TLP **16** (3): 296–324, 2016. © Cambridge University Press 2016
doi:10.1017/S1471068416000053

The rationale behind the concept of goal

GUIDO GOVERNATORI
Data61, CSIRO, NICTA, Australia
(*e-mail:* guidogovernatori@data61.csiro.au)

FRANCESCO OLIVIERI, SIMONE SCANNAPIECO
Department of Computer Science, University of Verona, Verona, Italy
(*e-mail:* francesco.olivieri@univr.it)

ANTONINO ROTOLO
CIRSFID, University of Bologna, Bologna, Italy
(*e-mail:* antonino.rotolo@unibo.it)

MATTEO CRISTANI
e-mail matteo.cristani@univr.it

submitted 31 December 2014; revised 21 May 2015; accepted 14 December 2015

Abstract

The paper proposes a fresh look at the concept of goal and advances that motivational attitudes like desire, goal and intention are just facets of the broader notion of (acceptable) outcome. We propose to encode the preferences of an agent as sequences of "alternative acceptable outcomes". We then study how the agent's beliefs and norms can be used to filter the mental attitudes out of the sequences of alternative acceptable outcomes. Finally, we formalise such intuitions in a novel Modal Defeasible Logic and we prove that the resulting formalisation is computationally feasible.

KEYWORDS: Agents, Defeasible logic, Desires, intentions, Goals, Obligations

1 Introduction and motivation

The core problem we address in this paper is how to formally describe a system operating in an environment, with some objectives to achieve, and trying not to violate the norms governing the domain in which the system operates.

To model such systems, we have to specify three types of information: (i) the environment where the system is embedded, i.e., how the system perceives the world, (ii) the norms regulating the application domain, and (iii) the system's internal constraints and objectives.

A successful abstraction to represent a system operating in an environment where the system itself must exhibit some kind of autonomy is that of BDI (Belief, Desire, Intention) architecture (Rao and Georgeff 1991) inspired by the work of Bratman (1987) on cognitive agents. In the BDI architecture, desires and intentions model the agent's mental attitudes and are meant to capture the objectives, whereas beliefs describe the environment. More precisely, the notions of belief, desire and intention represent respectively the informational, motivational and deliberative states of an agent (Wooldridge and Jennings 1995).

Over the years, several frameworks, either providing extensions of BDI or inspired by it, were given with the aim of extending models for cognitive agents to also cover normative aspects (see, among others, (Broersen *et al.* 2002; Thomason 2000; Governatori and Rotolo 2008)). (This is a way of developing normative agent systems, where norms are meant to ensure global properties for them (Andrighetto *et al.* 2013).) In such extensions, the agent behaviour is determined by the interplay of the cognitive component and the normative one (such as obligations). In this way, it is possible to represent how much an agent is willing to invest to reach some outcomes based on the states of the world (what we call beliefs) and norms. Indeed, beliefs and norms are of the utmost importance in the decision process of the agent. If the agent does not take beliefs into account, then she will not be able to plan what she wants to achieve, and her planning process would be a mere wishful thinking. On the other hand, if the agent does not respect the norms governing the environment she acts in, then she may incur sanctions from other agents (Bratman 1987).

The BDI approach is based on the following assumptions about the motivational and deliberative components. The agent typically defines *a priori* her desires and intentions, and only after this is done the system verifies their mutual consistency by using additional axioms. Such entities are therefore not interrelated with one another since "the notion of intention [. . .] has equal status with the notions of belief and desire, and cannot be reduced to these concepts" (Rao and Georgeff 1991). Moreover, the agent may consequently have intentions which are contradictory with her beliefs and this may be verified only *a posteriori*. Therefore, one of the main conceptual deficiencies of the BDI paradigm (and generally of almost all classical approaches to model rational agents) is that the deliberation process is bound to these mental attitudes which are independent and fixed *a priori*. Here, with the term independent, we mean that none of them is fully definable in terms of the others.

Approaches like the BOID (Belief, Obligation, Intention, Desire) architecture (Broersen *et al.* 2002) and Governatori and Rotolo (2008)'s system improve previous frameworks, for instance, by structurally solving conflicts between beliefs and intentions (the former being always stronger than any conflicting intention), while mental attitudes and obligations are just meant to define which kinds of agent (social, realistic, selfish, and so on) are admissible.

Unlike the BDI perspective, this paper aims at proposing a fresh conceptual and logical analysis of the motivational and deliberative components within a unified perspective.

Desideratum 1: A unified framework for agents' motivational and deliberative components. Goals, desires, and intentions are *different facets* of the *same phenomenon*, all of them being goal-like attitudes. This reduction into a unified perspective is done by resorting to the basic notion of *outcome*, which is simply something (typically, a state of affairs) that an agent expects to achieve or that can possibly occur.

Even when considering the vast literature on goals of the past decade, most of the authors studied the content of a goal (e.g., *achievement* or *maintenance* goals) and conditions under which a goal has to be either pursued, or dropped. This kind of (*a posteriori*) analysis results orthogonal to the one proposed hereafter, since we want to develop a framework that computes the agent's mental attitudes by combining her beliefs and the norms with her desires.

As we shall argue, an advantage of the proposed analysis is that it allows agents to compute different degrees of motivational attitudes, as well as different degrees of commitment that take into account other, external, factors, such as *beliefs* and *norms*.

Desideratum 2: Agents' motivations emerge from preference orderings among outcomes. The motivational and deliberative components of agents are generated from preference orderings among outcomes. As done in other research areas (e.g., rational choice theory), we move with the idea that agents have preferences and choose the actions to bring about according to such preferences. Preferences involve outcomes and are explicitly represented in the syntax of the language for reasoning about agents, thus following the logical paradigm initially proposed in Brewka *et al.* (2004); Governatori and Rotolo (2006).

The combination of an agent's mental attitudes with the factuality of the world defines her deliberative process, i.e., the objectives she decides to pursue. The agent may give up some of them to comply with the norms, if required. Indeed, many contexts may prevent the agent from achieving all of her objectives; the agent must then understand which objectives are mutually compatible with each other and choose which ones to attain the least of in given situations by ranking them in a preference ordering.

The approach we are going to formalise can be summarised as follows. We distinguish three phases an agent must pass through to bring about certain states of affairs: (i) The agent first needs to understand the environment she acts in; (ii) The agent deploys such information to deliberate which objectives to pursue; and (iii) The agent lastly decides how to act to reach them.

In the first phase, the agent gives a formal declarative description of the environment (in our case, a rule-based formalism). Rules allow the agent to represent relationships between pre-conditions and actions, actions and their effects (post-conditions), relationships among actions, which conditions trigger new obligations to come in force, and in which contexts the agent is allowed to pursue new objectives.

In the second phase, the agent combines the formal description with an input describing a particular state of affairs of the environment, and she determines which norms are actually in force along with which objectives she decides to commit to (by understanding which ones are attainable) and to which degree. The agent's decision is based on logical derivations.

Since the agent's knowledge is represented by rules, during the third and last phase, the agent combines and exploits all such information obtained from the conclusions derived in the second phase to select which activities to carry out in order to achieve the objectives. (It is relevant to notice that a derivation can be understood as a virtual simulation of the various activities involved.)

While different schemas for generating and filtering agents' outcomes are possible, the three phases described above suggest to adopt the following principles:

- When an agent faces alternative outcomes in a given context, these outcomes are ranked in preference orderings;
- Mental attitudes are obtained from a single type of rule (*outcome rule*) whose conclusions express the above mentioned preference orderings among outcomes;
- Beliefs prevail over conflicting motivational attitudes, thus avoiding various cases of wishful thinking (Thomason 2000; Broersen *et al.* 2002);
- Norms and obligations are used to filter social motivational states (*social intentions*) and compliant agents (Broersen *et al.* 2002; Governatori and Rotolo 2008);
- Goal-like attitudes can also be derived via a *conversion* mechanism using other mental states, such as beliefs (Governatori and Rotolo 2008). For example, believing that Madrid is in Spain may imply that the goal to go to Madrid implies the goal to go to Spain.

Our effort is finally motivated by computational concerns. The logic for agents' desires, goals, and intentions is expected to be computationally efficient. In particular, we shall prove that computing agents' motivational and deliberative components in the proposed unified framework has linear complexity.

2 The intuition underneath the framework

When a cognitive agent deliberates about what her outcomes are in a particular situation, she selects a set of *preferred* outcomes among a larger set, where each specific outcome has various alternatives. It is natural to rank such alternatives in a preference ordering, from the most preferred choice to the least objective she deems acceptable.

Consider, for instance, the following scenario. Alice is thinking what to do on Saturday afternoon. She has three alternatives: (i) she can visit John; (ii) she can visit her parents who live close to John's place; or (iii) she can watch a movie at home. The alternative she likes the most is visiting John, while watching a movie is the least preferred. If John is not at home, there is no point for Alice to visit him. In this case, paying a visit to her parents becomes the "next best" option. Also, if visiting her parents is not possible, she settles for the last choice, that of staying home and watching a movie.

Alice also knows that if John is away, the alternative of going to his place makes no sense. Suppose that, Alice knows that John is actually away for the weekend. Since the most preferred option is no longer available, she decides to opt for the now best option, namely visiting her parents.

To represent the scenario above, we need to capture the preferences about her alternatives, and her beliefs about the world. To model preferences among several options, we build a sequence of alternatives $A_1, \dots, A_n$ that are preferred when the previous choices are no longer feasible. Normally, each set of alternatives is the result of a specific context C determining under which conditions (premises) such a sequence of alternatives $A_1, \dots, A_n$ is considered.

Accordingly, we can represent Alice's alternatives with the notation

$$\textbf{If} \quad \textit{saturday}, \quad \textbf{then} \quad \textit{visit\_John},\ \textit{visit\_parents},\ \textit{watch\_movie}.$$

This intuition resembles the notion of contrary-to-duty obligations presented by Governatori and Rotolo (2006), where a norm is represented by an *obligation rule* of the type

$$r_1 : \textit{drive\_car} \Rightarrow_{\mathsf{OBL}} \neg \textit{damage} \odot \textit{compensate} \odot \textit{foreclosure}$$

where "$\Rightarrow_{\mathsf{OBL}}$" denotes that the conclusion of the rule will be treated as an obligation, and the symbol "$\odot$" replaces the symbol "," to separate the alternatives. In this case, each element of the chain is the reparative obligation that shall come in force in case the immediate predecessor in the chain has been violated. Thus, the meaning of rule r_1 is that, if an agent drives a car, then 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.

Following this perspective, we shall now represent the previous scenario with a rule introducing the outcome mode, that is an *outcome rule*:

$$r_2 : \textit{saturday} \Rightarrow_{\mathsf{OUT}} \textit{visit\_John} \odot \textit{visit\_parents} \odot \textit{watch\_movie}.$$

In both examples, the sequences express a preference ordering among alternatives. Accordingly, *watch_movie* and *foreclosure* are the last (and least) acceptable situations.

To model beliefs, we use *belief rules*, like

$$r_3 : \textit{John\_away} \Rightarrow_{\mathsf{BEL}} \neg \textit{visit\_John}$$

meaning that if Alice has the belief that John is not home, then she adds to her beliefs that it is not possible to visit him.

In the rest of the section, we shall illustrate the principles and intuitions relating sequences of alternatives (that is, outcome rules), beliefs, obligations, and how to use them to characterise different types of goal-like attitudes and degrees of commitment to outcomes: *desires*, *goals*, *intentions*, and *social intentions*.

2.1 Desires as acceptable outcomes

Suppose that an agent is equipped with the following outcome rules expressing two preference orderings:

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

and that the situations 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 all *acceptable outcomes*, including the incompatible outcomes b_1 and b'_1.

Desires are acceptable outcomes, independently of whether they are compatible with other expected or acceptable outcomes. Let us contextualise the previous example to better explain the notion of desire by considering the following setting.

Example 1

$F = \{saturday, John\_sick\}$ $\quad R = \{r_2, r_4 : John\_sick \Rightarrow_{\mathsf{OUT}} \neg visit\_John \odot short\_visit\}.$

The meaning of r_4 is that Alice would not visit John if he is sick, but if she does so, then the visit must be short.

Being the premises of r_2 and of r_4 the case, then both rules are activated, and the agent has both *visit_John* and its opposite as acceptable outcomes. Eventually, she needs to make up her mind. Notice that if a rule prevails over the other, then the elements of the weaker rule with an incompatible counterpart in the stronger rule are *not* considered desires. Suppose that, Alice has not visited John for a long time and she has recently placed a visit to her parents. Then, she prefers to see John instead of her parents despite John being sick. In this setting, r_2 prevails over r_4 ($r_2 > r_4$ in notation). Given that she explicitly prefers r_2 to r_4, her desire is to visit John (*visit_John*) and it would be irrational to conclude that she also has the opposite desire (i.e., $\neg visit\_John$).

2.2 Goals as preferred outcomes

We consider a *goal as the preferred desire in a chain.*

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. Therefore, the agent has to decide whether she prefers a state where b_1 holds to a state where b_1' (i.e., $\neg b_1$) does (or the other way around). 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.

Consider now the scenario where the agent establishes that the second rule overrides the first one ($s > r$). 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.

In the situation described by Example 1, *visit_John* is the goal according to r_2, while *short_visit* is the goal for r_4.

2.3 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, in other words, for the most preferred outcomes (goals).

We first consider the case where only rule r applies. Here, the agent should commit to the outcome she values the most, that is 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 the agent's resources; rationally, she should target the next best outcome b_2. Accordingly, the agent derives b_2 as her *intention. An intention is an acceptable outcome which does not conflict with the beliefs describing the environment*.

Suppose now that b_2 is *forbidden*, and that the agent is social (a social agent is an agent not knowingly committing to anything that is forbidden (Governatori and Rotolo 2008)). Once again, the agent has to lower her expectation and settle for b_3, which is one of her *social intentions. A social intention is an intention which does not violate any norm.*

To complete the analysis, consider the situation where both rules r and s apply and, again, 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. This time we assume that the agent knows she cannot achieve $\neg b_1$ (or equivalently, b_1 holds). If the agent is rational, she cannot commit to $\neg b_1$. Consequently, the best option for her is to commit to b'_2 and b_1 (both regarded as intentions and social intentions), where she is guaranteed to be successful.

This scenario reveals a key concept: there are situations where the agent's best choice is to commit herself to some outcomes that are not her preferred ones (or even to a choice 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 situation).

Example 2

$$F = \{saturday, John\_away, John\_sick\} \quad R = \{r_2, r_3, r_4\} \quad > = \{(r_2, r_4)\}.$$

Today, John is in rehab at the hospital. Even if Alice has the desire as well as the goal to visit John, the facts of the situation lead her to form the intention to visit her parents.

Consider now the following theory:

$$F = \{saturday, John\_home\_confined, third\_week\}$$
$$R = \{r_2, r_3, r_4, r_5 : John\_home\_confined, third\_week \Rightarrow_{\mathsf{OBL}} \neg visit\_John\}$$
$$> = \{(r_2, r_4)\}.$$

Unfortunately, John has a stream of bad luck. Now, he is not debilitated but has been home convicted for a minor crime. The law of his country states that during the first two months of his home conviction, no visits to him are allowed. This time, even if Alice knows that John is at home, norms forbid Alice to visit him. Again, Alice opts to visit her parents.

3 Logic

Defeasible Logic (DL) (Antoniou *et al.* 2001) is a simple, flexible, and efficient rule based non-monotonic formalism. Its strength lies in its constructive proof theory, which has an argumentation-like structure, and it allows us to draw meaningful conclusions from (potentially) conflicting and incomplete knowledge bases. Being non-monotonic means that more accurate conclusions can be obtained when more pieces of information are given (where some previously derived conclusions no longer follow from the knowledge base).

The framework provided by the proof theory accounts for the possibility of extensions of the logic, in particular extensions with modal operators. Several of such extensions have been proposed, which then resulted in successful applications in the area of normative reasoning (Governatori 2005), modelling agents (Governatori and Rotolo 2008; Governatori *et al.* 2009; Kravari *et al.* 2011), and business process compliance (Governatori and Sadiq 2008). A model theoretic possible world semantics for modal Defeasible Logic has been proposed in Governatori *et al.* (2012). In addition, efficient implementations of the logic (including the modal variants), able to handle very large knowledge bases, have been advanced in Lam and Governatori (2009); Bassiliades *et al.* (2006); Tachmazidis *et al.* (2012).

Definition 1 (Language)
Let PROP be a set of propositional atoms, and MOD $= \{\mathsf{B}, \mathsf{O}, \mathsf{D}, \mathsf{G}, \mathsf{I}, \mathsf{SI}\}$ the set of modal operators, whose reading is B for *belief*, O for *obligation*, D for *desire*, G for *goal*, I for *intention,* and SI for *social intention*. Let Lab be a set of arbitrary labels. The set Lit $=$ PROP $\cup \{\neg p | p \in$ PROP$\}$ denotes the set of *literals*. The *complement* 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 ModLit $= \{Xl, \neg Xl | l \in$ Lit$, X \in \{\mathsf{O}, \mathsf{D}, \mathsf{G}, \mathsf{I}, \mathsf{SI}\}\}$. We assume that modal operator "X" for belief B is the empty modal operator. Accordingly, a modal literal $\mathsf{B}l$ is equivalent to literal l; the complement of $\mathsf{B}{\sim}l$ and $\neg\mathsf{B}l$ is l.

Definition 2 (Defeasible theory)
A *defeasible theory* D is a structure $(F, R, >)$, where (1) $F \subseteq$ Lit $\cup$ ModLit is a set of *facts* or indisputable statements; (2) R contains three sets of *rules*: for beliefs, obligations, and outcomes; (3) $> \subseteq R \times R$ is a binary *superiority relation* to determine the relative strength of (possibly) conflicting rules. We use the infix notation $r > s$ to mean that $(r, s) \in >$. A theory is *finite* if the set of facts and rules are so.

Belief rules are used to relate the factual knowledge of an agent, that is to say, her vision of the environment she is situated in. Belief rules define 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 take the O modality.

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 goal-like attitudes: desires, goals, intentions, and social intentions.

Following ideas given in (Governatori and Rotolo 2006), rules can gain more expressiveness when a *preference operator* $\odot$ is adopted. 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:

(1) every literal is an $\odot$-expression;
(2) 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 the following properties:

(1) $a \odot (b \odot c) = (a \odot b) \odot c$ (associativity);
(2) $\bigodot_{i=1}^{n} a_i = (\bigodot_{i=1}^{k-1} a_i) \odot (\bigodot_{i=k+1}^{n} a_i)$ where there exists j such that $a_j = a_k$ and $j < k$ (duplication and contraction on the right).

Typically, $\odot$-expressions are given by the agent designer, or obtained through *construction rules* based on the particular logic (Governatori and Rotolo 2006).

In the present paper, we use the classical definition of *defeasible rule* in DL (Antoniou *et al.* 2001), while *strict rules* and *defeaters* are omitted[1].

Definition 3 (Defeasible rule)
A *defeasible rule* is an expression $r : A(r) \Rightarrow_X C(r)$, where (1) $r \in \text{Lab}$ is the name of the rule; (2) $A(r) = \{a_1, \ldots, a_n\}$, the *antecedent* (or *body*) of the rule, is the set of the premises of the rule. Each a_i is either in Lit or in ModLit; (3) $X \in \{\mathsf{B}, \mathsf{O}, \mathsf{U}\}$ represents the *mode* of the rule: $\Rightarrow_{\mathsf{B}}$, $\Rightarrow_{\mathsf{O}}$, $\Rightarrow_{\mathsf{U}}$ denote respectively rules for beliefs, obligations, and outcomes. From now on, we omit the subscript B in rules for beliefs, i.e., $\Rightarrow$ is used as a shortcut for $\Rightarrow_{\mathsf{B}}$;
$C(r)$ is the *consequent* (or *head*) of the rule, which is a single literal if $X = \mathsf{B}$, and an $\odot$-expression otherwise[2].

A defeasible rule is a rule that can be defeated by contrary evidence. The underlying idea is that if we know that the premises of the rule are the case, then we may conclude that the conclusion holds, unless there is evidence proving otherwise. Defeasible rules in our framework introduce modal literals; for instance, if we have rule $r : A(r) \Rightarrow_{\mathsf{O}} c$ and the premises denoted by $A(r)$ are the case, then r can be used to prove $\mathsf{O}c$.

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

Notice that 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 conversion* and *conflict resolution* (Governatori and Rotolo 2008).

In the remainder, we shall define a completely new inference machinery that takes this into account by adding preferences and dealing with a larger set of modalised conclusions, which are not necessarily obtained from the corresponding rules but also by using other rule types. For instance, we argued in Section 2 that a goal can

[1] The restriction does not result in any loss of generality: (i) the superiority relation does not play any role in proving definite conclusions, and (ii) for defeasible conclusions Antoniou *et al.* (2001) prove that it is always possible to remove strict rules from the superiority relation and defeaters from the theory to obtain an equivalent theory without defeaters and where the strict rules are not involved in the superiority relation.

[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.

be viewed as a preferred outcome and so the fact that a certain goal $\mathsf{G}p$ is derived depends on whether we can obtain p as a preferred outcome by using a rule for U.

3.1 Rule conversion

It is sometimes meaningful to use rules for a modality X as if they were for another modality Y, i.e., to convert one type of conclusion into a different one.

Formally, we define an asymmetric binary relation Convert $\subseteq$ MOD $\times$ MOD such that 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 inference schema:

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

In our framework obligations and goal-like attitudes cannot change what the agent believes or how she perceives the world, we thus consider only conversion from beliefs to the other modes (i.e., Convert(B, X) with $X \in \mathrm{MOD} \setminus \{\mathsf{B}\}$). Accordingly, we enrich the notation with $R^{\mathsf{B},X}$ for the set of belief rules that can be used for a conversion to mode $X \in \mathrm{MOD} \setminus \{\mathsf{B}\}$. The antecedent of all such rules is not empty, and does not contain any modal literal.

Example 3

$$F = \{saturday\} \quad R = \{r_2, r_6 : visit\_John \Rightarrow chocolate\_box\}$$

where we stipulate that Convert(B, D) holds.

Alice desires to visit John. John is a passionate of chocolate and, usually, when Alice goes to meet him at his place, she brings him a box of chocolate. Thus, we may state that her desire of visiting John implies the desire to bring him a box of chocolate. This is the case since we can use rule r_6 to convert beliefs into desires.

3.2 Conflict-detection/resolution

It is crucial to identify criteria for detecting and solving conflicts between different modalities. We define an asymmetric binary relation Conflict $\subseteq$ MOD $\times$ MOD such that Conflict(X, Y) means "modes X and Y are in conflict and mode X prevails over Y". In our framework, we consider conflicts between (i) beliefs and intentions, (ii) beliefs and social intentions, and (iii) obligations and social intentions. In other words, the agents are characterised by:

- Conflict(B, I), Conflict$(\mathsf{B}, \mathsf{SI})$ meaning that agents are realistic (Broersen *et al.* 2002);
- Conflict$(\mathsf{O}, \mathsf{SI})$ meaning that agents are social (Governatori and Rotolo 2008).

Consider the scenario of Example 2 with Conflict(B, I) and Conflict$(\mathsf{O}, \mathsf{SI})$. We recall that rule r_5 states the prohibition to visit John during the first month of his conviction. Thus, Alice has the intention to visit John, but she does not have the social intention to do so. This is due to rule r_5 that prevents through conflict to prove $\mathsf{SI}visit\_John$. At the end, it is up to the agent (or the designer of the agent) whether to comply with the obligation, or not.

The *superiority relation* $>$ among rules is used to define where one rule may override the (opposite) conclusion of another one. There are two applications of the superiority relation: the first considers rules of the same mode while the latter compares rules of different modes. Given $r \in R^X$ and $s \in R^Y$, $r > s$ 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:

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

where we stipulate that Convert(B, G) holds.

It is my parents' anniversary and they are going to celebrate it this August in Rome, which is the capital of Italy. Typically, I do not want to go to Italy in August since the weather is too hot and Rome itself is too crowded. Nonetheless, I have the goal to go to Italy this summer 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.

Aligning with (Cohen and Levesque 1990), Conflict and superiority relations narrow and regulate the intentionality of conclusions drawn by the Convert relation in such a way that "agents need not intend all the expected side-effects of their intentions". This also prevents the ill-famed "dentist problem" which brings counterintuitive consequences, as also pointed out by Kontopoulos *et al.* (2011). If I want to go to the dentist, either I know that the pain is a "necessary way" to get better, or I am a masochist. Either way, I intend to suffer some pain for getting some ends.

Definition 4 (Proof)
A *proof* P of *length* n is a finite sequence $P(1), \ldots, P(n)$ 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. Hereafter, the term *refuted* is a synonym of *not provable* and we use $D \vdash \pm\partial_X l$ iff there is a proof P in D such that $P(n) = \pm\partial_X l$ for an index n.

In order to characterise the notions of provability/refutability 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 notions of *body-applicable* and *body-discarded* must be introduced. A rule is *body-applicable* when each literal in its body is proved with the appropriate modality; a rule is *body-discarded* if (at least) one of its premises has been refuted.

Definition 5 (Body applicable)
Let P be a proof and $X \in \{\mathsf{O}, \mathsf{D}, \mathsf{G}, \mathsf{I}, \mathsf{SI}\}$. A rule $r \in R$ is *body-applicable* (at $P(n+1)$) iff for all $a_i \in A(r)$: (1) if $a_i = Xl$, then $+\partial_X l \in P(1..n)$, (2) if $a_i = \neg Xl$, then $-\partial_X l \in P(1..n)$, (3) if $a_i = l \in \text{Lit}$, then $+\partial_{\mathsf{B}} l \in P(1..n)$.

Definition 6 (Body discarded)
Let P be a proof and $X \in \{\mathsf{O}, \mathsf{D}, \mathsf{G}, \mathsf{I}, \mathsf{SI}\}$. A rule $r \in R$ is *body-discarded* (at $P(n+1)$) iff there is $a_i \in A(r)$ such that (1) $a_i = Xl$ and $-\partial_X l \in P(1..n)$, or (2) $a_i = \neg Xl$ and $+\partial_X l \in P(1..n)$, or (3) $a_i = l \in \text{Lit}$ and $-\partial_{\mathsf{B}} l \in P(1..n)$.

As already stated, belief rules allow us to derive literals with different modalities through the conversion mechanism. The applicability mechanism takes this constraint into account.

Definition 7 (Conv-applicable)
Let P be a proof. A rule $r \in R$ is *Conv-applicable* (at $P(n+1)$) for X iff (1) $r \in R^{\mathsf{B}}$, (2) $A(r) \neq \emptyset$, (3) $A(r) \cap \text{ModLit} = \emptyset$, and (4) $\forall a \in A(r)$, $+\partial_X a \in P(1..n)$.

Definition 8 (Conv-discarded)
Let P be a proof. A rule $r \in R$ is *Conv-discarded* (at $P(n+1)$) for X iff (1) $r \notin R^{\mathsf{B}}$, or (2) $A(r) = \emptyset$, or (3) $A(r) \cap \text{ModLit} \neq \emptyset$, or (4) $\exists a \in A(r)$ s.t. $-\partial_X a \in P(1..n)$.

Let us consider the following theory:

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

Rule r_1 is applicable while r_2 is not, given that c is not proved as a belief. Instead, r_2 is *Conv-applicable* for O, since $\mathsf{O}c$ is a fact and r_1 gives $\mathsf{O}b$.

The notion of applicability gives guidelines on how to consider the next element in a given chain. Given that a belief rule cannot generate reparative chains but only single literals, we conclude that the applicability condition for belief collapses into body-applicability. When considering obligations, each element before the current one must be a violated obligation. Concerning desires, given that each element in an outcome chain represents a possible desire, we only require the rule to be applicable either directly, or through the Convert relation. A literal is a candidate to be a goal only if none of the previous elements in the chain has been proved as such. An intention must pass the wishful thinking filter (that is, there is no factual knowledge for the opposite conclusion), while social intention is also constrained not to violate any norm.

Definition 9 (Applicable rule)
Given a proof P, $r \in R[q, i]$ is *applicable* (at index i and $P(n+1)$) for

(1) B iff $r \in R^{\mathsf{B}}$ and is body-applicable.
(2) O iff either (2.1) (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) $X \in \{\mathsf{G}, \mathsf{I}, \mathsf{SI}\}$ iff either (4.1) (4.1.1) $r \in R^{\mathsf{U}}$ and is body-applicable, and
(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 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 Conflict(B, I), and for SI we have Conflict$(\mathsf{B}, \mathsf{SI})$ and Conflict$(\mathsf{O}, \mathsf{SI})$.

Definition 10 (Discarded rule)
Given a proof P, $r \in R[q, i]$ is *discarded* (at index i and $P(n+1)$) for

(1) B iff $r \in R^{\mathsf{B}}$ or is body-discarded.
(2) O iff (2.1) (2.1.1) $r \notin R^{\mathsf{O}}$ or is body-discarded, or
(2.1.2) $\exists c_k \in C(r)$, $k < i$, s.t. $-\partial_{\mathsf{O}} c_k \in P(1..n)$ or $+\partial c_k \in P(1..n)$, and
(2.2) r is Conv-discarded.
(3) D iff (3.1) $r \notin R^{\mathsf{U}}$ or is body-discarded, and
(3.2) r is Conv-discarded.
(4) $X \in \{\mathsf{G}, \mathsf{I}, \mathsf{SI}\}$ iff (4.1) (4.1.1) $r \notin R^{\mathsf{U}}$ or is body-discarded, or
(4.1.2) $\exists c_k \in C(r)$, $k < i$, s.t. $-\partial_Y {\sim} c_k \in P(1..n)$ for all Y such that Conflict(Y, X) or $+\partial_X c_k \in P(1..n)$ and
(4.2) r is Conv-discarded.

For G there are no conflicts; for I we have Conflict(B, I), and for SI we have Conflict$(\mathsf{B}, \mathsf{SI})$ and Conflict$(\mathsf{O}, \mathsf{SI})$.

Notice that the conditions of Definition 10 are the *strong negation*[3] of those given in Definition 9. The conditions to establish a rule being discarded correspond to the constructive failure to prove that the same rule is applicable.

We are now ready to introduce the definitions of the proof conditions for the modal operators given in this paper. We start with that for desire.

Definition 11 (Defeasible provability for desire)
The proof conditions of *defeasible provability* for desire are
$+\partial_{\mathsf{D}}$: 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) $\exists r \in R[q, i]$ s.t. 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) $s \not> r$.

The above conditions determine when we are able to assert that q is a desire. Specifically, a *desire* is each element in a chain of an outcome rule for which there is no stronger argument for the opposite desire.

The negative counterpart $-\partial_{\mathsf{D}} q$ is obtained by the principle of strong negation.

[3] The strong negation principle is closely related to the function that simplifies a formula by moving all negations to an innermost position in the resulting formula, and replaces the positive tags with the respective negative tags, and the other way around. (See Antoniou *et al.* (2000); Governatori *et al.* (2009).)

Definition 12 (Defeasible refutability for desire)

The proof conditions of *defeasible refutability* for desire are
$-\partial_{\mathsf{D}}$: If $P(n+1) = -\partial_{\mathsf{D}} q$, then
(1) $\mathsf{D}q \notin \mathrm{F}$ and
(2) (2.1) $\neg\mathsf{D}q \in \mathrm{F}$, or
  (2.2) $\forall r \in R[q, i]$ either r is discarded for D, or
  (2.3) $\exists s \in R[\sim q, j]$ s.t. (2.3.1) s is applicable for D and (2.3.2) $s > r$.

The proof conditions for $+\partial_X$, with $X \in \mathrm{MOD} \setminus \{\mathsf{D}\}$ are as follows, provided that Y and T represent two arbitrary modalities in MOD:

Definition 13 (Defeasible provability for obligation, goal, intention, and social intention)

The proof conditions of *defeasible provability* for $X \in \mathrm{MOD} \setminus \{\mathsf{D}\}$ are
$+\partial_X$: If $P(n+1) = +\partial_X q$, then
(1) $Xq \in \mathrm{F}$ or
(2) (2.1) $\neg Xq \notin \mathrm{F}$ and $(Y{\sim}q \notin \mathrm{F}$ for $Y = X$ or Conflict$(Y, X))$ and
  (2.2) $\exists r \in R[q, i]$ s.t. r is applicable for X and
  (2.3) $\forall s \in R[\sim q, j]$ either
    (2.3.1) $\forall Y$ s.t. $Y = X$ or Conflict(Y, X), s is discarded for Y; or
    (2.3.2) $\exists T, \exists t \in R[q, k]$ s.t. 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 the modality X we have two choices: (1) the modal literal Xq is a fact; or (2) we need to argue using the defeasible part of D. For (2), we require that (2.1) a complementary literal (of the same modality, or of a conflictual modality) does not appear in the set of facts, and (2.2) there must be an applicable rule for X and q. Moreover, each possible attack brought by a rule s for $\sim q$ has to be either discarded for the same modality of r and for all modalities in conflict with X (2.3.1), 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 rules with conflictual modalities (clause 2.3.2.2). Trivially, if $X = \mathsf{B}$, then the proof conditions reduce to those of classical defeasible logic (Antoniou *et al.* 2001).

Again, conditions for $-\partial_X$ are derived by the principle of strong negation from that for $+\partial_X$ and are as follows.

Definition 14 (Defeasible refutability for obligation, goal, intention, and social intention)

The proof conditions of *defeasible refutability* for $X \in \{\mathsf{O}, \mathsf{G}, \mathsf{I}, \mathsf{SI}\}$ are

$-\partial_X$: If $P(n+1) = -\partial_X q$, then
(1) $Xq \notin$ F and either
(2) (2.1) $\neg Xq \in$ F or ($Y{\sim}q \in$ F for $Y = X$ or Conflict(Y, X)) or
    (2.2) $\forall r \in R[q, i]$ either r is discarded for X or
    (2.3) $\exists s \in R[{\sim}q, j]$ s.t.
        (2.3.1) $\exists Y$ s.t. ($Y = X$ or Conflict(Y, X)) and s is applicable for Y, and
        (2.3.2) $\forall T, \forall t \in R[q, k]$ either t is discarded for T, or
            (2.3.2.1) $t \not> s$ if $Y = T$, Convert(Y, T), or Convert(T, Y); and
            (2.3.2.2) not Conflict(T, Y).

To better understand how applicability and proof conditions interact to define the (defeasible) conclusions of a given theory, we consider the example below.

Example 4
Let D be the following modal theory:

$$F = \{a_1,\ a_2,\ \neg b_1,\ \mathsf{O}\neg b_2\} \qquad R = \{r : a_1 \Rightarrow_{\mathsf{U}} b_1 \odot b_2 \odot b_3 \odot b_4,\ \ s : a_2 \Rightarrow_{\mathsf{U}} b_4\}.$$

Here, r is trivially applicable for D and $+\partial_{\mathsf{D}} b_i$ holds, for $1 \leqslant i \leqslant 4$. Moreover, we have $+\partial_{\mathsf{G}} b_1$ and r is discarded for G after b_1. Due to $+\partial \neg b_1$, it follows that $-\partial_{\mathsf{I}} b_1$ holds (as well as $-\partial_{\mathsf{SI}} b_1$); the rule is applicable for I and b_2, and we are able to prove $+\partial_{\mathsf{I}} b_2$; the rule is thus discarded for I and b_3 as well as b_4. Due to $\mathsf{O}\neg b_2$ being a fact, r is discarded for SI and b_2 resulting in $-\partial_{\mathsf{SI}} b_2$, which in turn makes the rule applicable for SI and b_3, proving $+\partial_{\mathsf{SI}} b_3$. As we have argued before, this makes r discarded for b_4. Even if r is discarded for SI and b_4, we nonetheless have $D \vdash +\partial_{\mathsf{SI}} b_4$ due to s; specifically, $D \vdash +\partial_X b_4$ with $X \in \{\mathsf{D}, \mathsf{G}, \mathsf{I}, \mathsf{SI}\}$ given that s is trivially applicable for X.

For further illustrations of how the machinery works, the reader is referred to Appendix A.

The next definition extends the concept of complement for modal literals and is used to establish the logical connection among proved and refuted literals in our framework.

Definition 15 (Complement set)
The *complement set* of a given modal literal l, denoted by $\tilde{l}$, is defined as follows: (1) if $l = \mathsf{D}m$, then $\tilde{l} = \{\neg \mathsf{D}m\}$; (2) if $l = Xm$, then $\tilde{l} = \{\neg Xm, X{\sim}m\}$, with $X \in \{\mathsf{O}, \mathsf{G}, \mathsf{I}, \mathsf{SI}\}$; (3) if $l = \neg Xm$, then $\tilde{l} = \{Xm\}$.

The logic resulting from the above proof conditions enjoys properties describing the appropriate behaviour of the modal operators for consistent theories.

Definition 16 (Consistent defeasible theory)
A defeasible theory $D = (\mathrm{F}, R, >)$ is *consistent* iff $>$ is acyclic and F does not contain pairs of complementary literals, that is if F does not contain pairs like (i) l and ${\sim}l$, (ii) Xl and $\neg Xl$ with $X \in$ MOD, and (iii) Xl and $X{\sim}l$ with $X \in \{\mathsf{G}, \mathsf{I}, \mathsf{SI}\}$.

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

(1) $D \vdash +\partial_X l$ and $D \vdash -\partial_X l$ with $X \in \text{MOD}$;
(2) $D \vdash +\partial_X l$ and $D \vdash +\partial_X {\sim} l$ with $X \in \text{MOD} \setminus \{\mathsf{D}\}$.

All proofs of propositions, lemmas and theorems are reported in Appendix B and Appendix C. The meaning of the above proposition is that, for instance, it is not possible for an agent to obey something that is obligatory and forbidden (obligatory not) at the same time. On the other hand, an agent may have opposite desires given different situations, but then she will be able to plan for only one between the two alternatives.

Proposition 2 below governs the interactions between different modalities and the relationships between proved literals and refuted complementary literals of the same modality. Proposition 3 proves that certain (likely-expected) implications do no hold.

Proposition 2
Let D be a consistent defeasible theory. For any literal l, the following statements hold:

(1) if $D \vdash +\partial_X l$, then $D \vdash -\partial_X {\sim} l$ with $X \in \text{MOD} \setminus \{\mathsf{D}\}$;
(2) if $D \vdash +\partial l$, then $D \vdash -\partial_{\mathsf{I}} {\sim} l$;
(3) if $D \vdash +\partial l$ or $D \vdash +\partial_{\mathsf{O}} l$, then $D \vdash -\partial_{\mathsf{SI}} {\sim} l$;
(4) if $D \vdash +\partial_{\mathsf{G}} l$, then $D \vdash +\partial_{\mathsf{D}} l$;
(5) if $D \vdash -\partial_{\mathsf{D}} l$, then $D \vdash -\partial_{\mathsf{G}} l$.

Proposition 3
Let D be a consistent defeasible theory. For any literal l, the following statements *do not* hold:

(6) if $D \vdash +\partial_{\mathsf{D}} l$, then $D \vdash +\partial_X l$ with $X \in \{\mathsf{G}, \mathsf{I}, \mathsf{SI}\}$;
(7) if $D \vdash +\partial_{\mathsf{G}} l$, then $D \vdash +\partial_X l$ with $X \in \{\mathsf{I}, \mathsf{SI}\}$;
(8) if $D \vdash +\partial_X l$, then $D \vdash +\partial_Y l$ with $X = \{\mathsf{I}, \mathsf{SI}\}$ and $Y = \{\mathsf{D}, \mathsf{G}\}$;
(9) if $D \vdash -\partial_Y l$, then $D \vdash -\partial_X l$ with $Y \in \{\mathsf{D}, \mathsf{G}\}$ and $X \in \{\mathsf{I}, \mathsf{SI}\}$.

Parts (6) and (7) directly follow by Definitions from 9 to 14 and rely on the intuitions presented in Section 2. Parts from (7) to (9) reveal the true nature of expressing outcomes in a preference order: it may be the case that the agent desires something (may it be even her preferred outcome) but if the factuality of the environment makes this outcome impossible to reach, then she should not pursue such an outcome, and instead commit herself on the next option available. The statements of Proposition 3 exhibit a common feature which can be illustrated by the idiom: "What's your plan B?", meaning: even if you are willing for an option, if such an option is not feasible you need to strive for the plan B.

Proof
Example 2 in the extended version offers counterexamples showing the reason why the above statements do not hold.

$$F = \{saturday, John\_away, John\_sick\}$$

$$
\begin{aligned}
R = \{ & r_2 : saturday \Rightarrow_{\mathsf{U}} visit\_John \odot visit\_parents \odot watch\_movie \\
& r_3 : John\_away \Rightarrow_{\mathsf{B}} \neg visit\_John \\
& r_4 : John\_sick \Rightarrow_{\mathsf{U}} \neg visit\_John \odot short\_visit \} \\
& r_7 : John\_away \Rightarrow_{\mathsf{B}} \neg short\_visit \} \\
> = \{ & (r_2, r_4) \}.
\end{aligned}
$$

Given that $r_2 > r_4$, Alice has the desire to *visit_John*, and this is also her preferred outcome. Nonetheless, being *John_away* a fact, this is not her intention, while so are $\neg visit\_John$ and *visit_parents*. □

4 Algorithmic results

We now present procedures and algorithms to compute the *extension* of a *finite* defeasible theory (Subsection 4.2), in order to ascertain the complexity of the logic introduced in the previous sections. The algorithms are inspired to ideas proposed in Maher (2001); Lam and Governatori (2011).

4.1 Notation for the algorithms

From now on, $\blacksquare$ denotes a generic modality in MOD, $\Diamond$ a generic modality in MOD $\setminus \{\mathsf{B}\}$, and $\Box$ a fixed modality chosen in $\blacksquare$. Moreover, whenever $\Box = \mathsf{B}$ we shall treat literals $\Box l$ and l as synonyms. To accommodate the Convert relation to the algorithms, we recall that $R^{\mathsf{B},\Diamond}$ denotes the set of belief rules that can be used for a conversion to modality $\Diamond$. The antecedent of all such rules is not empty, and does not contain any modal literal.

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 the superiority relation $>$ by incorporating the Conflict relation into it:

$$>=> \cup \{(r,s) | r \in R^{\Box}[l], s \in R^{\blacksquare}[\sim l], \mathrm{Conflict}(\Box, \blacksquare)\}.$$

We also define:

(1) $r_{sup} = \{s \in R : (s,r) \in >\}$ and $r_{inf} = \{s \in R : (r,s) \in >\}$ for any $r \in R$;
(2) HB_D as the set of literals such that the literal or its complement appears in D, i.e., such that it is a sub-formula of a modal literal occurring in D;
(3) 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 17 (Defeasible extension)
Given a 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 $\Box \in \mathrm{MOD}$. Two defeasible theories D and D' are *equivalent* whenever they have the same extensions, i.e., $E(D) = E(D')$.

We introduce two operations that modify the consequent of rules used by the algorithms.

Definition 18 (*Truncation and removal*)
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 neither of them, and $c = c_1 \odot a_i \odot c_2$ is an $\odot$-expression. Let r be a rule with form $A(r) \Rightarrow_\diamond c$. We define the *truncation* of the consequent c at a_i as:

$$A(r) \Rightarrow_\diamond c!a_i = A(r) \Rightarrow_\diamond c_1 \odot a_i,$$

and the *removal* of a_i from the consequent c as:

$$A(r) \Rightarrow_\diamond c \ominus a_i = A(r) \Rightarrow_\diamond c_1 \odot c_2.$$

Notice that removal may lead to rules with empty consequent which strictly would not be rules according to the definition of the language. Nevertheless, we accept such expressions within the description of the algorithms but then such rules will not be in any $R[q,i]$ for any q and i. In such cases, the operation *de facto* removes the rules.

Given $\Box \in \mathrm{MOD}$, the sets $+\partial_\Box$ and $-\partial_\Box$ denote, respectively, the global sets of positive and negative defeasible conclusions (i.e., the set of literals for which condition $+\partial_\Box$ or $-\partial_\Box$ holds), while $\partial_\Box^+$ and $\partial_\Box^-$ are the corresponding temporary sets, that is the set computed at each iteration of the main algorithm. Moreover, to simplify the computation, we do not operate on outcome rules: for each rule $r \in R^{\mathsf{U}}$ we create instead a new rule for desire, goal, intention, and social intention (respectively, r^{D}, r^{G}, r^{I}, and r^{SI}). Accordingly, for the sake of simplicity, in the present section we shall use expressions like "the intention rule" as a shortcut for "the clone of the outcome rule used to derive intentions".

4.2 Algorithms

The idea of all the algorithms is to use the operations of truncation and elimination to obtain, step after step, a simpler but equivalent theory. In fact, proving a literal does not give local information regarding the element itself only, but rather reveals which rules should be discarded, or reduced, in their head or body. Let us assume that, at a given step, the algorithm proves literal l. At the next step,

(1) the applicability of any rule r with $l \in A(r)$ does not depend on l any longer. Hence, we can safely remove l from $A(r)$.
(2) Any rule s with $\widetilde{l} \cap A(s) \neq \emptyset$ is discarded. Consequently, any superiority tuple involving s is now useless and can be removed from the superiority relation.
(3) We can shorten chains by exploiting conditions of Definitions 9 and 10. For instance, if $l = \mathsf{O}m$, we can truncate chains for obligation rules at $\sim m$ and eliminate it as well.

Algorithm 1 DefeasibleExtension is the core algorithm to compute the extension of a defeasible theory. The first part of the algorithm (lines 1–5) sets up the data

Algorithm 1 DefeasibleExtension

1: $+\partial_{\blacksquare}, \partial^{+}_{\blacksquare} \leftarrow \emptyset$; $-\partial_{\blacksquare}, \partial^{-}_{\blacksquare} \leftarrow \emptyset$
2: $R \leftarrow R \cup \{r^{\Box} : A(r) \Rightarrow_{\Box} C(r) | r \in R^{\mathsf{U}}\}$, with $\Box \in \{\mathsf{D}, \mathsf{G}, \mathsf{I}, \mathsf{SI}\}$
3: $R \leftarrow R \setminus R^{\mathsf{U}}$
4: $R^{\mathsf{B},\Diamond} \leftarrow \{r^{\Diamond} : A(r) \hookrightarrow C(r) | r \in R^{\mathsf{B}}, A(r) \neq \emptyset, A(r) \subseteq \mathrm{Lit}\}$
5: $> \leftarrow > \cup \{(r^{\Diamond}, s^{\Diamond}) | r^{\Diamond}, s^{\Diamond} \in R^{\mathsf{B},\Diamond}, r > s\} \cup \{(r,s) | r \in R^{\blacksquare} \cup R^{\mathsf{B},\blacksquare}, s \in R^{\Diamond} \cup R^{\mathsf{B},\Diamond}, \mathrm{Conflict}(\blacksquare, \Diamond)\}$
6: **for** $l \in F$ **do**
7:     **if** $l = \Box m$ **then** Proved(m, $\Box$)
8:     **if** $l = \neg\Box m \wedge \Box \neq \mathsf{D}$ **then** Refuted(m, $\Box$)
9: **end for**
10: $+\partial_{\blacksquare} \leftarrow +\partial_{\blacksquare} \cup \partial^{+}_{\blacksquare}$; $-\partial_{\blacksquare} \leftarrow -\partial_{\blacksquare} \cup \partial^{-}_{\blacksquare}$
11: $R_{infd} \leftarrow \emptyset$
12: **repeat**
13:     $\partial^{+}_{\blacksquare} \leftarrow \emptyset$; $\partial^{-}_{\blacksquare} \leftarrow \emptyset$
14:     **for** $\Box l \in HB$ **do**
15:         **if** $R^{\Box}[l] \cup R^{\mathsf{B},\Box}[l] = \emptyset$ **then** Refuted(l, $\Box$)
16:     **end for**
17:     **for** $r \in R^{\Box} \cup R^{\mathsf{B},\Box}$ **do**
18:         **if** $A(r) = \emptyset$ **then**
19:             $r_{inf} \leftarrow \{r \in R : (r,s) \in >, s \in R\}$; $r_{sup} \leftarrow \{s \in R : (s,r) \in >\}$
20:             $R_{infd} \leftarrow R_{infd} \cup r_{inf}$
21:             Let l be the first literal of $C(r)$ in HB
22:             **if** $r_{sup} = \emptyset$ **then**
23:                 **if** $\Box = \mathsf{D}$ **then**
24:                     Proved(m, D)
25:                 **else**
26:                     Refuted($\sim l$, $\Box$)
27:                     Refuted($\sim l$, $\Diamond$) for $\Diamond$ s.t. Conflict($\Box$, $\Diamond$)
28:                     **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**
29:                         Proved(m, $\Box$)
30:                     **end if**
31:                 **end if**
32:             **end if**
33:         **end if**
34:     **end for**
35:     $\partial^{+}_{\blacksquare} \leftarrow \partial^{+}_{\blacksquare} \setminus +\partial_{\blacksquare}$; $\partial^{-}_{\blacksquare} \leftarrow \partial^{-}_{\blacksquare} \setminus -\partial_{\blacksquare}$
36:     $+\partial_{\blacksquare} \leftarrow +\partial_{\blacksquare} \cup \partial^{+}_{\blacksquare}$; $-\partial_{\blacksquare} \leftarrow -\partial_{\blacksquare} \cup \partial^{-}_{\blacksquare}$
37: **until** $\partial^{+}_{\blacksquare} = \emptyset$ and $\partial^{-}_{\blacksquare} = \emptyset$
38: **return** $(+\partial_{\blacksquare}, -\partial_{\blacksquare})$

structure needed for the computation. Lines 6–9 are to handle facts as immediately provable literals.

The main idea of the algorithm is to check whether there are rules with empty body: such rules are clearly applicable and they can produce conclusions with the right mode. However, before asserting that the first element for the appropriate modality of the conclusion is provable, we need to check whether there are rules for the complement with the appropriate mode; if so, such rules must be weaker than the applicable rules. The information about which rules are weaker than the applicable ones is stored in the support set R_{infd}. When a literal is evaluated to be provable, the algorithm calls procedure Proved; when a literal is rejected, procedure Refuted is invoked. These two procedures apply transformations to reduce the complexity of the theory.

A step-by-step description of the algorithm would be redundant once the concepts expressed before are understood. Accordingly, in the rest of the section we provide in depth descriptions of the key passage.

For every outcome rule, the algorithm makes a copy of the same rule for each mode corresponding to a goal-like attitude (line 2). At line 4, the algorithm creates a support set to handle conversions from a belief rule through a different mode. Consequently, the new $\Diamond$ rules have to inherit the superiority relation (if any) from the belief rules they derive from (line 5). Notice that we also augment the superiority relation by incorporating the rules involved in the Conflict relation. Given that facts are immediately proved literals, PROVED is invoked for positively proved modal literals (those proved with $+\partial_\Box$), and REFUTED for rejected literals (i.e., those proved with $-\partial_\Box$). The aim of the **for** loop at lines 14–16 is to discard any modal literal in HB for which there are no rules that can prove it (either directly or through conversion).

We now iterate on every rule that can fire (i.e., on rules with empty body, loop **for** at lines 17–34 and **if** condition at line 18) and we collect the weaker rules in the set R_{infd} (line 20). Since a consequent can be an $\odot$-expression, the literal we are interested in is the first element of the $\odot$-expression (line 21). If no rule stronger than the current one exists, then the complementary conclusion is refuted by condition (2.3) of Definition 14 (line 26). An additional consequence is that literal l is also refutable in D for any modality conflicting with $\Box$ (line 27). Notice that this reasoning does not hold for desires: since the logic allows to have $\mathsf{D}l$ and $\mathsf{D}{\sim}l$ at the same time, when $\Box = \mathsf{D}$ and the guard at line 22 is satisfied, the algorithm invokes procedure 2 PROVED (line 24) due to condition (2.3) of Definition 11.

The next step is to check whether there are rules for the complement literal of the same modality, or of a conflicting modality. The rules for the complement should not be defeated by applicable rules: such rules thus cannot be in R_{infd}. If all these rules are defeated by r (line 28), then conditions for deriving $+\partial_\Box$ are satisfied, and Algorithm 2 PROVED is invoked.

Algorithm 2 PROVED is invoked when literal l is proved with modality $\Box$, the key to which simplifications on rules can be done. The computation starts by updating the relative positive extension set for modality $\Box$ and, symmetrically, the local information on literal l (line 2); l is then removed from HB at line 3. Parts 1.–3. of Proposition 2 identifies the modalities literal ${\sim}l$ is refuted with, when $\Box l$ is proved (**if** conditions at lines 4–6). Lines 7 to 9 modify the superiority relation and the sets of rules R and $R^{\mathsf{B},\Box}$ accordingly to the intuitions given at the beginning of Section 4.2.

Depending on the modality $\Box$ of l, we perform specific operations on the chains (condition **switch** at lines 10–27). A detailed description of each **case** would be redundant without giving more information than the one expressed by conditions of Definitions 9 and 10. Therefore, we propose one significative example by considering the scenario where l has been proved as a belief (**case** at lines 11–14). First, conditions of Definitions 10 and 14 ensure that ${\sim}l$ may be neither an intention, nor a social intention. Algorithm 3 REFUTED is thus invoked at lines 5 and 6 which, in turn, eliminates ${\sim}l$ from every chain of intention and social intention rules (line 18 of Algorithm 3 REFUTED). Second, chains of obligation (resp. intention) rules can be truncated at l since condition (2.1.2) (resp. condition (4.1.2)) of Definition 10 makes

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

such rules discarded for all elements following l in the chain (line 12). Third, if $+\partial_{\mathsf{O}}\sim l$ has been already proved, then we eliminate $\sim l$ in chains of obligation rules since it represents a violated obligation (**if** condition at lines 13). Fourth, if $-\partial_{\mathsf{O}}\sim l$ is the case, then each element after l cannot be proved as a social intention (**if** condition at line 14). Consequently, we truncate chains of social intention rules at l.

Algorithm 3 REFUTED performs all necessary operations to refute literal l with modality $\Box$. The initialisation steps at lines 2–6 follow the same schema exploited at lines 2–9 of Algorithm 2 PROVED. Again, the operations on chains vary according to the current mode $\Box$ (**switch** at lines 7–19). For instance, if $\Box = \mathsf{B}$ (**case** at lines 8–11), then condition (4.1.2) for I of Definition 10 is 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 the obligation of l has been violated. Thus, l can be removed from all chains for obligations (line 10). If instead $-\partial_{\mathsf{O}} l$ holds, then the elements after $\sim l$ in chains for social intentions satisfy condition (4.1.2) of Definition 10, and the algorithm removes them (line 11).

4.3 Computational results

We now present the computational properties of the algorithms previously described. Since Algorithms 2 PROVED and 3 REFUTED are sub-routines of the main one, we shall exhibit the correctness and completeness results of these algorithms inside

Algorithm 3 REFUTED

procedure REFUTED($l \in \text{Lit}, \Box \in \text{MOD}$)
  $\partial^-_\Box \leftarrow \partial^-_\Box \cup \{l\}$; $l_\blacksquare \leftarrow l_\blacksquare \cup \{-\Box\}$
  $HB \leftarrow HB \setminus \{\Box l\}$
  $R \leftarrow \{r : A(r) \setminus \{\neg\Box l\} \hookrightarrow C(r) | \ r \in R, \ \Box l \notin A(r)\}$
  $R^{\mathsf{B},\Box} \leftarrow R^{\mathsf{B},\Box} \setminus \{r \in R^{\mathsf{B},\Box} : l \in A(r)\}$
  $> \leftarrow > \setminus \{(r,s),(s,r) \in > | \Box l \in A(r)\}$
  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

theorems for Algorithm 1 DEFEASIBLEEXTENSION. In order to properly demonstrate results on the complexity of the algorithms, we need the following definition.

Definition 19 (Size of a theory)
Given a finite defeasible theory D, the *size* S of D is the number of occurrences of literals plus the number of the rules in D.

For instance, the size of the theory

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

is equal to nine, since literal a occurs three times.

We also report some key ideas and intuitions behind our implementation.

(1) Each operation on global sets $\pm\partial_\blacksquare$ and $\partial^{\pm}_\blacksquare$ requires linear time, as we manipulate finite sets of literals;
(2) For each literal $\Box l \in HB$, we implement a hash table with pointers to the rules where the literal occurs in; thus, retrieving the set of rules containing a given literal requires constant time;
(3) The superiority relation can also be implemented by means of hash tables; once again, the information required to modify a given tuple can be accessed in constant time.

In Section 4, we discussed the main intuitions behind the operations performed by the algorithms, and we explained that each operation corresponds to a reduction that transforms a theory in an equivalent smaller theory. Appendix C exhibits a series of lemmas stating the conditions under which an operation that removes either rules or literals form either the head or rules or from the body results in an equivalent smaller theory. The Lemmas proved by induction on the length of derivations.

Theorem 4
Given a finite defeasible theory D with size S, Algorithms 2 Proved and 3 Refuted terminate and their computational complexity is $O(S)$.

Theorem 5
Given a finite defeasible theory D with size S, Algorithm 1 DefeasibleExtension terminates and its computational complexity is $O(S)$.

Theorem 6
Algorithm 1 DefeasibleExtension is sound and complete.

5 Summary and related work

This article provided a new proposal for extending DL to model cognitive agents interacting with obligations. We distinguished concepts of desire, goal, intention, and social intention, but we started from the shared notion of outcome. Therefore, such concepts spring from a single notion that becomes distinct based on the particular relationship with beliefs and norms. This reflects a more natural notion of mental attitude and can express the well-known notion of Plan B. When we consider the single chain itself, this justifies that from a single concept of outcome we can derive all the other mental attitudes. Otherwise we would need as many additional rules as the elements in the chain; this, in turn, would require the introduction of additional notions to establish the relationships with beliefs and norms. This adds to our framework an economy of concepts.

Moreover, since the preferences allow us to determine what preferred outcomes are adopted by an agent (in a specific scenario) when previous elements in sequences are no longer feasible, our logic provides an abstract semantics for several types of goal and intention reconsideration.

A drawback of our approach perhaps lies in the difficulty of translating a natural language description into a logic formalisation. This is a notoriously hard task. Even if the obstacle seems very difficult, the payoff is worthwhile. The first reason is due to the efficiency of the computation of the positive extension once the formalisation has been done (polynomial time against the majority of the current frameworks in the literature which typically work in exponential time). The second reason is that the use of rules (such as business rules) to describe complex systems is extremely common (Knolmayer *et al.* 2000). Future lines of research will then focus on developing such methods, by giving tools which may help the (business) analyst in writing such (business) rules from the declarative description.

The logic presented in this paper, as the vast majority of approaches to model autonomous agents, is propositional. The algorithms to compute the extension of theory relies on the theory being finite, thus the first assumption for possible first-order extensions would be to work on finite domains of individuals. Given this assumption, the algorithms can be still be used once a theory has been grounded. This means that the size of theory is in function of the size of the grounding. We expect that the size of the grounding depends on the cardinality of the domain of individuals and the length of the vector obtained by the join of the predicates occurring in the theory.

Our contribution has strong connections with those by Dastani *et al.* (2005); Governatori and Rotolo (2008); Governatori *et al.* (2009), but it completely rebuilds the logical treatment of agents' motivational attitudes by presenting significant innovations in at least two respects.

First, while in Dastani *et al.* (2005); Governatori and Rotolo (2008); Governatori *et al.* (2009) the agent deliberation is simply the result of the derivation of mental states from *precisely* the corresponding rules of the logic—besides conversions, intentions are derived using only intention rules, goals using goal rules, etc.—here, the proof theory is much more aligned with the BDI intuition, according to which intentions and goals are the results of the manipulation of desires. The conceptual result of the current paper is that this idea can be entirely encoded 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 richer than the one in BOID (Broersen *et al.* 2002), 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. The second type of preference is encoded via the superiority relation between rules: the superiority can work locally between single rules of the same or different types, or can work systematically by stating via Conflict(X, Y) that two different motivations X and Y collide, and X always overrides Y. The interplay between these two preference mechanisms can help us in isolating different and complex ways for deriving mental states, but the resulting logical machinery is still computationally tractable, as the algorithmic analysis proved.

Lastly, since the preferences allow us to determine what preferred outcomes are adopted by an agent when previous elements 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 (Bratman 1987; Cohen and Levesque 1990) since intentions obey the law of inertia and resist retraction or revision, but they can be reconsidered when new relevant information comes in Bratman (1987). Despite that, the problem of revising intentions in BDI frameworks has received little attention. A very sophisticated exception is that of van der Hoek *et al.* (2007), 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, Shapiro *et al.* (2012) discussed how to revise the commitments to planned activities because of mutually conflicting intentions, a contribution that 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.

Our framework shares the motivation with that of Winikoff *et al.* (2002), where the authors provide a logic to describe both the declarative and procedural nature of goals. The nature of the two approaches lead to conceptually different solutions. For instance, they require goals, as in Hindriks *et al.* (2000), "not to be entailed by beliefs, i.e., that they be unachieved", while our beliefs can be seen as ways to achieve goals. Other requirements such as persistence or dropping a goal when reached cannot be taken into account.

Shapiro *et al.* (2007) and Shapiro and Brewka (2007) deal with goal change. The authors consider the case where an agent readopts goals that were previously believed to be impossible to achieve up to revision of her beliefs. They model goals through an accessibility relation over possible worlds. This is similar to our framework where different worlds are different assignments to the set of facts. Similarly to us, they prioritise goals as a preorder $\leqslant$; an agent adopts a new goal unless another incompatible goal prior in the ordering exists. This is in line with our framework where if we change the set of facts, the algorithms compute a new extension of the theory where two opposite literals can be proved as D but only one as I. Notice also that the ordering used in their work is unique and fixed at design time, while in our framework chains of outcome rules are built through a context-dependent partial order which, in our opinion, models more realistic scenarios.

Dastani *et al.* (2006) present three types of declarative goals: perform, achievement, and maintenance goals. In particular, they define planning rules which relate configurations of the world as seen by the agent (i.e., her beliefs). A planning rule is considered *correct* only if the plan associated to the rule itself allows the agent to reach a configuration where her goal is satisfied. This is strongly connected to our idea of belief rules, which define a path to follow in order to reach an agent outcome. Notice that this kind of research based on temporal aspects is orthogonal to ours.

The unifying framework proposed by van Riemsdijk *et al.* (2008) and Dastani *et al.* (2011) specifies different facets of the concept of goal. However, several aspects make a comparative analysis between the two frameworks unfeasible. Their analysis is indeed merely taxonomical, and it does not address how goals are used in agent logics, as we precisely do here.

van Riemsdijk *et al.* (2009) share our aim to formalise goals in a logic-based representation of conflicting goals and propose two different semantics to represent *conditional* and *unconditional* goals. Their central thesis, supported by Prakken (2006), is that only by adopting a credulous interpretation is it possible to have conflicting goals. However, we believe that a credulous interpretation is not suitable if an agent has to deliberate what her primary goals are in a given situation. We opted to have a sceptical interpretation of the concepts we call goals, intentions, and social intentions, while we adopt a credulous interpretation for desires. Moreover, they do not take into account the distinction between goals and related motivational attitudes (as in van Riemsdijk *et al.* (2008); Dastani *et al.* (2006, 2011)). The characteristic property of intentions in these logics is that an agent may not drop intentions for arbitrary reasons, which means that intentions have a certain persistence. As such, their analysis results orthogonal to ours.

Vasconcelos *et al.* (2009) propose mechanisms for the detection and resolution of normative conflicts. They resolve conflicts by manipulating the constraints associated to the norms' variables, as well as through *curtailment*, that is reducing the scope of the norm. In other works, we dealt with the same problems in defeasible deontic logic (Governatori *et al*. 2013a). We found three problems in their solution: (i) the curtailing relationship ω is rather less intuitive than our preference relation $>$, (ii) their approach seems too convoluted in solving exceptions (and they do not provide any mechanism to handle reparative chains of obligations), and (iii) the space complexity of their *adoptNorm* algorithm is exponential.

The present framework is meant to be seen as the first step within a more general perspective of providing the business analyst with tools that allow the creation of a business process in a fully declarative manner (Olivieri *et al*. 2013). Another issue comes from the fact that, typically, systems implemented by business rules involve thousands of such rules. Again, our choice of Defeasible Logic allows to drastically reduce the number of rules involved in the process of creating, for example, a business process thanks to its exception handling mechanism. This is peculiarly interesting when dealing with the problem of visualising such rules. When dealing with a system with thousands of rules, understanding what they represent or what a group of rules stand for, may be a serious challenge. On the contrary, the model presented by Olivieri *et al*. (2013), once an input is given, allows for the identification of whether the whole process is compliant against a normative system and a set of goals (and if not, where it fails). To the best of our knowledge, no other system is capable of checking whether a process can start with its input requisites and reaches its final objectives in a way that is compliant with a given set of norms.

Acknowledgements

NICTA is funded by the Australian Government through the Department of Communications and the Australian Research Council through the ICT Centre of Excellence Program.

This paper is an extended and revised version of Governatori *et al*. (2013b) presented at the 7th International Symposium on Theory, Practice, and Applications of Rules on the Web (RuleML 2013). We thank all the anonymous reviewers for their valuable comments.

Supplementary materials

For supplementary material for this article, please visit http://dx.doi.org/10.1017/S1471068416000053

References

Andrighetto, G., Governatori, G., Noriega, P. and van der Torre, L. W. N., Ed. 2013. *Normative Multi-Agent Systems*, Vol. 4, *Dagstuhl Follow-Ups*, Schloss Dagstuhl – Leibniz-Zentrum fuer Informatik. ISBN 978-3-939897-51-4.

Antoniou, G., Billington, D., Governatori, G. and Maher, M. J. 2001. Representation results for defeasible logic. *ACM Transactions on Computational Logic 2*, (2), 255–287. ISSN 1529-3785.

Antoniou, G., Billington, D., Governatori, G., Maher, M. J. and Rock, A. 2000. A family of defeasible reasoning logics and its implementation. In *Proc. of ECAI 2000*, Horn, Werner, Ed. IOS Press, Amsterdam, 459–463.

Bassiliades, N., Antoniou, G. and Vlahavas, I. 2006. A defeasible logic reasoner for the semantic web. *International Journal on Semantic Web and Information Systems 2*, (1), 1–41.

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

Brewka, G., Benferhat, S. and Le Berre, D. 2004. Qualitative choice logic. *Artificial Intelligence 157*, (1-2), 203–237.

Broersen, J., Dastani, M., Hulstijn, J. and van der Torre, L. 2002. Goal generation in the BOID architecture. *Cognitive Science Quarterly 2*, (3-4), 428–447. URL `http://icr.uni.lu/leonvandertorre/papers/csqj02.ps.Z`.

Cohen, P. R. and Levesque, H. J. 1990. Intention is choice with commitment. *Artificial Intelligence 42*, (2-3), 213–261.

Dastani, M., Governatori, G., Rotolo, A. and van der Torre, L. 2005. Programming cognitive agents in defeasible logic. In *Proc. LPAR*, G. Sutcliffe and A. Voronkov, Ed. *LNAI*, Vol. 3835. Springer, Berlin, Heidelberg, 621–636.

Dastani, M., van Riemsdijk, M. B. and Meyer, J.-J. C. 2006. Goal types in agent programming. In *AAMAS*, H. Nakashima, M. P. Wellman, G. Weiss and P. Stone, Ed. ACM, New York, 1285–1287.

Dastani, M., van Riemsdijk, M. B. and Winikoff, M. 2011. Rich goal types in agent programming. In *AAMAS*, L. Sonenberg, P. Stone, K. Tumer and P. Yolum, Ed. IFAAMAS, 405–412.

Governatori, G. 2005. Representing business contracts in RuleML. *International Journal of Cooperative Information Systems 14*, (2-3), 181–216.

Governatori, G., Olivieri, F., Rotolo, A. and Scannapieco, S. 2013a. Computing strong and weak permissions in defeasible logic. *Journal of Philosophical Logic 42*, (6), 799–829, URL `http://dx.doi.org/10.1007/s10992-013-9295-1`.

Governatori, G., Olivieri, F., Rotolo, A., Scannapieco, S. and Cristani, M. 2013b. Picking up the best goal – an analytical study in defeasible logic. In *RuleML*, volume 8035, L. Morgenstern, P. S. Stefaneas, F. Lévy, A. Wyner and A. Paschke, Ed. *Lecture Notes in Computer Science*, Springer, Berlin, Heidelberg, 99–113. ISBN 978-3-642-39616-8.

Governatori, G., Padmanabhan, V., Rotolo, A. and Sattar, A. 2009. A defeasible logic for modelling policy-based intentions and motivational attitudes. *Logic Journal of the IGPL 17*, (3), 227–265.

Governatori, G. and Rotolo, A. 2006. Logic of violations: A Gentzen system for reasoning with contrary-to-duty obligations. *Australasian Journal of Logic 4*, 193–215. ISSN 1448-5052.

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

Governatori, G., Rotolo, A. and Calardo, E. 2012. Possible world semantics for defeasible deontic logic. In *DEON*, volume 7393, T. Ågotnes, J. Broersen and D. Elgesem, Ed. *Lecture Notes in Computer Science*, Springer, Berlin, Heidelberg, 46–60.

Governatori, G. and Sadiq, S. 2008. The journey to business process compliance. In *Handbook of Research on BPM*, Cardoso, Jorge and van der Aalst, Wil, Ed. IGI Global, Hershley, New York, 426–454.

Hindriks, K. V., de Boer, F. S., van der Hoek, W. and Meyer, J.-J. C. 2000. Agent programming with declarative goals. In *ATAL*, volume 1986, C. Castelfranchi and Y. Lespérance, Ed. *Lecture Notes in Computer Science*, Springer, Berlin, Heidelberg, 228–243.

Knolmayer, G., Endl, R. and Pfahrer, M. 2000. Modeling processes and workflows by business rules. In *Business Process Management*, Springer, van der Aalst, Wil and Desel, Jörg and Pfahrer, Marcel, Ed. Springer, Volume: 1806, *Lecture Notes in Computer Science*, Berlin, 16–29.

Kontopoulos, E., Bassiliades, N., Governatori, G. and Antoniou, G. 2011. A modal defeasible reasoner of deontic logic for the semantic web. *International Journal on Semantic Web and Information Systems 7*, (1), 18–43.

Kravari, K., Papatheodorou, C., Antoniou, G. and Bassiliades, N. 2011. Reasoning and proofing services for semantic web agents. In *IJCAI/AAAI*, T. Walsh, Ed. AAAI Press, Menlo Park, CA, 2662–2667.

Lam, H.-P. and Governatori, G. 2009. The making of SPINdle. In *Rule Representation, Interchange and Reasoning on the Web*, Vol. 5858, G. Governatori, J. Hall and A. Paschke, Ed. LNCS, Springer, Berlin, Heidelberg, 315–322.

Lam, H.-P. and Governatori, G. 2011. What are the necessity rules in defeasible reasoning? In *LPNMR-11*, J. Delgrande and W. Faber, Ed. Springer, Springer, Lecture Notes in Computer Science, Volume: 6645, Berlin Heidelberg, 187–192.

Maher, M. J. 2001. Propositional defeasible logic has linear complexity. *Theory and Practice of Logic Programming 1*, (6), 691–711.

Olivieri, F., Governatori, G., Scannapieco, S. and Cristani, M. 2013. Compliant business process design by declarative specifications. In *Proc. of PRIMA 2013: Principles and Practice of Multi-Agent Systems – 16th International Conference, Dunedin, New Zealand, December 1–6, 2013*, G. Boella, E. Elkind, B. T. R. Savarimuthu, F. Dignum and M. K. Purvis, Ed. *Lecture Notes in Computer Science*, Vol. 8291. Springer, Berlin, Heidelberg, 213–228. ISBN 978-3-642-44926-0. doi: 10.1007/978-3-642-44927-7URL `http://dx.doi.org/10.1007/978-3-642-44927-7`.

Prakken, H. 2006. Combining sceptical epistemic reasoning with credulous practical reasoning. In *COMMA*, Vol. 144, P. E. Dunne and T. Bench-Capon, Ed. *Frontiers in Artificial Intelligence and Applications*, IOS Press, Amsterdam, 311–322.

Rao, A. S. and Georgeff, M. P. 1991. Modeling rational agents within a bdi-architecture. In *Proc. of KR*, J. F. Allen, R. Fikes and E. Sandewall, Ed. Morgan Kaufmann, San Francisco, 473–484. ISBN 1-55860-165-1.

Shapiro. S. and Brewka, G. 2007. Dynamic interactions between goals and beliefs. In *IJCAI 2007*, M. Veloso, Ed. IOS Press, Amsterdam, 2625–2630.

Shapiro, S., Lespérance, Y. and Levesque, H. J. 2007. Goal change in the situation calculus. *Journal of Logic and Computation 17*, (5), 983–1018.

Shapiro, S., Sardina, S., Thangarajah, J., Cavedon, L. and Padgham, L. 2012. Revising conflicting intention sets in bdi agents. In *Proc. of the 11th International Conference on AAMAS - Volume 2*, AAMAS '12, IFAAMS, 1081–1088.

Tachmazidis, I., Antoniou, G., Flouris, G., Kotoulas, S. and McCluskey, L. 2012. Large-scale parallel stratified defeasible reasoning. In *ECAI*, Vol. 242, L. De Raedt, C. Bessière, D. Dubois, P. Doherty, P. Frasconi, F. Heintz and P. J. F. Lucas, Ed. *Frontiers in Artificial Intelligence and Applications*, IOS Press, Berlin, Heidelberg, 738–743.

Thomason, R. H. 2000. Desires and defaults: A framework for planning with inferred goals. In *Proc. of KR2000*, A. G. Cohn, F. Giunchiglia and B. Selman, Ed. Morgan Kaufmann, San Francisco, 702–713.

van der Hoek, W., Jamroga, W. and Wooldridge, M. 2007. Towards a theory of intention revision. *Synthese 155*, (2), 265–90.

van Riemsdijk, M. B., Dastani, M. and Winikoff, M. 2008. Goals in agent systems: A unifying framework. In *AAMAS (2)*, D. C. Padgham, L. Parkes, J. P. Müller and S. Parsons, Ed. IFAAMAS, 713–720.

van Riemsdijk, M. B., Dastani, M. and Meyer, J.-J. C. 2009. Goals in conflict: Semantic foundations of goals in agent programming. *Journal of Autonomous Agents and Multi-Agent Systems 18*, (3), 471–500.

Vasconcelos, W. W., Kollingbaum, M. J. and Norman, T. J. 2009. Normative conflict resolution in multi-agent systems. *Autonomous Agents and Multi-Agent Systems 19*, (2), 124–152.

Winikoff, M., Padgham, L., Harland, J. and Thangarajah, J. 2002. Declarative & procedural goals in intelligent agent systems. In *Proc. of KR*, D. Fensel, F. Giunchiglia, D. L. McGuinness and M. Williams, Ed. Morgan Kaufmann, San Francisco, 470–481.

Wooldridge, M. and Jennings, N. R. 1995. Agent theories, architectures, and languages: A survey. In *ECAI Workshop on Agent Theories, Architectures, and Languages*, Vol. 890, M. Wooldridge and N. R. Jennings, Ed. *Lecture Notes in Computer Science*, Berlin, Heidelberg, Springer, 1–39.