

# BIO Logical Agents: Norms, Beliefs, Intentions in Defeasible Logic

Guido Governatori and Antonino Rotolo

## Abstract

In this paper we follow the BOID (Belief, Obligation, Intention, Desire) architecture to describe agents and agent types in Defeasible Logic. We argue, in particular, that the introduction of obligations can provide a new reading of the concepts of intention and intentionality. Then we examine the notion of social agent (i.e., an agent where obligations prevail over intentions) and discuss some computational and philosophical issues related to it. We show that the notion of social agent either requires more complex computations or has some philosophical drawbacks.

## 1 Introduction and Motivation

Reasoning about mental attitudes is a traditional issue in philosophy and has been widely investigated in the field of AI. Some classical agent systems based on mental attitudes such as beliefs, desires and intentions are, for example, those presented in [6, 9, 33].

More recent works on cognitive agents tried combine two apparently independent perspectives [8, 18, 11, 10, 12, 13]: (a) a classical cognitive account of agents that specifies their mental attitudes; (b) modelling agents' behaviour by means of normative concepts. For the first approach, the background is basically the belief-desire-intention (BDI) architecture, where mental attitudes are taken as primitives to give rise to a set of Intentional Agent Systems [33, 6]. This view is interesting especially when the behaviour of agents is the outcome of a rational balance among their (possibly conflicting) mental states. The normative aspect is rather based on the assumption that normative concepts play a role to characterise the idea of social co-ordination of autonomous agents [32]. The nice result of this combination of perspectives is that of leading to an account of agents' deliberation and behaviour in terms of the interplay between mental attitudes and normative (external) factors such as obligations.

A crucial aspect in this recent trend is that reasoning about agents can be embedded in frameworks based on non-monotonic logics, as one the most interesting problems concerns the cases where the agent's mental attitudes are in conflict or when they are incompatible with obligations and other deontic provisions. In this specific perspective, the relation between mental attitudes and non-monotonicity should not sound surprising: works such as Thomason's [36] and on BOID [8] confirm this trend. Of particular interest is the BOID architecture, which in fact provides a number of strategies for solving conflicts among mental attitudes and obligations. BOID specifies logical criteria (i)

to retract agent’s attitudes with the changing environment, and so (ii) to settle conflicts by stating different general policies corresponding to the agent type considered. Agent types correspond to the different ways through which conflicts are detected and solved: a realistic agent thus corresponds to a conflict-resolution type in which beliefs override all other factors, while other agent types, such as simple-minded, selfish or social ones adopt different orders of overruling.

Following [18, 11, 10], in this paper we take advantage of this research line and discuss how the combination of mental attitudes and obligations can be framed in Defeasible Logic (DL). As is well-known, DL is based on a logic programming-like language and it is a simple, efficient but flexible non-monotonic formalism able to deal with many different intuitions of non-monotonic reasoning and recently applied in many fields. In addition, several efficient implementations have been developed [25, 4]. Here we discuss and extend some aspects of a non-monotonic logic of agency, based on the framework of [2], developed in [11, 10].

Why DL? Indeed, DL is one of the most expressive languages that allows for the definition of large sets of agent types. In particular, the aim of this article is to address the following issues:

1. We will devise an extension of DL able to cover a number of different agent types, but which, despite its expressiveness, is computationally feasible. We will prove that it is possible to compute the complete set of consequences of a given theory in linear time, thus preserving the nice computational features of standard DL.
2. On the other hand, we will argue that the notion of agent type can be problematic. The discussion will be devoted to some philosophical and computational aspects of the notion of “social agent”, by which we mean a norm-complying agent<sup>1</sup>. However, we will argue that similar considerations also apply to other agent types.

Our system, which considers here three components –Beliefs, Intentions, and Obligations (BIO agents)– has some substantial peculiarities that make it different from other frameworks such as BOID’s<sup>2</sup>. In particular,

- the system develops a constructive account of those modalities that correspond to mental states and obligations; rules are thus meant to devise suitable logical

---

<sup>1</sup>The term “social agent” is taken from previous literature, and in particular from works on the BOID architecture and some of our earlier papers. We preferred not to change this terminology to avoid confusion. Other terms could be used, such as “respectful” or “obedient”. However, “social” does not have here any moral connotations and does not imply that a norm-complying agent has some positive attitude towards others. (Of course this may be the case, but we do not necessarily suggest this intuitive reading for norm-complying agents.) The term “social” simply stresses the contrast with other agent types, such as “selfish agent” (internal vs. external motivations). Analogously, the term “deviant” does not have, too, any moral connotation: it is taken from social sciences to denote precisely what we describe in our paper.

<sup>2</sup>The choice of excluding desires is only motivated by offering a simpler presentation of the logic. Social agents, according to the previous literature (e.g., [8, 11, 10]), are minimally those for which obligations override conflicting intentions. The inclusion of desires would not substantially change the inference mechanism presented in Section 3. In addition, adding desires would not affect the computational results presented in the paper. All results applicable to the relation between intentions and obligations hold for a similar relation between obligations and desires.

conditions for introducing modalities; if so, rules may also contain modalised literals;

- possible conversions of a modality into another can be accepted, as when the applicability of rule leading to derive, for example,  $OBL p$  ( $p$  is obligatory) may permit, under appropriate conditions, to obtain  $INT p$  ( $p$  is intended).

We believe that both these aspects are necessary to account for some relatively simple, but important reasoning patterns. In particular, we maintain that conversions are required to capture some aspects of agents' rationality. In fact, conversions permit to derive, for example, intentions from beliefs. As we shall see, these reasoning patterns are suitable for modelling the so-called side-effect problem. Usually, side effects are not considered as a part of the intentional sphere of agents, but this analysis is not always satisfactory when we have to check effects against obligations and possible normative violations. In presence of obligations regulating agent's behaviour, a satisfactory model for agent's rationality should suggest that some side effects (but not all) are intended. Indeed, conversions correspond to an inferential mechanism that is precisely meant to capture this fact. But, at this point, the conclusion will be that the notion of social agent gets problematic (point 2 above).

The layout of the paper is as follows. Section 2 provides the theoretical background of our system. In particular, since the concept of social agent focuses on the interplay between obligations and intentions, we will discuss which kind of intentions have to be considered in this regard. Section 3 will present our logical framework, based on DL, which will embed our intuitions and permit to deal with BIO agents. Section 4 presents a first discussion of the notion of agent type; in particular, we will argue that conversions, too, are relevant in identifying specific cognitive profiles for the agents; the section ends with an open problem concerning the feasibility of agent types based on the strategies for solving conflicts. Section 5 deals with the computational complexity of social agency. A concluding section on related work completes the paper.

## 2 Norms, Beliefs and Intentions

The focus of this paper is on the so-called policy-based attitudes. The term was coined by Bratman [7] with specific reference to the idea of intention. For example, I have a policy to patch up and reboot the Unix server in the department once every month. This morning, on the basis of this policy, I form the intention to reboot the machine at 7.00 PM in the evening. My intention this morning to reboot the machine this evening is a *policy-based* intention. This specific intention will play a major part in my planning process for the day, as it will pose problems about means and constrain my other options.

Hence, intentions of this type concern potentially recurring circumstances in an agent's life. A policy-based intention is such that it is not simply a case of retaining an intention previously formed. Neither is it based on a full-blown deliberation where an attempt is made to weigh pros and cons for and against conflicting options. It also differs from an intention in favour of necessary means, i.e., an intention in favour of a specific end, in the sense that the defeasibility of general policies makes it possible to

*block* the application of the policy to the particular case without *abandoning* the policy. Otherwise one could abandon the intention in favour of the end. The peculiarity policy-based intentions is that in each case the policy concerns a kind of circumstance that is expected to recur in the agent loop and in each case the agent might well have a general intention to act in the particular circumstances. Whether the agent is able to perform that action or not depends on the circumstances.

As argued in detail elsewhere [16], it may happen that a policy-based intention needs to be *re-considered* if not *blocked* for the application to particular cases. But this does not mean that the agent should know all such conditions in a scenario, but only those she considers necessary for the intended outcome and that she is not confident of their being satisfied. To intend the necessary consequence the agent has to make sure that all the evidence to the contrary has been defeated, which is basically a defeasible conclusion.

The starting point of this paper is to extend the policy-based approach to other attitudes and motivational factors such as beliefs and obligations (see [20] for a similar idea). In this way, all motivational factors are naturally represented within a rule-based system: intentions and beliefs are viewed as constituting the internal constraints (based on policies) of an agent while obligations are her external constraints (based on rules). As constraints they are defeasible. Notice, in particular, that such an extension to obligations can capture the well-known defeasible character of deontic reasoning. In this last case, a policy-based obligation –conceived of as an external motivational attitude– turns out to be simply a conditional obligation, namely, a rule that allows for the inference of an obligation whenever the antecedent of this rule holds [30, 34].

## 2.1 Expected Side Effects and Agents' Rationality

As we mentioned in Section 1, a satisfactory model for agent's rationality sometimes requires that a side effect should be intended, even though we cannot properly say it was directly wanted by the agent.

It is quite common to distinguish between actions performed intentionally and unintentionally. But it is philosophically hard to explain what the distinction precisely amounts to. For instance, some philosophers argued that an action is not properly intentional if the agent does not have the intention to perform it [1, 26]. On the other hand, it is somehow reasonable to say that an agent, who did not *specifically* intend to perform an action, intentionally performed it. This idea is more clear when we just have a look at the philosophical debate on the problem of side effects. Imagine an agent does *A* to achieve *B* and knows that *A* will also produce some other result *C*. If the agent's motivation is only the desire to obtain *B*, can we say that the agent's action was fully intentional with respect to *C*? Some philosophers argued that, insofar as the side effect is viewed as dangerous or somehow unpleasant ("it is harming the environment"), people are psychologically inclined to think that the agent acted intentionally; when the side effect is "helping the environment", most people are inclined to think the contrary ([21]; for a general discussion see, e.g., [27, 35]). Of course, it is outside the scope of this paper to provide the reader with a philosophical answer to these thorny questions. But there is a lesson that we can learn from this debate: we may sometimes have good reasons to include in the intentional sphere of agents some side effects. In

the remainder, we will not offer substantial criteria to establish what effects should be accepted as intentional (a problem for which no real consensus has emerged among philosophers). We will simply offer a logical analysis of this idea and justify the inclusion of side effects when an agent is acting in an environment where obligations regulate her behaviour.

Let us examine Michael Bratman's idea of agent's rationality. Bratman plays an important role in our case: Not only he was the first who introduced the notion of policy-based intention, but, also, his theory is traditionally considered as one of the main philosophical references for modelling cognitive agents in MAS. As is well known, Bratman admits that, in some cases, an action is intentionally performed even though the agent did not specifically intend to perform it. On the other hand, according to him, rational agents can be basically modelled as follows [7]:

- agents are goal-directed without being necessarily aware of their activity;
- intentions are used to choose partial plans for the realisation of a goal;
- not all consequences are intended but only some initial intentions and the goal as a result of the plan; if some side-effects occur, they are never intended.

According to this view, side effects should be in principle excluded from the intentional sphere of goal-directed agents. From the logical point of view, this idea makes it necessary to avoid several variants of logical omniscience: omniscience arises when the agent is required to know all the truths defined by her logic, or when the logic that depicts the agent automatically includes all the logical truths of classical logic, or, finally, if the agent knows all the logical consequences of the known propositions [15]. In this perspective, the *expected side-effects* problem seems to depend on the interactions between the reasoning mechanism for the propositional inferences and the mechanism ruling the introduction and the behaviour of the modal operators representing mental states. A simple and rather unsatisfactory solution would be to consider two completely unrelated consequence relations, one for the propositional part and the second one for the modal operators. The consequence relation for a modal operator is meant to give the condition under which one can prove a modal formula. For example the pair  $\Gamma \sim_X \alpha$ , where  $X$  is a modal operator, means that if we can prove all the formulas in  $\Gamma$  then we can deduce  $X\alpha$ . In what follows we will develop a system for mental states and motivational attitudes based on this idea. However, we will allow the consequence relation for intentions and obligations to interact with the propositional module and we will also consider possible interactions between the modal operators. To this end we have to show that the expected side-effects phenomenon is not a drawback for policy-based agents: such a kind of agents must accept the expected-side effects unless they have some reasons to reject the consequences corresponding to them.

In effect, though our proposed theory does not entertain many of the properties leading to logical omniscience, some aspects of the side-effects problem are accepted. Consider

$$\text{INTGoToDentist}, \text{GoToDentist} \Rightarrow \text{Pain} \vdash \sim \text{INTPain} \quad (1)$$

$$\text{INTGoToRome}, \text{GoToRome} \Rightarrow \text{GoToItaly} \vdash \sim \text{INTGoToItaly} \quad (2)$$

The first inference says that, if the agent intends to go to the dentist, and going to the dentist will cause pain, then the agent intends to have pain. The second inference states that, if the agent intends to go to Rome, and Rome is in Italy, then the agent intends to go to Italy. Actually, whereas the first case is clearly unacceptable, the second should be accepted by a rational agent. In this perspective the side-effects problem is similar to the substitution of indiscernible in opaque contexts. An agent may have the intention to visit Rome and not to visit Italy. But if the agent knows that Rome is the capital of Italy then it would be irrational for the agent not to have the intention to go to Italy given the intention to visit Rome.

Accordingly, some cases of the side-effects problem are not necessarily a weakness of a theory. This holds in particular if we assume that our agents are *aware* of their activities, which means that they know (or believe in) the policies regulating their own deliberation. This implies that awareness is nothing but a form of epistemic introspection: accordingly, the example above can be reframed as follows:

$$\text{INTGoToDentist}, \text{BEL}(\text{GoToDentist} \Rightarrow \text{Pain}) \sim \text{INTPain} \quad (3)$$

$$\text{INTGoToRome}, \text{BEL}(\text{GoToRome} \Rightarrow \text{GoToItaly}) \sim \text{INTGoToItaly} \quad (4)$$

In our view, modelling rational agents corresponds to the following assumptions:

- agents are aware of their activities, of their policies;
- some cases of the side-effects problem can be accepted;
- if a case has to be rejected this means that some of its consequences should not be intended;
- when some consequences are not intended, this only means that they are blocked by conflicting attitudes or facts.

The theory an agent is equipped with can be understood as the specification of the behaviour of the agent. If the agent is *aware* that  $B$  is an unavoidable/indisputable consequence of  $A$  and the agent intends  $A$ , then  $B$  is a consequence of the agent's intentions and the agent must accept it as part of her intentions. Suppose we have that "raising one's hand at an auction counts as making a bid". Thus if the agent (aware of this policy) intends to raise her hand, then she intends to bid in the auction, and her action will be understood as making a bid. In other words, in our system we will try to balance and moderate some unpleasant aspects of the side-effects problem with the equally important need for modelling rational agents. Of course, according to our view, we may have that something is intended even if it is causally distant with respect to the original derived intentions. But this is not necessarily a drawback if we conceive agents as rational and, as such, being aware of the policies which are related with the environment and with their interests: even a causally distant behaviour can be rationally intended unless it is removed in the meantime from deliberation. But this case is indeed considered within our analysis because we may have concrete contexts in which some policy-based intentions, as soon as they are applicable, turn out to be overridden by other policies: we may have reasons to argue that, if an agent intends  $A$  and believes

that  $B$  is a consequence of  $A$ , this is not a reason for necessarily intending  $B$ ; in fact, the derivation of  $B$  as an intention may be blocked, in our view, by competing attitudes or made non-applicable by concrete facts.

According to the previous discussion it should be clear that, though inspired by Bratman's [7] analysis, the notion of intention we study in this paper is slightly different, as it focuses on the idea of *intentionality*. In Bratman's view intentions are used to choose partial plans for the realisation of a goal; in this way they have a close relation to means-ends. In our view intentions should be related not only to means-ends but also to their consequences.

This concept of intention is particularly relevant in conjunction with deontic and normative notions, for example if we want to say that an agent is legally or morally responsible for  $A$  if the agent did  $A$  with the intention to do  $A$ . In other words, that an agent can be qualified as responsible for a normative violation (i.e., to act in contrast with some normative provisions) requires that agent's behaviour is performed intentionally (see the discussion in [7, 27]). Of course, we may identify different degrees of responsibility by checking, for instance, whether the agent has a direct, or only an indirect, control over the consequences of her actions. But, in the legal perspective, in particular, responsibility is usually associated with some minimal form of intentionality: only in case of an agent acting completely unintentionally, the law usually excludes that the agent is responsible. In other words, we have the following options:

$$Th = \{\text{INT}a, \text{BEL}(a \Rightarrow b), \text{OBL}\neg b\}$$

**responsible**

$$Th \vdash \text{INT}b$$

**not responsible**

$$Th \not\vdash \text{INT}b$$

In one option the agent is not responsible, as we do not derive any conflicting intention with respect to the prohibition to do  $b$ ; in the other case, the agent is responsible, as she intends to do  $b$ . On account of this logical analysis, since we have sometimes to accept some side effects, the agent has to include in the set of her intentions not only her intentions in Bratman's sense but also some of their consequences. It is worth noting that our intuition is compatible with von Wright's [37] classical theory of normative actions. Von Wright's problem is to identify what should be the content of norms. He argues that norms should deal with actions. Roughly, actions can be described in terms of state transitions and as the sets of all changes of world that follow from them. It is not our purpose discussing here von Wright's theory of action. It should be noted, however, that he considers the related problem of intentions. On the one hand, von Wright is clear when he says that any action may have an arbitrary number of consequences and not all of them are intended. On the other hand, he provides a very broad concept of action, according to which all actions in norms, strictly speaking, are intentional. If so, what are the boundaries of intentions to be considered when they interplay with obligations?

Let us see how to recast Bratman's Strategic Bomber scenario [7] in this perspective. The basic scenario runs as follows: Strategic Bomber intends to bomb a munition plant of the enemy being aware that the resulting explosion will kill innocent children in a nearby school. Bratman argues that Strategic Bomber does not have the intention

to kill the children. Formally, Bratman’s scenario can be represented as follows: if

$$Th = \{INTbomb, BEL(bomb \Rightarrow kill), INTbomb \Rightarrow INT\neg kill, OBL\neg kill\}$$

then

$$Th \not\vdash INTkill$$

Let us expand the scenario by supposing that despite the bombing, Strategic Bomber loses the war, and that there is a process for war crimes against him. Civil casualties are a sad but almost unavoidable consequence of war, but usually the killing of civilians does not constitute a war crime if there was no intention to kill. According to Bratman, Strategic Bomber did not commit a war crime since he did not have such an intention (see above). However, let us assume that Strategic Bomber did not do anything to prevent or minimise civil casualties (let us say by a movement of troops that might have resulted in an evacuation of the area surrounding the munition plant). In this extended scenario the killing of children is brought about by a (successful) intentional act of Strategic Bomber. Accordingly, he must be held responsible for the killing of innocent civilians.

Formally, this means that the intention to minimise civil casualties is a condition for not obtaining the intention to kill innocents and, conversely, intending not to minimise casualties permits to derive the intention to kill. Hence, if

$$Th' = \{INTbomb, INT\neg minimise, OBL\neg kill, BEL(\neg minimise \Rightarrow kill), INTbomb, INTbomb \Rightarrow INT\neg kill\}$$

then

$$Th' \vdash INTkill$$

Clearly, under this reading we assume that we can derive the intention to kill from the fact that the agent intended not to minimise civil casualties. Accordingly, the agent is responsible for the killing. Note that this can be captured only if we adopt defeasible reasoning:  $Th'$  would prima facie also imply  $INT\neg kill$ , a conclusion which should be blocked in this context. Given this interpretation of intentions, we will see in the rest of this paper that some standard accounts of agent types, and of social agents in particular, are not satisfactory.

### 3 BIO Agents in Defeasible Logic

#### 3.1 Basics of Defeasible Logic

Defeasible Logic (DL) was originally proposed by Nute [29, 28] with a particular concern about computational efficiency and developed over the years notably by [5, 3, 2]. DL is suitable for implementations [24], is flexible [2] (it has a constructively defined and easy to use proof theory), and it is modular [3] (it can be easily extended to cover different logical components: besides the current contribution, see, e.g., [11, 10]). In addition, DL is efficient: it is possible to compute the complete set of consequences of a given theory in linear time [23]. As we will see, this result also applies to the logical framework presented in this paper.

Knowledge in DL can be represented in two ways: facts and rules.

*Facts* are indisputable statements and are represented by predicates. We only use a propositional language. Facts containing free variables are interpreted as the set of their variable-free instances. For example, “the price of the spam filter is \$50” is represented by  $Price(SpamFilter, 50)$ .

A *rule*, on the other hand, describes the relationship between a set of literals (premises) and a literal (conclusion), and we can specify how strong the relationship is. As usual, rules allow us to derive new conclusions given a set of premises. As far as the strength of rules is concerned we distinguish between *strict rules*, *defeasible rules* and *defeaters*.

Strict rules, defeasible rules and defeaters are represented, respectively, by expressions of the form  $A_1, \dots, A_n \rightarrow B$ ,  $A_1, \dots, A_n \Rightarrow B$  and  $A_1, \dots, A_n \rightsquigarrow B$ , where  $\{A_1, \dots, A_n\}$  is a possibly empty set of prerequisites and  $B$  is the conclusion of the rule<sup>3</sup>. We only consider rules that are essentially propositional. Rules containing free variables are interpreted as the set of their ground instances.

*Strict rules* are rules in the classical sense: whenever the premises are indisputable then so is the conclusion. Thus they can be used for definitional clauses. An example of a strict rule is “A ‘Premium Customer’ is a customer who has spent \$10,000 on goods”:

$$TotalExpense(x, 10000) \rightarrow PremiumCustomer(x).$$

*Defeasible rules* are rules that can be defeated by contrary evidence. An example of such a rule is “Premium Customer are entitled to a 5% discount”:

$$PremiumCustomer(x) \Rightarrow Discount(x).$$

The idea is that if we know that someone is a Premium Customer, then we may conclude that she is entitled to a discount *unless there is other evidence suggesting that she may not be* (for example if she buys a good in promotion).

*Defeaters* are a special kind of rules. They are used to prevent conclusions not to support them. For example:

$$SpecialOrder(x), PremiumCustomer(x) \rightsquigarrow \neg Surcharge(x).$$

This rule states that premium customers placing special orders might be exempt from the special order surcharge. This rule can prevent the derivation of a “surcharge” conclusion. On the other hand it cannot be used to support a “not surcharge” conclusion.

DL is a “skeptical” non-monotonic logic, meaning that it does not support contradictory conclusions. Instead DL seeks to resolve conflicts. In cases where there is some support for concluding  $A$  but also support for concluding  $\neg A$ , DL does not conclude either of them (thus the name “skeptical”). If the support for  $A$  has priority over the support for  $\neg A$  then  $A$  is concluded.

As we have alluded to above, no conclusion can be drawn from conflicting rules in DL unless these rules are prioritised. The *superiority relation* among rules is used to

---

<sup>3</sup>We will drop set notation for the antecedents of rules

define priorities among rules, that is, where one rule may override the conclusion of another rule. For example, given the defeasible rules

$$\begin{aligned} r &: \text{PremiumCustomer}(x) \Rightarrow \text{Discount}(x) \\ r' &: \text{SpecialOrder}(x) \Rightarrow \neg \text{Discount}(x) \end{aligned}$$

which contradict one another, no conclusive decision can be made about whether a Premium Customer who has placed a special order is entitled to the 5% discount. But if we introduce a superiority relation  $>$  with  $r' > r$ , then we can indeed conclude that special orders are not subject to discount.

Informally, conclusions can be drawn in DL according to the following intuition. Let  $D$  be a theory in DL (i.e., a collection of facts, rules and a superiority relation over the set of rules). A *conclusion* of  $D$  is a tagged literal and can have one of the following four forms:

- $+\Delta q$  meaning that  $q$  is definitely provable in  $D$  (i.e., using only facts and strict rules).
- $-\Delta q$  meaning that we have proved that  $q$  is not definitely provable in  $D$ .
- $+\partial q$  meaning that  $q$  is defeasibly provable in  $D$ .
- $-\partial q$  meaning that we have proved that  $q$  is not defeasibly provable in  $D$ .

Strict derivations are obtained by forward chaining of strict rules, while a defeasible conclusion  $p$  can be derived if there is a rule whose conclusion is  $p$ , whose prerequisites (antecedent) have either already been proved or given in the case at hand (i.e., facts), and any stronger rule whose conclusion is  $\neg p$  has prerequisites that fail to be derived. In other words, a conclusion  $p$  is derivable when:

- $p$  is a fact; or
- there is an applicable strict or defeasible rule for  $p$ , and either
  - all the rules for  $\neg p$  are discarded (i.e., not applicable) or
  - every applicable rule for  $\neg p$  is weaker than an applicable strict<sup>4</sup> or defeasible rule for  $p$ .

In the next sections we will see how the basic machinery of DL can be extended to deal with the multi-modal logic required to model BIO agents.

### 3.2 Modal Defeasible Logic

Our purpose is to account for policy-based motivations of BIO agents, which requires to capture at least some basic facets of the modal notions of belief, intention, and obligation.

Usually modal logics are extensions of classical propositional logic with some intensional operators. Thus any modal logic should account for two components: (1) the

<sup>4</sup>Notice that a strict rule can be defeated only when its antecedent is defeasibly provable.

underlying logical structure of the propositional base and (2) the logic behaviour of the modal operators. Alas, as is well-known, classical propositional logic is not well suited to deal with real life scenarios. The main reason is that the descriptions of real-life cases are, very often, partial and somewhat unreliable. Our discussion in Section 2 is in line with this intuition as far as agents' motivational attitudes are concerned. Accordingly, in such circumstances classical propositional logic might produce counterintuitive results insofar as it requires complete, consistent and reliable information. Hence any modal logic based on classical propositional logic is doomed to suffer from the same problems.

On the other hand the logic should specify how modalities can be introduced and manipulated. Some common rules for modalities are, e.g.,

$$\frac{\vdash \varphi}{\vdash \Box \varphi} \text{ Necessitation} \quad \frac{\vdash \varphi \supset \psi}{\vdash \Box \varphi \supset \Box \psi} \text{ RM}$$

Both dictates conditions to introduce modalities based purely on the derivability and structure of the antecedent. These inference rules are related to the problem of logical omniscience: if  $\Box$  corresponds either to INT, BEL, or OBL, they put unrealistic assumptions on the cognitive capabilities of an agent. In effect, although some aspects of the expected side-effects problem should be accepted in modelling rational agents, rules such as Necessitation and RM are clearly too demanding: both in general permit to derive that an agent believes or intends something, or that something is obligatory for her, assuming that she knows all the truths defined by her logic, or that the logic that depicts her behaviour automatically includes all the logical truths of classical logic, or that she knows all the logical consequences of known propositions.

The point is thus avoid these difficulties by only admitting the side effects for which no contrary reason can be advanced. Our strategy is twofold. First, we take a constructive interpretation of  $\Box$ : we have that if an agent can build a derivation of  $\varphi$  then she can build a derivation of  $\Box \varphi$ . We want to maintain this intuition, but also to replace derivability in classical logic with a practical and feasible notion like derivability in DL. Thus the intuition behind this work is that we are allowed to derive  $\Box p$  if we can prove  $p$  with the mode  $\Box$  in DL.

To extend DL with modal operators we have two options: 1) to use the same inferential mechanism as basic DL and to represent explicitly the modal operators in the conclusion of rules [31]; 2) introduce new types of rules for the modal operators to differentiate between modal and factual rules.

For example the “deontic” statement “The Purchaser shall follow the Supplier price lists” can be represented as

$$\text{AdvertisedPrice}(X, Y) \Rightarrow \text{OBL}_{\text{purchaser}} \text{Pay}(X, Y)$$

if we follow the first option and

$$\text{AdvertisedPrice}(X, Y) \Rightarrow_{\text{OBL}_{\text{purchaser}}} \text{Pay}(X, Y)$$

according to the second option, where  $\Rightarrow_{\text{OBL}_{\text{purchaser}}}$  denotes a new type of defeasible rule relative to the modal operator  $\text{OBL}_{\text{purchaser}}$ . In both cases the meaning of the rule is

that given that the price of the item  $X$  advertised by the supplier is  $Y$ , then the purchaser has the obligation to pay  $Y$  for item  $X$ .

The differences between the two approaches, besides the fact that in the first approach there is only one type of rules while the second accounts for factual and modal rules, is that the first approach has to introduce an additional machinery for introducing and reasoning with modal operators. Hence, explicitly representing the modal operators in the conclusion of rules does not follow the basic intuition we have suggested above. In fact, in this case we would have to provide a definition of  $p$ -incompatible literals (i.e., a set of literals that cannot be hold when  $p$  holds.) for every literal  $p$ . For example we can have a modal logic where  $\Box p$  and  $\neg p$  cannot be both true at the same time. Moreover the first approach is less flexible than the second: in particular in some cases it must account for rules to derive  $\Diamond p$  from  $\Box p$ ; similarly conversions –which permit to use a rule for a certain modality as it were for another modality (see *infra*)– require additional operational rules in a theory, thus the second approach seems to offer a more conceptual tool than the first one. It seems that the second approach can use different proof conditions based on the modal rules to offer a more fine grained control over the modal operators and it allows for interaction between modal operators.

If we label the arrows of the rules (i.e., agent’s policies) of our rule-based system by the different modalities we want to deal with, then this solution leads to distinguishing different modes through which the literals can be derived using rules. How such types of derivation are related to the introduction of the corresponding modalised literals can be expressed as follows: if  $X \in \{\text{BEL}, \text{INT}, \text{OBL}\}$ , then

$$\frac{\Gamma \quad \Gamma \Rightarrow_X \psi}{\Gamma \vdash_X \psi} \text{ MI}$$

As we will see, we do make an exception when rules for belief are concerned since we will state that  $X \in \{\text{INT}, \text{OBL}\}$ . The reason for this is that we assume that beliefs are conceived of as the knowledge the agent has of the environment, and so they are used by the agent to make inferences about how the world is: in this perspective, belief conclusions correspond to factual knowledge and do not need to be modalised. But besides this exception, which can be removed if required, schema MI captures the basic logical behaviour of our modal rules.

However, if nothing is done besides labelling the rules of DL, what we have in our hands is nothing but a simple treatment of modalities: what we obtain is that the conditions for introducing modalities (and in particular intentions and obligations) collapse into those for deriving literals in standard DL. Hence, the next step is to allow the consequence relations to interact with the propositional module and with each other. Indeed, we could in theory define sets of many interaction patterns, but what we need for the purposes of our paper are only two interaction strategies: one that permits to use rules for a modality  $X$  as they were for another modality  $Y$  (*rule conversions*), and one that considers conflicts between rules (*conflict-detection* and *conflict-resolution*).

**Rule Conversions** The notion of *rule conversion* allows us to model peculiar interactions between different modal operators. In general, notice that in many formalisms it is possible to convert from one type of conclusion into a different one. Take for example the right weakening rule of non-monotonic consequence relations (see, for example

[22])

$$\frac{B \vdash C \quad A \vdash B}{A \vdash C}$$

which allows the combination of non-monotonic and classical consequences.

Suppose that a rule of a specific type is given and also suppose that all the literals in the antecedent of a rule are provable in one and the same modality. If so, is it possible to argue that the conclusion of the rule inherits the modality of the antecedent? To give an example, suppose we have that  $\psi \Rightarrow_{\text{BEL}} \phi$  and that we derive  $\psi$  using a rule labelled by INT. Can we conclude  $\text{INT}\phi$ ? If the answer is positive, on the basis of MI this can be represented as follows:

$$\frac{\Gamma \vdash \text{INT}\psi \quad \psi \Rightarrow_{\text{BEL}} \phi}{\Gamma, \text{INT}\psi \vdash \text{INT}\phi} \text{ Conversion}$$

In many cases this is a reasonable conclusion to obtain. Indeed, this is the inference pattern we discussed in Section 2: if an agent believes to visit Italy if she visits Rome, and she has the intention to visit Rome, then it seems rational that she has the intention to visit Italy. Thus, conversions are ways through which some rational side effects can be derived. An additional example can help us illustrate the notion of conversion. Consider the following formalisation of the Yale Shooting Problem.<sup>5</sup>

$$\text{load\_live\_ammo, shoot} \Rightarrow_{\text{BEL}} \text{kill}$$

This rule encodes the knowledge of an agent that knows that loading the gun with live ammunitions, and then shooting will kill her friend. This example clearly shows that the qualification of the conclusions depends on the modalities relative to the individual acts “load” and “shoot”. In particular, if we obtain that the agent intends to load and to shoot the gun ( $\text{INT}(\text{load}), \text{INT}(\text{shoot})$ ), then, since she knows that the consequence of these actions is the death of her friend, she intends to kill him. However, if shooting was not intended, then we have *prima facie* to say that killing, too, was not intentional.

To define the admitted conversions we introduce a binary relation “Convert” over the modalities of the language. When we write  $\text{Convert}(\text{BEL}, \text{INT})$  this means that a belief rule  $r$  can be used to derive an intention (of course, provided that all its antecedents are derived as intentions):  $r$  can thus be converted into a rule for intention. Notice that we do not impose any specific constraint on Convert. In particular, we do *not* require Convert to be irreflexive. In fact, rule conversions can be viewed as corresponding, in a multi-modal setting, to the following inference schema:

$$\frac{X\psi \quad Y(\psi \rightarrow \phi)}{X\phi} \quad (5)$$

If we have  $\text{Convert}(X, Y)$  and  $X = Y$ , we do not obtain something necessarily odd. As is well-known, in deontic logic, for example, this inference pattern corresponds to the so-called deontic detachment:

$$\frac{\text{OBL}\psi \quad \text{OBL}(\psi \rightarrow \phi)}{\text{OBL}\phi} \quad (6)$$

<sup>5</sup>Here we will ignore all temporal aspects and we will assume that the sequence of actions is done in the correct order.

Although (6) is far from being uncontroversial, it seems that the same philosophical reasons that lead to accept it may support, for example, the adoption of its counterpart for intentions. Thus, even though we do not want in general to accept (5) when  $X = Y$ , we believe that this case cannot be excluded, and so, a fortiori, that  $\text{Convert}(X, X)$  be always rejected.

**Conflicts** As was mentioned in the previous sections, conflict-detection and conflict-resolution play an important role in the current context. It is in fact crucial to establish criteria for detecting and solving conflicts between the different components which characterise the cognitive profiles of agent’s deliberation. In a multi-modal setting, we can establish which modalities can be incompatible with each other, and, also, we can impose various forms of consistency, such as the following:

$$X\phi \rightarrow \neg Y\neg\phi \quad (7)$$

$$(X\phi \wedge Y\neg\phi) \rightarrow \neg Z\neg\phi \quad (8)$$

Criteria for conflict-detection and -resolution in DL can capture the rationale of schemata such as (7) and (8). However, their precise definition makes it necessary to take care of the peculiar approach adopted. In particular, various forms of consistency between agents’ motivations require to define incompatibility relations between the modalities by referring to rule types as well as to specific methods to solve conflicts between the rules. Many complex conflict patterns can be identified [18, 10, 11]. For the purpose of this paper, we introduce a binary and asymmetric relation  $\text{Conflict}$  over the set of modalities that defines which types of rules are in conflict and which are the stronger ones (the formal definition of  $\text{Conflict}$  is given in Section 3.4). Suppose, for example, that we have

$$\begin{aligned} r &: a \Rightarrow_{\text{BEL}} q \\ s &: b \Rightarrow_{\text{OBL}} \neg q \\ t &: c \Rightarrow_{\text{INT}} q \end{aligned}$$

If we only have  $\text{Conflict}(\text{BEL}, \text{OBL})$ , this means that rule  $r$  is in conflict with rule  $s$  and that  $r$  is stronger than  $s$ : for this reason, if applicable,  $r$  will defeat  $s$ . Suppose now to drop  $r$ . Nothing is said about the relation between obligations and intentions, and so about rules  $s$  and  $t$ . This means that there is no incompatibility relation between INT and OBL and we are free to derive both  $\text{INT}q$  and  $\text{OBL}\neg q$ .

The relation  $\text{Conflict}$  is explicitly linked to that of agent type. Classically, agent types are characterised by stating conflict resolution types in terms of orders of overruling between rules [8, 18, 10, 11]. In this perspective, agent types are meaningful within a non-monotonic setting and are nothing but general strategies to detect and solve conflicts between the different components of the cognitive profiles of agent’s deliberation. In [8] 24 possible types are identified while, in [11], based on a different framework, 20 combinations are proposed. Typically, rational agents are assumed to be at least *realistic*: a realistic agent, in fact, is such that rules for beliefs override all other components, as beliefs correspond to agent’s account of how the environment is. If the realistic condition is abandoned, we may have situations where intentions and desires override beliefs, thus leading to various forms of wishful thinking. Given the

minimal assumption that a rational agent should be realistic, we may further constrain agent’s deliberation in order not to violate obligations: a *social agent* type requires that obligations are stronger than the other motivational components with the exception of beliefs. Other agent types can be specified, for which see Section 4.

### 3.3 A Summary of Our Intuitions

Before providing a detailed presentation of the logical system for BIO agents, let us briefly summarise the logical intuitions previously presented and check them against the conceptual discussion we have developed in Section 2.

The main intuitions characterising our logical approach are the following:

1. The rules aim to capture policy-based motivations. For example, the rule  $\neg SunShining \Rightarrow_{BEL} Raining$  says that, if the sun is not shining, then the agent believes that it is raining; the rule  $SunShining \Rightarrow_{INT} Jogging$  says that the agent intends to do jogging if the sun is shining; the rule  $Order \Rightarrow_{OBL} Pay$  says that, if the agent sends a purchase order, then she will be obliged to pay.
2. Rules for intention and obligation are meant to introduce modalities: for example, if we have  $a \Rightarrow_{INT} b$  and we derive  $a$ , then we obtain  $INTb$ .
3. Rules labelled with BEL are an exception to the intuition under point 2 above. In the perspective of a single agent, agent’s beliefs describe how things effectively stand in the world. Hence, they are taken as true beliefs, and so we do not need to derive in this case modalised literals. For instance, if we have  $a \Rightarrow_{BEL} b$  and derive  $a$ , then we simply get  $b$ .
4. For the sake of simplicity, modal literals can only occur in the antecedent of rules. This is in line with our idea that the applicability of rules labelled with a modality  $X$  is the condition for deriving literals modalised with  $X$ . In other words, we do not admit rules such as  $a \Rightarrow_{OBL} INTb$ .
5. We introduce conversions, which allow to derive modalised literals using rules labelled with different modalities. For example, if we have  $a \Rightarrow_{BEL} b$ , derive  $INTa$ , and  $Convert(BEL, INT)$  holds, then we obtain  $INTb$ .
6. We devise methods (in particular, the relation Conflict) for detecting and solving conflicts between rules. This is in the spirit of standard DL, but here conflict resolution has a peculiar role, given the specific defeasible nature of policy-based motivations and the possibility of identifying different agent types.

It is worth noting that INT and OBL are not simple labels: they are modalities. In fact, in contrast with BEL, we model INT and OBL as non-reflexive modalities<sup>6</sup>. In addition, conversions provide complex interaction patterns between modalities which regulate various form of modal detachment.

<sup>6</sup>As is well-known, in a non-reflexive modal logic  $a$  does not follow from  $Xa$ , where  $X$  is a modal operator.

Secondly, we do not admit iterated modalities. Clearly, this is a simplification aimed at keeping the system manageable, but it does not pose severe limits for our purposes<sup>7</sup>. Since literals modalised with BEL never occur, we only fail to treat structures such as  $\text{INT}(\text{INT}a)$ ,  $\text{OBL}(\text{OBL}a)$ ,  $\text{OBL}(\text{INT}a)$ , and  $\text{INT}(\text{OBL}a)$ . While iterations of the same modality have a little significance, the last two structures express cases which are not needed for our discussion:  $\text{INT}(\text{OBL}a)$  makes sense when the agent is a sort of law-giver;  $\text{OBL}(\text{INT}a)$  establishes the obligation to intend, which is something that normative systems usually do not state (especially in the law).

How do these intuitions match with the conceptual points discussed in Section 2? Two questions are worthy of comment here: the fact that an agent is assumed to be aware of her policies, and the side-effect problem.

nAs regards the first question, awareness can be modelled as a kind of epistemic introspection. In other words, for any policy such as  $a \Rightarrow b$ , the agent is aware of it if we have  $\text{BEL}(a \Rightarrow b)$ . However, rules are implicitly assumed to be believed, exactly as assume that non-modal literals derived via belief rules are also believed (remember that we take beliefs as true beliefs). Making explicit modalities for belief, in our framework, would be strictly necessary only if we considered more than one agent: only in this case, we would need to mark the fact that some  $a$  is believed by one or another agent. On the other hand, it is worth noting that a reasoning schema such as

$$\frac{\text{INT}a \quad \text{BEL}(a \Rightarrow b)}{\text{INT}b} \quad (9)$$

is captured in our framework by stating that, if we have  $a \Rightarrow_{\text{BEL}} b$ , derive  $\text{INT}a$  and  $\text{Convert}(\text{BEL}, \text{INT})$  holds, then we obtain  $\text{INT}b$ .

This reconstruction of (9) allows us to account for the inclusion of some side effects. Let us provide a possible formalisation of the revised Strategic Bomber scenario we discussed in Section 2.1:

$$\begin{aligned} Th' = \{ & \text{INT}b_{\text{bomb}}, \text{INT}\neg_{\text{minimise}}, \text{OBL}\neg_{\text{kill}}, \\ & \text{INT}b_{\text{bomb}} \Rightarrow_{\text{INT}} \neg_{\text{kill}}, \\ & \neg_{\text{minimise}} \Rightarrow_{\text{BEL}} \text{kill} \} \end{aligned}$$

By default it is assumed that the intention to bomb does not imply the intention to kill innocents. But, if we assume that  $\text{Convert}(\text{BEL}, \text{INT})$  holds and the agent is realistic (beliefs override all other factors), since we derive  $\text{INT}\neg_{\text{minimise}}$  (it is a fact), through the belief rule in  $Th'$  we obtain  $\text{INT}kill$ . In other words, under this interpretation, the side effect *kill* is intended.

### 3.4 The Language of Modal Defeasible Logic

The inference process derives factual knowledge (through belief rules), intentions and obligations based on existing facts, intentions and obligations. Thus, rules allow for the

<sup>7</sup>However, notice that it does not seem hard to extend the framework of Sections 3.4 and 3.5 to cover nested modalities. It is sufficient to modify some language definitions, revise the definition of conflict, and make a few changes in the proof conditions. For a treatment in DL of nested modalities, even though applied to the logic of agency, see [17].

derivation of new motivational factors of an agent. As was mentioned, we divide the rules into rules for beliefs, intentions, and obligations. Provability for beliefs does not generate modalised literals, since in our view beliefs concern the knowledge an agent has about the world and corresponds to the basic inference mechanism of the agent.

A defeasible agent theory consists of a set of *facts* or indisputable statements, three sets of rules for beliefs, intentions, and obligations, a set of *conversions* saying when a rule of one type can be used also as another type, a set of *conflict relations* saying when two rule types can be in conflict and which rule type prevails, and a *superiority relation*  $>$  among rules saying when a single rule may override the conclusion of another rule. For  $X \in \{\text{BEL}, \text{INT}, \text{OBL}\}$ , we have that  $\phi_1, \dots, \phi_n \rightarrow_X \psi$  is a *strict rule* such that whenever the premises  $\phi_1, \dots, \phi_n$  are indisputable so is the conclusion  $\psi$ .  $\phi_1, \dots, \phi_n \Rightarrow_X \psi$  is a *defeasible rule* that can be defeated by contrary evidence.  $\phi_1, \dots, \phi_n \rightsquigarrow_X \psi$  is a *defeater* that is used to defeat some defeasible rules by producing evidence to the contrary. It is worth noting that modalised literals can occur only in the antecedent of rules: the reason of this is that the rules are used to derive modalised conclusions while we do not conceptually need to iterate modalities. This limitation makes the system more manageable.

**DEFINITION 1 (Language).** *Let PROP be a set of propositional atoms, MOD = {BEL, INT, OBL} be the set of modal operators, and Lab be a set of labels. The sets below are the smallest sets closed under the following rules:*

**Literals**

$$\text{Lit} = \text{PROP} \cup \{\neg p \mid p \in \text{PROP}\}$$

*If  $q$  is a literal,  $\sim q$  denotes the complementary literal (if  $q$  is a positive literal  $p$  then  $\sim q$  is  $\neg p$ ; and if  $q$  is  $\neg p$ , then  $\sim q$  is  $p$ );*

**Modal literals**

$$\text{ModLit} = \{Xl, \neg Xl \mid l \in \text{Lit}, X \in \{\text{INT}, \text{OBL}\}\};$$

**Rules**  $\text{Rule} = \text{Rule}_s \cup \text{Rule}_d \cup \text{Rule}_{dfi}$ , where for  $X \in \text{MOD}$

$$\text{Rule}_s = \{r : \phi_1, \dots, \phi_n \rightarrow_X \psi \mid r \in \text{Lab}, A(r) \subseteq \text{Lit} \cup \text{ModLit}, \psi \in \text{Lit}\}$$

$$\text{Rule}_d = \{r : \phi_1, \dots, \phi_n \Rightarrow_X \psi \mid r \in \text{Lab}, A(r) \subseteq \text{Lit} \cup \text{ModLit}, \psi \in \text{Lit}\}$$

$$\text{Rule}_{dfi} = \{r : \phi \rightsquigarrow_X \psi \mid r \in \text{Lab}, A(r) \subseteq \text{Lit} \cup \text{ModLit}, \psi \in \text{Lit}\}$$

*We use some obvious abbreviations, such as superscript for mental attitude, subscript for type of rule, and  $\text{Rule}[\phi]$  for rules whose consequent is  $\phi$ , for example:*

$$\text{Rule}^{\text{BEL}} = \{r : \phi_1, \dots, \phi_n \triangleright_{\text{BEL}} \psi \mid (r : \phi_1, \dots, \phi_n \triangleright_{\text{BEL}} \psi) \in \text{Rule}, \triangleright \in \{\rightarrow, \Rightarrow, \rightsquigarrow\}\}$$

$$\text{Rule}_s[\psi] = \{\phi_1, \dots, \phi_n \rightarrow_X \psi \mid \{\phi_1, \dots, \phi_n\} \subseteq \text{Lit} \cup \text{ModLit}, \psi \in \text{Lit}, X \in \text{MOD}\}$$

*We use  $A(r)$  to denote the set  $\{\phi_1, \dots, \phi_n\}$  of antecedents of the rule  $r$ , and  $C(r)$  to denote the consequent  $\psi$  of the rule  $r$ .*

DEFINITION 2 (Conversion and Conflict Relations). *The conversion relation Convert is defined as follows:*

$$\text{Convert} \subseteq \text{MOD} \times \text{MOD}$$

*The conflict relation Conflict  $\subseteq \text{MOD} \times \text{MOD}$  is such that*

$$\forall X, Y \in \text{MOD}, \text{Conflict}(X, Y) \Rightarrow \neg(\text{Conflict}(Y, X)) \text{ (asymmetry)}$$

DEFINITION 3 (Defeasible Agent Theory). *A defeasible agent theory is a structure*

$$D = (F, R^{\text{BEL}}, R^{\text{INT}}, R^{\text{OBL}}, >, \mathcal{C}, \mathcal{V})$$

where

- $F \subseteq \text{Lit} \cup \text{ModLit}$  is a finite set of facts;
- $R^{\text{BEL}} \subseteq \text{Rule}^{\text{BEL}}, R^{\text{INT}} \subseteq \text{Rule}^{\text{INT}}, R^{\text{OBL}} \subseteq \text{Rule}^{\text{OBL}}$  are three finite sets of rules such that each rule has a unique label;
- The superiority relation  $>$  is such that  $> = >^{\text{sm}} \cup >^{\text{Conflict}}$ , where  $>^{\text{sm}} \subseteq R^X \times R^X$  such that if  $r > s$ , then if  $r \in \text{Rule}^X[p]$  then  $s \in \text{Rule}^X[\sim p]$  and  $>$  is acyclic; and  $>^{\text{Conflict}}$  is such that

$$\forall r \in \text{Rule}^X[p], \forall s \in \text{Rule}^Y[\sim p], \text{if } \text{Conflict}(X, Y), \text{ then } r >^{\text{Conflict}} s$$

- $\mathcal{C} \subseteq \{\text{Convert}(X, Y) \mid X, Y \in \text{MOD}\}$  is a set of conversions;
- $\mathcal{V} \subseteq \{\text{Conflict}(X, Y) \mid X, Y \in \text{MOD}\}$  is a set of conflict relations.

The construction of the superiority relation combines two components: the first  $>^{\text{sm}}$  considers pairs of rules of the same mode. This component is usually given by the designer of the theory and capture the meaning of the single rules, and thus encodes the domain knowledge of the designer of the theory. The second component,  $>^{\text{Conflict}}$  is obtained from the rules in a theory and depends on the meaning of the modalities.

The following running example illustrates the defeasible agent theory.

EXAMPLE 1. (RUNNING EXAMPLE). Frodo, our Tolkienian agent, is entrusted by Elrond to be the bearer of the ring of power, a ring forged by the dark lord Sauron. Frodo has the task to bring the ring to Mordor, the realm of Sauron, and to destroy it by throwing it into the fires of Mount Doom. However, Frodo loves the place where he was born, the Shire, and intends to go there.

$$\begin{aligned} F &= \{\text{INTGoToShire}, \text{EntrustedByElrond}\} \\ R &= \{r_1 : \text{EntrustedByElrond} \Rightarrow_{\text{BEL}} \text{RingBearer} \\ &\quad r_2 : \text{RingBearer} \Rightarrow_{\text{OBL}} \text{DestroyRing} \\ &\quad r_3 : \text{INTGoToShire} \Rightarrow_{\text{INT}} \neg \text{GoToMordor} \\ &\quad r_4 : \neg \text{GoToMordor} \Rightarrow_{\text{BEL}} \neg \text{DestroyRing}\} \\ > &= \{r_4 > r_2\} \\ \mathcal{C} &= \{\text{Convert}(\text{BEL}, \text{INT})\} \\ \mathcal{V} &= \{\text{Conflict}(\text{BEL}, \text{OBL})\} \end{aligned}$$

### 3.5 Inferences with BIO Agents

Proofs are sequences of literals and modal literals together with so-called proof tags  $+\Delta$ ,  $-\Delta$ ,  $+\partial$  and  $-\partial$ . Given a defeasible agent theory  $D$ ,  $+\Delta_X q$  means that literal  $q$  is provable in  $D$  using only facts and strict rules for modality  $X$ ,  $-\Delta_X q$  means that it has been proved in  $D$  that  $q$  is not definitely provable in  $D$ ,  $+\partial_X q$  means that  $q$  is defeasibly provable in  $D$ , and  $-\partial_X q$  means that it has been proved in  $D$  that  $q$  is not defeasibly provable in  $D$ .

DEFINITION 4. *Given an agent theory  $D$ , a proof in  $D$  is a linear derivation, i.e. a sequence of labelled formulas of the type  $+\Delta_X q$ ,  $-\Delta_X q$ ,  $+\partial_X q$  and  $-\partial_X q$ , where the proof conditions defined in the rest of this section hold.*

We start with some terminology. As was explained, the following definition states the special status of belief rules, and that the introduction of a modal operator corresponds to being able to derive the associated literal using the rules for the modal operator.

DEFINITION 5. *Let  $\# \in \{\Delta, \partial\}$ , and  $P = (P(1), \dots, P(n))$  be a proof in  $D$ . A (modal) literal  $q$  is  $\#$ -provable in  $P$  if there is a line  $P(m)$  of  $P$  such that either*

1.  $q$  is a literal and  $P(m) = +\#_{\text{BEL}} q$  or
2.  $q$  is a modal literal  $Xp$  and  $P(m) = +\#_X p$  or
3.  $q$  is a modal literal  $\neg Xp$  and  $P(m) = -\#_X p$ .

A literal  $q$  is  $\#$ -rejected in  $P$  if there is a line  $P(m)$  of  $P$  such that

1.  $q$  is a literal and  $P(m) = -\#_{\text{BEL}} q$  or
2.  $q$  is a modal literal  $Xp$  and  $P(m) = -\#_X p$  or
3.  $q$  is a modal literal  $\neg Xp$  and  $P(m) = +\#_X p$ .

The definition of  $\Delta_X$  describes just forward chaining of strict rules:

- $+\Delta_X$ : If  $P(n+1) = +\Delta_X q$  then
- (1)  $q \in F$  if  $X = \text{BEL}$  or  $Xq \in F$  or
  - (2)  $\exists r \in R_S^X[q] : \forall a \in A(r) a$  is  $\Delta$ -provable or
  - (3)  $\exists r \in R_S^Y[q] : \text{Convert}(Y, X) \in \mathcal{C}, \forall a \in A(r) Xa$  is  $\Delta$ -provable.
- $-\Delta_X$ : If  $P(n+1) = -\Delta_X q$  then
- (1)  $q \notin F$  if  $X = \text{BEL}$  and  $Xq \notin F$  and
  - (2)  $\forall r \in R_S^X[q] \exists a \in A(r) : a$  is  $\Delta$ -rejected and
  - (3)  $\forall r \in R_S^Y[q] : \text{if } \text{Convert}(Y, X) \in \mathcal{C} \text{ then } \exists a \in A(r) Xa$  is  $\Delta$ -rejected.

For a literal  $q$  to be definitely provable with the mode  $X$  we need to find a strict rule for  $X$  with head  $q$ , whose antecedents have all been definitely proved previously. And to establish that  $q$  cannot be definitely proven we must establish that for every strict rule with head  $q$  there is at least one antecedent which has been shown to be non-provable. Condition (3) says that a rule for  $Y$  can be used as a rule for a different modal operator

$X$  in case all literals in the body of the rule are modalised with the modal operator we want to prove. For example, given the rule  $p, q \rightarrow_{\text{BEL}} s$ , we can derive  $+\Delta_{\text{INT}}s$  if we have  $+\Delta_{\text{INT}}p$ ,  $+\Delta_{\text{INT}}q$ , and the conversion  $\text{Convert}(\text{BEL}, \text{INT})$  holds in the theory.

Conditions for  $\partial_X$  are more complicated. We define when a rule is applicable or discarded. A rule for a belief is applicable if all the literals in the antecedent of the rule are provable with the appropriate modalities, while the rule is discarded if at least one of the literals in the antecedent is not provable. As before, for the other types of rules we have to take conversions into account. We have thus to determine conditions under which a rule for  $Y$  can be used to directly derive a literal  $q$  modalised by  $X$ . Roughly, the condition is that all the antecedents  $a$  of the rule are such that  $+\partial_X a$ .

**DEFINITION 6.** *Given a derivation  $P$ ,  $P(1..n)$  denotes the initial part of the derivation of length  $n$ . Let  $X, Y, Z \in \text{MOD}$ .*

- A rule  $r \in R_{sd}$  is applicable in the proof condition for  $\pm\partial_X$  iff
  1.  $r \in R^X$  and  $\forall a \in A(r)$ ,  $+\partial_{\text{BEL}}a \in P(1..n)$  and  $\forall Za \in A(r)$ ,  $+\partial_Za \in P(1..n)$ , or
  2.  $r \in R^Y$ ,  $\text{Convert}(Y, X) \in \mathcal{C}$ , and  $\forall a \in A(r)$ ,  $+\partial_Xa \in P(1..n)$ .
- A rule  $r$  is discarded in the condition for  $\pm\partial_X$  iff
  1.  $r \in R^X$  and  $\exists a \in A(r)$  such that  $-\partial_{\text{BEL}}a \in P(1..n)$  or  $\exists Za \in A(r)$  such that  $-\partial_Za \in P(1..n)$ ; or
  2.  $r \in R^Y$  and, if  $\text{Convert}(Y, X)$ , then  $\exists a \in A(r)$  such that  $-\partial_Xa \in P(1..n)$ , or
  3.  $r \in R^Z$  and either  $-\text{Convert}(Z, X)$  or  $-\text{Conflict}(Z, X)$ .

**EXAMPLE 2.** The rule  $a, \text{INT}b \Rightarrow_{\text{BEL}} c$  is applicable if we can prove both  $+\partial_{\text{BEL}}a$  and  $+\partial_{\text{INT}}b$ .

**EXAMPLE 3.** If we have a type of agent that allows a deontic rule to be converted into a rule for intention,  $\text{Convert}(\text{OBL}, \text{INT})$ , then the definition of applicable in the condition for  $\pm\partial_{\text{INT}}$  is as follows: a rule  $r \in R_{sd}[q]$  is applicable iff (1)  $r \in R^{\text{INT}}$  and  $\forall a \in A(r)$ ,  $+\partial_{\text{BEL}}a \in P(1..n)$  and  $\forall Xa \in A(r)$ ,  $+\partial_Xa \in P(1..n)$ , (2) or  $r \in R^{\text{OBL}}$  and  $\forall a \in A(r)$ ,  $+\partial_{\text{INT}}a \in P(1..n)$ . In this second case, for example, given the rule  $p, q \Rightarrow_{\text{OBL}} s$ , we can derive  $+\partial_{\text{INT}}s$  if we have  $+\partial_{\text{INT}}p$  and  $+\partial_{\text{INT}}q$ .

As a corollary of the definition of applicability, we can establish when a literal is supported (see Section 5.2 for the use of this notion):

**DEFINITION 7.** *Given a theory  $D$ , a literal  $l$  is supported in  $D$  iff there exists a rule  $r \in R[l]$  such that  $r$  is applicable, otherwise  $l$  is not supported. For  $X \in \text{MOD}$  we use  $+\Sigma_X l$  and  $-\Sigma_X l$  to indicate that  $l$  is supported / not supported by rules for  $X$ .*

We are now ready to provide proof conditions for  $\pm\partial_X$ :

- $+\partial_X$ : If  $P(n+1) = +\partial_X q$  then
- (1)  $+\Delta_X q \in P(1..n)$  or
  - (2) (2.1)  $-\Delta_X \sim q \in P(1..n)$  and
    - (2.2)  $\exists r \in R_{sd}[q]$  such that  $r$  is applicable, and
    - (2.3)  $\forall s \in R[\sim q]$  either  $s$  is discarded, or
      - (2.3.1)  $\exists t \in R[q]$  such that  $t$  is applicable and  $t > s$ , and either  $t, s \in R^Z$ , or  $\text{Convert}(Y, X)$  and  $t \in R^Y$

- $-\partial_X$ : If  $P(n+1) = -\partial_X q$  then
- (1)  $-\Delta_X q \in P(1..n)$  and either
    - (2.1)  $+\Delta_X \sim q \in P(1..n)$  or
    - (2.2)  $\forall r \in R_{sd}[q]$ , either  $r$  is discarded, or
    - (2.3)  $\exists s \in R[\sim q]$ , such that  $s$  is applicable, and
      - (2.3.1)  $\forall t \in R[q]$  either  $t$  is discarded, or  $t \not> s$ , or  $t \in R^Z, s \in R^{Z'}, Z \neq Z'$  and, if  $t \in R^Y$  then  $\neg \text{Convert}(Y, X)$ .

To show that  $q$  is defeasibly provable we have two choices: (1) We show that  $q$  is already definitely provable; or (2) we need to argue using the defeasible part of a theory  $D$ . For this second case, three (sub)conditions must be satisfied. First, we need to consider possible reasoning chains in support of  $\sim q$ , and show that  $\sim q$  is not definitely provable (2.1). Second, we require that there must be a strict or defeasible rule for  $q$  which can be applied (2.2). Third, we must consider the set of all rules which are not known to be inapplicable and which permit to get  $\sim q$  (2.3). Essentially, each such a rule  $s$  attacks the conclusion  $q$ . For  $q$  to be provable,  $s$  must be counterattacked by a rule  $t$  for  $q$  with the following properties: (i)  $t$  must be applicable, and (ii)  $t$  must be stronger than  $s$ . Thus each attack on the conclusion  $q$  must be counterattacked by a stronger rule. In other words,  $r$  and the rules  $t$  form a team (for  $q$ ) that defeats the rules  $s$ . However, since we can have rules for different modes, we have to ensure we have the appropriate relationships among the rules. Thus clause (2.3.1) prescribes that either the rule that attacks the conclusion we want to prove ( $s$ ) and the rule used to counterattack it (i.e.,  $t$ ) have the same mode (i.e.,  $s, t \in R^Z$ ), or that  $t$  can be used to produce a conclusion of the mode we want to prove (i.e.,  $t \in R^Y$  and  $\text{Convert}(Y, X)$ ).

$-\partial_X q$  is defined in an analogous manner.

EXAMPLE 4. (RUNNING EXAMPLE; CONTINUED). Below is the set  $C$  of all conclusions we get using the rules in  $R$ :

$$C = \{RingBearer, \text{INT}\neg GoToMordor, \text{INT}\neg DestroyRing\}$$

As facts, we know that Frodo has the primitive intention to go to the Shire and that he has been entrusted by Elrond. These facts make applicable rules  $r_3$  and  $r_1$ , which permit to derive that Frodo is the ring bearer and that he has the intention not to go to Mordor. At this point we have a conflict, as we have  $\text{Conflict}(\text{BEL}, \text{OBL})$  and  $\text{Convert}(\text{BEL}, \text{INT})$ . In effect, given the conversion,  $r_4$  permits to derive that Frodo has the intention not to destroy the ring while rule  $r_2$  should lead to the obligation to destroy it. However,  $r_4$  is stronger than  $r_2$  and so we only get  $+\partial_{\text{INT}\neg DestroyRing}$ .

## 4 Agent Types

Classically, agent types are characterized by stating conflict resolution types in terms of orders of overruling between rules [8, 18]. For example, an agent is *realistic* when rules for beliefs override all other components; she is *social* when obligations are stronger than the other motivational components with the exception of beliefs, etc.

As suggested in [10, 11], agent types can be characterised in DL as follows:

DEFINITION 8 (Agent Type (1)). *An agent type is defined by a set of pairs  $(X, Y)$ ,  $X, Y \in \{\text{BEL}, \text{OBL}, \text{INT}\}$ , such that for every  $r$  and  $r'$  such that  $r \in R^X[q]$  and  $r' \in R^Y[\sim q]$ , we have that  $r > r'$ .*

For example, while realistic agents are such that  $X = \text{BEL}$  and  $Y \in \{\text{INT}, \text{OBL}\}$ , social agents are such that  $X = \text{OBL}$  and  $Y = \text{INT}$ . It is clear that the notion of agent type is defined in terms of the relation Conflict we have previously introduced.

Let us see the agent types that can be identified in the framework we have defined so far. Table 1 shows all possible cases and, for each kind of rule, indicates all attacks on it. It should be read as follows. Each of the three main columns identifies a possible kind of conflict between two types  $X, Y$  of applicable rules that would permit to infer the literals  $p$  and  $\sim p$  labelled by  $X$  and  $Y$  respectively. The first row from the top in the three main columns specifies the case where both literals are derived (i.e., there is no conflict, which indeed corresponds to the case where the modalities involved are not in Conflict); the second and third rows from top identify the cases where we have a conflict and one rule prevails over the other. The third sub-column in each main column defines the agent type for which each conflict-detection and -resolution policy is appropriate. (To save space, in Table 1 “indep.” abbreviates “independent”, “wish. th.” “wishful thinking”, and “real.” “realistic”).

$\Rightarrow_{\text{OBL}} p / \Rightarrow_{\text{INT}} \sim p$			$\Rightarrow_{\text{OBL}} p / \Rightarrow_{\text{BEL}} \sim p$			$\Rightarrow_{\text{INT}} p / \Rightarrow_{\text{BEL}} \sim p$		
$+\partial_{\text{OBL}} p$	$+\partial_{\text{INT}} \sim p$	indep.	$+\partial_{\text{OBL}} p$	$+\partial_{\text{BEL}} \sim p$	wish. th.	$+\partial_{\text{INT}} p$	$+\partial_{\text{BEL}} \sim p$	wish. th.
$+\partial_{\text{OBL}} p$	$-\partial_{\text{INT}} \sim p$	social	$+\partial_{\text{OBL}} p$	$-\partial_{\text{BEL}} \sim p$	wish. th.	$+\partial_{\text{INT}} p$	$-\partial_{\text{BEL}} p$	wish. th.
$-\partial_{\text{OBL}} p$	$+\partial_{\text{INT}} \sim p$	deviant	$-\partial_{\text{OBL}} p$	$+\partial_{\text{BEL}} \sim p$	real.	$-\partial_{\text{INT}} p$	$+\partial_{\text{BEL}} \sim p$	real.

Table 1: Conflict: Agent Types

Independent agents are free to adopt intentions for  $p$  in presence of derivations for  $\text{OBL} \sim p$ . This is possible in our framework when we have that  $\neg \text{Conflict}(\text{OBL}, \text{INT})$  and  $\neg \text{Conflict}(\text{INT}, \text{OBL})$ : this means that the system admits both conclusions, as they are not in conflict. As expected, for social agents obligations override intentions and so  $\text{Conflict}(\text{OBL}, \text{INT})$ ; the opposite case is when an agent is deviant and her intentions override the obligations,  $\text{Conflict}(\text{INT}, \text{OBL})$ . Where beliefs are defeated either by obligations or by intentions we have classical examples of wishful thinking. Notice that in Table 1 also the cases  $+\partial_{\text{OBL}} p / +\partial_{\text{BEL}} \sim p$  and  $+\partial_{\text{INT}} p / +\partial_{\text{BEL}} \sim p$  have been classified as wishful thinking, given the basic nature of beliefs we adopted in our framework. However, we are aware that this reading is debatable: in effect, if we can derive both conclusions, this means that there is no real conflict. Last, it is worth noting that we do not consider here the case where  $-\partial_X p / -\partial_Y \sim p$ : here we would have that  $X$  and  $Y$  are incompatible, but that it is not possible to establish what rule is the strongest

Convert(BEL, OBL)	c-realistic	Convert(INT, OBL)	c-deviant
Convert(BEL, INT)	c-realistic	Convert(OBL, BEL)	NO
Convert(OBL, INT)	c-social	Convert(INT, BEL)	NO

Table 2: Conversions

one, thus leading to a mutual defeating of the rules involved. This case –which is discussed in [18, 10, 11] and permits to identify other agent types– is excluded here, as the relation Conflict both identifies conflicts and solves them by establishing what rule type must prevail.

It is possible to integrate the above classifications by referring to the notion of conversion. Conversions do not have a direct relation with conflict resolution because they simply affect the condition of applicability of rules. However, they indeed contribute to define the cognitive profile of agents because they allow to obtain conclusions modalised by a certain  $X$  through the application of rules which are not modalised by  $X$ . Table 2 shows the conversions and specify new agent types with respect to which each conversion seems to be appropriate.

A preliminary remark before commenting Table 2. We do not consider here conversions  $\text{Convert}(X, Y)$  where  $X = Y$ . In fact, even though they can be admitted, they do not seem to characterise a specific cognitive profile for the agents. Consider  $\text{Convert}(\text{BEL}, \text{OBL})$  and  $\text{Convert}(\text{BEL}, \text{INT})$ . Both seem appropriate for some types of realistic agent. Indeed, for a realistic agent beliefs correspond to her basic reasoning mechanism. Accordingly, if we have

$$\begin{array}{l}
r : \neg \text{open\_umbrella} \Rightarrow_{\text{BEL}} \text{wet} \\
+ \partial_{\text{INT}} \neg \text{open\_umbrella} \qquad \qquad \qquad + \partial_{\text{OBL}} \neg \text{open\_umbrella}
\end{array}$$

it is reasonable to derive both that the agent has the intention to be wet, and that it is obligatory for her to be wet.

Other conversions look more appropriate for other agent types. For example, we may have agent types for which  $\text{Convert}(\text{OBL}, \text{INT})$  holds. This means that from

$$r : \text{kill} \Rightarrow_{\text{OBL}} \text{kill\_gently} \quad + \partial_{\text{INT}} \text{kill}$$

we can derive that the agent has the intention to kill gently. But this derivation is conceptually meaningful only if we assume a kind of norm regimentation, by which we impose that all agents intend what is prescribed by deontic rules.

The peculiarity of  $\text{Convert}(\text{INT}, \text{OBL})$  is that the simple fact that something is derived as obligatory can permit to obtain through a rule for intention that something else is obligatory as well. Consider this case:

$$r : \text{help\_needy\_people} \Rightarrow_{\text{INT}} \text{save\_money} \quad + \partial_{\text{OBL}} \text{help\_needy\_people}$$

If  $\text{Convert}(\text{INT}, \text{OBL})$  holds, then we can derive that it is obligatory for the agent to save money: an intention supports the derivation of an obligation. In other papers [18], this case has been classified as an example of an agent legislator. Here, we prefer to consider it as a case of a deviant agent [11], due to its structural similarity to  $\text{Conflict}(\text{INT}, \text{OBL})$ .

Finally, notice that the conversions  $\text{Convert}(\text{OBL}, \text{BEL})$  and  $\text{Convert}(\text{INT}, \text{BEL})$ , which are marked in the table by a “NO”, seem meaningless. They say that a rule for obligation and for intention may respectively be used to derive a belief. This sounds odd, at least adopting the interpretation of beliefs of this paper. In fact, since the belief modality captures the basic knowledge the agent has about the environment, it is treated as its logic were reflexive (namely, that  $\text{BEL}\psi \rightarrow \psi$  holds). Consider, for example, the following:

$$r : \text{help\_needy\_people} \Rightarrow_{\text{INT}} \text{save\_money} + \partial_{\text{BEL}} \text{help\_needy\_people}$$

If  $\text{Convert}(\text{INT}, \text{BEL})$  holds, then we obtain that the agent in fact saves money, which is odd: beliefs, according to our interpretation should be independent from agent’s deliberation, even though they are used to derive motivational attitudes such as intentions and obligations. In addition, adopting both  $\text{Convert}(\text{OBL}, \text{BEL})$  and  $\text{Convert}(\text{INT}, \text{BEL})$  would determine a collapse of our logic, as we could dispense with explicit modalities in the antecedent of rules.

Since our logic system is characterised by  $\text{Conflict}$  as well as by  $\text{Convert}$ , and conversions indeed contribute to define the cognitive profile of an agent, it seems that an agent type should take both parameters into account:

**DEFINITION 9 (Agent Type (2)).** *An agent type is defined by a pair  $(\Gamma, \Delta)$ , where  $\Gamma \subseteq \{\text{Conflict}(X, Y) \mid X, Y \in \text{MOD}\}$  and  $\Delta \subseteq \{\text{Convert}(Z, W) \mid Z, W \in \text{MOD}\}$ .*

It is easy to see that the notion of agent type of Definition 8 (proposed in [10, 11]) is captured by Definition 9.

This completes our picture of the notion of agent types. However, a serious difficulty is around the corner when we focus on the notion of agent type based on defining criteria for conflict-detection and -resolution. Are we sure that this view is sufficient, given the account of policy-based attitudes we previously discussed? In the reminder we will consider only the interaction between intentions and obligations, even though similar remarks can be easily extended to all other agent types presented in Table 1. But, even confining the problem to these components, the question at stake is: How to deal with social agents? The simplest solution is the classical one, corresponding to adopting schema (7) and that we have adopted so far: when we have two rules, one leading to  $\text{INT}\phi$  and the other to  $\text{OBL}\sim\phi$ , the former is blocked. As we shall see, this strategy is not enough.

## 5 Social Agents

### 5.1 The Problem

The idea of social agent based on the intuition of Definition 8—which is also adopted in [8]—does not guarantee that agent’s deliberation is oriented to fully complying with obligations. The same holds when Definition 9 is used. In effect, to our view a social agent can be defined by the following pair

$$\begin{aligned} & (\{\text{Conflict}(\text{BEL}, \text{INT}), \text{Conflict}(\text{BEL}, \text{OBL}), \text{Conflict}(\text{OBL}, \text{INT})\}, \\ & \{\text{Convert}(\text{BEL}, \text{OBL}), \text{Convert}(\text{BEL}, \text{INT})\}) \end{aligned}$$

according to which the agent is realistic (beliefs override the other components, and the appropriate conversions hold) and obligations prevail over conflicting intentions.

In both cases, the drawback is mainly due to the introduction of conversions. Since conversions allow to obtain conclusions modalised by a certain  $X$  through the application of rules which are not modalised by  $X$ , they are fundamental in order to capture the fact that some side-effects should be accepted insofar as they are consequences of policies of which the agent is aware. Moreover, some conversions seem useful to integrate the basic idea of social agency.

It is clear that our system admits three different types of intentions and obligations. First, we have *primitive* intentions and obligations when these are facts of the theory. But we can also have what we may call *primary* and *secondary* intentions and obligations, depending on whether we accept at least basic conversions via belief rules.

Let us consider Example 1.  $\text{INTGoToShire}$  is a primitive intention. On the other hand,  $\text{OBLDestroyRing}$  –if it were derived from rule  $r_2$ – and  $\text{INT}\neg\text{GoToMordor}$  are primary obligations and intentions as they would be obtained without the use of conversions (see Example 4). Finally,  $\text{INT}\neg\text{DestroyRing}$  is a secondary intention because it is obtained from the rule  $r_4 : \neg\text{GoToMordor} \Rightarrow_{\text{BEL}} \neg\text{DestroyRing}$  and from  $+\partial_{\text{INT}}\neg\text{GoToMordor}$  (again, see Example 4). It should be noted that  $\text{OBLDestroyRing}$  cannot be derived because  $r_4 > r_2$ , but this just amounts to assuming that the agent is realistic:  $r_4$  is a belief rule whereas  $r_2$  is a deontic rule. In other words, when we have in general that

$$\begin{array}{ll} a \Rightarrow_{\text{OBL}} q & b \Rightarrow_{\text{BEL}} \sim q \\ +\partial_{\text{BEL}} a & +\partial_{\text{INT}} b \end{array}$$

we are doomed to have social agents who cannot be truly social since some of their (primitive) intentions lead to behaviours against what would be otherwise obligatory for the agents. However, this issue is not a matter of a direct conflict between rules for intentions and obligations. Thus, to deal with norm-complying agents in these scenarios and to restore their sociality we are required to change the notion of agent type. We cannot anymore define it in terms of an order of overruling between rules, but we have to focus on how the conflicting literals are derived during the proof. Indeed, this is feasible, but has a high computational cost, and even then we cannot guarantee the sociality of an agent.

## 5.2 The Cost of Social Agents

In this section we investigate the complexity of the defeasible logic for BIO agents where we assume the conversions  $\text{Convert}(\text{BEL}, \text{OBL})$  and  $\text{Convert}(\text{BEL}, \text{INT})$  and then we turn our attention to the complexity of social agents. We first introduce some notions to make precise the definition of the issues at hand.

**DEFINITION 10.** *Let  $\#$  be one of the proof tags. Given a theory  $D$ ,  $D \vdash \pm\#p$  iff there is a derivation  $P$  in  $D$  such that for some  $n$   $P(n) = \pm\#p$ .*

**DEFINITION 11.** *Given a theory  $D$ , the universe of  $D$  ( $U^D$ ) is the set of all the atoms occurring in  $D$ ; the extension of  $D$  ( $E^D$ ), is defined as follows:*

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

where for  $X \in \{\text{BEL}, \text{INT}, \text{OBL}\}$

$$\begin{aligned}\Delta^+ &= \{Xl : D \vdash +\Delta_X l\}; \\ \Delta^- &= \{Xl : D \vdash -\Delta_X l\}; \\ \partial^+ &= \{Xl : D \vdash +\partial_X l\}; \\ \partial^- &= \{Xl : D \vdash -\partial_X l\}.\end{aligned}$$

Two theories  $D$  and  $D'$  are *equivalent* if and only if they have the same extension, namely  $D \equiv D'$  iff  $E^D = E^{D'}$ .

We now prove the main theorem about the complexity of our defeasible logic. We show that the logic has linear complexity if we compute the whole set of conclusions, i.e., the extension, of a given theory.

**THEOREM 1.** *For every theory  $D$ ,  $E^D$  can be computed in time linear to the size of the theory, i.e.,  $O(|U^D| * |R|)$ .*

**PROOF.** The proof is based on a modification of the algorithm given by Maher [23] to show that propositional defeasible logic has linear complexity.

The main idea of the proof is to build appropriate data structure to implement a series of transformations reducing the complexity of the rules, and where each literal and modal literal is examined only once. The focal point of the transformations is based on the following properties:

- Let  $D \vdash +\partial p$  then

$$D \cup \{r : p_1, \dots, p_n, p \Rightarrow q\} \equiv D \cup \{r : p_1, \dots, p_n \Rightarrow q\}.$$

- Let  $D \vdash -\partial p$  then  $D \cup \{r : p_1, \dots, p_n, p \Rightarrow q\} \equiv D$ .

The properties allow us (1) to remove already proved literals from the body of rules and (2) to remove rules which have been discarded.

The algorithm has three phases. (1) A pre-processing phase where we use similar transformations to those given in [3] to transform a theory into an equivalent theory without superiority relation and defeaters; the transformation is linear. We will propose two linear transformations, one to empty the superiority relation and one to remove defeaters. We will show that these transformations are correct, that is, they produce the same sets of conclusions in the language of the theory they transform (Theorem 2 and Theorem 3). (2) A *rule loader* that parses the theory obtained in the first phase and generates the data structure that encodes the theory. (3) The *inference engine* applies transformations to the data structure, where at every step it reduces the complexity of the data structure.

**(1) Transformations** Theory transformations are an important tools to study properties of defeasibly logic. In [3] we extensively used transformations to show under which conditions it is possible to simplify the presentation of basic defeasible logic by dispensing defeaters and the superiority relation. In what follows we are going to give transformations that allow us to remove defeaters and the superiority relation from modal defeasible theories for BIO agents.

DEFINITION 12. Let  $\#$  be one of the proof tags. Two modal defeasible theories  $D_1$  and  $D_2$  are equivalent (written  $D_1 \equiv D_2$ ) iff  $\forall p, D_1 \vdash \#p$  iff  $D_2 \vdash \#p$ , i.e., they have the same consequences. Similarly  $D_1 \equiv_{\Sigma} D_2$  means that  $D_1$  and  $D_2$  have the same consequences in the language  $\Sigma$ .

DEFINITION 13. A transformation is a mapping from modal defeasible theories to modal defeasible theories. A transformation  $T$  is correct iff for all modal defeasible theories  $D$ ,  $D \equiv_{\Sigma} T(D)$  where  $\Sigma$  is the language of  $D$ .

DEFINITION 14. Let  $A = \{l_1, \dots, l_n\} \subseteq \text{Lit}$  and  $X \in \text{MOD}$ , then  $XA = \{Xl_i : l_i \in A\}$ .

DEFINITION 15. Let  $D = (F, R, >)$  be a defeasible theory such that  $R_{dfi} = \emptyset$ . Let  $\Sigma$  be the language of  $D$ . Define  $\text{elimsup}(D) = (F, R', \emptyset)$ , where

$$R' = \{\neg \text{inf}(r) \Rightarrow_{\text{BEL}} \text{inf}(s) : (r, s) \in >\} \bigcup_{r \in R} \text{elimsup}(r)$$

and

$$\text{elimsup}(r) = \{A(r) \leftrightarrow_{\text{BEL}} \neg \text{inf}(r), \neg \text{inf}(r) \leftrightarrow_X C(r) : A(r) \leftrightarrow_X C(r) \in R_{sd}\}$$

For each rule  $r \in R$ ,  $\text{inf}(r)$  is a new atom, i.e., they do not appear in  $\Sigma$ . Furthermore all new atoms generated are distinct.

THEOREM 2. The transformation  $\text{elimsup}$  is correct.

PROOF. The proof by induction on the length of derivations is similar to that given in [3]. Here we give in full the case of strict derivations and we outline the main part of the case of defeasible derivations.

Case if  $D \vdash +\Delta_X p$  then  $\text{elimsup}(D) \vdash +\Delta_X p$ . For a proof of length 1 of  $+\Delta_X p$ , i.e.,  $P(1) = +\Delta_X p$ , then we have two cases: (1)  $Xp \in F$ , (2)  $\exists r \in R_s^X[p], A(r) = \emptyset$ . The first case is trivial since  $F$  is the same in  $D$  and  $\text{elimsup}(D)$ . For (2) we have that  $\text{elimsup}(D)$  contains the rules  $r^a : \rightarrow_{\text{BEL}} \neg \text{inf}(r)$ , and  $r^c : \neg \text{inf}(r) \rightarrow p$ .  $r^a$  is applicable, so we have  $+\Delta_{\text{BEL}} \neg \text{inf}(r)$ , then this makes  $r^c$  applicable and then we have  $\text{elimsup}(D) \vdash +\Delta_X p$ .

For the inductive step, we assume as usual that the property holds for proofs whose length is up to  $n$ , and then we consider  $P(n+1) = +\Delta_X p$ . Beside the cases for the inductive base, we have two additional cases to consider here: (a)  $\exists r \in R_s^X[p], \forall a \in A(r), +\Delta_Y a \in P(1..n)$ ; (b)  $\text{Convert}(X, Y)$  and  $\exists s \in R_s^Y, +\Delta_X a \in P(1..n)$ .

For (a) by inductive hypothesis,  $\forall a \in A(r)$ ,  $\text{elimsup}(D) \vdash +\Delta a$ , thus the rule  $r^a : A(r) \rightarrow_{\text{BEL}} \neg \text{inf}(r)$  is applicable, thus  $\text{elimsup}(D) \vdash +\Delta_{\text{BEL}} \neg \text{inf}(r)$ , which makes rule  $r^c : \neg \text{inf}(r) \rightarrow_X p$  applicable as well, and we can conclude  $\text{elimsup}(D) \vdash +\Delta_X p$ .

For (b) by inductive hypothesis  $\forall a \in A(r)$ ,  $\text{elimsup}(D) \vdash +\Delta_X a$ , thus we can use the rule  $s^a : A(r) \rightarrow_{\text{BEL}} \neg \text{inf}(s)$  to derive  $+\Delta_X \neg \text{inf}(s)$ . Since we have  $\text{Convert}(Y, X)$ , we can apply conversion to the rule  $s^c : \neg \text{inf}(s) \rightarrow_Y p$  to derive  $\text{elimsup}(D) \vdash +\Delta_X p$ .

For the other direction, i.e.,  $\text{elimsup}(D) \vdash +\Delta_X p$  (for  $p \in \Sigma$ ) then  $D \vdash +\Delta_X p$ , the proof is again by induction on the length of derivations.

The inductive base is trivial since the only possible derivation for a modal literal  $Xp$  in  $\Sigma$  is only when  $Xp \in F$ , and thus  $Xp$  is also a fact in  $D$ ,

For the inductive bases,  $P(n+1) = +\Delta_X p$  we have that for every literal in  $\Sigma$  which is not a fact,  $\exists r \in R$  such that either (i)  $\neg inf(r) \rightarrow_X p$  or (ii)  $\neg inf(r) \rightarrow_Y p$  is in  $elimsup(D)$ . In addition we have a rule  $A(r) \rightarrow_{BEL} \neg inf(r)$ .

For (i) in the proof we have  $\forall a \in A(r), +\Delta a \in P(1..n)$ , thus by inductive hypothesis  $D \vdash +\Delta a$ , which makes applicable the rule  $r : A(r) \rightarrow_X p$ . For (ii) to derive  $+\Delta_X p$  from  $\neg inf(r) \rightarrow_Y p$ , we must have  $+\Delta_X \neg inf(r) \in P(1..n)$ , which means that we have  $+\Delta_X a \in P(1..n)$  for all  $a \in A(r)$ . Again by inductive hypothesis we have  $D \vdash +\Delta_X a$  for all  $a \in A(r)$ , and  $r$  is  $A(r) \rightarrow_Y p$  were  $Convert(Y, X)$ . Therefore  $D \vdash +\Delta_X p$ .

The proof for  $-\Delta_X$  is analogous and uses the same ideas of conversion from the case for  $+\Delta$  and the basic structure from the proof for the transformation that removes the superiority relation from [3].

The proof of the case for  $+\partial$  is essentially the same as that given in [3]. The only difference is in the iterative construction of the sets of maximal applicable rules, the existence of such sets is guaranteed by the clause of the proof conditions saying  $t > s$ . If a rule  $r$  is maximal applicable then either  $\forall a \in A(r), +\partial a$  or  $\forall a \in A(r), +\partial_X a$  (applicable condition), and there is no applicable rule  $s$  such that  $s > r$ . Thus all rules  $\neg inf(s) \leftrightarrow_{BEL} inf(r)$  are discarded while the rule  $A(r) \leftrightarrow_{BEL} \neg inf(r)$  is applicable, thus we prove either  $+\partial_{BEL} \neg inf(r)$  or  $+\partial_X \neg inf(r)$ . Thus every rule  $\neg inf(r) \Rightarrow_{BEL} inf(t)$ , is applicable, this means that we prove  $-\partial_Z \neg inf(t)$  for all  $Z \in MOD$ . The main points here is that BEL converts universally and that there are conflict between all pairs of modalities. Accordingly the rule  $t^c$ , attacking a rule for  $p$  is discarded. Using all rules in the maximal applicable sets we can show that all rules attacking  $p$  are discarded, and that we have at least one applicable rule for  $p$ . The proof for  $-\partial$  has the same structure of that given in [3] for the same case and the construction just outlined for the case  $+\partial$ . ■

DEFINITION 16. Let  $D = (F, R, >)$  be a modal defeasible theory, and  $\Sigma$  be the language of  $D$ . Define  $elimdft = (F, R', >')$  where

$$R' = \bigcup_{r \in R} elimdft(r)$$

and

$$elimdft(r) = \begin{cases} \{r^+ : A(r) \leftrightarrow_{BEL} p^+, r^- : A(r) \leftrightarrow_{BEL} \neg p^-, r : p^+ \leftrightarrow_X p\} & r \in R_{sd}^X[p] \\ \{r^- : A(r) \leftrightarrow_{BEL} p^-, r^+ : A(r) \leftrightarrow_{BEL} \neg p^+, r : p^- \leftrightarrow_X \neg p\} & r \in R_{sd}^X[\neg p] \\ \{r : A(r) \Rightarrow_{BEL} \neg p^-\} & r \in R_{dft}[p] \\ \{r : A(r) \Rightarrow_{BEL} \neg p^+\} & r \in R_{dft}[\neg p] \end{cases}$$

and the superiority relation  $>'$  is defined by the following conditions:

$$\forall r', s' \in R' (r' >' s' \iff \exists r, s \in R : r' \in elimdft(r), s' \in elimdft(s), r > s)$$

where  $r$  and  $s$  are conflicting.

For each atom  $p \in \Sigma$ ,  $p^+$  and  $p^-$  are new atoms, i.e., they do not appear in  $\Sigma$ . Furthermore all new atoms generated are distinct.

THEOREM 3. The transformation  $elimdft$  is correct.

PROOF. Notice that the transformation *elimdft* is essentially the same transformation as that given in [3]. The only difference is that the rules  $p^+ \leftrightarrow_X p$  and  $p^- \leftrightarrow_X \neg p$  are modalised with  $X$  instead of BEL. However, this difference is flattened by the definition of social agents, where BEL converts universally and the all modalities are involved in conflicts. ■

**(2) Rule Loader** The rule loader builds a data structure as follows: for every atom  $\alpha \in U^D$  we create three entries  $\alpha$ , INT $\alpha$  and OBL $\alpha$ . Each entry has associated to it a list of hash tables:

For  $\alpha$  we have

- $+h$  is a list of (pointers to) rules in  $R^{\text{BEL}}$  where  $\alpha$  appears in the head;
- $-h$  is the list of rules in  $R^{\text{BEL}}$  where  $\sim\alpha$  appears in the head;
- $+b$  is the list of rules in  $R$  where  $\alpha$  occurs in the body;
- $-b$  is the list of rules in  $R$  where  $\sim\alpha$  occurs in the body.

For  $X\alpha$ ,  $X \in \{\text{INT}, \text{OBL}\}$  we have

- $+h$  is a list of rules in  $R^X$  where  $\alpha$  appears in the head;
- $-h$  is the list of rules in  $R^X$  where  $\sim\alpha$  appears in the head;
- $+h^B$  is a list of rules in  $R^{\text{BEL}}$  where  $\alpha$  appears in the head;
- $-h^B$  is a list of rules in  $R^{\text{BEL}}$  where  $\sim\alpha$  appears in the head;
- $+b$  is the list of rules in  $R$  where  $X\alpha$  occurs in the body;
- $-b$  is the list of rules in  $R$  where  $X\sim\alpha$  occurs in the body.
- $+b_\sim$  is the list of rules in  $R$  where  $\sim X\alpha$  occurs in the body;
- $-b_\sim$  is the list of rules in  $R$  where  $\sim X\sim\alpha$  occurs in the body.

To each rule in  $R^X$ ,  $X \neq \text{BEL}$ , we associate a structure consisting of a (modal) literal (the head of the rule) and a set of pointers to the modal literals in the body of the rule, implemented as an hash table; while for belief rules we create the same structure as the other types of rules plus two other structures one for INT and one for OBL, the single pointer refers to the modal literal and the set of pointers corresponds to the literals in the body modalised, respectively, with INT and OBL.

**(3) The Inference Engine** The Inference Engine is based on an extension of the *Delores* algorithm/implementation proposed in [25] as a computational model of Basic Defeasible Logic. In turn

1. It asserts each fact (as an atom) as a conclusion and removes the atom from the rules where the atom occurs positively in the body, and it “deactivates” the rules where either the atom occurs negatively in the body, or incompatible modal literals occur in the body.

2. It scans the list of active rules for rules where the body is empty. It takes head and searches for rule (of the appropriate type) where the head is the negation of the atom or a modal literal incompatible with it. If there are no such rules then, the atom is appended to the list of facts, and removed from the rules.
3. It repeats the first step.
4. The algorithm terminates when one of the two steps fails. On termination the algorithm outputs the set of conclusions.<sup>8</sup>

It is immediate to see that the algorithm runs in linear time. Each (modal) atom/literal in a theory is processed exactly once and every time we have to scan the set of rules, thus the complexity of the above algorithm is  $O(|U^D| * |R|)$ . ■

Given the above result it might seem that social agents are computationally feasible. However, as we have seen in the previous sections there are situations (let us call them deviant situations) where social agents do not behave as expected. First of all, we have to identify when we have a deviant situation and what are the reasons why we have them, and what kind of control an agent has over them. Here we assume that a deviant situation depends on some primitive intentions of an agent (i.e., intentions given as facts). Since these intentions are independent of the policy the theory describe the only alternative a social agent has is to give up some of them. In the rest of the section we study whether this is possible and what price an agent has to pay to be social. The answer is negative; we will provide a theory that is essentially deviant, and we will show that social agents are (computationally) expensive.

A precise definition of the problem is provided in the next section.

### 5.3 Restoring Sociality Problem

INSTANCE:

Let  $I$  be a finite set of primitive intentions,  $OBLp$  a primary obligation, and  $D$  a theory such that  $I \subseteq F$ ,  $D \vdash -\partial_{OBL}p$ ,  $D \vdash -\Sigma_{OBL}\sim p$ ,  $D \vdash +\partial_{INT}\sim p$ ,  $D \vdash +\Sigma_{OBL}p$  and  $D \vdash -\Sigma_{BEL}\sim p$ .

QUESTION:

Is there a theory  $D'$  equal to  $D$  apart from containing only a proper subset  $I'$  of  $I$  instead of  $I$ , such that  $\forall q$  if  $D \vdash +\partial_{OBL}q$  then  $D' \vdash \partial_{OBL}q$  and  $D' \vdash +\partial_{OBL}p$ ?

The specification of the problem is meant to formalise the situation we have described in the previous sections. The combination of the proof tags in the specification of the instance is only possible in case there is an applicable deontic rule for  $p$  ( $+\Sigma_{OBL}p$ ) such that (i) would be otherwise unchallenged –i.e., there are no deontic rules to support  $\sim p$  ( $-\Sigma_{OBL}\sim p$ ); (ii) there are no reasons to believe the opposite of the conclusion of the deontic rule; (iii) but the deontic rule is defeated, against the sociality

<sup>8</sup>This algorithm outputs  $\partial^+$ ;  $\partial^-$  can be computed by an algorithm similar to this with the “dual actions”. For  $\Delta^+$  we have just to consider similar constructions where we examine only the first parts of step 1 and 2.  $\Delta^-$  follows from  $\Delta^+$  by taking the dual actions.

of the agent, by the intentionality of  $\sim p$  obtained as a consequence of an intention of the agent (this means it has been obtained by converting a belief rule into an intention rule). In other terms a potentially valid obligation is blocked by a consequence of an intentional behaviour.

EXAMPLE 5. Let us consider the theory consisting of

$$\begin{aligned} F &= \{\text{INT}p, \text{INT}s\} \\ R &= \{r_1 : p, s \Rightarrow_{\text{BEL}} q \quad r_2 : \Rightarrow_{\text{OBL}} \sim q \quad r_3 : \Rightarrow_{\text{BEL}} s\} \\ &> = \{r_1 > r_2\} \end{aligned}$$

$r_1$  is a belief rule and so the rule is stronger than the deontic rule  $r_2$ . In addition we have that the belief rule is not applicable (i.e.,  $-\Sigma_{\text{BEL}}q$ ) since there is no way to prove  $+\partial_{\text{BEL}}p$ . There are no deontic rules for  $q$ , so  $-\partial_{\text{OBL}}q$ . However, rule  $r_1$  behaves as an intention rule since all its antecedent can be proved as intentions, i.e.,  $+\partial_{\text{INT}}p$  and  $+\partial_{\text{INT}}s$ . Hence, since  $r_1$  is stronger than  $r_2$ , the derivation of  $+\partial_{\text{OBL}}\sim q$  is prevented against the sociality of the agent.

The related decision problem is whether it is possible to avoid the “deviant” behaviour by giving up some primitive intentions, retaining all the (primary) obligations, and maintaining a set of primitive intentions as close as possible to the original set of intentions.

EXAMPLE 5. (CONTINUED). When we examine the theory we notice that both primitive intentions concur to the prevention of the derivation of  $+\partial_{\text{OBL}}\sim q$ . These intentions are under the control of the agent. The agent has the opportunity to avoid the deviant behaviour if she gives up at least one of her primitive intentions. Accordingly, the agent has three alternatives: to give up  $\text{INT}p$ , to give up  $\text{INT}s$ , or to give up both. The first two options minimise the difference between the original theory and the resulting theory.

There could be cases where, no matter what intentions are removed, the theory will result in a deviant situation. The simplest case is where there are intentions that are at the same time primitive and primary.

EXAMPLE 6. Let the theory  $D$  be

$$\begin{aligned} F &= \{\text{INT}p\} \\ R &= \{r_1 : \Rightarrow_{\text{INT}} p \quad r_2 : p \Rightarrow_{\text{BEL}} q \quad r_3 : \Rightarrow_{\text{OBL}} \sim q\} \\ &> = \{r_2 > r_3\} \end{aligned}$$

In this theory we have only one primitive intention and therefore the only way to see whether it is possible to avoid the problem is to give up that intention. However, we have that  $r_1$  is an intention rule for  $p$ , and thus we can use it to derive  $+\partial_{\text{INT}}p$ , which allows  $r_2$  to be used to derive an intention instead of a belief, and consequently to prevent the derivation of an obligation against the sociality of the agent.

Notice that, given the non-monotonic nature of defeasible logic, it is possible that a solution to the problem is given by a superset of the original set of intentions instead of a subset.

EXAMPLE 7. Given a theory  $D$  as follows

$$\begin{aligned}
F &= \{\text{INT}a, \text{INT}b\} \\
R^{\text{BEL}} &= \{r_1 : \text{INT}a \Rightarrow_{\text{BEL}} d, \quad r_2 : \text{INT}b \Rightarrow_{\text{BEL}} d, \\
&\quad r_3 : \text{INT}c \Rightarrow_{\text{BEL}} \sim d, \quad r_4 : d \Rightarrow_{\text{BEL}} e\} \\
R^{\text{INT}} &= \{r_5 : \Rightarrow_{\text{INT}} a, \quad r_6 : \Rightarrow_{\text{INT}} b\} \\
R^{\text{OBL}} &= \{r_7 : \Rightarrow_{\text{OBL}} \sim e\} \\
>= &= \{r_3 > r_1, r_3 > r_2, r_4 > r_7\}
\end{aligned}$$

As we have seen in the previous example, throwing away the two primitive intentions is of no avail, they are reinstated by the intention rules  $r_5$  and  $r_6$ . However, to block the side effect  $d$  of the two intentions we can introduce a further primitive intention,  $\text{INT}c$ .

If we replace the theory  $D$  by a theory  $D'$  obtained from  $D$  by emptying the set of intention rules, then we have two alternatives to avoid the deviance. The first is to drop both the primitive intentions  $\text{INT}a$  and  $\text{INT}b$ , or we can form a new primitive intention  $\text{INT}c$ . In this case the theory obtained from adding the new intention is, intuitively, more similar to the original theory than the theory obtained from dropping the two primitive intentions.

Variations of the problem can be obtained by changing other parameters of the specification. Some of these can define new types of agents. For example a *pro-active social agent* might try to recover from a deviant situation by changing the raw facts (facts that are neither primitive intentions nor primitive obligations). Thus a pro-active social agent tries to adapt the environment to her goals (intentions). A legalistic social agent, on other the hand, might change the set of primitive obligations, while a cheating social agent might change the rules. However, it is important to realise that all these variations have a structure isomorphic to the specification we discuss in this paper. In addition it is possible to generalise the problem to the case of multiple deviant behaviours.

THEOREM 4. *The Restoring Sociality Problem is NP-complete.*

PROOF. We have to show that the problem is both NP and NP-hard. For the NP part all we have to do is to notice that we can guess a theory, we compute the extension of the theory in linear time (Theorem 1) and then verify in linear time whether the restore conditions are satisfied.

For the NP-hard part we have to map a known NP-complete problem to the Restoring Sociality Problem. Here we use the *knapsack problem* [14, Problem MP9].

### Knapsack Problem

INSTANCE:

Given a finite set  $U$ , for each  $u \in U$  a size  $s(u) \in \mathbb{Z}^+$  and a value  $v(u) \in \mathbb{Z}^+$ , and integer  $B$  and  $K$ .

QUESTION:

Is there a subset  $U' \subseteq U$  such that  $\sum_{u \in U'} s(u) \leq B$  and  $\sum_{u \in U'} v(u) \geq K$ ?

The knapsack problem is encoded by a defeasible theory  $D$  where  $R$  is as follows:

- $\text{INTload}(u) \Rightarrow_{\text{BEL}} \text{load}(u)$  for each  $u \in U$ .
- $\sum_{s(u):D \vdash +\partial_{\text{BEL}} \text{load}(u)} s(u) > B \Rightarrow_{\text{INT}} \text{overload}$
- $\sum_{s(u):D \vdash +\partial_{\text{BEL}} \text{load}(u)} v(u) < K \Rightarrow_{\text{INT}} \text{undervalue}$
- $\text{overload} \Rightarrow_{\text{BEL}} \neg \text{good}$
- $\text{undervalue} \Rightarrow_{\text{BEL}} \neg \text{good}$
- $\Rightarrow_{\text{OBL}} \text{good}$

$F$  is given by the relationship  $\text{INTload}(u) \in F$  iff  $u \in U'$ .

The theory of the above construction has several interesting properties. First of all  $D \vdash +\partial_{\text{BEL}} \text{load}(u)$  iff  $\text{INTload}(u) \in F$ , which means  $u \in U'$ ; then  $D \vdash +\partial_{\text{OBL}} \text{good}$  iff either of the two conditions of the knapsack problem are satisfied; notice that since there are no literals for  $\neg \text{load}(u)$ , the computation of the rule  $\text{INTload}(u) \Rightarrow_{\text{BEL}} \text{load}(u)$  can be computed independently of the rest of the theory thanks to the modularity of DL [3], thus the sums in the antecedent of the second and third rule can be considered as “facts” in the theory. In case one of the condition of the knapsack problem is not satisfied we have exactly a deviant situation as in the restoring sociality problem. The encoding of the knapsack problem in DL is clearly linear, thus any algorithm that solves the restoring sociality problem in polynomial time will solve the knapsack problem in polynomial time. Therefore the restoring sociality problem is NP-complete. ■

## 5.4 Revising Deviant Situations

In this paper we focused on what we called social agents, i.e., agents who refrain from planning activities which may result in a violation of existing obligations. However, we would like to stress out that the so called “restoring sociality problem”, and the computational complexity results associated with it, is not specific to social agents, but it depends on the structure of an agent type. In particular any agent type defined by the following parameters

$$\text{Convert}(X, Y), \text{Conflict}(X, Z), \text{Conflict}(X, Y), \text{Conflict}(Z, Y)$$

suffers from the same problem (of course with a different intuitive reading of the problem).

In a similar way the transformations to remove defeaters and to empty the superiority relation, as well as the general complexity result for the logic obtain for all agent types (modal defeasible logic variants) isomorphic to social agents.

A first solution to the complexity of social agents is to avoid conversions. However, we believe that this is a rather unsatisfactory approach for agents with both internal (intentions) and external (obligations) motivational attitudes. It is not possible to capture the notion of intentionality which is of paramount importance when we deal with agents situated in normative contexts.

A second solution would be to assume that belief rules behaving as intention rules (i.e., obtained from the conversion  $\text{Convert}(\text{BEL}, \text{INT})$ ) are always weaker than deontic rules or belief rules behaving as deontic rules (i.e., where the conversion  $\text{Convert}(\text{BEL}, \text{OBL})$  applies). In this case the problem is with theory like

$$\begin{array}{ll} r_1 : a \Rightarrow_{\text{BEL}} q & r_2 : b \Rightarrow_{\text{BEL}} \sim q \\ +\partial_{\text{INT}} a & +\partial_{\text{OBL}} b \\ r_1 > r_2 \end{array}$$

where  $r_1$  is at the same time stronger and weaker than  $r_2$ .

## 6 Related Work

This article provides an extensive proposal of how DL can be extended to model cognitive agents interacting with obligations. In this sense, it is the final result of a series of earlier works [18, 10, 11, 19]. In [10, 11] DL is extended by introducing the  $\otimes$  operator to represent explicit violations and contrary-to-duty reasoning. Cognitive profiles of agents are characterised by their beliefs, desires, intentions and obligations. A large agent type classification is developed accordingly. In [18] a similar picture of agents is presented but desires are defined as literals supported by (but not necessarily derived from) reasoning chains of rules for intention. In addition, an operator for intentional and successful action is introduced. In all these works conflicts are simply modelled by using standard superiority relation ( $>$ ) of DL, and so it is never admitted that we can derive, for example,  $\text{INT}a$  and  $\text{OBL}\neg a$ . Moreover, conversions are not discussed in connection with the problem of side effects and no complexity result is offered. In [19], too, conflicts are simply modelled via the standard superiority relation. However, some preliminary discussion on the side-effect problem is developed and complexity results about the logic and social agents are sketched. Hence, the present article directly extends the analysis of [19]: it offers a more comprehensive discussion on intentional side effects and conversions, adopts a different and more general method for dealing with conflicts, proposes full proofs for complexity results, and suggests hints about how the problem of social agents also concerns other agent types.

Reasoning about mental attitudes is a central issue in philosophy and AI. Despite the plethora of proposals devoted to this topic, the related work that is directly relevant for this paper is mainly the BOID architecture. In fact, the basic calculation scheme used in BOID [8] is similar to the one proposed in this paper: as done in BOID, we distinguish conflicts between rules for the same modality and for different modalities. In the second case, the relation  $\text{Conflict}(X, Y)$  assumes that  $X$  rules are always stronger than  $Y$ 's.

The BOID framework has four components representing respectively the beliefs (B), obligations (O), intentions (I) and desires (D) of the agent. The behaviour of each component is specified by sets of propositional logical formulas often in the form of defeasible rules. BOID identifies two general types of conflicts that could arise either within each component (*internal conflicts*) or between the components (*external conflicts*). These two types of general conflicts are further subdivided into different subtypes which gives rise to several possible conflicts among the mental attitudes. In

order to solve possible conflicts among the attitudes an ordering function ( $\rho$ ) is defined on rules based on the *agent type*. An agent type is determined by allowing one component to overrule others. For example, a *realistic* agent type can be defined by having an ordering in which the belief component overrules any other component (BOID, BODI, BDIO etc.). This means that in BOID a conflict resolution type is an order of overruling and in general the order of derivation can be used to identify different types of agents. Agent types like *simple-minded* (agent type where prior intentions overrule desires and obligations), *social* (agent type where obligations overrule desires) etc. could be defined in a similar manner. Formally an agent type is defined as a function,  $\rho$  that assigns a unique integer to each rule. It should be noted that the ordering function  $\rho$  assigns unique values to the rules of all components such that the values of all rules from one component are either smaller or greater than the values of all rules from another component.

Besides the specific result discussed in Section 5.3, the general aspects that differentiate the current framework from BOID's are the following:

- our proof conditions permit to derive modalised literals; accordingly, in addition to labelling rules by the elements of MOD, modalities are also made explicit in rule antecedents, thus enriching the expressive power of the logic;
- conversions are introduced to capture some fundamental reasoning patterns which, in most cases, should be admitted or which may in any case contribute to characterise agent types;
- we admit that Conflict may cover only some modalities; this makes it possible that, for any rule types  $X$  and  $Y$  that are not covered by Conflict, we can obtain  $+\partial_X p$  and  $+\partial_Y \sim p$ ;
- our logic for BIO agents has linear complexity, whereas to our knowledge there is no analogous result for BOID.

## 7 Summary

In conclusion, let us summarise step by step the aims and results of this article.

Our preliminary step was to describe agent's deliberation by considering her policy-based motivations, which are triggered by potentially recurring circumstances in agent's life. In particular, we extended Bratman's model of policy-based intentions to also cover beliefs and obligations. It turned out that this type of motivations are easily captured by a rule-based approach to cognitive agents.

Secondly, on account of this definition of motivations, we discussed some aspects of the so-called side-effect problem. In contrast with the idea that side effects are never intended, we argued that there are conceptual reasons for arguing that some side effects should be intended, at least according a realistic model of agent's rationality. We maintained that the inclusion of some side-effects in the intentional sphere of agents does not endanger the logical analysis but, on the contrary, is beneficial to explain notions, such as intentionality and responsibility, of paramount importance for agents situated in normative and legal contexts.

Thirdly, the logical framework has been presented. Before providing a rigorous definition of the formal language and proof conditions, we informally introduced the concept of rule conversion, according to which we can derive some motivations by using rules devised for inferring different motivations. In addition, we discussed methods for dealing with rule conflicts. An intermediate summary was provided to outline our logical intuitions and match them with the conceptual issues regarding the side-effect problem. We argued that conversions are a natural way to include or exclude side effects.

Fourthly, we illustrated the notion of agent type. Classically, agent types are characterised by stating conflict resolution types in terms of orders of overruling between rules. For example, an agent is realistic when rules for beliefs override all other components; she is social when obligations are stronger than the other motivational components with the exception of beliefs. We argued that agent types are not only useful in devising mechanisms for solving conflicts, but are of theoretical interest, as they define the cognitive profile of agents. We focused in particular on social agents.

Fifthly, we investigated the computational properties of our logical framework. First of all, we showed the computational feasibility of the logic: we have demonstrated that it has linear complexity. As far as we know this is the first result of this kind for cognitive agents. We then moved to critically examining the concept of social agent, but we argued that our considerations can be easily applied to the other agent types: in fact, the analysis was mainly formal and independent of what motivational factors (such as intentions and obligations) are considered. In particular, we proved that the classical notion of agent type is not satisfactory: in presence of conversions, which seem necessary to deal with the side-effect problem, conflict resolutions cannot be limited to examining pairs of rules having complementary literals in their heads, but we need to consider all possible reasoning chains supporting conclusions. This problem turned out to be very expensive from the computational point of view. Again, this is the first result of this kind we are aware of. In addition, although we showed that this difficulty formally holds for DL only, we also argued that similar problems affect any rule-based defeasible formalism which incorporates conversions or analogous inferential mechanisms.

## References

- [1] Frederick Adams. Intention and intentional action: The simple view. *Mind and Language*, 1:281–301, 1986.
- [2] Grigoris Antoniou, David Billington, Guido Governatori, and Michael J. Maher. A flexible framework for defeasible logics. In *Proc. American National Conference on Artificial Intelligence (AAAI-2000)*, pages 401–405, Menlo Park, CA, 2000. AAAI/MIT Press.
- [3] Grigoris Antoniou, David Billington, Guido Governatori, and Michael J. Maher. Representation results for defeasible logic. *ACM Transactions on Computational Logic*, 2(2):255–287, 2001.

- [4] Nick Bassiliades, Grigoris Antoniou, and Ioannis Vlahavas. DR-DEVICE: A defeasible logic system for the Semantic Web. In Hans Jürgen Ohlbach and Sebastian Schaffert, editors, *2nd Workshop on Principles and Practice of Semantic Web Reasoning*, number 3208 in LNCS, pages 134–148. Springer, 2004.
- [5] David Billington. Defeasible logic is stable. *Journal of Logic and Computation*, 3, 1993.
- [6] M.E. Bratman, D.J. Israel, and M.E Pollack. Plans and resource-bounded practical reasoning. *Computational Intelligence*, 4:349–355, 1988.
- [7] Michael E. Bratman. *Intentions, Plans and Practical Reason*. Harvard University Press, Cambridge, MA, 1987.
- [8] Jan Broersen, Mehdi Dastani, Joris Hulstijn, and Leendert van der Torre. Goal generation in the BOID architecture. *Cognitive Science Quarterly*, 2(3-4):428–447, 2002.
- [9] P. Cohen and H. Levesque. Intention is choice with commitment. *Artificial Intelligence*, 42:213–261, 1990.
- [10] Mehdi Dastani, Guido Governatori, Antonino Rotolo, and Leendert van der Torre. Preferences of agents in defeasible logic. In S. Zhang and R. Jarvis, editors, *Proc. Australian AI05*, volume 3809 of *LNAI*, pages 695–704. Springer, 2005.
- [11] Mehdi Dastani, Guido Governatori, Antonino Rotolo, and Leendert 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.
- [12] Frank Dignum. Autonomous agents with norms. *Artificial Intelligence and Law*, 7(1):69–79, 1999.
- [13] Frank Dignum, David Morley, Liz Sonenberg, and Lawrence Cavedon. Towards socially sophisticated BDI agents. In *ICMAS (4th International Conference on Multi-Agent Systems)*, pages 111–118, 2000.
- [14] Michael Garey and David Johnson. *Computers and Intractability: A Guide to the Theory of NP-Completeness*. W. H. Freeman and Company, 1979.
- [15] Rod Girle. *Modal Logic and Philosophy*. Acumen, Teddington, 2000.
- [16] Guido Governatori and Vineet Padmanabhan. A defeasible logic of policy-based intention. In *Proceedings of AI 2003*. Springer Verlag, 2003.
- [17] Guido Governatori and Antonino Rotolo. A computational framework for institutional agency. *Artificial Intelligence and Law*.
- [18] Guido Governatori and Antonino Rotolo. Defeasible logic: Agency, intention and obligation. In Alessio Lomuscio and Donald Nute, editors, *Deontic Logic in Computer Science*, number 3065 in *LNAI*, pages 114–128, Berlin, 2004. Springer-Verlag.

- [19] Guido Governatori, Antonino Rotolo, and Vineet Padmanabhan. The cost of social agents. In *5th International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS06)*, New York, 2006. ACM.
- [20] Lalana Kagal and Tim Finin. Modeling conversation policies using permissions and obligations. *Journal of Autonomous Agents and Multi-Agent Systems*, 14(2):187–206, 2007.
- [21] Joshua Knobe. Intentional action and side effects in ordinary language. *Analysis*, 63:190–193, 2003.
- [22] Sarit Kraus, Daniel Lehmann, and Menachem Magidor. Nonmonotonic reasoning, preferential models and cumulative logics. *Artificial Intelligence*, 44:167–207, 1990.
- [23] Michael J. Maher. Propositional defeasible logic has linear complexity. *Theory and Practice of Logic Programming*, 1(6):691–711, 2001.
- [24] Michael J. Maher, Andrew Rock, Grigoris Antoniou, David Billington, and Timothy Miller. Efficient defeasible reasoning systems. *International Journal of Artificial Intelligence Tools*, 10(4), 2001.
- [25] Michael J. Maher, Andrew Rock, Grigoris Antoniou, David Billington, and Tristan Miller. Efficient defeasible reasoning systems. *International Journal of Artificial Intelligence Tools*, 10(4):483–501, 2001.
- [26] Hugh McCann. Rationality and the range of intention. *Midwest Studies in Philosophy*, 10:191–211, 1986.
- [27] Alfred Mele and Steven Sverdlik. Intention, intentional action, and moral responsibility. *Philosophical Studies*, 82:265–287, 1996.
- [28] Donald Nute. Defeasible logic. In *Handbook of Logic in Artificial Intelligence and Logic Programming*, volume 3. Oxford University Press, 1987.
- [29] Donald Nute. Defeasible reasoning. In *Proceedings of 20th Hawaii International Conference on System Science*. IEEE press, 1987.
- [30] Donald Nute, editor. *Defeasible Deontic Logic*. Kluwer, Dordrecht, 1997.
- [31] Donald Nute. Norms, priorities, and defeasible logic. In Paul McNamara and Henry Prakken, editors, *Norms, Logics and Information Systems*, pages 201–218. IOS Press, Amsterdam, 1998.
- [32] J. Pitt, editor. *Open Agent Societies*. Wiley, Chichester, 2005.
- [33] Anand S. Rao and Michael P. Georgeff. Modelling rational agents within a BDI-architecture. In *Proceedings of the Second International Conference on Principles of Knowledge Representation and Reasoning (KR'91)*, pages 473–484. Morgan Kaufmann, 1991.

- [34] Giovanni Sartor. *Legal Reasoning: A Cognitive Approach to the Law*. Springer, Dordrecht, 2005.
- [35] John Searle, editor. *Intentionality*. Cambridge University Press, Cambridge, 1983.
- [36] Richmond H. Thomason. Desires and defaults: A framework for planning with inferred goals. In Anthony G. Cohn, Fausto Giunchiglia, and Bart Selman, editors, *KR2000*, San Francisco, 2000. Morgan Kaufmann.
- [37] Georg Henrik von Wright. *Norm and Action*. Routledge, London, 1963.