA is funded by the Australian Government as represented by the Department of Broadband, Communications and the Digital Economy and the Australian Research Council through the ICT Centre of Excellence program.

References

1. Bratman, M.E., Israel, D.J., Pollack, M.E.: Plans and resource-bounded practical reasoning. Computational Intelligence **4** (1988) 349–355
2. Cohen, P.R., Levesque, H.J.: Intention is choice with commitment. Artificial Intelligence **42**(2-3) (1990) 213–261
3. Rao, A.S., Georgeff, M.P.: Modeling rational agents within a BDI-architecture. In Allen, J.F., Fikes, R., Sandewall, E., eds.: KR, Kaufmann, M. (1991) 473–484
4. Rao, A.S., Georgeff, M.P.: Decision procedures for bdi logics. Journal of Logic and Computation **8**(3) (1998) 293–342
5. Dastani, M., Governatori, G., Rotolo, A., van der Torre, L.: Programming cognitive agents in defeasible logic. In Sutcliffe, G., Voronkov, A., eds.: LPAR 2005. Volume 3835 of LNAI., Springer (2005) 621–636

6. Governatori, G., Rotolo, A.: BIO logical agents: Norms, beliefs, intentions in defeasible logic. Journal of Autonomous Agents and Multi-Agent Systems **17**(1) (2008) 36–69
7. Governatori, G., Padmanabhan, V., Rotolo, A., Sattar, A.: A defeasible logic for modelling policy-based intentions and motivational attitudes. Logic Journal of the IGPL **17**(3) (2009) 227–265
8. Thomason, R.H.: Desires and defaults: A framework for planning with inferred goals. In Cohn, A.G., Giunchiglia, F., Selman, B., eds.: KR2000, Morgan Kaufmann (2000)
9. Broersen, J., Dastani, M., Hulstijn, J., van der Torre, L.: Goal generation in the BOID architecture. Cognitive Science Quarterly **2**(3-4) (2002) 428–447
10. Governatori, G., Rotolo, A.: Logic of violations: A Gentzen system for reasoning with contrary-to-duty obligations. Australasian Journal of Logic **4** (2006) 193–215
11. Antoniou, G., Billington, D., Governatori, G., Maher, M.J.: Representation results for defeasible logic. ACM Transactions on Computational Logic **2**(2) (2001) 255–287
12. Governatori, G.: Representing business contracts in RuleML. International Journal of Cooperative Information Systems **14**(2-3) (2005) 181–216
13. Kravari, K., Papatheodorou, C., Antoniou, G., Bassiliades, N.: Reasoning and proofing services for semantic web agents. In Walsh, T., ed.: IJCAI 2011. (2011) 2662–2667
14. Governatori, G., Sadiq, S.: The journey to business process compliance. In: Handbook of Research on BPM. IGI Global (2008) 426–454
15. Lam, H.P., Governatori, G.: The making of SPINdle. In Governatori, G., Hall, J., Paschke, A., eds.: RuleML 2009. Number 5858 in LNCS, Springer (2009) 315–322
16. Bassiliades, N., Antoniou, G., Vlahavas, I.: A defeasible logic reasoner for the semantic web. Int. J. Semantic Web Inf. Syst. **2**(1) (2006) 1–41
17. Tachmazidis, I., Antoniou, G., Flouris, G., Kotoulas, S., McCluskey, L.: Large-scale parallel stratified defeasible reasoning. In De Raedt, L., Bessière, C., Dubois, D., Doherty, P., Frasconi, P., Heintz, F., Lucas, P.J.F., eds.: ECAI 2012, IOS Press (2012) 738–743
18. Antoniou, G., Billington, D., Governatori, G., Maher, M.J., Rock, A.: A family of defeasible reasoning logics and its implementation. In: ECAI 2000. (2000) 459–463
19. Maher, M.J.: Propositional defeasible logic has linear complexity. Theory and Practice of Logic Programming **1**(6) (2001) 691–711
20. Lam, H.P., Governatori, G.: What are the Necessity Rules in Defeasible Reasoning? In Delgrande, J., Faber, W., eds.: LPNMR-11. Springer (2011) 187–192
21. Governatori, G., Olivieri, F., Rotolo, A., Scannapieco, S.: Computing strong and weak permissions in defeasible logic. CoRR **abs/1212.0079** (2012)
22. Bratman, M.E.: Intentions, Plans and Practical Reason. Harvard University Press (1987)
23. van der Hoek, W., Jamroga, W., Wooldridge, M.: Towards a theory of intention revision. Synthese **155**(2) (2007) 265–90
24. Shapiro, S., Sardina, S., Thangarajah, J., Cavedon, L., Padgham, L.: Revising conflicting intention sets in BDI agents. In: AAMAS 2012, IFAAMS (2012) 1081–1088
25. Winikoff, M., Padgham, L., Harland, J., Thangarajah, J.: Declarative & procedural goals in intelligent agent systems. In Fensel, D., Giunchiglia, F., McGuinness, D.L., Williams, M., eds.: KR-02, Morgan Kaufmann (2002) 470–481
26. Dastani, M., van Riemsdijk, M.B., Meyer, J.J.C.: Goal types in agent programming. In Nakashima, H., Wellman, M.P., Weiss, G., Stone, P., eds.: AAMAS, ACM (2006) 1285–1287
27. van Riemsdijk, M.B., Dastani, M., Meyer, J.J.C.: Goals in conflict: Semantic foundations of goals in agent programming. Journal of Autonomous Agents and Multi-Agent Systems **18**(3) (2009) 471–500