1

The Rationale behind the Concept of Goal

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

[1] *NICTA, Australia*
[2] *Department of Computer Science, Verona, Italy*
[3] *CIRSFID, University of Bologna, Bologna, Italy*

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

2

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 $A_1, \ldots, A_n$ 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 $A_1, \ldots, A_n$ is considered.

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

> **If**   *saturday*   **then**   *visit_John, visit_parents, 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

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

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*:

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

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

4

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

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

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*.

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

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

and that the situations described by $a_1, \ldots, a_n$ and $a'_1, \ldots, a'_n$ are mutually compatible but b_1 and b'_1 are not, namely $b_1 = \neg b'_1$. In this case $b_1, \ldots, b_m, b'_1, \ldots, b'_k$ 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., $\neg visit\_John$).

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.

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

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

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, Kravari et al., 2011, Governatori et al., 2009], 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 = {B, O, D, G, I, 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, $X \in$ {O, D, G, I, SI}}. We assume that modal operator "X" for belief B is the empty modal operator. Accordingly, a modal literal Bl is equivalent to literal l; the complement of B$\sim l$ and $\neg$Bl 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 expres-

siveness 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) $r \in$ 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 $\Rightarrow_\mathsf{B}$; (4) $C(r)$ 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 $1 \leq i \leq n$.

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,

---

[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.

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

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:

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

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 $X \in$ MOD $\setminus \{B\}$). Accordingly, we enrich the notation with $R^{B,X}$ for the set of belief rules that can be used for a conversion to mode $X \in$ MOD $\setminus \{B\}$. 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.

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(B, SI) meaning that agents are realistic [Broersen et al., 2002];
- Conflict(O, SI) meaning that agents are social [Governatori and Rotolo, 2008].

Consider the scenario of Example 2 with Conflict(B, I) and Conflict(O, 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 $SIvisit\_John$. 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

$$F = \{go\_to\_Rome,\ parent\_anniversary,\ August\}$$
$$R = \{r_1 : go\_to\_Rome \Rightarrow go\_to\_Italy$$
$$\quad\quad r_2 : parent\_anniversary \Rightarrow_\cup go\_to\_Rome$$
$$\quad\quad r_3 : August \Rightarrow_\cup \neg go\_to\_Italy\}$$
$$> = \{(r_1, r_3)\}$$

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 ($\pm\partial_B$), obligations ($\pm\partial_O$), desires ($\pm\partial_D$), goals ($\pm\partial_G$), intentions ($\pm\partial_I$) and social intentions ($\pm\partial_{SI}$), 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 $X \in \{O, D, G, I, SI\}$. 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 $a_i = l \in$ Lit then $+\partial_B l \in P(1..n)$.

Definition 6 *(Body discarded)*
Let P be a proof and $X \in \{O, D, G, I, SI\}$. A rule $r \in R$ is *body-discarded* (at $P(n + 1)$)

10

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 $-\partial_B l \in P(1..n)$.

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) $r \in R^B$, (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) $r \notin R^B$, 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 $r \in R^B$ 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 $\text{Conflict}(Y, X)$ 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 $\text{Conflict}(\mathsf{B}, \mathsf{I})$, and for SI we have $\text{Conflict}(\mathsf{B}, \mathsf{SI})$ and $\text{Conflict}(\mathsf{O}, \mathsf{SI})$.

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) $\mathsf{D}q \in F$ or
(2) (2.1) $\neg \mathsf{D}q \notin F$ 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.

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) $\mathsf{D}q \notin F$ and
(2) (2.1) $\neg \mathsf{D}q \in F$, 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:

---

[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 13 (*Defeasible provability for obligation, goal, intention and social intention*)
The proof conditions of *defeasible provability* for $X \in \text{MOD} \setminus \{\mathsf{D}\}$ 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 $\text{Conflict}(Y, X)$) 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 $\text{Conflict}(Y, X)$, 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 $Y = T$, $\text{Convert}(Y, T)$, or $\text{Convert}(T, Y)$; or
         (2.3.2.2) $\text{Conflict}(T, Y)$.

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) $Xq \notin \mathrm{F}$ and either
(2) (2.1) $\neg Xq \in \mathrm{F}$ or ($Y{\sim}q \in \mathrm{F}$ for $Y = X$ or $\text{Conflict}(Y, X)$) 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 $\text{Conflict}(Y, X)$) 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 $Y = T$, $\text{Convert}(Y, T)$, or $\text{Convert}(T, Y)$; and
         (2.3.2.2) not $\text{Conflict}(T, Y)$.

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 $1 \le i \le 4$. 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 $-\partial_{\text{SI}} b_1$); the rule is applicable for I and b_2, and we are able to prove $+\partial_{\text{I}} b_2$; the rule is thus discarded for I and b_3 as well as b_4. Due to $O \neg b_2$ being a fact, r is discarded for SI and b_2 resulting in $-\partial_{\text{SI}} b_2$, which in turn makes the rule applicable for SI and b_3, proving $+\partial_{\text{SI}} b_3$. 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 $D \vdash +\partial_{\text{SI}} b_4$ due to s; specifically, $D \vdash +\partial_X b_4$ with $X \in \{D, G, I, SI\}$ 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 $l = Dm$, then $\tilde{l} = \{\neg Dm\}$; (2) if $l = Xm$, then $\tilde{l} = \{\neg Xm, X{\sim}m\}$, with $X \in \{O, G, I, SI\}$; (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 $D = (F, R, >)$ 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 $X \in \{G, I, SI\}$.

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 $X \in$ MOD;
2. $D \vdash +\partial_X l$ and $D \vdash +\partial_X {\sim} l$ with $X \in$ MOD $\setminus \{D\}$.

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 $X \in$ MOD $\setminus \{D\}$;
2. if $D \vdash +\partial_{\text{I}} l$, then $D \vdash -\partial_{\text{I}} {\sim} l$;
3. if $D \vdash +\partial l$ or $D \vdash +\partial_O l$, then $D \vdash -\partial_{\text{SI}} {\sim} l$;
4. if $D \vdash +\partial_G l$, then $D \vdash +\partial_D l$;
5. if $D \vdash -\partial_D l$, then $D \vdash -\partial_G l$.

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

14

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

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.

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 \ {B}, and $\square$ a fixed modality chosen in $\blacksquare$. Moreover, whenever $\square = $ B we shall treat literals $\square l$ 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 $\pm\square \in l_\blacksquare$ iff $D \vdash \pm\partial_\square l$. Given a modal defeasible theory D, a set of rules R, and a rule $r \in R^\square[l]$, we expand the superiority relation $>$ by incorporating the Conflict relation into it:

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

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 $HB = \{\square l | \square \in \text{MOD}, l \in HB_D\}$.

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

$$E(D) = (+\partial_\square, -\partial_\square),$$

where $\pm\partial_\square = \{l \in HB_D : D \vdash \pm\partial_\square l\}$ with $\square \in \text{MOD}$. 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 $\square \in \text{MOD}$, the sets $+\partial_\square$ and $-\partial_\square$ denote, respectively, the global sets of positive and negative defeasible conclusions (i.e., the set of literals for which condition $+\partial_\square$ or $-\partial_\square$ holds), while $\partial_\square^+$ and $\partial_\square^-$ 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 $l = \bigcirc m$, 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 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

16

Algorithm 1 DEFEASIBLEEXTENSION

```
 1: +∂■, ∂■⁺ ← ∅; −∂■, ∂■⁻ ← ∅
 2: R ← R ∪ {r□ : A(r) ⇒□ C(r)|r ∈ RU}, with □ ∈ {D, G, I, SI}
 3: R ← R \ RU
 4: RB,◇ ← {r◇ : A(r) ↪ C(r)|r ∈ RB, A(r) ≠ ∅, A(r) ⊆ Lit}
 5: >↩> ∪{(r◇, s◇)|r◇, s◇ ∈ RB,◇, r > s} ∪ {(r, s)|r ∈ R■ ∪ RB,■, s ∈ R◇ ∪ RB,◇, Conflict(■, ◇)}
 6: for l ∈ F do
 7:     if l = □m then PROVED(m, □)
 8:     if l = ¬□m ∧ □ ≠ D then REFUTED(m, □)
 9: end for
10: +∂■ ← +∂■ ∪ ∂■⁺; −∂■ ← −∂■ ∪ ∂■⁻
11: Rinfd ← ∅
12: repeat
13:     ∂■⁺ ← ∅; ∂■⁻ ← ∅
14:     for □l ∈ HB do
15:         if R□[l] ∪ RB,□[l] = ∅ then REFUTED(l, □)
16:     end for
17:     for r ∈ R□ ∪ RB,□ do
18:         if A(r) = ∅ then
19:             rinf ← {r ∈ R : (r, s) ∈>, s ∈ R}; rsup ← {s ∈ R : (s, r) ∈>}
20:             Rinfd ← Rinfd ∪ rinf
21:             Let l be the first literal of C(r) in HB
22:             if rsup = ∅ then
23:                 if □ = D then
24:                     PROVED(m, D)
25:                 else
26:                     REFUTED(∼l, □)
27:                     REFUTED(∼l, ◇) for ◇ s.t. Conflict(□, ◇)
28:                     if R□[∼l] ∪ RB,□[∼l] ∪ R■[∼l] \ Rinfd ⊆ rinf, for ■ s.t. Conflict(■, □) then
29:                         PROVED(m, □)
30:                     end if
31:                 end if
32:             end if
33:         end if
34:     end for
35:     ∂■⁺ ← ∂■⁺ \ +∂■; ∂■⁻ ← ∂■⁻ \ −∂■
36:     +∂■ ← +∂■ ∪ ∂■⁺; −∂■ ← −∂■ ∪ ∂■⁻
37: until ∂■⁺ = ∅ and ∂■⁻ = ∅
38: return (+∂■, −∂■)
```

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 ◇ 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 $+\partial_\square$), and REFUTED for

rejected literals (i.e., those proved with $-\partial_\square$). 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 $\square$ (line 27). Notice that this reasoning does not hold for desires: since the logic allows to have Dl and D$\sim l$ at the same time, when $\square = $ 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 $+\partial_\square$ are satisfied, and Algorithm 2 PROVED is invoked.

Algorithm 2 PROVED

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

Algorithm 2 PROVED is invoked when literal l is proved with modality $\square$, the key to which simplifications on rules can be done. The computation starts by updating the relative positive extension set for modality $\square$ 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 $\square l$ is proved (**if** conditions at lines 4–6). Lines 7 to 9 modify

the superiority relation and the sets of rules R and $R^{\mathsf{B},\square}$ accordingly to the intuitions given at the beginning of Section 4.2.

Depending on the modality $\square$ 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 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

```
1:  procedure REFUTED(l ∈ Lit, □ ∈ MOD)
2:      ∂□⁻ ← ∂□⁻ ∪ {l};  l■ ← l■ ∪ {−□}
3:      HB ← HB \ {□l}
4:      R ← {r : A(r) \ {¬□l} ↪ C(r)| r ∈ R, □l ∉ A(r)}
5:      R^{B,□} ← R^{B,□} \ {r ∈ R^{B,□} : l ∈ A(r)}
6:      > ← > \{(r, s), (s, r) ∈ > |□l ∈ A(r)}
7:      switch (□)
8:          case B:
9:              R^I ← {A(r) ⇒_I C(r)!∼l |r ∈ R^I[∼l, n]}
10:             if +O ∈ l■ then R^O ← {A(r) ⇒_O C(r) ⊖ l |r ∈ R^O[l, n]}
11:             if −O ∈ l■ then R^{SI} ← {A(r) ⇒_{SI} C(r)!∼l |r ∈ R^{SI}[∼l, n]}
12:         case O:
13:             R^O ← {A(r) ⇒_O C(r)!l ⊖ l |r ∈ R^O[l, n]}
14:             if −B ∈ l■ then R^{SI} ← {A(r) ⇒_{SI} C(r)!∼l |r ∈ R^{SI}[∼l, n]}
15:         case D:
16:             R^X ← {A(r) ⇒_X C(r) ⊖ l |r ∈ R^X[l, n]} with X ∈ {D, G}
17:         otherwise:
18:             R^□ ← {A(r) ⇒_□ C(r) ⊖ l |r ∈ R^□[l, n]}
19:     end switch
20: end procedure
```

Algorithm 3 REFUTED performs all necessary operations to refute literal l with modality $\square$. 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 $\square$ (**switch** at lines 7–19). For instance, if $\square = \mathsf{B}$ (**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).

19

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 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

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

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 $\partial_\blacksquare^\pm$ 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 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 ⊙ 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 $\leq$; 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 trough 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., 2011, 2006]). 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.

References

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

G. Antoniou, D. Billington, G. Governatori, M. J. Maher, and A. Rock. A family of defeasible reasoning logics and its implementation. In *ECAI 2000*, pages 459–463, 2000.

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

N. Bassiliades, G. Antoniou, and I. Vlahavas. A defeasible logic reasoner for the semantic web. *Int. J. Semantic Web Inf. Syst.*, 2(1):1–41, 2006.

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

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

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

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

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

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

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

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

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

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

G. Governatori and S. Sadiq. The journey to business process compliance. *Handbook of Research on BPM*, pages 426–454, 2008.

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

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

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

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

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

G. Knolmayer, R. Endl, and M. Pfahrer. Modeling processes and workflows by business rules. In *Business Process Management*, pages 16–29. Springer, 2000.

E. Kontopoulos, N. Bassiliades, G. Governatori, and G. Antoniou. A modal defeasible reasoner of deontic logic for the semantic web. *Int. J. Semantic Web Inf. Syst.*, 7(1):18–43, 2011.

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

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

H.-P. Lam and G. Governatori. What are the Necessity Rules in Defeasible Reasoning? In J. Delgrande and W. Faber, editors, *LPNMR-11*, pages 187–192. Springer, 2011.

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

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

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

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

S. Shapiro and G. Brewka. Dynamic interactions between goals and beliefs. In G. Bonanno, J. P. Delgrande, J. Lang, and H. Rott, editors, *Formal Models of Belief Change in Rational Agents*, volume 07351 of *Dagstuhl Seminar Proceedings*. Internationales Begegnungs- und Forschungszentrum fuer Informatik (IBFI), Schloss Dagstuhl, Germany, 2007.

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

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

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

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

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

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

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

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

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

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

Appendix A Inferential mechanism example

This appendix is meant to offer more details for illustrating the inference mechanism proposed in this paper, and to consider Definition 13 more carefully. Therefore, first we report in Table A 1 the most interesting scenarios, where a rule r proves $+\partial_{\mathsf{SI}} q$ when attacked by an applicable rule s, which in turn is successfully counterattacked by an applicable rule t. Lastly, we end this appendix by reporting an example. The situation described there starts from a natural language description and then shows how it can be formalised with the logic we proposed.

For the sake of clarity, notation $\mathsf{B}, \square$ (with $\square \in \{\mathsf{O}, \mathsf{SI}\}$) represents belief rules which are Conv-applicable for mode $\square$.

For instance, the sixth row of the table denotes situations like the following:

$$F = \{a, b, \mathsf{O}c\}$$
$$R = \{r : a \Rightarrow_{\mathsf{U}} q,$$
$$s : b \Rightarrow_{\mathsf{O}} \neg q,$$
$$t : c \Rightarrow q\}$$
$$> = \{(t, s)\}.$$

26

The outcome rule r for q is applicable for SI according to Definition 9. Since in our framework we have Conflict(O, SI), the rule s for $\neg q$ (which is applicable for O) does not satisfy condition (2.3.1) of Definition 13. As a result, s represents a valid attack to r. However, since we have Convert(B, O), rule t is Conv-applicable for O by Definition 7, with $t > s$ by construction. Thus, t satisfies condition (2.3.2.1) of Definition 13 and successfully counterattacks s. Consequently, r is able to conclude $+\partial_{\mathsf{SI}} q$.

Mode of r	**Mode of s**	**Mode of t**	$+\partial_{\mathsf{SI}} q$ **because...**
U applicable for SI	U applicable for SI	U applicable for SI	$t > s$
U applicable for SI	U applicable for SI	O	Conflict(O, SI)
U applicable for SI	U applicable for SI	B, SI	$t > s$
U applicable for SI	U applicable for SI	B, O	Conflict(O, SI)
U applicable for SI	O	O	$t > s$
U applicable for SI	O	B, O	$t > s$
U applicable for SI	B, SI	U applicable for SI	$t > s$
U applicable for SI	B, SI	O	Conflict(O, SI)
U applicable for SI	B, SI	B, SI	$t > s$
U applicable for SI	B, SI	B, O	Conflict(O, SI)
U applicable for SI	B, O	B, O	$t > s$
B, SI	U applicable for SI	U applicable for SI	$t > s$
B, SI	U applicable for SI	O	Conflict(O, SI)
B, SI	U applicable for SI	B, SI	$t > s$
B, SI	U applicable for SI	B, O	Conflict(O, SI)
B, SI	O	O	$t > s$
B, SI	O	B, O	$t > s$
B, SI	B, SI	U applicable for SI	$t > s$
B, SI	B, SI	O	Conflict(O, SI)
B, SI	B, SI	B, SI	$t > s$
B, SI	B, SI	B, O	Conflict(O, SI)
B, SI	B, O	B, O	$t > s$

Table A 1. *Definition 13: Attacks and counterattacks for social intention*

Example 5

PeoplEyes is an eyeglasses manufacturer. Naturally, its final goal is to produce cool and perfectly assembled eyeglasses. The final steps of the production process are to shape the lenses to glasses, and mount them on the frames. To shape the lenses, *PeoplEyes* uses a very innovative and expensive laser machine, while for the final mounting phase two different machines can be used. Although both machines work well, the first and newer one is more precise and faster than the other one; *PeoplEyes* thus prefers to use the first machine as much as possible. Unfortunately, a new norm comes in force stating that no laser technology can be used, unless human staff wears laser-protective goggles.

If *PeoplEyes* has both human resources and raw material, and the three machines are fully working, but it has not yet bought any laser-protective goggles, all its goals would be

achieved but it would fail to comply with the applicable regulations, since the norm for the no-usage of laser technology is violated and not compensated.

If *PeoplEyes* buys the laser-protective goggles, their entire production process also becomes norm compliant. If, at some time, the more precise mounting machine breaks, but the second one is still working, *PeoplEyes* can still reach some of its objectives since the usage of the second machine leads to a state of the world where the objective of mounting the glasses on the frames is accomplished. Again, if *PeoplEyes* has no protective laser goggles and both the mounting machines are out of order, *PeoplEyes'* production process is neither norm, nor outcome compliant.

The following theory is the formalisation into our logic of the above scenario.

$$F = \{lenses,\ frames,\ new\_safety\_regulation\}$$

$$R = \{r_1 :\Rightarrow_U eye\_Glasses$$

$$r_2 :\Rightarrow laser$$

$$r_3 : lenses, laser \Rightarrow glasses$$

$$r_4 :\Rightarrow mounting\_machine1$$

$$r_5 :\Rightarrow mounting\_machine2$$

$$r_6 : mounting\_mach1 \Rightarrow \neg mounting\_machine2$$

$$r_7 : frames, glasses, mounting\_machine1 \Rightarrow eye\_Glasses$$

$$r_8 : frames, glasses, mounting\_machine2 \Rightarrow eye\_Glasses$$

$$r_9 : new\_safety\_regulation \Rightarrow_O \neg laser \otimes goggles$$

$$r_{10} :\Rightarrow_U mounting\_machine1 \oplus mounting\_machine2\}$$

$$>^{sm} = \{r_6 > r_5\}.$$

We assume *PeoplEye* has enough resources to start the process by setting *lenses* and *frames* as facts. Rule r_1 states that producing *eye_Glasses* is the main objective ($+\partial_I eye\_Glasses$, we choose intention as the mental attitude to comply with/attain to); rules r_2, r_4 and r_5 describe that we can use, respectively, the laser and the two mounting machineries. Rule r_3 is to represent that, if we have lenses and a laser machinery available, then we can shape glasses; in the same way, rules r_7 and r_8 describe that whenever we have glasses and one of the mounting machinery is available, then we obtain the final product. Therefore, the positive extension for belief $+\partial$ contains *laser*, *glasses*, *mounting_machine1* and *eye_Glasses*. In that occasion, rule r_6 along with $>$ prevent the using of both machineries at the same time and thus $-\partial mounting\_machine2$ (we assumed, for illustrative purpose even if unrealistically, that a parallel execution is not possible). When a new safety regulation comes in force (r_9), the usage of the laser machinery is forbidden, unless protective goggles are worn ($+\partial_O \neg laser$ and $+\partial_O \neg goggles$). Finally, rule r_{10} is to describe the preference of using *mounting_machine1* instead of *mounting_machine2* (hence we have $+\partial_I mounting\_machine1$ and $-\partial_I mounting\_machine2$).

Since there exists no rule for goggles, the theory is outcome compliant (that is, it reaches some set of objectives), but not norm compliant (given that it fails to meet some obligation rules without compensating them). If we add *goggles* to the facts and we substitute r_2 with

$$r_2' : O goggles \Rightarrow laser$$

then we are both norm and outcome compliant, as well as if we add

$$r_{11} : mounting\_machine1\_broken \Rightarrow \neg mounting\_machine1$$

to R and *mounting_machine1_broken* to F. Notice that, with respect to *laser*, we are intention compliant but *not* social intention compliant (given $O\neg lenses$). This is a key characteristic of our logic: The system is informed that the process is compliant but some violations have occurred.

Appendix B Proofs of Propositions in Section 3

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 $X \in \text{MOD} \setminus \{\mathsf{D}\}$.

Proof

1. (Coherence of the logic) The negative proof tags are the strong negation of the positive ones, and so are the conditions of a rule being discarded (Definition 10) for a rule being applicable (Definition 9). Hence, when the conditions for $+\partial_X$ hold, those for $-\partial_X$ do not.

2. (Consistency of the logic) We split the proof into two cases: (i) at least one of Xl and $X\sim l$ is in F, and (ii) neither of them is in F. For (i) the proposition immediately follows by the assumption of consistency. In fact, suppose that $Xl \in$ F. Then clause (1) of $+\partial_X$ holds for l. By consistency $X\sim l \notin$ F, thus clause (1) of Definition 13 does not hold for $\sim l$. Since $Xl \in$ F, also clause (2.1) is always falsified for $\sim l$, and the thesis is proved.

For (ii), let us assume that both $+\partial_X l$ and $+\partial_X \sim l$ hold in D. A straightforward assumption derived by Definitions 9 and 10 is that no rule can be at the same time applicable and discarded for X and l for any literal l and its complement. Thus, we have that there are applicable rules for X and l, as well as for X and $\sim l$. This means that clause (2.3.2) of Definition 13 holds for both l and $\sim l$. Therefore, for every applicable rule for l there is an applicable rule for $\sim l$ stronger than the rule for l. Symmetrically, for every applicable rule for $\sim l$ there is an applicable rule for l stronger than the rule for $\sim l$. Since the set of rules in D is finite by construction, this situation is possible only if there is a cycle in the transitive closure of the superiority relation, which is in contradiction with the hypothesis of D being consistent. $\qquad\square$

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 $D \vdash -\partial_\mathsf{G} l$.

Proof

For part 1., let D be a consistent defeasible theory, and $D \vdash +\partial_X l$. Literal $\sim l$ can be in only one of the following, mutually exclusive situations: (i) $D \vdash +\partial_X \sim l$; (ii) $D \vdash -\partial_X \sim l$;

(iii) $D \nvdash \pm\partial_X \sim l$. Part 2 of Proposition 1 allows us to exclude case (i), since $D \vdash +\partial_X l$ by hypothesis. Case (iii) denotes situations where there are loops in the theory involving literal $\sim l$,[4] but inevitably this would affect also the provability of Xl, i.e., we would not be able to give a proof for $+\partial_X l$ as well. This is in contradiction with the hypothesis. Consequently, situation (ii) must be the case.

Parts 2. and 3. directly follow by Definitions 9 and 10, while Definitions 9 and 13 justify part 4., given that G is not involved in any conflict relation.

Part 5. Trivially, from part 4. $\qquad\square$

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 $X \in \{\mathsf{I}, \mathsf{SI}\}$.

Proof

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

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

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 ¬*visit_John* and *visit_parents*. $\qquad\square$

Appendix C Correctness and Completeness of DEFEASIBLEEXTENSION

In this appendix we give proofs of the lemmas used by Theorem 6 for the soundness and completeness of the algorithms proposed.

We recall that the algorithms in Section 4 are based on a series of transformations that reduce a given theory into an equivalent, simpler one. Here, equivalent means that the two theories have the same extension, and simpler means that the size of the target theory is smaller than that of the original one. Remember that the size of a theory is the number of instances of literals occurring in the theory plus the number of rules in the theory. Accordingly, each transformation either removes some rules or some literals from rules

---

[4] For example, situations like $X \sim l \Rightarrow_X \sim l$, where the proof conditions generate a loop without introducing a proof.

(specifically, rules or literals we know are no longer useful to produce new conclusions). There is an exception. At the beginning of the computation, the algorithm creates four rules (one for each type of goal-like attitude) for each outcome rule (and the outcome rule is then eliminated). The purpose of this operation is to simplify the transformation operations and the bookkeeping of which rules have been used and which rules are still able to produce new conclusions (and the type of conclusions). Alternatively, one could implement flags to achieve the same result, but in a more convoluted way. A consequence of this operation is that we no longer have outcome rules. This implies that we have (i) to adjust the proof theory, and (ii) to show that the adjusted proof theory and the theory with the various goal-like rules are equivalent to the original theory and original proof conditions.

The adjustment required to handle the replacement of each outcome rule with a set of rules of goal-like modes (where each new rule has the same body and consequent of the outcome rule it replaces) is to modify the definition of being applicable (Definition 9) and being discarded (Definition 10). Specifically, we have to replace

- $r \in R^U$ in clause 3 of Definition 9 with $r \in R^D$;
- $r \notin R^U$ in clause 3 of Definition 10 with $r \notin R^D$;
- $r \in R^U$ in clause 4.1.1 of Definition 9 with $r \in R^X$; and
- $r \notin R^U$ in clause 4.1.1 of Definition 10 with $r \notin R^X$.

Given a theory D with goal-like rules instead of outcome rules we will use $E_3(D)$ to refer to the extension of D computed using the proof theory obtained from the proof theory defined in Section 3 with the modified versions of the notions of applicable and discarded just given.

Lemma 7

Let $D = (F, R, >)$ be a defeasible theory. Let $D' = (F, R', >')$ be the defeasible theory obtained from D as follows:

$$R' = R^B \cup R^O \cup \{r_X : A(r) \hookrightarrow_X C(r) | r : A(r) \hookrightarrow_U C(r) \in R, X \in \{D, G, I, SI\}\}$$

$$>' = \{(r, s) | (r, s) \in >, \ s, r \in R^B \cup R^O\} \cup \{(r_X, s_Y) | (r, s) \in >, r, s \in R^U\} \cup$$

$$\{(r_X, s) | (r, s) \in >, r \in R^U, s \in R^B \cup R^O\} \cup \{(r, s_X) | (r, s) \in >, r \in R^B \cup R^O, s \in R^U\}$$

Then, $E(D) = E_3(D')$.

Proof

The differences between D and D' are that each outcome-rule in D corresponds to four rules in D' each for a different mode and all with the same antecedent and consequent of the rule in D. Moreover, every time a rule r in D is stronger than a rule s in D, then any rule corresponding to r in D' is stronger than any rule corresponding to s in D'.

The differences in the proof theory for D and that for D' is in the definitions of applicable for X and discarded for X. It is immediate to verify that every time a rule r is applicable (at index n) for X, then r_X is applicable (at index n) for X (and the other way around). $\quad\square$

Given the functional nature of the transformations involved in the algorithms, we shall refer to the rules in the target theory with the same labels as the rules in the source theory. Thus, given a rule $r \in D$, we will refer to the rule corresponding to it in D' (if it exists) with the same label, namely r.

In the algorithms, belief rules may convert to another mode $\diamond$ only through the support

set $R^{B,\Diamond}$. Definition 7 requires $R^{B,\Diamond}$ to be initialised with a modal version of each belief rule with *non-empty* antecedent, such that every literal a in the antecedent is replaced by the corresponding modal literal $\Diamond a$.

In this manner, rules in $R^{B,\Diamond}$ satisfy clauses 1 and 2 of Definitions 7 and 8 by construction, while clauses 3 of both definitions are satisfied iff these new rules for $\Diamond$ are body-applicable (resp. body-discarded). Therefore, conditions for rules in $R^{B,\Diamond}$ to be applicable/discarded collapse into those of Definition 5 and 6, and accordingly these rules are applicable for mode $\Diamond$ only if they satisfy clauses clauses (2.1.1), (3.1), or (4.1.1) of Definitions 9 and 10, based on how $\Diamond$ is instantiated. That is to say, during the execution of the algorithms, we can empty the body of the rules in $R^{B,\Diamond}$ by iteratively proving all the modal literals in the antecedent to decide which rules are applicable at a given step.

Before proceeding with the demonstrations of the lemmas, we recall that in the formalisation of the logic in Section 3, we referred to modes with capital roman letters (X, Y, T) while the notation of the algorithms in Section 4 proposes the variant with $\Box$, $\blacksquare$ and $\Diamond$ since it was needed to fix a given modality for the iterations and pass the correct input for each call of a subroutine. Therefore, being that the hypotheses of the lemmas refer to the operations performed by the algorithms, while the proofs refer to the notation of Definitions 5–15, in the following the former ones use the symbol $\Box$ for a mode, the latter ones the capital roman letters notation.

Lemma 8

Let $D = (F, R, >)$ be a defeasible theory such that $D \vdash +\partial_\Box l$ and $D' = (F, R', >')$ be the theory obtained from D where

$$R' = \{r : A(r) \setminus \{\Box l, \neg\Box\sim l\} \hookrightarrow C(r)|\ r \in R,\ A(r) \cap \widetilde{\Box l} = \emptyset\}$$

$$R'^{B,\Box} = \{r : A(r) \setminus \{\Box l\} \hookrightarrow C(r)|r \in R^{B,\Box},\ A(r) \cap \widetilde{\Box l} = \emptyset\}$$

$$>' = > \setminus\{(r, s), (s, r) \in >|\ A(r) \cap \widetilde{\Box l} \neq \emptyset\}.$$

Then $D \equiv D'$.

Proof

The proof is by induction on the length of a derivation P. For the inductive base, we consider all possible derivations for a literal q in the theory.

$P(1) = +\partial_X q$, *with* $X \in \text{MOD} \setminus \{D\}$. This is possible in two cases: (1) $Xq \in F$, or (2) $\widetilde{Yq} \cap F = \emptyset$, for $Y = X$ or Conflict(Y, X), and $\exists r \in R^X[q, i]$ that is applicable in D for X at i and $P(1)$, and every rule $s \in R^Y[\sim q, j]$ is either (a) discarded for X at j and $P(1)$, or (b) defeated by a stronger rule $t \in R^T[q, k]$ applicable for T at k and $P(1)$ (T may conflict with Y).

Concerning (1), by construction of D', $Xq \in F$ iff $Xq \in F'$, thus if $+\partial_X q$ is provable in D then is provable in D', and vice versa.

Regarding (2), again by construction of D', $\widetilde{Yq} \cap F = \emptyset$ iff $\widetilde{Yq} \cap F' = \emptyset$. Moreover, r is applicable at $P(1)$ iff $i = 1$ (since lemma's operations do not modify the tail of the rules) and $A(r) = \emptyset$. Therefore, if $A(r) = \emptyset$ in D then $A(r) = \emptyset$ in D'. This means that if a rule is applicable in D at $P(1)$ then is applicable in D' at $P(1)$. In the other direction, if r is applicable in D' at $P(1)$, then either (i) $A(r) = \emptyset$ in D, or (ii) $A(r) = \{\Box l\}$, or

$A(r) = \{\neg\Box\sim l\}$. For (i), r is straightforwardly applicable in D, as well as for (ii) since $D \vdash +\partial_\Box l$ by hypothesis.

When we consider possible attacks to rule r, namely $s \in R^Y[\sim q, j]$, we have to analyse cases (a) and (b) above.

(a) Since we reason about $P(1)$, it must be the case that no such rule s exists in R, and thus s cannot be in R' either. In the other direction, the difference between D and D' is that in R we have rules with $\widetilde{\Box}l$ in the antecedent, and such rules are not in R'. Since $D \vdash +\partial_\Box l$ by hypothesis, all rules in R for which there is no counterpart in R' are discarded in D.

(b) We modify the superiority relation by only withdrawing instances where one of the rules is discarded in D. But only when t is applicable then is active in the clauses of the proof conditions where the superiority relation is involved, i.e., (2.3.2) of Definition 13. We have just proved that if a rule is applicable in D then is applicable in D' as well, and if is discarded in D then is discarded in D'. If s is not discarded in D for Y at 1 and $P(1)$, then there exists an applicable rule t in D for q stronger than s. Therefore t is applicable in D' for T and $t >' s$ if $T = Y$, or $\mathrm{Conflict}(T, Y)$. Accordingly, $D' \vdash +\partial_X q$. The same reasoning applies in the other direction. Consequently, if we have a derivation of length 1 of $+\partial q$ in D', then we have a derivation of length 1 of $+\partial q$ in D as well.

Notice that in the inductive base by their own nature rules in $R^{\mathsf{B}, \diamond}$, even if can be modified or erased, cannot be used in a proof of length one.

$P(1) = +\partial_\mathsf{D} q$. The proof is essentially identical to the inductive base for $+\partial_X q$, with some slight modifications dictated by the different proof conditions for $+\partial_\mathsf{D}$: (1) $\mathsf{D}q \in \mathsf{F}$, or (2) $\neg\mathsf{D}q \notin \mathsf{F}$, and $\exists r \in R^\mathsf{D}[q, i]$ that is applicable for D at 1 and $P(1)$ and every rule $s \in R^\mathsf{D}[\sim q, j]$ is either (a) discarded for D at 1 and $P(1)$, or (b) s is not stronger than r.

$P(1) = -\partial_X q$ with $X \in \mathrm{MOD}$. Clearly conditions (1) and (2.1) of Definition 14 hold in D iff they do in D', given that $\mathsf{F} = \mathsf{F}'$. The analysis for clause (2.2) is the same of case (a) of $P(1) = +\partial_X q$, while for clause (2.3.1) the reader is referred to case (2), where in both cases r and s change their role. For condition (2.3.2) if $X = \mathsf{D}$, then $s > r$. Otherwise, either there is no $t \in R^T[q, k]$ in D (we recall that at $P(1)$, t cannot be discarded in D because that would imply a previous step in the proof), or $t \not\succ s$ and not $\mathrm{Conflict}(T, Y)$. Therefore $s \in R'$ by construction, and conditions on the superiority relation between s and t are preserved. Hence, $D' \vdash -\partial_X q$. For the other direction, we have to consider the case of a rule s in R but not in R'. As we have proved above, all rules discarded in D' are discarded in D, and all rules in R for which there is no corresponding rule in R' are discarded in D as well, and we can process this case with the same reasoning as above.

For the inductive step, the property equivalence between D and D' is assumed up to the n-th step of a generic proof for a given literal p.

$P(n + 1) = +\partial_X q$, with $X \in \mathrm{MOD}$. Clauses (1) and (2.1) follow the same conditions treated in the inductive base for $+\partial_X q$. As regards clause (2.2), we distinguish if $X = \mathsf{B}$, or not. In the former case, if there exists a rule $r \in R[q, i]$ applicable for B in D, then clauses 1.–3. of Definition 5 are all satisfied. By inductive hypothesis, we conclude that the clauses are satisfied by r in D' as well no matter whether $\Box l \in A(r)$, or not.

Otherwise, there exists a rule r applicable in D for X at $P(n + 1)$ such that r is either in $R^X[q, i]$, or $R^{B, X}[q, 1]$. By inductive hypothesis, we can conclude that: (i) if $r \in R^X[q, i]$ then r is body-applicable and the clauses of Definition 5 are satisfied by r in D' as well; (ii) if $r \in R^{B, X}[q, 1]$ then r is Conv-applicable and the clauses of Definition 7 are satisfied by r in D' as well. As regards conditions (2.1.2) or (4.1.2), the provability/refutability of the elements in the chain prior to q is given by inductive hypothesis. The direction from rule applicability in D' to rule applicability in D follows the same reasoning and so is straightforward.

Condition (2.3.1) states that every rule $s \in R^Y[\sim q, j] \cup R^{B, Y}[\sim q, 1]$ is discarded in D for X at $P(n + 1)$. This means that there exists an $a \in A(s)$ satisfying one of the clauses of Definition 6 if $s \in R^{B, Y}[\sim q, 1]$, or Definition 10 if $s \in R^Y[\sim q, j]$. Two possible situations arise. If $a \in \widetilde{\Box}l$, then $s \notin R'$; otherwise, by inductive hypothesis, either a satisfies Definition 6 or 8 in D' depending on $s \in R^Y[\sim q, j]$ or $s \in R^{B, Y}[\sim q, 1]$. Hence, s is discarded in D' as well. The same reasoning applies for the other direction. The difference between D and D' is that in R we have rules with elements of $\widetilde{\Box}l$ in the antecedent, and these rules are not in R'. Consequently, if s is discarded in D', then is discarded in D and all rules in R for which there is no corresponding rule in R' are discarded in D since $D \vdash +\partial_\Box l$ by hypothesis.

If $X \neq D$, then condition (2.3.2) can be treated as case (b) of the corresponding inductive base except clause (2.3.2.1) where if $t > s$ then either: (i) $Y = T$, (ii) $s \in R^{B, T}[\sim q]$ and $t \in R^T[q]$ (Convert(Y, T)), or (iii) $s \in R^Y[\sim q]$ and $t \in R^{B, Y}[q]$ (Convert(T, Y)). Instead if $X = D$, no modifications are needed.

$P(n + 1) = -\partial_X q$, *with* $X \in$ MOD. The analysis is a combination of the inductive base for $-\partial_X q$ and inductive step for $+\partial_X q$ where we have already proved that a rule is applicable (discarded) in D iff is so in D' (or it is not contained in R'). Even condition (2.3.2.1) is just the strong negation of the reason in the above paragraph. $\qquad\square$

Lemma 9
Let $D = (F, R, >)$ be a defeasible theory such that $D \vdash -\partial_\Box l$ and $D' = (F, R', >')$ be the theory obtained from D where

$$R' = \{r : A(r) \setminus \{\neg \Box l\} \hookrightarrow C(r) | r \in R, \Box l \notin A(r)\}$$
$$R'^{B, \Box} = \{r \in R^{B, \Box} | \Box l \notin A(r)\}$$
$$>' = > \setminus \{(r, s), (s, r) \in > | \Box l \in A(r)\}.$$

Then $D \equiv D'$.

Proof
We split the proof in two cases, depending on if $\Box \neq D$, or $\Box = D$.

As regards the former case, since Proposition 2 states that $+\partial_X m$ implies $-\partial_X \sim m$ then modifications on R', $R'^{B, \Box}$, and $>'$ represent a particular case of Lemma 8 where $m = \sim l$.

We now analyse the case when $\Box = D$. The analysis is identical to the one shown for the inductive base of Lemma 8 but for what follows.

$P(1) = +\partial_X q$. Case (2)–(ii): $A(r) = \{\neg \Box l\}$ and since $D \vdash -\partial_\Box l$ by hypothesis, then if r is applicable in D' at $P(1)$ then is applicable in D at $P(1)$ as well.

Case (2)–(a): the difference between D and D' is that in R we have rules with $\Box l$ in the

antecedent, and such rules are not in R'. Since $D \vdash -\partial_\Box l$ by hypothesis, all rules in R for which there is no counterpart in R' are discarded in D.

The same modification happens in the inductive step $P(n + 1) = +\partial_X q$, where also the sentence 'If $a \in \widetilde{\Box} l$, then $s \notin R'$' becomes 'If $a = \Box l$, then $s \notin R'$'.

Finally, the inductive base and inductive step for the negative proof tags are identical to ones of the previous lemma. $\qquad\square$

Hereafter we consider theories obtained by the transformations of Lemma 8. This means that all applicable rules are such because their antecedents are empty and every rule in R appears also in R' and vice versa, and there are no modifications in the antecedent of rules.

Lemma 10

Let $D = (F, R, >)$ be a defeasible theory such that $D \vdash +\partial l$ and $D' = (F, R', >)$ be the theory obtained from D where

$$R'^{\mathsf{O}} = \{A(r) \Rightarrow_{\mathsf{O}} C(r)!l \mid r \in R^{\mathsf{O}}[l, n]\} \tag{C1}$$

$$R'^{\mathsf{I}} = \{A(r) \Rightarrow_{\mathsf{I}} C(r)!l \mid r \in R^{\mathsf{I}}[l, n]\} \cup$$

$$\{A(r) \Rightarrow_{\mathsf{I}} C(r) \ominus \sim l \mid r \in R^{\mathsf{I}}[\sim l, n]\} \tag{C2}$$

$$R'^{\mathsf{SI}} = \{A(r) \Rightarrow_{\mathsf{SI}} C(r) \ominus \sim l \mid r \in R^{\mathsf{SI}}[\sim l, n]\}. \tag{C3}$$

Moreover,

- if $D \vdash +\partial_{\mathsf{O}} \sim l$, then instead of (C1)

$$R'^{\mathsf{O}} = \{A(r) \Rightarrow_{\mathsf{O}} C(r)!l \mid r \in R^{\mathsf{O}}[l, n]\} \cup$$

$$\{A(r) \Rightarrow_{\mathsf{O}} C(r) \ominus \sim l \mid r \in R^{\mathsf{O}}[\sim l, n]\}. \tag{C1}$$

- if $D \vdash -\partial_{\mathsf{O}} \sim l$, then instead of (C3)

$$R'^{\mathsf{SI}} = \{A(r) \Rightarrow_{\mathsf{SI}} C(r) \ominus \sim l \mid r \in R^{\mathsf{SI}}[\sim l, n]\} \cup$$

$$\{A(r) \Rightarrow_{\mathsf{SI}} C(r)!l \mid r \in R^{\mathsf{SI}}[l, n]\}. \tag{C3}$$

Then $D \equiv D'$.

Proof

The demonstration follows the inductive base and inductive step of Lemma 8 where we consider the particular case $\Box = \mathsf{B}$. Since here operations to obtain D' modify only the consequent of rules, verifying conditions when a given rule is applicable/discarded reduces to clauses (2.1.2) and (4.1.2) of Definitions 9–10, while conditions for a rule being body-applicable/discarded are trivially treated. Moreover, the analysis is narrowed to modalities O, I, and SI since rules for the other modalities are not affected by the operations of the lemma. Finally, notice that the operations of the lemma do not erase rules from R to R' but it may be the case that, given a rule r, if removal or truncation operate on an element c_k in $C(r)$, then $r \in R[l]$ while $r \notin R'[l]$ for a given literal l (removal of l or truncation at c_k).

$P(1) = +\partial_X q$, *with* $X \in \{\mathsf{O}, \mathsf{I}, \mathsf{SI}\}$. We start by considering condition (2.2) of Definition 13 where a rule $r \in R^X[q, i]$ is applicable in D at $i = 1$ and $P(1)$. In both cases when $q = l$ or $q \neq l$, q is the first element of $C(r)$ since either we truncate chains at l, or we remove $\sim l$ from them. Therefore, r is applicable in D' as well. In the other direction, if r is applicable

35

in D' at 1 and $P(1)$, then $r \in R$ has either q as the first element, or only $\sim l$ precedes q. In the first case r is trivially applicable, while in the second case the applicability of r follows from the hypothesis that $D \vdash +\partial l$ and $D \vdash +\partial_O \sim l$ if $r \in R^O$, or $D \vdash +\partial l$ and $D \vdash -\partial_O \sim l$ if $r \in R^{SI}$.

Concerning condition (2.3.1) of Definition 13 there is no such rule s in R, hence s cannot be in R' (we recall that at $P(1)$, s cannot be discarded in D because that would imply a previous step in the proof). Regarding the other direction, we have to consider the situation where there is a rule $s \in R^Y[\sim q, j]$ which is not in $R'^Y[\sim q]$. This is the case when the truncation has operated on $s \in R^Y[\sim q, j]$ since l preceded $\sim q$ in $C(s)$, making s discarded in D as well (either when (i) $Y = O$ or $Y = I$, or (ii) $D \vdash -\partial_O \sim l$ and $Y = SI$).

For (2.3.2) the reasoning is the same of the equivalent case in Lemma 8 with the additional condition that rule t may be applicable in D' at $P(1)$ but q appears at index 2 in $C(t)$ in D.

$P(n + 1) = +\partial_X q$, with $X \in \{O, I, SI\}$. Again, let us suppose $r \in R[q, i]$ to be applicable in D for X at i and $P(n + 1)$. By hypothesis and clauses (2.1.2) or (4.1.2) of Definition 9, we conclude that $c_k \neq l$ and $q \neq \sim l$ (Conflict(B, I) and Conflict(B, SI)). Thus, r is applicable in D' by inductive hypothesis. The other direction sees $r \in R'[q, i]$ applicable in D' and either $\sim l$ preceded q in $C(r)$ in D, or not. Since in the first case, the corresponding operation of the lemma is the removal of $\sim l$ from $C(r)$, while in the latter case no operations on the consequent are done, the applicability of r in D at $P(n + 1)$ is straightforward.

For condition (2.3.1), the only difference between the inductive base is when there is a rule s in $R^Y[\sim q, j]$ but $s \notin R'^Y[\sim q, k]$. This means that l precedes $\sim q$ in $C(s)$ in D, and thus by hypothesis s is discarded in D. Notice that if $q = l$, then $R'^Y[\sim l, k] = \emptyset$ for any k by the removal operation of the lemma, and thus condition (2.3.1) is vacuously true.

$P(1) = -\partial_X q$ and $P(n + 1) = -\partial_X q$, with $X \in \text{MOD}$. They trivially follow from the inductive base and inductive step. $\qquad\square$

Lemma 11
Let $D = (F, R, >)$ be a defeasible theory such that $D \vdash -\partial l$ and $D' = (F, R', >)$ be the theory obtained from D where

$$R'^I = \{A(r) \Rightarrow_I C(r)!\sim l \mid r \in R^I[\sim l, n]\}.$$

Moreover,

- if $D \vdash +\partial_O l$, then

$$R'^O = \{A(r) \Rightarrow_O C(r) \ominus l \mid r \in R^O[l, n]\};$$

- if $D \vdash -\partial_O l$, then

$$R'^{SI} = \{A(r) \Rightarrow_{SI} C(r)!\sim l \mid r \in R^{SI}[\sim l, n]\}.$$

Then $D \equiv D'$.

Proof
The demonstration is a mere variant of that of Lemma 10 since: (i) Proposition 2 states that $+\partial_X m$ implies $-\partial_X \sim m$ (mode D is not involved), and (ii) operations of the lemma are a subset of those of Lemma 10 where we switch l with $\sim l$, and the other way around. $\qquad\square$

Lemma 12

Let $D = (F, R, >)$ be a defeasible theory such that $D \vdash +\partial_O l$ and $D' = (F, R', >)$ be the theory obtained from D where

$$R'^O = \{A(r) \Rightarrow_O C(r)!{\sim}l \ominus {\sim}l \mid r \in R^O[{\sim}l, n]\} \tag{C1}$$

$$R'^{SI} = \{A(r) \Rightarrow_{SI} C(r) \ominus {\sim}l \mid r \in R^{SI}[{\sim}l, n]\}. \tag{C2}$$

Moreover,

- if $D \vdash -\partial l$, then instead of (C1)

$$R'^O = \{A(r) \Rightarrow_O C(r)!{\sim}l \ominus {\sim}l \mid r \in R^O[{\sim}l, n]\} \cup$$
$$\{A(r) \Rightarrow_O C(r) \ominus l \mid r \in R^O[l, n]\}; \tag{C1}$$

- if $D \vdash -\partial{\sim}l$, then instead of (C2)

$$R'^{SI} = \{A(r) \Rightarrow_{SI} C(r) \ominus {\sim}l \mid r \in R^{SI}[{\sim}l, n]\} \cup$$
$$\{A(r) \Rightarrow_{SI} C(r)!l \mid r \in R^{SI}[l, n]\}. \tag{C2}$$

Then $D \equiv D'$.

Proof

Again, the proof is a variant of that of Lemma 10 that differs only when truncation and removal operate on a consequent at the same time.

A CTD is relevant whenever its elements are proved as obligations. Consequently, if D proves Ol, then $O{\sim}l$ cannot hold. If this is the case, then $O{\sim}l$ cannot be violated and elements following ${\sim}l$ in obligation rules cannot be triggered. Nonetheless, the inductive base and inductive step do not significantly differ from those of Lemma 10. In fact, even operation (1) involving truncation and removal of ${\sim}l$ does not affect the equivalence of conditions for being applicable/discarded between D and D'. $\qquad\square$

Proofs for Lemmas 13–17 are not reported. As stated for Lemma 12, they are variants of that for Lemma 10 where the modifications concern the set of rules on which we operate. The underlying motivation is that truncation and removal operations affect when a rule is applicable/discarded as shown before where we have proved that, given a rule s and a literal ${\sim}q$, it may be the case that ${\sim}q \notin C(s)$ in R' while the opposite holds in R. Such modifications reflect only the nature of the operations of truncation and removal while they do not depend on the mode of the rule involved.

Lemma 13

Let $D = (F, R, >)$ be a defeasible theory such that $D \vdash -\partial_O l$ and $D' = (F, R', >)$ be the theory obtained from D where

$$R'^O = \{A(r) \Rightarrow_O C(r)!l \ominus l \mid r \in R^O[l, n]\}.$$

Moreover,

- if $D \vdash -\partial l$, then

$$R'^{SI} = \{A(r) \Rightarrow_{SI} C(r)!{\sim}l \mid r \in R^{SI}[{\sim}l, n]\}.$$

Then $D \equiv D'$.

Lemma 14

Let $D = (F, R, >)$ be a defeasible theory such that $D \vdash +\partial_D l$, $D \vdash +\partial_D \sim l$, and $D' = (F, R', >)$ be the theory obtained from D where

$$R'^G = \{A(r) \Rightarrow_G C(r)!l \ominus l \mid r \in R^G[l, n]\} \cup$$
$$\{A(r) \Rightarrow_G C(r)!\sim l \ominus \sim l \mid r \in R^G[\sim l, n]\}.$$

Then $D \equiv D'$.

Lemma 15

Let $D = (F, R, >)$ be a defeasible theory such that $D \vdash -\partial_D l$ and $D' = (F, R', >)$ be the theory obtained from D where

$$R'^D = \{A(r) \Rightarrow_D C(r) \ominus l \mid r \in R^D[l, n]\}$$
$$R'^G = \{A(r) \Rightarrow_G C(r) \ominus l \mid r \in R^G[l, n]\}.$$

Then $D \equiv D'$.

Lemma 16

Let $D = (F, R, >)$ be a defeasible theory such that $D \vdash +\partial_X l$, with $X \in \{G, I, SI\}$, and $D' = (F, R', >)$ be the theory obtained from D where

$$R'^X = \{A(r) \Rightarrow_X C(r)!l \mid r \in R^X[l, n]\} \cup$$
$$\{A(r) \Rightarrow_X C(r) \ominus \sim l \mid r \in R^X[\sim l, n]\}.$$

Then $D \equiv D'$.

Lemma 17

Let $D = (F, R, >)$ be a defeasible theory such that $D \vdash -\partial_X l$, with $X \in \{G, I, SI\}$, and $D' = (F, R', >)$ be the theory obtained from D where

$$R'^X = \{A(r) \Rightarrow_X C(r) \ominus l \mid r \in R^X[l, n]\}.$$

Then $D \equiv D'$.

Lemma 18

Let $D = (F, R, >)$ be a defeasible theory and $l \in$ Lit such that (i) $Xl \notin F$, (ii) $\neg Xl \notin F$ and $Y\sim l \notin F$ with $Y = X$ or $\text{Conflict}(Y, X)$, (iii) $\exists r \in R^X[l, 1] \cup R^{B,X}[l, 1]$, (iv) $A(r) = \emptyset$, and (v) $R^X[\sim l] \cup R^{B,X}[\sim l] \cup R^Y[\sim l] \setminus R_{infd} \subseteq r_{inf}$, with $X \in \text{MOD} \setminus \{D\}$. Then $D \vdash +\partial_X l$.

Proof

To prove Xl, Definition 13 must be taken into consideration: since hypothesis (i) falsifies clause (1), then clause (2) must be the case. Let r be a rule that meets the conditions of the lemma. Hypotheses (iii) and (iv) state that r is applicable for X. In particular, if $r = s^\diamond \in R^{B,X}$ then s is Conv-applicable. Finally, for clause (2.3) we have that all rules for $\sim l$ are inferiorly defeated by an appropriate rule with empty antecedent for l, but a rule with empty body is applicable. Consequently, all clauses for proving $+\partial_X$ are satisfied. Thus, $D \vdash +\partial_X l$. $\square$

Lemma 19

Let $D = (F, R, >)$ be a defeasible theory and $l \in$ Lit such that (i) $Dl \notin F$, (ii) $\neg Dl \notin F$, (iii) $\exists r \in R^D[l, 1] \cup R^{B,D}[l, 1]$, (iv) $A(r) = \emptyset$, and (v) $r_{sup} = \emptyset$. Then $D \vdash +\partial_D l$.

Proof

The demonstration is analogous to that for Lemma 18 since all lemma's hypotheses meet clause (2) of Definition 11. □

Lemma 20

Let $D = (F, R, >)$ be a defeasible theory and $l \in$ Lit such that l, $Xl \notin F$ and $R^X[l] \cup R^{B,X}[l] = \emptyset$, with $X \in$ MOD. Then $D \vdash -\partial_X l$.

Proof

Conditions (1) and (2.2) of Definitions 12 and 14 are vacuously satisfied with the same comment for $R^{B,X}$ in Lemma 18. □

Lemma 21

Let $D = (F, R, >)$ be a defeasible theory and $l \in$ Lit such that (i) $X \sim l \notin F$, (ii) $\neg X \sim l \notin F$ and $Yl \notin F$ with $Y = X$ or Conflict(Y, X), (iii) $\exists r \in R^X[l, 1] \cup R^{B,X}[l, 1]$, (iv) $A(r) = \emptyset$, and (v) $r_{sup} = \emptyset$, with $X \in$ MOD. Then $D \vdash -\partial_X \sim l$.

Proof

Let r be a rule in a theory D for which the conditions of the lemma hold. It is easy to verify that clauses (1) and (2.3) of Definitions 12 and 14 are satisfied for $\sim l$. □

Theorem 4

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

Proof

Every time Algorithms 2 Proved or 3 Refuted are invoked, they both modify a subset of the set of rules R, which is finite by hypothesis. Consequently, we have their termination. Moreover, since $|R| \in O(S)$ and each rule can be accessed in constant time, we obtain that 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 $O(S)$.

Proof

The most important part to analyse concerning termination of Algorithm 1 DefeasibleExtension is the **repeat/until** cycle at lines 12–37. Once an instance of the cycle has been performed, we are in one of the following, mutually exclusive situations:

1. No modification of the extension has occurred. In this case, line 37 ensures the termination of the algorithm;
2. The theory has been modified with respect to a literal in HB. Notice that the algorithm takes care of removing the literal from HB once the suitable operations have been performed (specifically, at line 3 of Algorithm 2 Proved and 3 Refuted). Since this set is finite, the process described above eventually empties HB and, at the next iteration of the cycle, the extension of the theory cannot be modified. In this case, the algorithm ends its execution as well.

Moreover, Lemma 4 proved the termination of its internal sub-routines.

In order to analyse complexity of the algorithm, it is of the utmost importance to

correctly comprehend Definition 19. Remember that the size of a theory is the number of *all occurrences* of each literal in every rule plus the number of the rules. The first term is usually (much) bigger than the latter. Let us examine a theory with x literals and whose size is S, and consider the scenario when an algorithm A, looping over all x literals of the theory, invokes an inner procedure P which selectively deletes a literal given as input from all the rules of the theory (no matter to what end). A rough computational complexity would be $O(S^2)$, given that, when one of the $x \in O(S)$ literal is selected, P removes all its occurrences from every rule, again $O(S)$.

However, a more fined-grained analysis shows that the complexity of A is lower. The mistake being to consider the complexity of P separately from the complexity of the external loop, while instead they are strictly dependent. Indeed, the overall number of operations made by the sum of all loop iterations cannot outrun the number of occurrences of the literals, $O(S)$, because the operations in the inner procedure directly decrease, iteration after iteration, the number of the remaining repetitions of the outmost loop, and the other way around. Therefore, the overall complexity is not bound by $O(S) \cdot O(S) = O(S^2)$, but by $O(S) + O(S) = O(S)$.

We can now contextualise the above reasoning to Algorithm 1 DEFEASIBLEEXTENSION, where D is the theory with size S. The initialisation steps (lines 1–5 and 10–11) add an $O(S)$ factor to the overall complexity. The main cycle at lines 12–37 is iterated over HB, whose cardinality is in $O(S)$. The analysis of the preceding paragraph implies that invoking Algorithm 2 PROVED at lines 7 and 29 as well as invoking Algorithm 3 REFUTED at lines 8, 15, 26 and 27 represent an additive factor $O(S)$ to the complexity of **repeat/until** loop and **for** cycle at lines 6–9 as well. Finally, all operations on the set of rules and the superiority relation require constant time, given the implementation of data structures proposed. Therefore, we can state that the complexity of the algorithm is $O(S)$. □

Theorem 6

Algorithm 1 DEFEASIBLEEXTENSION is sound and complete.

Proof

As already argued at the beginning of the section, the aim of Algorithm 1 DEFEASIBLEEXTENSION is to compute the defeasible extension of a given defeasible theory D through successive transformations on the set of facts and rules, and on the superiority relation: at each step, they compute a simpler theory while retaining the same extension. Again, we remark that the word 'simpler' is used to denote a theory with fewer elements in it. Since we have already proved the termination of the algorithm, it eventually comes to a fixed-point theory where no more operations can be made.

In order to demonstrate the soundness of Algorithm 1 DEFEASIBLEEXTENSION, we show in the list below that all the operations performed by the algorithm are justified by Proposition 2 and described in Lemmas 7–21, where we prove the soundness of each operation involved.

1. Algorithm 1 DEFEASIBLEEXTENSION:

   - Lines 2–3 and 5: Lemma 7;
   - Line 7: item 2. below;
   - Line 8: item 3. below;
   - Line 15: Lemma 20 and item 3. below;

- Line 24: Lemma 19 and item 2. below;
- Lines 26–27: Lemma 21 and item 3. below;
- Line 29: Lemma 18 and item 2. below;

2. Algorithm 2 PROVED:

  - Line 4: Lemma 21 and item 3. below;
  - Line 5: Part 2. of Proposition 2 and item 3. below;
  - Line 6: Part 3. of Proposition 2 and item 3. below;
  - Lines 7–9: Lemma 8;
  - CASE B at lines 11–14: Lemma 10;
  - CASE O at lines 15–18: Lemma 12;
  - CASE D at lines 19–23: Lemma 14;
  - OTHERWISE at lines 24–26: Lemma 16;

3. Algorithm 3 REFUTED:

  - Lines 4–6: Lemma 9;
  - CASE B at lines 8–11: Lemma 11;
  - CASE O at lines 12–14: Lemma 13;
  - CASE D at lines 15–16: Lemma 15;
  - OTHERWISE at lines 17–18: Lemma 17;

The result of these lemmas is that whether a literal is defeasibly proved or not in the initial theory, so it will be in the final theory. This proves the soundness of the algorithm.

Moreover, since (i) all lemmas show the equivalence of the two theories, and (ii) the equivalence relation is a bijection, this also demonstrates the completeness of Algorithm 1 DEFEASIBLEEXTENSION. □