A Defeasible Logic for Modelling Policy-based Intentions and Motivational Attitudes

Guido Governatori[1], Vineet Padmanabhan[2],
Antonino Rotolo[3], Abdul Sattar[4]
[1]NICTA, Queensland Research Laboratory, Australia
[2]Information Sciences, University of Hyderabad, India
[3]CIRSIFD and Law Faculty, University of Bologna, Italy
[4]IIIS, Griffith University, Queensland, Australia.

Abstract

In this paper we show how *defeasible logic* could formally account for the non-monotonic properties involved in motivational attitudes like intention and obligation. Usually, *normal* modal operators are used to represent such attitudes wherein classical logical consequence and the rule of necessitation comes into play, i.e., $\vdash A / \vdash \Box A$, that is from $\vdash A$ derive $\vdash \Box A$. This means that such formalisms are affected by the *Logical Omniscience* problem. We show that policy-based intentions exhibit non-monotonic behaviour which could be captured through a non-monotonic system like defeasible logic. To this end we outline a defeasible logic of intention that specifies how modalities can be introduced and manipulated in a non-monotonic setting without giving rise to the problem of logical omniscience. In a similar way we show how to add deontic modalities defeasibly and how to integrate them with other motivational attitudes like beliefs and goals. Finally we show that the basic aspect of the BOID architecture is captured by this extended framework.

1 Introduction

Representing and reasoning about motivational attitudes like intention and obligation is crucial in the modelling of autonomous agents exhibiting cognitive and social characteristics [22, 55, 57, 26, 27]. An important logical tool used to formalise these attitudes is that of *Normal Multimodal Logics* (NML)[1] wherein intention/obligation are formalised as *normal* modal operators on the framework of Kripkean possible worlds semantics. Normal modal operators have certain *monotonic* properties that are occasionally considered undesirable for the common sense notions that they are intended to formalise (they suffer in particular

[1] General modal systems with an arbitrary set of normal modal operators all characterised by the axiom **K**: $\Box(\varphi \to \psi) \to (\Box\varphi \to \Box\psi)$ and the necessitation rule, i.e., $\vdash \varphi / \vdash \Box\varphi$.

This is the preprint version of the article.
The article has been accepted for publication in the *Logic Journal of the IGLP*
© Oxford University Press, 2009
The original publication is available at Oxford University Press.

from the *logical omniscience* problem [39, 63]). For instance, when formalising intention, one of the undesired properties we get is that of closure under logical consequence: given a normal modal operator $\mathbf{X}$ representing intention, if $\mathbf{X}\varphi$ is true, and $\varphi \models \psi$ (φ logically implies ψ), it follows that $\mathbf{X}\psi$ is true. Logically, consequential closure cannot be assumed for intention even as an idealisation, as not all the consequences of an agent's intention are intentions of the agent even if he/she has anticipated the consequences[2]. There are several other closely related problems of logical omniscience that plague formal models of intention based on NML and we discuss them in detail in Section 3. In the case of obligation, formalisation in terms of NML leads to paradoxes or counter-intuitive interpretations as shown in works like [20, 62]. Similar objections to NML's are also found when reasoning about knowledge and belief [28, 64, 65].

In this article we deviate from the NML approach and address the problem of formalising motivational attitudes like intention and obligation in a non-monotonic logic like *defeasible logic*. In order to achieve this end we take the following steps;

- Based on Bratman's classification of intention we outline a formal theory of *policy-based intention* which is non-monotonic (defeasible) in nature;
- We discuss different variants of logical omniscience with respect to intention and obligation and address these issues in terms of our defeasible framework;
- We provide a defeasible logic of policy-based intention and relate it with an intentional agent system like BDI;
- We outline an original analysis of interaction among mental attitudes (belief, intention, obligation) in defeasible logic.

The need to frame logics for cognitive states and motivational attitudes within a non-monotonic setting has been recently acknowledged. Here we have to mention in particular the BOID architecture [15, 17, 18], since this approach, as we shall see, has a number of similarities when compared to our system. The BOID approach is rule-based: motivational attitudes are represented as rules and a number of strategies are provided for detecting and solving conflicts among them. The BOID approach refers to Thomason's BDP logic [61], which in turn is based on Reiter's default logic. Roughly speaking, the basic idea behind BOID is to settle conflicts by stating different general policies corresponding to the agent type considered [16, 33]. In this perspective, a realistic agent corresponds to a conflict-resolution type in which beliefs override all other components, while other agent types, such as simple-minded, selfish or social ones adopt different orders of overruling. BOID system also incorporates obligations. This is in line with recent works on agents and their societies, in which it is assumed that as in human societies, also in artificial societies normative

[2] An agent that intends its teeth to be restored will in general not intend for the pain that inevitably accompanies such a restoration.

concepts may play a decisive role, allowing for the flexible co-ordination of intelligent autonomous agents [23, 54]. This intuition is adopted here too; on this basis we can get cognitive states to interact with one another and with obligations as well. This paper does not focus on the well-known problems that regard specifically defeasible normative reasoning [52]. However, it is worth noting that the adoption of a framework based on defeasible logic seems to be a promising approach to deal with obligations [53, 37] and, more generally with normative reasoning [3].

The paper is mainly restricted to the issues related with the idea of intention and its connection with other motivational attitudes such as obligations, beliefs and goals. As such, the paper will focus directly on the concept of intention and not provide an extensive account, for example, of the idea of obligation in itself, an idea that has received a great and specific attention in the recent literature on non-monotonic reasoning (for an overview, see, e.g., [52, 56]). For the same reason, we will not deal with the role played by action and commitment pertaining to agency.

The layout of the paper is as follows. Section 2 considers the case for a non-monotonic theory of intention based on Bratman's classification of intention. In Section 3 we discuss the problem of logical omniscience and, in Section 4, we briefly deal with some issues on how obligations may be incorporated in our framework. Section 5 gives an overview of defeasible logic. A new framework for a defeasible logic of intention and obligation is developed in Sections 6 and 7. Section 8 extends the discussion of Section 7 with some notes on the notion of permission and its interplay with mental states. Although we do not touch upon the issue of reconsideration, Section 9 outlines the main differences between the non-monotonicity behind reconsideration of deliberative intentions and the non-monotonicity determined by the defeasibility of policy-based intentions. Section 10 shows that the basic aspect of the BOID architecture is a particular case of our framework. Finally, in Section 11 we suggest and summarise the findings and point to several directions in which the current work could be extended.

2 Policy-based Intentions and Defeasibility

In his effort to provide an *extended planning theory* of intention Bratman [13, 14] coined the concept of a policy-based attitude with specific reference to the idea of intention. The intuition behind policy-based intention is based on a distinction between *future-directed* intentions and *general* intentions. The former is concerned with *intentions to act in a certain way on a certain relatively specific occasion* (I may intend to go to Boston *tomorrow* or *some time next month*). The latter is more *general with respect to their occasions of execution*, i.e., an intention to act in a certain way whenever a certain type of occasion is present or an intention to act in a certain way on a *regular* basis. For example, I have a general policy to patch up and reboot the Unix server, *hobbit* 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. Bratman terms such general intentions as *personal policies* and argues that an extended planning theory of intention must include such personal policies [14]. In this paper we provide a formal framework based on defeasible logic to capture the properties involved in *policy-based* intentions.

Before getting into the details of the formal framework, we need to outline some of the properties characteristic of policy-based intention. Based on the distinction between future-directed and general intention as outlined above, Bratman classifies intention into three categories; *deliberative, non-deliberative* and *policy-based.* An intention held at a given moment is "deliberative" or "non-deliberative" depending on whether it was formed out of immediately prior deliberation or at some time in the past. For instance, when an agent i has an intention of the form $\mathrm{INT}_i^{t_1}\varphi, t_2$ (read as *agent i intends at t_1 to φ at t_2*) as a process of *present* deliberation, then we have a *deliberative intention.* On the other hand, if the agent comes to have such an intention not on the basis of present deliberation, but at some earlier time t_0 and has retained it from t_0 to t_1 without reconsidering it, then this intention is called *non-deliberative.* The third case arises when intentions are general and concern potentially recurring circumstances in an agent's life and is defined as follows:

> *when the agent i has a general intention/personal policy to φ in circumstances of type ψ and i notes at t_1 that i is (will be) in a ψ-type circumstance at t_2, and thereby i arrives at an intention to φ at t_2.*

From the above definition it can be noted that a *policy-based intention* is not a non-deliberative intention because it is not simply a case of retaining an intention previously formed. Neither is it a deliberative intention since it is not based on 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., 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[3]. The difference here is that in each case the policy concerns not just a single future situation, but 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. When dealing with such personal policies we have to take into account two cases. Policy-based intentions could be either (1) *periodic* or (2)

[3] *If I begin with the intended end of running today, but balk at some necessary means to so acting I will be rationally obliged to give up my intended end. This contrasts with a case in which one does not apply a policy to a particular case due to that policy's defeasibility; for in this latter case the fact that one does not settle on a specification of the policy in the particular case does not force one to give up the policy. In special circumstances I might block the application of my policy of* buckling up while driving *in an emergency situation without abandoning that policy; for my policy is defeasible* [12].

circumstance-triggered. They are *periodic* in the sense that their occasion for execution is guaranteed by the mere passage of a specific interval of time and they are circumstance triggered as their execution requires that certain specific circumstances obtain.

Example 1. Consider a TV agent whose main objective is to make recommendations of TV programs that the user is likely to want to watch given the user's interests and preferences. The agent must be able to decide whether or not to recommend a particular TV program to the user for a specific duration. The TV agent has access via the Internet to a TV guide (for the next week or so) for all Australian channels. Then the *weekly update of the TV guide* could be considered as a policy-based intention that is `periodic` for the agent. In contrast to this, a policy-based intention could be `circumstance triggered` as is the case when the TV agent has a personal policy to *recommend TV-programs the user likes.* Its occasion is not guaranteed by the mere passage of time but requires that certain specific circumstances obtain.

In both cases one can find that the policy-based intention has an underlying defeasible nature in the sense that the agent does not intend so to act *no matter what.* For instance, the agent cannot do the weekly update of the TV guide if the network is down due to some unforeseen circumstances. As for *circumstance-triggered* policy-based intention, suppose that the agent finds two programs the user likes which are scheduled at the same time. Then he/she has to recommend one of them on the basis of the ratings given in the TV guide or based on other preferences of the user. In such scenarios the personal policies needs to be *reconsidered* if not *blocked* for application to particular cases. Faced with new information about "User preferences" the agent might need to only block the application of the policy to the particular case.

Having outlined the defeasible character of personal policies/policy-based intentions we now move on to discuss different variants of the logical omniscience problem that plague logical models of agents with specific reference to models of intention. We show that in the case of policy-based intention by adopting a non-monotonic framework we can avoid some unwanted properties associated with logical omniscience. At the same time we also show that retaining the *side-effect problem* (a variant of the logical omniscience problem) is not a weakness of our framework.

3 Logical Omniscience and Non-Monotonicity

3.1 Three Faces of Logical Omniscience

Any logic based agent model has to face the (logical) omniscience problem. Traditionally this was considered to be only an epistemic problem, but it also extends to cover other mental and motivational attitudes such as intentions and obligations. Girle [30] proposes a threefold partition of the omniscience problem: *logical omniscience*, *deductive omniscience* and *factual omniscience*. Logical omniscience in turn is divided in two sub-cases: strong logical omniscience and weak logical omniscience.

An agent is *factually omniscient* if for every proposition A she knows whether A or $\neg A$. So the agent has a complete knowledge of the environment. However, this is a too strong and unrealistic assumption.

In this work we are going to use policies to describe the external and internal motivational attitudes as well as the mental states of an agent. Policies are represented by rules, that, in general, represent the expected normal behaviour of an agent, and in some cases, the agent can override the rule without having to reconsider its whole policy. This means that there are cases where a rule may be overridden by more specific rules. In other words rules are defeasible and admit exceptions. In addition there can be several different exceptions to one policy. In this perspective a policy and its exceptions can be represented by the following set of rules

$$A_1 \wedge \ldots \wedge A_n \rightarrow B, \tag{1}$$

$$A_1 \wedge \ldots \wedge A_n \wedge E_1 \rightarrow \neg B, \tag{2}$$

$$\vdots$$

$$A_1 \wedge \ldots \wedge A_n \wedge E_m \rightarrow \neg B. \tag{3}$$

Hence if the model is based on any extension of classical propositional logic (or on any monotonic logic), we obtain an inconsistency as soon as one of the exceptions is applicable. Indeed the applicability of one of the exceptions implies that the main rule (1) is also applicable. A way to obviate this issue is to rewrite the main rule to take into account the possible exceptions. Thus the first rule must be rewritten as

$$A_1 \wedge \ldots \wedge A_n \wedge \bigwedge_{i=1}^{m} \neg E_i \rightarrow B \tag{4}$$

The above representation still suffers from several drawbacks. To apply this rule the agent has to know for each E whether it is true or false. This means that if the agent only has partial knowledge of the environment she cannot draw conclusions about her mental and motivational attitudes, and in the extreme case this approach would require factually omniscient agents. A possible solution to this issue is to base the logic for the motivational attitudes of an agent on formalisms designed to solve it, for example, on non-monotonic logics. Here we propose Defeasible Logic, a computationally oriented non-monotonic system, as the logical base for our logic on intentions and obligations, as well as the logic for representing the knowledge (belief) of the agent.

An agent is *strongly logical omniscient* when she knows all the truths defined by her logic and *weakly logical omniscient* in case the logic that depicts the agent automatically includes all the logical truths of classical (predicate) logic. As we have already argued an agent based on classical logic is deemed to be inconsistent if she is rule-based and has to reason with partial information.

An agent is *deductively omniscient* if she knows all the logical consequences of known propositions.

According to the above definitions logical omniscience is a semantic problem, and it includes deductive omniscience in the sense that if the the logic of the agent is consistent and the known propositions are assumed to be true, then the deductive consequences will be true. However, there could be true formulas that do not depend on the known propositions. A strongly logical omniscient agent has to accept them, while this is not the cases for a deductively omniscient agent. As we will see logical omniscience will collapse into deductive omniscience in the defeasible logic of intentions and obligations we are going to develop in the rest of the paper, since all the true formulas of the logic depend on the basic knowledge of the agent. Hence our focus will be deductive omniscience.

3.2 Side-effect problems of Logical Omniscience

Given a logic, a theory is the closure of the base of the theory (a set of formulas, in this case the set of known propositions or the set of appropriate mental and motivational attitudes) under the logic. The first issue here is that, unless we unrealistically assume agents with unlimited computational power and capacity, the agent is not able to compute the extension of the theory. The second issue is whether and how the basic inferential mechanism interacts with the modalities. To obviate the first problem we will use defeasible logic, a simple, efficient and flexible non-monotonic logic as the basic inferential engine of our logic. The extension of a theory in defeasible logic is finite and it can be computed in linear time. Of course this implies a trade-off between the expressive power of the logic and its computational complexity. However it is possible to argue that the limitations are not so severe and several works propose defeasible logic as representation and reasoning formalism in many application areas (see for example [38, 31, 37, 8, 24]).

To deal with the second part of the problem, i.e., the interactions between the inferential mechanism and the modalities, we have to take into consideration several variants of logical omniscience that usually accompany normal modal operators and are commonly referred to as the *expected side-effects* problem [13]. Often termed as the (in)famous problems of logical omniscience [63] they can be formulated as in Table 1 (**X** represents motivational attitudes like intention (INT) or obligation (OBL)). Before getting into the details of each property we want to sketch our intuition regarding the consequence relations for the expected side-effects problem.

The side-effects problem depends on the interactions between the reasoning mechanism for the propositional inferences and the mechanism ruling the introduction and the behaviour of the modal operators. Accordingly 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 $\Gamma \mathrel{|\!\!\sim}_{\mathbf{X}} \alpha$, where **X** is a modal operator, means that if we can prove all

	Property	Problem
(A)	$\mathbf{X}\varphi \wedge \mathbf{X}(\varphi \rightarrow \psi) \Rightarrow \mathbf{X}\psi$	*side-effect problem*
(B)	$\varphi \rightarrow \psi \Rightarrow \mathbf{X}\varphi \rightarrow \mathbf{X}\psi$	*side-effect problem*
(C)	$\varphi \Leftrightarrow \psi \Rightarrow \mathbf{X}\varphi \Leftrightarrow \mathbf{X}\psi$	*side-effect problem*
(D)	$\varphi \Rightarrow \mathbf{X}\varphi$	*transference-problem*
(E)	$(\mathbf{X}\varphi \wedge \mathbf{X}\psi) \rightarrow \mathbf{X}(\varphi \wedge \psi)$	*unrestricted combining*
(F)	$\mathbf{X}\varphi \rightarrow \mathbf{X}(\varphi \vee \psi)$	*unrestricted weakening*
(G)	$\neg(\mathbf{X}\varphi \wedge \mathbf{X}\neg\varphi)$	

Table 1: Problems related to logical-omniscience

the formulas in Γ then we can deduce $\mathbf{X}\alpha$. In the rest of the paper 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; actually we will argue that policy-based agents must accept the expected-side effects unless they have some reasons to reject the consequences corresponding to them.

Most of the properties mentioned in Table 1, except for (G), must not be valid when we take intention into consideration. For example, suppose that the TV agent wants to recommend $program_1$ and also recommend $program_2$ because it has a policy-based intention to recommend programs the user likes. Hence according to (E) it could be formally given as:

$$\mathrm{INT}_{tv}(program_1) \wedge \mathrm{INT}_{tv}(program_2) \rightarrow \mathrm{INT}_{tv}(program_1 \wedge program_2) \quad (5)$$

But this personal policy is *defeasible* in the sense that the TV agent might find that the user does not have enough time to watch both the programs and hence drops the intention to recommend both of them and now only intends to recommend one which has a better rating in the TV guide. An intention model based on normal modal logic fails to account for such type of reasoning. Though there are alternatives to using NML for modelling intention [21, 41, 59], as we outlined in section 2, policy-based intentions are defeasible in nature and hence a non-monotonic reasoning system would be more efficient to represent and reason about them. Consider, for example, the following scenario (set up in a non-monotonic formalism, i.e., Defeasible Logic (DL))[4] as

$$\begin{array}{ll} (a) & program_1(X) \Rightarrow recommendable(X) \\ (b) & program_2(X) \Rightarrow recommendable(X) \\ (c) & low\_rating(X) \rightsquigarrow \neg recommendable(X) \end{array}$$

[4] In DL notation, rules with the conditional link $\Rightarrow$ are defeasible while rules with $\rightsquigarrow$ are defeaters, namely, rules used to defeat some defeasible rules by producing evidence to the contrary. For the formal meaning of the following example see Section 5.

where (a) and (b) are premises which reflect the agent's personal policy of recommending $program_1$ and $program_2$ unless there is other evidence like (c) suggesting that it may not be able to recommend. When intention is formalised in the background of NMLs it is often the case that the agent has to have a complete description of the environment before-hand –in other terms the agent must be factually omniscient. Classically the logical omniscience problem amounts to saying that an agent has to compute all consequences of its own theory. It is obvious that some of the consequences are not intended as shown above. Moreover in classical NML the set of consequences is infinite. Hence we need a system like Defeasible Logic where the set of consequences consists of the set of literals occurring in the agent theory, i.e., in the knowledge base, which is finite.

It should be noted that, though our proposed theory does not entertain properties from (D) to (F), the side-effect problems ((A)–(C)) are accepted. Actually this is not a weakness of our theory because we assume that our agents are *aware*[5] of their activities. In this perspective the side-effect problem is similar to the substitution of indiscernible in opaque contexts. It may be possible for an agent to have the intention to visit the capital of Honduras and not to visit Tegucigalpa. But if the agent knows that Tegucigalpa is the capital of Honduras then it would be irrational of the agent not to have the intention to visit Tegucigalpa given the intention to visit the capital of Honduras.

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 auctions, 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. Analogously, we may have of course 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 is not necessary insofar as it may be blocked, in our view, by competing attitudes or made non-applicable by concrete facts.

[5] The set of formulas an agent is aware of is just a *list* of these formulas (the knowledge base of the agent), i.e., a string of symbols and nothing more. Thus, it is possible to be aware of some formula φ and not be aware of another formula ψ, even if φ and ψ are logically equivalent [28, 29].

3.3 Intention vs. Intentionality

One important point to be noted is that the notion of intention we study in this paper is slightly different from Bratman's [13] because 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 and are often termed as *goal-directed*. In our view intentions should be related not only to means-ends but also to their consequences. This notion of intention is particularly relevant in conjunction with deontic and normative notions, for example if we want to say that an agent is legally responsible for A if the agent did A with the intention to do A. In such cases the agent has to include in the set of her intentions not only her intentions in Bratman's sense but also their consequences.

Let us see how to recast Bratman's Strategic Bomber scenario [13] 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. Let us expand the scenario by supposing that despite the bombing Strategic Bomber lose the war, and that there is a process for war crime 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. 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. For further discussion on the issue of the difference between intention and intentionality see [34].

4 Obligations and Mental States

Our logical framework incorporates obligations. Two questions may be decisive in this regard. First, it would be of great importance to recast the logical nature of obligations and second to investigate how defeasible logic, as described in the following sections, might capture the well-known defeasible character of deontic reasoning. The analysis of this issue is outside the scope of the paper, even if it may be seen when the use of the superiority relation and the interplay between indisputable and defeasible conclusions are relevant in treating different kinds of normative conflict.

With regard to the logical nature of obligations, however, it is at least worth mentioning here that our framework avoids a well-known difficulty that is historically recognised in the deontic literature. The source of this difficulty, which is roughly the deontic counterpart of the problem of logical omniscience in epistemic contexts, is the closure, accepted in Standard Deontic Logic, of the obligation operator under logical consequence. As is well known, this principle is the origin of a number of paradoxes in deontic logic [19]. For example, Ross' paradox corresponds to obtaining

$$\vdash \mathrm{OBL}\phi \rightarrow \mathrm{OBL}(\phi \vee \psi) \tag{6}$$

This looks odd since we cannot derive that it is obligatory to mail a letter or to burn it from the obligation to mail such a letter. For similar reasons, the Good Samaritan paradox consists of having

$$\vdash \mathrm{OBL}(\phi \wedge \psi) \rightarrow \mathrm{OBL}\psi \tag{7}$$

If it is assumed that "Guido helps Nino who has had an accident" corresponds to saying "Guido helps Nino and Nino has had an accident", then we could conclude that it is obligatory that Nino has an accident.

These problems have been largely analysed in the deontic literature (for an overview, see, among others, [19, 7, 47]). We simply point out that the difficulties implied by the principle of closure under logical consequence are avoided here by developing a suitable notion of logical derivation of obligations.

The main issue here concerns rather the relationship between obligations and mental states. As it is pointed out [17], a number of possible approaches are available in this regard. However, and despite this variety, it is possible to identify some minimal principles that may regulate the interaction between normative and mental components. Here we focus shortly on the minimal principles that emerge from the agent specification approach considered in [18]. In particular, as argued there, we may adopt, for example, the following schemata[6]:

$$\mathrm{OBL}_i\phi \rightarrow \mathrm{BEL}_i\phi \tag{8}$$

This axiom is the strong version of epistemic *norm regimentation* since it does not simply prescribe the consistency between obligations and beliefs but states the inclusion of the former in the latter ones. This of course means that what is not believed is also not obligatory.

$$\mathrm{OBL}_i\phi \rightarrow \neg\mathrm{BEL}_i\neg\phi \tag{9}$$

$$\mathrm{OBL}_i\phi \rightarrow \neg\mathrm{INT}_i\neg\phi \tag{10}$$

$$\mathrm{OBL}_i\phi \rightarrow \neg\mathrm{GOAL}_i\neg\phi \tag{11}$$

These principles correspond to weak forms of norm regimentation with regard to agent's mental states. In this sense, they express hard constraints on agent systems. Notice that the strongest principle of norm regimentation is $\mathrm{OBL}_i\phi \rightarrow \phi$: It states agent's compliance with the norms characterising the system but

[6] As usual in recent literature, the obligation operator is labelled by an agent i to mean that the obligation is directed towards i. In other words, an expression like $\mathrm{OBL}_i\phi$ means that i is under the obligation to do ϕ, or that ϕ is obligatory for i.

does not refer directly to its mental states. This principle is not adopted here. On the other hand, schemes 9, 10 and 11 deal with conflicts between internal (mental states) and external motivations (obligations). Suppose I have been asked to give a seminar on defeasible logic sometimes next week and I have accepted without looking in my diary. My acceptance generates the obligation on me to give the seminar, and then I do not believe that my timetable is such that I cannot give the seminar. In other terms 9 prevents the system from getting my belief as prevailing in the conflict. Analogously, if I am under the obligation to write two sections of this paper but I was planning (or I desire) to have a long holiday, the system again will guarantee that the obligation overrides the other conflicting motivations. A different principle that regulates the interaction between obligations and desires is the following:

$$(\mathrm{OBL}_i\phi \wedge \mathrm{GOAL}_i\neg\phi) \rightarrow \neg\mathrm{INT}_i\neg\phi \tag{12}$$

which avoids that the output of a conflict between an obligation and a desire is that of adopting a plan for obtaining what is desired.

In general, conflict detection and resolution, which involve motivational attitudes, may be treated by using the notion of agent type. Agent types are usually characterised by stating conflict resolution types in terms of orders of overruling between the types of rules assigned to represent the different motivational attitudes. In [16] 24 possible types are identified while, in [33], based on a different framework, 20 combinations are proposed. For example, an agent will be *realistic* when rules for beliefs override all other components; she will be *social* when obligations are stronger than the other motivational components with the exception of beliefs. Stable and selfish agents are those for which, respectively, intentions override desires or the opposite. In this paper we will not discuss directly this notion, for which we refer the reader to [16, 33]. Rather, in the following sections we will give some hints on how to express in defeasible logic some of the principles 8-12 just recalled.

5 Overview of Defeasible Logic

As shown in the previous section, reasoning about general intention has a defeasible nature (in the sense that when we add further premises we may retract conclusions derived without them) and hence we need an efficient and easily implementable system to capture the required defeasible instances. Defeasible logic, as developed by Nute [51, 50] with a particular concern about computational efficiency and developed over the years by [9, 5, 4] is our choice. The reason being ease of implementation [46], flexibility [4] (it has a constructively defined and easy to use proof theory), modularity [5] and it is efficient: it is possible to compute the complete set of consequences of a given theory in linear time [44]. We do not address any semantic issues in this paper but the *argumentation semantics* as given in [32] could be straightforwardly extended to the present case.

We begin by presenting the basic ingredients of DL. A defeasible theory contains five different kinds of knowledge: facts, strict rules, defeasible rules, defeaters, and a superiority relation. We consider only essentially propositional rules. Rules containing free variables are interpreted as the set of their variable-free instances.

Facts are indisputable statements, for example, "Soccer is a $TV\_sports\_programme$". In the logic, this might be expressed as $TV\_sports\_programme(Soccer)$.

Strict rules are rules in the classical sense: whenever the premises are indisputable (e.g., facts) then so is the conclusion. An example of a strict rule is "Every $TV\_sports\_programme$ is a $sports\_channel\_programme$" Written formally:

$$TV\_sports(X) \rightarrow sports\_channel\_programme(X)$$

Defeasible rules are rules that can be defeated by contrary evidence. An example of such a rule is "sports channel programmes are usually recommendable"; written formally:

$$sports\_channel\_programme(X) \Rightarrow recommendable(X).$$

The idea is that if we know that something is a sports channel programme, then we may conclude that it is recommendable, *unless there is other evidence suggesting that it may not be recommendable.*

Defeaters are rules that cannot be used to draw any conclusions. Their only use is to prevent some conclusions. In other words, they are used to defeat some defeasible rules by producing evidence to the contrary. An example is "If a sports channel programme is free style wrestling then it might not be recommendable". Formally:

$$free\_style\_wrestling(X) \leadsto \neg recommendable(X).$$

The main point is that the information that a programme is free-style wrestling is not sufficient evidence to conclude that it is not recommendable. It is only evidence that the programme *may* not be able to be recommended. In other words, we do not wish to conclude $\neg recommendable$ if $free\_style\_wrestling$, we simply want to prevent a conclusion $recommendable$.

The *superiority relation* among rules is used to define priorities among rules, that is, where one rule may override the conclusion of another rule. For example, given the defeasible rules

$$\begin{array}{lrl} r: & wrestling(X) & \Rightarrow recommendable(X) \\ r': & free\_style\_wrestling(X) & \Rightarrow \neg recommendable(X) \end{array}$$

which contradict one another, no conclusive decision can be made about whether a wrestling programme which includes free-style can be recommended. But if we introduce a superiority relation $>$ with $r' > r$, then we can indeed conclude that the free-style wrestling cannot be recommended. The superiority relation is required to be acyclic. It turns out that we only need to define the superiority relation over rules with contradictory conclusions.

The language of Defeasible Logic consists of a set of atomic propositions, the negation sign $\neg$, and the rule signs $\rightarrow$ (for strict rules), $\Rightarrow$ (for defeasible rules), and $\leadsto$ (for defeaters). As usual a *literal* is either an atomic proposition or the negation of an atomic proposition.

A *rule* r consists of its *antecedent* (or *body*) $A(r)$ ($A(r)$ may be omitted if it is the empty set) which is a finite set of literals, an arrow, and its *consequent* (or *head*) $C(r)$ which is a literal. Given a set R of rules, we denote the set of all strict rules in R by R_s, the set of strict and defeasible rules in R by R_{sd}, the set of defeasible rules in R by R_d, and the set of defeaters in R by R_{dft}. $R[q]$ denotes the set of rules in R with consequent q. 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).

A *defeasible theory* D is a triple $(F, R, >)$ where F is a finite set of facts (a set of literals), R a finite set of rules, and $>$ a superiority relation on R.

A *conclusion* of D is a tagged literal and can have one of the following four forms:

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

Provability is based on the concept of a *derivation* (or proof) in $D = (F, R, >)$. A derivation is a finite sequence $P = (P(1), \ldots, P(n))$ of tagged literals satisfying four conditions (which correspond to inference rules for each of the four kinds of conclusion). $P(1..n)$ denotes the initial part of the sequence P of length n.

Given a theory $D = (F, R, >)$, we will use $D \vdash A$ to denote that there is a derivation of A in D, this means, that given a derivation in D, there is some n, such that $P(n) = A$.

$+\Delta$: If $P(n+1) = +\Delta q$ then
(1) $q \in F$ or
(2) $\exists r \in R_s[q]\ \forall a \in A(r) : +\Delta a \in P(1..n)$.

$-\Delta$: If $P(n+1) = -\Delta q$ then
(1) $q \notin F$ and
(2) $\forall r \in R_s[q]\ \exists a \in A(r) : -\Delta a \in P(1..n)$.

The definition of Δ describes just forward chaining of strict rules. For a literal q to be definitely provable we need to find a strict rule with head q, of which all antecedents have been definitely proved previously. And to establish that q cannot be proven definitely 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.

$+\partial$: If $P(n+1) = +\partial q$ then either
(1) $+\Delta q \in P(1..n)$ or
(2) (2.1) $\exists r \in R_{sd}[q] \forall a \in A(r) : +\partial a \in P(1..n)$ and
(2.2) $-\Delta{\sim}q \in P(1..n)$ and
(2.3) $\forall s \in R[{\sim}q]$ either
(2.3.1) $\exists a \in A(s) : -\partial a \in P(1..n)$ or
(2.3.2) $\exists t \in R_{sd}[q]$ such that
$\forall a \in A(t) : +\partial a \in P(1..n)$ and $t > s$.

Let us work through this condition. To show that q is provable defeasibly we have two choices: (1) We show that q is already definitely provable; or (2) we need to argue using the defeasible part of D as well. In particular, we require that there must be a strict or defeasible rule with head q which can be applied (2.1). But now we need to consider possible "attacks", that is, reasoning chains in support of ${\sim}q$. To be more specific: to prove q defeasibly we must show that ${\sim}q$ is not definitely provable (2.2). Also (2.3) we must consider the set of all rules which are not known to be inapplicable and which have head ${\sim}q$ (note that here we consider defeaters, too, whereas they could not be used to support the conclusion q; this is in line with the motivation of defeaters given earlier). Essentially each such rule s attacks the conclusion q. For q to be provable, each such rule s must be counterattacked by a rule t with head q with the following properties: (i) t must be applicable at this point, and (ii) t must be stronger than s. Thus each attack on the conclusion q must be counterattacked by a stronger rule. Notice that the stronger rule, i.e., t, can be a rule different from r, thus t and r can form a team to defeat for q to defeat the rule s for ${\sim}q$. In an analogous manner we can define $-\partial q$ as

$-\partial$: If $P(n+1) = -\partial q$ then
(1) $-\Delta q \in P(1..n)$ and
(2) (2.1) $\forall r \in R_{sd}[q]\ \exists a \in A(r) : -\partial a \in P(1..n)$ or
(2.2) $+\Delta{\sim}q \in P(1..n)$ or
(2.3) $\exists s \in R[{\sim}q]$ such that
(2.3.1) $\forall a \in A(s) : +\partial a \in P(1..n)$ and
(2.3.2) $\forall t \in R_{sd}[q]$ either
$\exists a \in A(t) : -\partial a \in P(1..n)$ or $t \not> s$.

The purpose of the $-\partial$ inference rules is to establish that it is not possible to prove $+\partial$. This rule is defined in such a way that all the possibilities for proving $+\partial q$ (for example) are explored and shown to fail before $-\partial q$ can be concluded. Thus conclusions tagged with $-\partial$ are the outcome of a constructive proof that the corresponding positive conclusion cannot be obtained.

Sometimes all we want to know is whether a literal is *supported*, that is if there is a chain of reasoning that would lead to a conclusion in absence of conflicts. This notion is captured by the following proof conditions:

$+\Sigma$: if $P(n+1) = +\Sigma p$ then
(1) $+\Delta p \in P(1..n)$ or
(2) $\exists r \in R_{sd}[p] : \forall a \in A(r) + \Sigma a \in P(1..n)$.

and

$-\Sigma$: if $P(n+1) = -\Sigma p$ then
  (1) $-\Delta p \in P(1..n)$ and
  (2) $\forall r \in R_{sd}[p] \exists a \in A(r) : -\Sigma a \in P(1.i)$.

The notion of support corresponds to monotonic proofs using both the monotonic (strict rules) and non-monotonic (defeasible rules) parts of defeasible theories.

In general the inference conditions for a negative proof tag (i.e., $-\Delta$, $-\partial$, $-\Sigma$) explore all the possibilities to derive a literal (with a given proof strength) before stating that the literal is not provable (with the same proof strength). Thus conclusions with these tags are the outcome of a constructive proof that the corresponding positive conclusion cannot be obtained. As a result, there is a close relationship between the inference rules for $+\partial$ and $-\partial$, (and also between those for $+\Delta$ and $-\Delta$, and $+\Sigma$ and $-\Sigma$). The structure of the inference rules is the same, but the conditions are negated in some sense. To be more precise the inference conditions for a negative proof tag are derived from the inference conditions for the corresponding positive proof tag by applying the Principle of Strong Negation introduced in [4]. The strong negation of a formula is closely related to the function that simplifies a formula by moving all negations to an innermost position in the resulting formula and replaces the positive tags with the respective negative tags and vice-versa. Formally it is defined as follows.

$$\begin{array}{lcll}
sneg(+\#p \in X) & = & -\#p \in X & \\
sneg(-\#p \in X) & = & +\#p \in X & \\
sneg(A \wedge B) & = & sneg(A) \vee sneg(B) & \\
sneg(A \vee B) & = & sneg(A) \wedge sneg(B) & \\
sneg(\exists x\ A) & = & \forall x\ sneg(A) & \\
sneg(\forall x\ A) & = & \exists x\ sneg(A) & \\
sneg(\neg A) & = & \neg sneg(A) & \\
sneg(A) & = & \neg A & \text{if } A \text{ is a pure formula.}
\end{array}$$

A pure formula is a formula that does not contain a tagged literal and # ranges over the set of proof tags. The strong negation of the applicability condition of an inference rule is a constructive approximation of the conditions where the rule is not applicable.

This feature allows us to prove some properties showing the well behaviour of defeasible logic.

Theorem 1. *[4] Let L be a defeasible logic where all proof tags are defined according to the principle of strong negation. Let $+\#$ and $-\#$ be two proof tags in L and D be a defeasible theory. There is no literal p such that $D \vdash_L +\#p$ and $D \vdash_L -\#p$.*

Proof. The proof is by induction of the complexity of the proof conditions (i.e., the number of times proof tags appear the clauses defining a proof tag, and the length of a demonstration).

For the inductive base we have only one proof tag, and the length of P is 1. In this case the definition of $+\#$ if a pure formula A, and either the D and P satisfy it or not. But the proof condition for $-\#$ is $\neg A$ (we can reverse the argument and consider $+\#$ to be $\neg A$ and $-\#$ to be A), but then for no formula A, both A and $\neg A$ are both satisfied by D and P.

For the inductive steps, we have 3 cases: (i) we assume that the property holds for proof conditions with only one occurrence of a proof tag in it and for proof of length up to n. Again the proof condition turns out to be a pure formula and we can repeat the argument used for the inductive base. (ii) We consider derivations of length 1 and we assume that the property holds for definition of proof tags with up to n occurrences of proof tags in it. However, since we have proofs of only one single line, conditions with $\forall$ referring to previous lines in the proof are vacuously true and conditions with $\exists$ are false, and a $\forall$ is converted into $\exists$ and $\exists$ into $\forall$ for proof tags defined according to the principle. In addition, conditions for $\pm\#q \in P(1..n)$ are always false. Thus the only conditions that can be evaluated as true are pure formulas, and again we can repeat the same argument as before. Finally for (iii) we assume that the theorem holds for up to n occurrences of proof tags in the conditions and for proofs of length up to n. Now we have to consider the case of a condition where in a proof tag we have a condition $\pm\#'q \in P(1..n)$, but for the other proof tag we have $\mp\#'q \in P(1..n)$, and now we can use the inductive hypothesis to state that at most only one of them can hold. So one of the proof conditions for proof tags mutually defined according to the principle of strong negation holds. □

Intuitively the above theorem states that no literal is simultaneously provable and demonstrably unprovable (with the same strength), thus it establishes the coherence of the defeasible logic presented in this paper (and in general of all the defeasible logics that satisfy the principle of strong negation).[7]

In the rest of the paper our focus is on giving proof conditions for a defeasible logic incorporating different kinds of mental attitudes, mostly intentions and obligations. The aim is to provide a logic avoiding the logical omniscience problem and at the same time able to account for (intentional) side-effects. In addition we want out logic to be realistic, an agent does not intend something the agent knows cannot be realised.

6 Defeasible Logic for Intentions

6.1 The System

As we have seen in Section 3 NMLs have been put forward to capture the intensional nature of mental attitudes such as intention. Usually modal logics are extensions of classical propositional logic with some intensional operators. Thus any classical (normal) modal logic should account for two components:

[7] Unfortunately the property that, given a proof tag $\#$ and a theory D, either $D \vdash +\#p$ or $D \vdash -\#p$ for any literal p, does not hold in general. See [6] for the conditions for this to hold.

(1) the 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. 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. Common rules for modalities are

$$\frac{\vdash \varphi}{\vdash \Box\varphi} \quad \text{Necessitation}$$

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

Consider the necessitation rule of normal modal logic which dictates the condition that an agent knows all the valid formulas and thereby all the tautologies. As was seen, such a formalisation does not seem suitable for intentions. Moreover an agent need not be intending all the consequences of a particular action it does. It might be the case that it is not confident of them being successful. Thus the two rules are not appropriate for a logic of intention.

A logic of policy-based intentions should take care of the underlying principles governing such intentions. It should have a notion of the direct and indirect knowledge of the agent, where the former relates to facts as literals whereas the latter to that of the agent's theory of the world in the form of rules. Similarly the logic should also be able to account for general intentions as well as the policy-based (derived ones) intentions of the agent.

The first step is to extend the language to incorporate the modal operator for intention INT. Thus if l is a literal then INTl and $\neg$INTl are modalised literals. The next step is to define what a knowledge based for this logic is. Accordignly, a defeasible intention theory is a structure $(F, R^K, R^I, >)$ where, as usual F is a set of facts, R^K is a set of rules for knowledge (i.e., $\rightarrow_K$, $\Rightarrow_K$, $\leadsto_K$), R^I is a set of rules for intention (i.e., $\rightarrow_I$, $\Rightarrow_I$, $\leadsto_I$), and $>$, the superiority relation, is a binary relation over the set of rules (i.e., $> \subseteq (R^K \cup R^I)^2$).

Intuitively, given an agent, F consists of the information the agent has about the world and its immediate intentions; R^K corresponds to the agent's theory of the world, while R^I encodes its policy and $>$ its strategy (or its preferences). The policy part of a defeasible theory capture both intentions and goals. The main difference is the way the agent perceives them: goals are possible outcomes of a given context while intentions are the actual goals the agent tries to achieve in the actual situation. In other words goals are the choices an agent has and intentions are the chosen goals; in case of conflicting goals (policies) the agent has to evaluate the pros and cons and then decide according to its aims (preferences), which are encoded by the superiority relation.

Notice that the main function of rules is to allow for the derivation of new conclusions and those conclusions can be new pieces of knowledge, new intentions and, as we will see in Section 7, new obligations. Accordingly we have divided the rules in rules for K, rules for I and rules for obligations. For example, the application of $p \Rightarrow_I q$ permits to infer $\mathrm{INT}q$. However, if the purpose of rules is to provide special conditions to derive modalised literals, we will make an exception with regard to rules for knowledge, which, as we alluded to, are meant to constitute the reasoning core of the system. Rules for K, in other words, will express agents' factual knowledge about the world.

To make this precise we impose some restriction of the format of rules. Specifically the antecedent of a rule is a set (possibly the empty set) of literals and modalised literals, while the conclusion of a rule is a literal.

In what follows we provide the appropriate inference rules for intentions, and we identify strong intentions – i.e., intentions for which there are no alternatives – using $\pm\Delta_I$; goals using $\pm\Sigma_I$, and intentions using $\pm\partial_I$. This distinction, which emerges from the three levels of provability in Defeasible Logic, has intuitively some conceptual ground. In fact, even though we argued about the defeasible nature of policy-based intentions (a nature that is primarily incarnated by $\pm\partial_I$), nothing prevents from having also policies that are indisputable. This is so exactly for the same reasons, as we shall see in Section 7, for which we are used to say that the conditional obligations are defeasible but despite which we cannot reject a priori the possibility that strict obligations occur in agent's theory. As for strong intentions, for example, an agent may have the policy that, if it is summertime, her intention will be indisputably that of visiting Australia, and this despite any possibly conflicting motivation. This is indeed a possibility that we cannot drop from our formalism[8]. On the other hand, as we have maintained above, if intentions are the actual goals that the agent obtains in the actual situation, goals in themselves may be viewed just as the possible intentional outcomes of a given context. Goals, in other words, correspond to possible intention patterns of the agent, which have to be selected on the basis of concrete contexts. For example, we may have two potentially applicable rules such as

$$July \rightarrow_K Summertime$$
$$Summertime \Rightarrow_I VisitSpain$$

and, despite the fact that the theory may include a third rule such as

$$Summertime,\ Hot \Rightarrow_I \neg VisitSpain$$

nothing prevents to say that $VisitSpain$ can be an option potentially considered by the agent.

[8] It is also worth noting that the distinction between indisputable and defeasible conclusions, if applied to rules for K, may be used to capture the difference between knowledge and belief. In fact, derivations via strict rules are indisputable, they must be necessarily true: at the worst, it can happen that we may have unsolvable inconsistencies. On the other hand, when conclusions are defeasible, they are assumed to be true unless (stronger) evidence of the contrary is obtained.

In order to correctly capture the notion of intention we extend the signature of the logic with the modal operator INT. However we impose some restrictions on the form of the rules: modal literals can only occur in the antecedents of rules for intention.

This restriction is motivated from the fact that such rules are meant to introduce the modalities and in the present context sequences of modalities are not particularly meaningful.[9]

Derivability for knowledge ($\pm\Delta_K$, $\pm\partial_K$) has the same conditions as those given for derivability in Section 5. It is true that the complete and accurate definition of the inference conditions is cumbersome but the intuition is natural and easy to understand. The conditions for deriving an intention are as follows:

$+\Delta_I$: if $P(n+1) = +\Delta_I p$ then
(1) $\text{INT}p \in F$ or
(2) $\exists r \in R_s^K[p] \forall a \in A(r) : +\Delta_I a \in P(1..n)$ or
(3) $\exists r \in R_s^I[p]$ such that
(3.1) $\forall \text{INT}a \in A(r) : +\Delta_I a \in P(1..n)$ and
(3.2) $\forall a \in A(r) : +\Delta_K a \in P(1..n)$.

To prove a strong intention, we need either that the intention is unconditional (1), or that we have a strict rule for intention (an irrevocable policy) whose antecedent is indisputable (3). However we have another case (2): if an agent knows that B is an indisputable consequence of A, and it strongly intends A, then it must intend B. This is in contrast with the NML interpretation whereby the agent has to intend all the consequences of his/her intention.

$-\Delta_I$: if $P(n+1) = -\Delta_I p$ then
(1) $\text{INT}p \notin F$ and
(2) $\forall r \in R_s^K[p]\ \exists a \in A(r) : -\Delta_I a \in P(1..n)$; and
(3) $\forall r \in R_s^I[p]$ either
(3.1) $\exists \text{INT}a \in A(r) : -\Delta_I a \in P(1..n)$ or
(3.2) $\exists a \in A(r) : -\Delta_K a \in P(1..n)$.

To prove that a strong intention A does not hold ($-\Delta_I A$), first, A should not be a basic intention (1); then we have to discard all possible reasons in favour of it. If A is a definite consequence of B, that is $B \rightarrow_K A \in R^K$, we can disprove it if we can show that B is not strongly intended (i.e., $-\Delta_I B$). In case of strict policies for A (3), such as, for example the strict rule for intention $\text{INT}B, C \rightarrow_I A$, we have to show that either B is not strongly intended (3.1), or the fact triggering the policy is not the case (3.2).

At the other extreme we have goals: literals supported by evidence and basic intentions.

$+\Sigma_I$: if $P(n+1) = +\Sigma_I p$ then
(1) $\text{INT}p \in F$ or

[9] If we drop this constraint then rules such as $p, q \Rightarrow_K \text{INT}r$ are permitted. In the rest of the paper we will argue that the intended meaning of this rule is better represented by a rule for an intention, namely $p, q \Rightarrow_I r$.

(2) $\exists r \in R^K_{sd}[p] \forall a \in A(r) : +\Sigma_I a \in P(1..n)$ or
(3) $\exists r \in R^I_{sd}[p]$ such that
(3.1) $\forall \text{INT} a \in A(r) : +\Sigma_I a \in P(1..n)$ and
(3.2) $\forall a \in A(r) : +\Sigma_K a \in P(1..n)$.

$-\Sigma_I$: if $P(n+1) = -\Sigma_I p$ then
(1) $\text{INT} p \notin F$ and
(2) $\forall r \in R^K_{sd}[p]\ \exists a \in A(r) : -\Sigma_I a \in P(1..n)$; and
(3) $\forall r \in R^I_{sd}[p]$ either
(3.1) $\exists \text{INT} a \in A(r) : -\Sigma_I a \in P(1..n)$ or
(3.2) $\exists a \in A(r) : -\Sigma_K a \in P(1..n)$.

The inference conditions for goals are very similar to those for strong intentions; essentially they are monotonic proofs using both the monotonic part (strict rules) and the supportive non-monotonic part (defeasible rules) of a defeasible theory.

On the other hand to capture intentions we have to use the superiority relations to resolve conflicts. Thus we can give the following definition for the inference rules for $\pm\partial_I$.

$+\partial_I$: if $P(n+1) = +\partial_I p$ then
(1) $+\Delta_I p \in P(1..n)$ or
(2) (2.1) $-\Delta_K {\sim} p, -\Delta_I {\sim} p \in P(1..n)$ and
(2.2) either
(2.2.1) $\exists r \in R^K_{sd}[p]\ \forall a \in A(r) : +\partial_I a \in P(1..n)$, or
(2.2.2) $\exists r \in R^I_{sd}[p]\ \forall \text{INT} a \in A(r) : +\partial_I a \in P(1..n)$ and
$\forall a \in A(r) : +\partial_K a \in P(1..n)$; and
(2.3) $\forall s \in R[{\sim} p]$ either
(2.3.1) if $s \in R^K[{\sim} p]$ then
$\exists a \in A(s) : -\partial_I a \in P(1..n)$ and
$\exists b \in A(s) : -\partial_K b \in P(1..n)$; and
if $s \in R^I[{\sim} p]$ then either
$\exists \text{INT} a \in A(s) : -\partial_I a \in P(1..n)$ or
$\exists a \in A(s) : -\partial_K a \in P(1..n)$; or
(2.3.2) $\exists t \in R[p]$ such that $t > s$ and
if $t \in R^K[p]$ then $\forall a \in A(t) : +\partial_K a \in P(1..n)$ or
$\forall a \in A(t) : +\partial_I a \in P(1..n)$; and
if $t \in R^I[p]$ then $\forall a \in A(t) : +\partial_K a \in P(1..n)$ and
$\forall \text{INT} a \in A(t) : +\partial_I a \in P(1..n)$.

The conditions for proving defeasible intentions are essentially the same as those given for defeasible derivations in Section 5. The only difference is that at each stage we have to check for two cases, namely: (1) the rule used is a rule for an intention; (2) the rule is a rule for knowledge. In the first case we have to verify that factual antecedent are defeasibly proved/disproved using knowledge ($\pm\partial_K$), and intentional antecedent are defeasibly proved/disproved using intention ($\pm\partial_I$). In the second case we have to remember that a conclusion of a factual rule can be transformed in an intention if all the literals in the antecedent are defeasibly intended.

$$
\begin{array}{l}
-\partial_I\text{: if } P(n+1) = -\partial_I p \text{ then}\\
(1)\ -\Delta_I p \in P(1..n) \text{ and}\\
(2)\ (2.1)\ +\Delta_K{\sim}p \text{ or } +\Delta_I{\sim}p \in P(1..n) \text{ or}\\
\quad (2.2) \text{ both}\\
\quad\quad (2.2.1)\ \forall r \in R^K_{sd}[p]\ \exists a \in A(r) : -\partial_I a \in P(1..n)\text{; and}\\
\quad\quad (2.2.2)\ \forall r \in R^I_{sd}[p]\ \exists \text{INT}a \in A(r) : -\partial_I a \in P(1..n) \text{ or}\\
\qquad\qquad\qquad\qquad\qquad \exists a \in A(r) : -\partial_K a \in P(1..n)\text{; or}\\
\quad (2.3)\ (2.3.1)\ \exists s \in R^K[{\sim}p]\ \forall a \in A(s) : +\partial_K a \in P(1..n) \text{ or}\\
\qquad\qquad\qquad\qquad\qquad \forall a \in A(s) : +\partial_I a \in P(1..n)\text{, or}\\
\qquad\qquad\qquad \exists s \in R^K[{\sim}p]\ \forall a \in A(s) : +\partial_K a \in P(1..n) \text{ and}\\
\qquad\qquad\qquad\qquad\qquad \forall \text{INT}a \in A(s) : +\partial_I a \in P(1..n)\text{; and}\\
\qquad\quad (2.3.2)\ \forall t \in R[p] \text{ such that } t > s \text{ and}\\
\qquad\qquad\qquad \text{if } t \in R^K[p] \text{ then } \forall a \in A(t) : +\partial_K a \in P(1..n) \text{ or}\\
\qquad\qquad\qquad\qquad\qquad \forall a \in A(t) : +\partial_I a \in P(1..n)\text{; and}\\
\qquad\qquad\qquad \text{if } t \in R^I[p] \text{ then } \forall a \in A(t) : +\partial_K a \in P(1..n) \text{ or}\\
\qquad\qquad\qquad\qquad\qquad \forall \text{INT}a \in A(t) : +\partial_I a \in P(1..n).
\end{array}
$$

The intuition behind the definition of $-\partial_I$ is a combination of the motivation for $-\partial$ and the intuition of $-\Delta_I$.

We want to illustrate some of the aspects of derivability by means of examples. If it does not rain we intend to play cricket, and if we intend to play cricket we intend to stay outdoor. This example can be formalised as follows

$$
\begin{array}{c}
\neg \mathit{rain} \Rightarrow_I \mathit{cricket}\\
\text{INT}\mathit{cricket} \Rightarrow_I \mathit{outdoor}
\end{array}
$$

Once the fact $\neg\mathit{rain}$ is supplied we can derive $+\partial_I \mathit{cricket}$, and then the intention of staying outdoor ($+\partial_I \mathit{outdoor}$). However the same intention cannot be derived if the fact *cricket* is given.

If Vineet intends to travel to Italy then he intends to travel to Europe since Italy is in Europe. This argument can be formalised as

$$
\mathit{Italy} \rightarrow_K \mathit{Europe}
$$

plus the basic intention INT*Italy*. The conclusion $+\Delta_I \mathit{Europe}$ follows from clause (2) of $+\Delta_I$.

6.2 Defeasible Logic of Intention and BDI

In this Section we will show some properties of the Defeasible Logic of intention we have developed so far and we relate it with some of the basic intuitions of BDI. The first thing we have to do is to relate the degree of provability in defeasible logic and the mode of the conclusion with the modality used in BDI systems. Thus if D is a defeasible theory we establish the following relationships

1. $D \vdash \text{BEL}\phi$ iff $D \vdash +\partial_K\phi$, and $D \vdash \neg\text{BEL}\phi$ iff $D \vdash -\partial_K\phi$;

2. $D \vdash \text{INT}\phi$ iff $D \vdash +\partial_I\phi$, and $D \vdash \neg\text{INT}\phi$ iff $D \vdash -\partial_I\phi$;

3. $D \vdash \text{GOAL}\phi$ iff $D \vdash +\Sigma_I\phi$, and $D \vdash \neg\text{GOAL}\phi$ iff $D \vdash -\Sigma_I\phi$.

The purpose of the $-\Delta$ and $-\partial$ inference rules is to establish that it is not possible to prove a corresponding tagged literal. These rules are defined in such a way that all the possibilities for proving $+\partial p$ (for example) are explored and shown to fail before $-\partial p$ can be concluded. Thus conclusions with these tags are the outcome of a constructive proof that the corresponding positive conclusion cannot be obtained.

As a result, there is a close relationship between the inference rules for $+\partial$ and $-\partial$, (and also between those for $+\Delta$ and $-\Delta$, and $+\Sigma$ and $-\Sigma$). The structure of the inference rules is the same, but the conditions are negated in some sense. This feature allows us to prove some properties showing that defeasible logic is well behaved.

Theorem 2. *Let $\# \in \{\Delta_K, \partial_K, \Sigma_K, \Delta_I, \partial_I, \Sigma_I\}$, and D be a defeasible theory. There is no literal p such that $D \vdash +\#p$ and $D \vdash -\#p$.*

Proof. This Theorem is an immediate consequence of the *Principle of Strong Negation* and Theorem 2 of [4]. Indeed it is straightforward to prove that the conditions for $+\#$ are the strong negation of those for $-\#$, and vice-versa. □

Intuitively the above theorem states that no literal is simultaneously provable and demonstrably unprovable, thus it establishes the coherence of the defeasible logic presented in this paper.

Theorem 3. *Let D be a defeasible theory, and $M \in \{K, I\}$. $D \vdash +\partial_M p$ and $D \vdash +\partial_M {\sim} p$ iff $D \vdash +\Delta_M p$ and $D \vdash +\Delta_M {\sim} p$.*

Proof. For $M = K$ see Proposition 3.3 in [5], and the proof for $M = I$ is analogous. □

This theorem gives the consistency of defeasible logic. In particular it affirms that it is not possible to obtain conflicting intentions ($+\partial_I p$ and $+\partial_I {\sim} p$) unless the information given about the environment is itself inconsistent. Notice, however, that the theorem does not cover goals (Σ_I). Indeed, it is possible to have conflicting goals.

Let D be a defeasible theory. With Δ_K^+ we denote the set of literals strictly provable using the epistemic (knowledge) part of D, i.e., $\Delta_K^+ = \{p : D \vdash +\Delta_K p\}$. Similarly for the other proof tags.

Theorem 4. *For every defeasible theory D, and $M \in \{K, I\}$*

1. $\Delta_M^+ \subseteq \partial_M^+ \subseteq \Sigma_M^+$*;*

2. $\Sigma_M^- \subseteq \partial_M^- \subseteq \Delta_M^-$*.*

Proof. We prove 1, and 2 is a consequence of 1 and the principle of strong negation [4]. For $M = K$ see [5], since the conditions for knowledge are those for derivability in DL. The inclusion $+\Delta_I \subseteq +\partial_I$ is immediate from condition 1 of $+\partial_I$. For the other inclusion, i.e., $+\partial_I \subseteq +\Sigma_I$ we notice that if we restrict ourselves to strict rules in $+\Sigma_I$ we obtain clause 1 of $+\partial_I$, the basic case of the recursive definition of derivation in DL. Moreover clause 2 and 3 of $+\Sigma_I$ correspond to clause 2.2.2 and 2.2.3 of $+\partial_I$. □

This theorem states that strict intentions are intentions ($\Delta_I^+ \subseteq \partial_I^+$), and intentions are goals ($\partial_I^+ \subseteq \Sigma_I^+$), which corresponds to the well-known BDI principle

$$\text{INT}\phi \rightarrow \text{GOAL}\phi.$$

At the same time, we have that $\Delta_K^+ \subseteq \partial_K^+$. Thus if we assume that Δ_K corresponds to knowledge and ∂_K corresponds to belief[10] we obtain

$$\text{KNOW}\phi \rightarrow \text{BEL}\phi,$$

the standard BDI axiom relating the two epistemic notions.

The proposed theory of intention satisfies many of the properties outlined by Bratman in [13]. The role of intention as a *conduct-controlling* pro-attitude rather than *conduct-influencing* is clearly illustrated in the elaborate proof-theory outlined for the types of intention. The proposed theory supports the fact that the rationality of an agent for his intention depends on the rationality of the relevant processes leading to that intention where the relevant processes includes using superiority relations to resolve conflicts as well as satisfying the rules of inclusion as shown in Theorem 3. The new approach provides a good formalisation as to the relation between *guiding intention* and *intentional action* termed as *historical principle of policy-based rationality* in [13]. The problem in general is to account for the rationality of an agent in performing a particular policy-based intention from a general policy. In our approach the defeasibility of general policies makes it possible to block/not block the application of the policy to the particular case without abandoning the policy. This idea is explained further in the next section.

7 How Intentions and Obligations Interact

7.1 The System Extended

In the previous section we outlined a defeasible logic to reason about the mental attitudes employed in the BDI framework. As we have seen in Section 4 there have been some attempts to supplement BDI like framework with the notion of

[10] One of the main differences between knowledge and belief is that knowledge is true, while beliefs can be false. Defeasibility is used to encode exception, thus it is possible that the opposite of a defeasible conclusion might be true under difference circumstances we are not aware of and then we assert the conclusion as a belief. This is not the case for definite conclusions.

obligation. It is well known that, by their own nature, deontic and, in general, normative notions are defeasible. Thus we believe that the present approach offers a suitable environment for the combination of policy-based intentions and obligations. In the rest of this section we will examine some design choices for the combination of these notions.

The notion of obligation is independent from that of knowledge and intention, hence we have to introduce a new set of (defeasible) rules for the introduction of obligation in agreement with the general principles outlined in Section 6.

A defeasible theory is now a structure

$$D = (F, R^K, R^I, R^O, >)$$

where F, R^K, R^I are as before and R^O is a set of rules (strict, defeasible and defeaters) for obligation, and the superiority relation $>$ is now defined over $R^K \cup R^I \cup R^O$. Moreover if l is a literal then $\mathrm{OBL}l$ and $\neg\mathrm{OBL}l$ are modal literals. Also, as is the case for intention where modal literals are not allowed as consequent of rules for intention, modal literals are not permitted in consequents of rules for obligations.[11]

The intuition behind the interpretation of the rules for the deontic component is that strict rules express hard constraints while defeasible rules represent soft constraints.[12] A hard constraint is a condition that cannot be violated, while soft constraints can be violated in exceptional situations and thus they are open to exceptions. Defeaters, as we have seen in Section 5, cannot be used to support a conclusion (obligation), thus in some way a defeater encodes a particular notion of weak permission (lack of an obligation for the opposite).

The relationships between conclusions labelled with obligation and derivability of deontic literals (i.e., modal literals whose modal operator is OBL) is as follows:

$$D \vdash \mathrm{OBL}\phi \text{ iff } D \vdash +\partial_O \phi.$$

The above establishes the conditions under which we can prove an obligation in defeasible logic.

In the simplest and less interesting case we can assume that there are no interactions between mental attitudes and obligations. In this case the inference conditions for $\pm\Delta_O$, $\pm\partial_O$, and $\pm\Sigma_O$ are the same as those for normal derivability in Defeasible Logic (see Section 5). Here we give only the condition for $\pm\Delta_O$ and the inference conditions for the other proof tags can be obtained in a similar way.

$+\Delta_O$: If $P(n+1) = +\Delta_O p$ then either:
(1) $\mathrm{OBL}p \in F$ or

[11] In general we could relax these restrictions by allowing OBL-literals as consequents of rules for intention and INT-literals as consequents of rules for obligation. In any case it seems to us that in the present context nested modalities are not very meaningful, and thus we refrain from relaxing the constraints.

[12] For a detailed analysis of and relationships between hard and soft constraints and Deontic Logic see, among others, [49].

(2) $\exists r \in R_s^O[p]$:
(2.1) $\forall \mathrm{INT} a \in A(r) + \Delta_I a \in P(1..n)$ and
(2.2) $\forall \mathrm{OBL} a \in A(r) + \Delta_O a \in P(1..n)$ and
(2.3) $\forall a \in A(r) + \Delta_K a \in P(1..n)$.

Moreover we have to add conditions about the derivability of OBL-literals in the definitions of the tags for Knowledge and Intention. For example the following condition is needed in clause 3 of the definition of $+\Delta_I$ and in clause 2 of $+\Delta_K$

$$\forall \mathrm{OBL} a \in A(r) + \Delta_O a \in P(1..n).$$

Similar additions are required for the other proof tags. Notice that in this case pairs of conclusions such as $+\#_O p$ and $+\#_K \neg p$ do not generate a conflict. In Section 4 we have described some plausible interaction axioms for the modalities at hand and some of these axioms aim to avoid some kind of conflicts.

In what follows we show how to reformulate the ideas represented by those axioms in the present framework.

Let us consider the axiom (8)

$$\mathrm{OBL}\phi \rightarrow \mathrm{BEL}\phi$$

This is a simple interaction axiom which states that every obligation is also a belief. Accordingly the relationship between BEL and OBL described by axiom (8) can be achieved by the following proof condition[13]

$+\#_K^O$: If $P(n+1) = +\#_K^O p$ then either
(1) $+\#_O p \in P(1..n)$ or
(2) $+\#_K p \in P(1..n)$.

The above condition tells us that we can introduce at any step of a derivation $\#_K^O p$ if we have either proved p using a derivation whose final step was obtained via a rule in R_O or via a rule for R_K. In addition we have to modify the condition to derive a BEL literal. The new condition is

$$D \vdash \mathrm{BEL}\phi \text{ iff } D \vdash +\partial_K^O \phi.$$

The advantage of this construction is that it allows modular definition of the proof tags for the various kind of rules.

We turn now our attention to the interaction axioms (10) and (11), namely

$$\mathrm{OBL}\phi \rightarrow \neg\mathrm{BEL}\neg\phi \qquad \mathrm{OBL}\phi \rightarrow \neg\mathrm{INT}\neg\phi$$

Those two axioms are equivalent to

$$\mathrm{OBL}\phi \wedge \mathrm{BEL}\neg\phi \rightarrow \bot \qquad \mathrm{OBL}\phi \wedge \mathrm{INT}\neg\phi \rightarrow \bot$$

[13] In this section we will state only the proof condition for some positive proof tag. The condition for the corresponding negative proof tag can be derived from it using the principle of strong negation. The other proof tags require trivial modifications along the line of that of the conditions given in this section.

which state, respectively, that obligations and beliefs should not conflict, and obligations and intentions should not conflict.

A possible way to solve conflicts in defeasible logic is to consider the types of rules that possibly attack a particular rule. In the first case we have that a rule for knowledge relative to $\neg p$ attacks a rule for obligation relative to p; in the second case rules for obligations are attacked by rules for intentions. Formally this can be achieved by replacing the quantification of condition (2.3) of the proof condition for $+\partial_O$ (i.e., $\forall r \in R[\sim p]$) with

$$\forall r \in R^X[\sim p]$$

such that $O \in X$, where $R^X[\sim p]$ is the set of rules for $\sim p$ defined as follows:

$$R^X = \{r \in R : r \in R^O \text{ if } O \in X \text{ or } r \in R^K \text{ if } K \in X \text{ or } r \in R^I \text{ if } I \in X\}$$

Notice that if we add only the above condition, we do avoid conflicts between obligations and mental attitudes by giving precedence to mental attitudes. For example if $X = \{O, I\}$, then, from the theory

$$\Rightarrow_O p \qquad \Rightarrow_I \neg p$$

we obtain $-\partial_O^X p$ and $+\partial_I \neg p$. Thus intentions attack obligations but not the other way around. To restore symmetry we have to propagate the modifications to the mental attitudes affected by obligation: in particular we have to replace the quantification of the attack clause of $+\partial_Y$, where $Y \in X$ with

$$\forall r \in R^{\{Y,O\}}[\sim p].$$

The conditions for proving GOALp are much weaker than those for intentions. In fact for a goal all we have to do is to verify that the goal is indeed reachable, while for an intention we have to evaluate the reasons for its negation. Hence a simple way to encode the interaction axiom (11) is to think it as a form of inclusion axiom (actually an exclusion one), and to formulate the proof conditions accordingly:

$-\Sigma_I^O$: If $P(n+1) = -\Sigma_I^O p$ then either
(1) $+\partial_O{\sim}p \in P(1..n)$ or
(2) $-\Sigma_I p \in P(1..n)$.

Let us illustrate the interaction between intentions and obligations by looking at a couple of examples.

Example 2. Let us consider a theory where $F = \{a, b\}$ and $R = \{a, b \Rightarrow_K c,\ c, b \Rightarrow_I d,\ \text{INT}d \Rightarrow_O e\}$. Here we can prove $+\partial_K a$, $+\partial_K b$ since they are facts. Then the first rule is applicable and we can derive $+\partial_K c$, and now the second rule is applicable and we obtain $+\partial_I d$; this last conclusion allows for applying the third rule, thus obtaining $+\partial_O e$. Let us comment the theory with the help of a concrete example. A drunk surgeon operates a patient. The surgeon is aware that operating under the influence of alcohol will result in a failure and that failing and being drunk means intending to cause damages. Moreover the legal system under which the surgeon operates prescribes that people causing damages as a result of negligence must be sanctioned. Thus the three rules can be rewritten, respectively as

$$\begin{gathered} \mathit{operate},\ \mathit{drunk} \Rightarrow_K \mathit{fail} \\ \mathit{fail},\ \mathit{drunk} \Rightarrow_I \mathit{damages} \\ \mathit{fail},\ \mathrm{INT}\mathit{damages} \Rightarrow_O \mathit{sanction} \end{gathered}$$

The conclusion is that the surgeon has to be sanctioned, because the damages are the result of an intentional negligence. What about when a surgeon, not on duty and being the only person able to complete the required medical procedure, is drunk and the patient will die without the operation? The surgeon knows that the patient will suffer damages as a result of the operation, but he operates anyway. In this case we have to change the second rule into $\mathit{fail},\ \mathit{drunk} \Rightarrow_K \mathit{damages}$. Here we derive $+\partial_K \mathit{damages}$ instead of $+\partial_I \mathit{damages}$, and thus we block the application of the third rule. Hence we cannot conclude that the surgeon is subject to a sanction.

Example 3. To illustrate the potential conflicts between obligations and intentions we examine a version of the well-known prisoner dilemma. Two people are arrested for a major crime, however the police does not have enough evidence to incriminate them, but they can be charged with and convicted for a minor crime. However if one of them confesses the crime she will be sentenced to one year and the other to twenty-five years. If both confess they will be imprisoned for ten years each. Finally if none of them confesses then they have to serve for three years each. The two criminals are part of a criminal organisation renowned for its code of honour that prescribes to not betray your fellows. The best individual outcome is to confess the crime, while the best outcome according to the organisation code is not confessing it. Hence this situation can be represented by the following theory:

$$\Rightarrow_I \mathit{confess} \qquad \Rightarrow_O \neg \mathit{confess}$$

A "selfish criminal" (the intention overrides the obligation) will intend to confess ($+\partial_I \mathit{confess}$, $-\partial_O \neg \mathit{confess}$), giving thus priority to his welfare, while a "social criminal" (the obligation overrides the intention) will stick with the code of honour and will not confess the crime ($+\partial_O \neg \mathit{confess}$, $-\partial_I \mathit{confess}$).

Another interesting feature that could be explained using our formalism is that of *rule conversion*. Indeed, this feature allows us to model the interaction between mental attitudes and obligations. 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 [42])

$$\frac{B \vdash C \quad A \mathrel{|\!\sim} B}{A \mathrel{|\!\sim} C}$$

which allows the combination of non-monotonic consequence with classical consequences. While not every combination of obligations and mental attitudes will produce meaningful results for the conversion, some of them can prove useful in the present context. For instance, suppose that a rule of a specific type is given and also suppose that all the literals in the antecedent of the rule are provable in one and the same modality. If so, it is possible to argue that the conclusion of the rule inherits the modality of the antecedent. To give an example, suppose we have that $p, q \Rightarrow_O r$ and that we obtain $+\partial_I p$ and $+\partial_I q$. Can we conclude $\mathrm{INT}(r)$? Here we should be careful but, if we aim to model agents that intend to comply with norms (namely, we want to model social agents)[14], then this conversion is indeed appropriate. Let us formally illustrate this case. As we explained, here we have to determine conditions under which a rule for obligation can be used to directly derive an intention. The condition we are after is that all the antecedents on the rule can be shown to be intentions. Formally we have thus to add (disjunctively) in the support phase of the proof condition for ∂_I the following clause

$$\exists r \in R^O[p] : \forall a \in A(r) + \partial_I a \in P(1..n).$$

This conversion configures a weak form of normative regimentation. In fact, assume that it is summer and that, if it is summertime then I have the intention to visit Australia

$$\mathit{Summertime} \Rightarrow_I \mathit{VisitAustralia}$$

and, again, that I am aware that Australian legal system forbids to bring with me fresh food from abroad,

$$\mathit{VisitAustralia} \Rightarrow_O \neg \mathit{BringFreshFood}$$

If so, this conversion will state that I have also the intention not to bring with me any fresh food. Thus, roughly speaking, it can be intuitively viewed as the non-monotonic (and so weaker) version of the principle according to which $\mathrm{INT}p$ and $\mathrm{OBL}(p \rightarrow q)$ entail $\mathrm{INT}q$. Let us see with a concrete example the meaning of another type of conversion, based on rules for K. The Yale Shooting Problem can be described as follows[15]

$$\mathit{liveAmmo}, \mathit{load}, \mathit{shoot} \Rightarrow_K \mathit{kill}$$

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 the agent intends to load and shoot the gun ($\mathrm{INT}(\mathit{load})$, $\mathrm{INT}(\mathit{shoot})$), then, since she knows that the consequence of these actions is the death of her friend, she

[14] See [33] for a discussion and definitions of agent types based on possible combinations of obligations and mental attitudes.

[15] Here we will ignore all temporal aspects and we will assume that the sequence of actions is done in the correct order.

intends to kill him ($+\partial_I kill$). However, in the case she has the intention to load the gun ($+\partial_I load$) and then for some reason shoot it ($shoot$), then the friend is still alive ($-\partial_K kill$).

8 Permission

In the previous section we considered the interplay between agents' mental states and obligations. But obligations do not exhaust the possible deontic qualifications that normative systems can represent. In this section we provide a brief note on the notion of permission. As done with conditional obligations, we do not enter here in a discussion of the philosophical meaning of permissions, too: we simply offer three different technical options to characterise permissive provisions within the framework previously described. Of course, our proposals roughly follow some basic ideas in the literature of deontic logic (such as the distinction between strong and weak permissions). However, these notions can be problematic, and only a few extensive studies of them have been proposed so far (an exception is the fine investigation developed in [48]).

Approach I A first way to define permissions in DL is by simply stating that the complement of what is permitted is not provable as obligatory:

Definition 1. *Let D be a defeasible agent theory. For any literal q, permission is characterised as follows:*

$$\begin{aligned} D \vdash -\partial_{\mathrm{OBL}}{\sim}q \;\Rightarrow\; D \vdash +\Delta_{\mathrm{PERM}}q \\ D \vdash -\Delta_{\mathrm{OBL}}{\sim}q \;\Rightarrow\; D \vdash +\partial_{\mathrm{PERM}}q \end{aligned}$$

This is the most direct way to define the idea of weak permission, according to which something is permitted by a code iff it is not prohibited by that code. In the formal language, this possibility consists in adding the modality PERM without having specific rules for it. We can only have modal literals, such as PERMq in antecedents of rules, and tagged literals, such as $+\partial_{\mathrm{PERM}}q$, in derivations that are based on Definition 1. To reflect this we can devise the following special proof conditions for PERM:

$+\Delta_{\mathrm{PERM}}$: If $P(n+1) = +\Delta_{\mathrm{PERM}}q$ then
(1) PERM$q \in F$ or
(2) $-\partial_{\mathrm{OBL}}{\sim}q \in P(1..n)$

$+\partial_{\mathrm{PERM}}$: If $P(n+1) = +\partial_{\mathrm{PERM}}q$ then
(1) $-\Delta_{\mathrm{OBL}}{\sim}q \in P(1..n)$.

Approach II The second approach to permission is based on introducing the modality PERM but also specific rules for it. This means that any $\phi_1, \ldots, \phi_n \rhd_{\mathrm{PERM}} \psi$, where $\rhd \in \{\rightarrow, \Rightarrow, \rightsquigarrow\}$, is a rule. Proof tags and conditions are characterised as follows:

$+\Delta_{\mathrm{PERM}}$: If $P(n+1) = +\Delta_{\mathrm{PERM}}q$ then

(1) $\mathrm{PERM}q \in F$ or

(2) $\exists r \in R_s^{\mathrm{PERM}}[q]$ such that

(2.1) $\forall \mathrm{INT}a \in A(r) + \Delta_I a \in P(1..n)$ and

(2.2) $\forall \mathrm{OBL}a \in A(r) + \Delta_O a \in P(1..n)$ and

(2.3) $\forall \mathrm{PERM}a \in A(r) + \Delta_{\mathrm{PERM}} a \in P(1..n)$ and

(2.4) $\forall a \in A(r) + \Delta_K a \in P(1..n)$, or

(3) $\exists r \in R_s^K[q]$ such that

(3.1) $\forall a \in A(r) + \Delta_{\mathrm{PERM}} a \in P(1..n)$.

$+\partial_{\mathrm{PERM}}$: If $P(n+1) = +\partial_{\mathrm{PERM}}q$ then

(1)$+\Delta_{\mathrm{PERM}}q \in P(1..n)$ or

(2) (2.1) $-\Delta_{\mathrm{OBL}}{\sim}q \in P(1..n)$ and

(2.2) $\exists r \in R_{sd}^{\mathrm{PERM}}[q]$ such that

$\forall \mathrm{INT}a \in A(r) + \partial_I a \in P(1..n)$ and

$\forall \mathrm{OBL}a \in A(r) + \partial_O a \in P(1..n)$ and

$\forall \mathrm{PERM}a \in A(r) + \partial_{\mathrm{PERM}} a \in P(1..n)$ and

$\forall a \in A(r) + \partial_K a \in P(1..n)$, or

(2.3) $\exists r \in R_{sd}^K[q]$ such that

$\forall a \in A(r) + \partial_{\mathrm{PERM}} a \in P(1..n)$, and

(2.4) $\forall s \in R[{\sim}q]$ either

(2.4.1) if $s \in R^O[{\sim}q]$ then

$\exists \mathrm{INT}a \in A(s) - \partial_I a \in P(1..n)$ or

$\exists \mathrm{OBL}a \in A(s) - \partial_O a \in P(1..n)$ or

$\exists \mathrm{PERM}a \in A(s) - \partial_{\mathrm{PERM}} a \in P(1..n)$ or

$\exists a \in A(s) - \partial_K a \in P(1..n)$, or

if $s \in R^K[{\sim}q]$ then

$\exists a \in A(s) - \partial_O a \in P(1..n)$, or

(2.4.2) $\exists t \in R[q]$ such that $t > s$ and

if $t \in R^{\mathrm{PERM}}[q]$ then

$\forall \mathrm{INT}a \in A(t) + \partial_I a \in P(1..n)$ and

$\forall \mathrm{OBL}a \in A(t) + \partial_O a \in P(1..n)$ and

$\forall \mathrm{PERM}a \in A(t) + \partial_{\mathrm{PERM}} a \in P(1..n)$ and

$\forall a \in A(t) + \partial_K a \in P(1..n)$; and

if $t \in R^K[q]$ then $\forall a \in A(t) + \partial_{\mathrm{PERM}} a$.

First of all, note that we admit conversions for permission, too. Suppose we have a rule saying that, if I get wet under the rain, then I get the flu. If we can prove that it is permitted to get wet, then we have reasons to state that it is permitted to get the flu as well. Second, no mental state can attack a permission. Suppose it is permitted to smoke in public spaces and the agent intends not to smoke: there is nothing contradictory in this situation. The only rules that can attack permission are those for obligation that allow for proving the opposite conclusion of the permission we want to derive, or, via conversion, those for knowledge that allow for deriving the opposite prohibition. In other words, two rules like $\Rightarrow_{\mathrm{OBL}} {\sim}q$ and $\Rightarrow_{\mathrm{PERM}} q$ are in conflict, as the former states that that q is forbidden whereas the latter that q is permitted. Analogously, two rules like $a \Rightarrow_K {\sim}q$ and $b \Rightarrow_{\mathrm{PERM}} q$ are in conflict if we can prove $+\partial_O a$.

Approach III The third approach is a special variation of the first one. Here, defeaters for OBL specifically express the idea of permission. In fact, if we have that $a \rightsquigarrow_{\mathrm{OBL}} q$, this means that a is a reason for blocking the derivation of $+\partial_{\mathrm{OBL}}{\sim}q$, which is the prohibition for q. Since in our logic defeaters are in general weaker than strict rules, this approach can only provide a defeasible notion of permission. Here is the proof condition that expresses this intuition:

$+\partial_{\mathrm{PERM}}$: If $P(1..n) = +\partial_{\mathrm{PERM}} q$, then
(1) $-\Delta_{\mathrm{OBL}}{\sim}p \in P(1..n)$; and
(2) (2.1.) $\exists r \in R^{\mathrm{OBL}}_{dft}[q]$ such that
$\forall \mathrm{INT}a \in A(r) + \partial_I a \in P(1..n)$ and
$\forall \mathrm{OBL}a \in A(r) + \partial_O a \in P(1..n)$ and
$\forall \mathrm{PERM}a \in A(r) + \partial_{\mathrm{PERM}} a \in P(1..n)$ and
$\forall a \in A(r) + \partial_K a \in P(1..n)$, or
(2.2) $\exists r \in R^{K}_{dft}[q]$ such that
$\forall a \in A(r) + \partial_{\mathrm{PERM}} a \in P(1..n)$, and
(3) $\forall s \in R[{\sim}q : t]$ either
(3.1) if $s \in R^{O}[{\sim}q]$ then
$\exists \mathrm{INT}a \in A(s) - \partial_I a \in P(1..n)$ or
$\exists \mathrm{OBL}a \in A(s) - \partial_O a \in P(1..n)$ or
$\exists \mathrm{PERM}a \in A(s) - \partial_{\mathrm{PERM}} a \in P(1..n)$ or
$\exists a \in A(s) - \partial_K a \in P(1..n)$, and
(3.2) if $s \in R^{K}[{\sim}q]$ then
$\exists a \in A(s) - \partial_O a \in P(1..n)$, or
(3.3) $r > s$.

Discussion In standard deontic logic PERM is defined as the dual of OBL. Something similar to this has been also adopted here as regards the first and the third approaches.

In the third approach it suffices to understand a literal p to be permitted if there is a defeater with the head p such that this defeater overrides all obligation rules that allow to infer $\neg p$. In this case, to derive a permission, such as PERMq, we have to show that the derivation of the corresponding prohibition (OBL$\neg p$) fails. In the logic we have developed, rules are meant to introduce a positive modal conclusion, thus it is not possible to prove directly $\neg$OBL$\neg p$. There could be two reasons why $\neg$OBL$\neg p$ fails to be provable: the theory does not have enough resources to prove it, i.e., there are no applicable obligation rules for $\neg p$, or the obligation rules for $\neg p$ are all defeated. As we have seen, only strict and defeasible rules can be used to support a conclusion, but any rule can be used to prevent the derivation of a conclusion. Thus if we do not want to obtain the stronger conclusion that p is permitted because it is obligatory we have to use a defeater to defeat a rule for $\neg p$.

The first and the third approaches are based on a similar intuition, but with a substantial difference. Since mental states, under certain conditions, can defeat rules for obligations, the reasons that lead to derive that q is permitted may depend, in the first approach, on rules, for example, for knowledge or for other mental states. To understand the difference between these two approaches, let us consider the following theory:

$$\begin{aligned}
F &= \{\mathit{Weekend},\ \mathit{AirPollution}\,, \mathit{Emergency}\}\\
R &= \{r_1 : \mathit{Weekend}, \mathit{AirPollution} \Rightarrow_{\mathrm{OBL}} \neg\, \mathit{UseCar}\\
&\qquad r_2 : \mathit{Weekend}, \mathit{Emergency} \Rightarrow_{\mathrm{INT}}\ \mathit{UseCar}\}\\
&>=\{r_2 > r_1\}
\end{aligned}$$

Rule r_1 states that on weekends it is forbidden to use private cars if certain air pollution limit values are exceeded. Rule r_2 says that the agent intends to use the car on weekends in case of emergency. The peculiarity of the first approach is that we can derive permissions without having any explicit normative provision in such sense. Thus, since r_2 attacks and defeats r_1, then $-\partial_{\mathrm{OBL}}\neg\mathit{UseCar}$, which implies $+\Delta_{\mathrm{PERM}}\mathit{UseCar}$. Adopting the third approach, however, this conclusion is not obtained: a defeasible permission $+\partial_{\mathrm{PERM}}\mathit{UseCar}$ can be derived if, for example, we change r_2 into

$$r_2 : \mathit{Weekend}, \mathit{Emergency} \leadsto_{\mathrm{OBL}}\ \mathit{UseCar}$$

This suggests that the first approach is unsatisfactory in many cases where we allow for the interaction between normative provisions and mental states, and that the third approach should be preferred.

Consider now the following theory.

$$\begin{aligned}
F &= \{\mathit{Weekend},\ \mathit{AirPollution}\,, \mathit{Emergency}\}\\
R &= \{r_1 : \mathit{Weekend}, \mathit{AirPollution} \Rightarrow_{\mathrm{OBL}} \neg\, \mathit{UseCar}\\
&\qquad r_2 : \mathit{Emergency} \leadsto_{\mathrm{OBL}}\ \mathit{UseCar}\\
&\qquad r_3 : \mathit{Weekend} \Rightarrow_{\mathrm{INT}} \neg\, \mathit{UseCar}\}\\
&>= \{r_2 > r_1\}
\end{aligned}$$

Let us adopt the third approach. Rules r_3 and r_2 attack and defeat each other, and so we cannot derive that it is permitted to use the car. However, this does not happen if we reframe this theory according to the second approach and thus change r_2 into

$$r_2 : \mathit{Emergency} \Rightarrow_{\mathrm{PERM}}\ \mathit{UseCar}$$

In fact, in this case r_3 does not attack r_2. This suggests that in this case the second approach should be preferred.

9 Intention Reconsideration and Defeasibility

One of the issues in the design of agents based on the models of intention is that of when to reconsider intentions. An agent cannot simply maintain an intention, once adopted, without ever stopping to reconsider. It is necessary from time to time for an agent to check whether the intention has been achieved or whether it is no longer achievable.

Most of the existing models of intentional systems are based on deliberative or non-deliberative theory of intention where reconsideration of intention is a costly computational process. The decision making procedure in such agents are composed of two main activities like *deliberation* (deciding *what intentions* to achieve) and *means/ends reasoning* (deciding *how* to achieve these intentions). Deliberation itself is a computationally costly process and requires an appropriate intention reconsideration policy which helps the agent to deliberate only when necessary which as already mentioned is itself a potentially costly computational process. Whereas in the case of our policy-based intention framework the defeasibility of general intentions makes it possible to block the application of the intention to the particular case without abandoning/reconsidering the intention. This is in agreement with Bratman's view of policy-based intentions where the defeasibility of general intentions makes it possible to block the application of the intention to the particular case without abandoning the intention. For instance,

> *In an emergency, one does not give up the general policy of buckling up while driving, but will block the application to the particular case [13].*

This way the amount of deliberation required for intention re-consideration can be minimised to some extent.

A distinction between deliberative and policy based on one hand and non-deliberative intentions is that the first two can be modified when the conditions leading to their formation have changed while non-deliberative intentions stay the same. Thus it seems that contrary to non-deliberative intentions both deliberative and policy-based intentions are subject to non-monotonicity. However, this is not the case. Deliberative intentions can be changed only if the conditions leading to their adoption have changed or the theory of the agent has been modified (for example using belief revision). On the other hand a policy-based intention can be defeated if new pieces of information have been supplied. Of course we give up a previously adopted policy-based intention if the conditions leading to it are no longer valid.

The above distinction is closely related to strong intentions and weak intentions, which, as we have seen, are related to the four types of conclusions available in DL. To give up a strong intention we have to change (revise) the theory (i.e., we have to modify the strict rules), and thus we have to reconsider our policies, while we can abandon a weak intention if we have an exception to it without having to change the theory. To further illustrate this point let us consider the following theory where we have

$$r_1 : a \rightarrow_I b \tag{13}$$

$$r_2 : c \rightarrow_I \neg b \tag{14}$$

and a second theory where the same connections are expressed as defeasible connections

$$r'_1 : a \Rightarrow_I b \tag{15}$$

$$r'_2 : c \Rightarrow_I \neg b \tag{16}$$

$$r'_2 > r'_1 \tag{17}$$

Obviously in both cases we obtain INTb given a as a fact. However if both a and c are given then from the first theory we get an inconsistency (see 3) and we have to revise the theory. If we use belief revision to change the theory then we have to remove r_1 from the theory (if we use the information similar to the superiority relation of the second to avoid an empty theory as result of the belief revision operation)[16]. A consequence of this operation is that we are no longer able to derive INTb from a. An alternative would be to use base revision (see, for example [25]) instead of belief revision. If this strategy is taken then r_1 is changed into

$$r''_1 : a, \neg c \rightarrow_I b$$

Again, it is not possible to obtain INTb from a. To derive it we have to supplement the theory with the information whether c or $\neg c$ is definitely the case, increasing then the cognitive burden on the agent and it may lead to factual omniscience.

If the same information were encoded as weak intention, as in the second theory, then we would not suffer from the above drawback, since the second theory prevents the conclusion of an inconsistency (in case we do not specify that r'_2 is stronger than r_1, we are not able to conclude INTb nor INT$\neg b$).

10 Related Work

Reasoning about mental attitudes is a central issue in modelling agents, multi-agents, and normative multi-agent systems. For a survey of current research on modelling interactions between mental attitudes for (normative) multi-agents see [11]. Despite the plethora of proposals devoted to this topic, the related work that is directly relevant for this paper is mainly the BOID architecture for the interactions of mental attitudes and [40] for minimising the effects of logical omniscience using rule base formalisms for agents.

The approach of [40] is somehow similar to ours. The key idea is to model the epistemic state of an agent using rules. Thus the logical omniscience problem is avoided by considering a sentential semantics, instead of a semantics based on possible worlds and then propositions. The central point is that sentences must not be closed under logical inferences. A limitation of this approach is that by eliminating the omniscience problem it also eliminates side-effects. But as we have discussed, side effects are important when one discusses the interactions

[16] For Belief Revision in the spirit of AGM [1] for Defeasible Logic, see [10], and also [2]. However, proper modelling of norm and policy dynamic could be problematic for AGM based approaches, see [35].

among mental attitudes. Furthermore [40] is restricted to beliefs only, and it is not clear how to extend to other modalities, since rules are not used to obtain modal consequences.

The basic calculation scheme used in BOID [16] 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. But the major difference in our approach is in the way we use the superiority relation ($>$). In what follows we show how BOID is a particular case of our framework.

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 (B, O, I^{-}, D) often in the form of defeasible rules. I^{-} is used to represent the fact that set I contains *prior* intentions. 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 (ρ) 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, and this ordering in turn determines the agent type.

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, $\rho : B \cup O \cup I^{-} \cup D \rightarrow N$, that assigns a unique integer to each rule from $B \cup O \cup I^{-} \cup D$. It should be noted that the ordering function ρ 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. Hence a social simple-minded agent in BOID could be characterised as

$$\rho(r_b) < \rho(r_{i^-}) < \rho(r_o) < \rho(r_d)$$

where r_b, r_{i^-}, r_o, r_d etc. denotes the set of rules for belief, prior intentions obligations and desires.

Contrast this with our formalism wherein the superiority relation depends on the individual rules rather than on the type of rules. For instance, in our model the superiority relation ($>$) is a binary relation over the set of rules, i.e., $> \subseteq R^K \cup R^I$. Defining the superiority relation in this manner has several advantages. For example, modelling orders of overruling by means of individual relations permits not only to capture all basic agent type strategies but also to devise specific policies in which different characterisations of the same agent may be adopted with regard to different components or facts. Moreover it is easy to show that original BOID system for generating goals is a particular case of our theory. For instance, based on our proposed theory we can have 3 types of rules, $\Rightarrow_K, \Rightarrow_O, \Rightarrow_I$, respectively for belief, obligation and intention. Knowledge and belief can be expressed by using $\Rightarrow_K$ and intentions and goals by $\Rightarrow_I$. The concept of a derivation for obligation is the same as that given in Section 7. In order to reconstruct BOID all we have to do is to define that all the rules for K are stronger than the rules for O which in turn are stronger than the rules for I. We can switch between O and I according to the type of agent we want to model.

It is worth noticing that in the BOID architecture there are two phases. The goal generation and the goal selection phase. In our approach these two phases are combined in the same mechanism, the derivation of conclusions from a defeasible theory. Defeasible logic is a skeptical non-monotonic formalism, where, in case of a conflict between possible conclusions, if the conflict cannot be resolved by the superiority relation, none of the conclusions involved in the conflict is supported. BOID instead generates two conflicting goals and two corresponding extensions. After that the goal selection phase must be called to identify the selection an agent has to commit to for the generation of plans.

The order on the rules in BOID is given by an architectural design to determine classes of agents. The same approach was taken in [33] where we use a meta-programming approach to define classes of defeasible logic to model agents types. We discussed how to modify parameters in the meta-program to capture conflict between mental attitudes and then sets of parameters define agent types. The framework in [33] allows us to model a large class of agent types, including all agent types covered by the original BOID architecture.

Unfortunately, an architectural approach, as BOID, can lead to some counterintuitive results when one considers interactions among mental attitudes. A social agent is an agent such that obligations prevail over intentions. Thus a social agent is an agent that in case of a conflict between an obligation and an intention, gives up the intention. Similarly, a realistic agent is an agent who prefers beliefs over obligations. This mean that in case we have a conflict between an obligation and a belief, the agent deems the fulfilment of the obligation no longer achievable. In [36, 34] we show that for realistic social agents there are situations where, despite of the sociality of agents, some intentions prevail over obligations, contrary to the sociality of the agent. In [36, 34] we also show that the complexity of the reasoning mechanism to derive conclusion for defeasible logic with mental attitudes can be computed in time linear to the size of a defeasible theory, but that restoring sociality is an NP-complete problem. This shows that, while the architectural approach taken by BOID can be helpful for simple design of agents, it might not appropriate when interactions among modalities corresponding to principle used in real life, as the side-effects related to intention and thus responsibility in some legal doctrine, have to be modelled.

11 Discussion and Conclusion

Based on Bratman's classification of intention, we have outlined a formal theory of *policy-based* intention which differs from the usual NML-based approaches in the sense of having a non-monotonic nature. We provide further support to the idea that policy-based intentions are non-monotonic, and we argue that the notion of intention proposed by Bratman is not appropriate when agents are situated in an environment where they are subject to obligations. In fact Bratman's account fails to reconciliate with the commonly accepted doctrine of intention in legal theory where an agent is held responsible for the consequences of her intentions. To this end we advance that the so called side-effects are not a drawback of a theory of intention but an important feature of it. However, side-effects are closely related to the problem of logical omniscience. We examine three different degrees of logical omniscience and we discuss whether and to what extent they are acceptable for a theory of intention for rational agents. We advance that a limited form of logical omniscience is useful for the representation of agents in terms of a belief-desire-intention-obligation architecture. On the technical side we show how to implement the the above features (non-monotonicity, restricted side-effects and logical omniscience) in a computationally oriented non-monotonic formalism (defeasible logic) that accounts for a constructive view of the modal operators corresponding to the intensional notions at hand. In addition the proposed logic is flexible enough to deal with different intuitions about the interactions of the internal and external motivational attitudes.

The approach outlined in this paper could be extended in at least two different directions.

The first is in agreement with the work done in [58, 43]. Here they outline a policy description language called *PDL* and use logic programs to reason about the policies. The main concern in that work is in tracing the *event* history that gives rise to an *action* history based on stable model semantics. In a similar manner our approach could be developed using the appropriate semantics (Kunen [45] or argumentation [32]) and developed from a logic programming point of view. The advantage in our approach is the use of the superiority relation (>) whereby we can mention a hierarchy between the rules and this is absent in other works.

The second direction in which our work could be extended is to define various rules required for constructing goals from beliefs, intentions from goals, intentions from beliefs etc. and giving a superiority relation among these rules. The recent work on BDI [60] seems to take this direction. On the other hand many new applications in emerging information technologies have advanced needs for managing relations such as authorisation, trust and control among interacting agents (humans or artificial). This necessitates new models and mechanisms for structuring and flexible management of those relations. The issues of automated management of organisations in terms of policies and trust relations in highly dynamic and decentralised environments has become the focus in recent years.

Finally, as we have alluded to many semantics have been devised for defeasible logic and can be adapted straightforwardly to the extension proposed here. The method developed in [45] gives a set-theoretic fixed-point construction for $\Delta^+, \partial^+, \ldots$, which leads to a logic programming characterisation of defeasible logic. Programs corresponding to defeasible theories are sound and complete wrt Kunen semantics. The same technique is applicable in the present case with the obvious adjustments; however, it does not offer further insights on defeasible logic for BDI, because of the almost one-to-one correspondence between the inference conditions and the steps of the fixed-point construction. However semantics for defeasible BDI logic remains an interesting technical problem.

Acknowledgements

We would like to thank the anonymous referees for their valuable comments on early drafts of this paper.

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

This work was partially supported by the Australian Research Council under the Discovery Project DP0452628.

References

[1] Carlos E. Alchourrón, Peter Gardenfors, and David Makinson. On the logic of theory change: partial meet contraction and revision functions. *Journal of Symbolic Logic*, 50:510–530, 1985.

[2] Natasha Alechina, Mark Jago, and Brian Logan. Preference-based belief revision for rule-based agents. *Synthese*, 165(2/3):159–177, 2008.

[3] Grigoris Antoniou, David Billington, Guido Governatori, and Michael J. Maher. On the modeling and analysis of regulations. In *Proceedings of the Australian Conference Information Systems*, pages 20–29, 1999.

[4] 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.

[5] 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.

[6] Grigoris Antoniou, David Billington, Guido Governatori, and Michael J. Maher. Embedding defeasible logic into logic programming. *Theory and Practice of Logic Programming*, 6(6):703–735, 2006.

[7] L. Åqvist. Deontic logic. In Dov Gabbay and Franz Guentner, editors, *Handbook of Philosophical Logic (2nd edition), vol. 8*, pages 147–264. Kluwer, Dordrecht, 2000.

[8] N. Bassiliades, G. Antoniou, and I. Vlahavas. A defeasible logic reasoner for the Semantic Web. *International Journal on Semantic Web and Information Systems*, 2:1–41, 2006.

[9] David Billington. Defeasible logic is stable. *Journal of Logic and Computation*, 3:370–400, 1993.

[10] David Billington, Grigoris Antoniou, Guido Governatori, and Michael J. Maher. Revising nonmonotonic belief sets: The case of defeasible logic. In Wolfram Burgard, Thomas Christaller, and Armin B. Cremers, editors, *KI-99: Advances in Artificial Intelligence*, volume 1701 of *LNAI*, pages 101–112, Berlin, 1999. Springer-Verlag.

[11] Guido Boella, Leendert van der Torre, and Harko Verhagen. Introduction to the special issue on normative multiagent systems. *Journal of Autonomous Agents and Multi-Agent Systems*, 17(1):1–10, 2008.

[12] M.E. Bratman, D.J. Israel, and M.E Pollack. Plans and resource-bounded practical reasoning. *Computational Intelligence*, 4:349–355, 1988.

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

[14] Michael E. Bratman and Israel. Intention and personal policies. *Philosophical Perspectives*, 3:443–469, 1989.

[15] Jan Broersen, Mehdi Dastani, Joris Hulstijn, Zisheng Huang, and Leendert van der Torre. The BOID architecture: Conflicts between beliefs, obligations, intentions and desires. In *AGENTS '01: Proceedings of the fifth international conference on Autonomous agents*, pages 9–16. ACM Press, 2001.

[16] 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.

[17] Jan Broersen, Mehdi Dastani, and Leendert van der Torre. Resolving Conficts between Beliefs, Obligations, Intentions, and Desires. In S. Benferhat and P. Besnard, editors, *Symbolic and Quantitative Approaches to Reasoning with Uncertainty, 6th European Conference, ECSQARU 2001, Toulouse, France, September 19-21, 2001, Proceedings*, pages 568–579. Springer Verlag, 2001.

[18] Jan Broersen, Mehdi Dastani, and Leendert van der Torre. BDIO$_{CTL}$: Obligations and the specification of agent behavior. In Georg Gottlob and

Toby Walsh, editors, *Proceedings of Eighteenth International Joint Conference on Artificial Intelligence (IJCAI-03)*, pages 1389–1390. Morgan Kaufmann, 9-15 August 2003.

[19] José Carmo and Andrew J.I. Jones. Deontic logic and contrary-to-duties. In Dov Gabbay and Franz Guentner, editors, *Handbook of Philosophical Logic (2nd edition), vol. 8*, pages 265–343. Kluwer, Dordrecht, 2000.

[20] Brian F. Chellas. *Modal Logic, An Introduction.* Cambridge University Press, Cambridge, 1980.

[21] Xiaoping Chen and Guiquan Liu. A logic of intention. In *Proceedings of the Sixteenth International Joint Conference on Artificial intelligence (IJCAI-99)*, pages 172–179, 1999.

[22] Philip R. Cohen and Hector J. Levesque. Intention is choice with commitment. *Artificial Intelligence*, 42(2-3):213–261, 1990.

[23] Rosaria Conte and Chrysanthos Dellarocas. *Social Order in Multiagent Systems.* Kluwer Academic Publishers, Boston, 2001.

[24] 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.

[25] Paolo Di Giusto and Guido Governatori. A new approach to base revision. In Pedro Barahona and José Júlio Alferes, editors, *Progress in Artificial Intelligence*, volume 1695 of *LNAI*, pages 327–341, Berlin, 1999. Springer-Verlag.

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

[27] 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.

[28] Ronald Fagin, Joseph halpern, Yoram Moses, and Moshe Vardi. *Reasoning about Knowledge.* MIT Press, 1995.

[29] Ronald Fagin and Joseph Y. Halpern. Belief awareness and limited reasoning. *Artificial Intelligence Journal*, 534:39–76, 1988.

[30] Rod Girle. *Modal Logic and Philosophy.* Acumen, Teddington, 2000.

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

[32] Guido Governatori and Michael J. Maher. An argumentation-theoretic characterisation of defeasible logic. In *Proceedings of the 14th european conference on artificial intelligence (ECAI-2000)*, pages 469–473, 2000.

[33] 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.

[34] Guido Governatori and Antonino Rotolo. BIO logical agents: Norms, beliefs, intentions in defeasible logic. *Journal of Autonomous Agents and Multi Agent Systems*, 17(1):36–69, 2008.

[35] Guido Governatori and Antonino Rotolo. Changing legal systems: Abrogation and annulment. part i: Revision of defeasible theories. In Ron van der Meyden and Leon van der Torre, editors, *9th International Conference on Deontic Logic in Computer Science (DEON2008)*, pages 3–18, Berlin, 2008.

[36] Guido Governatori, Antonino Rotolo, and Vineet Padmanabhan. The cost of social agents. In *5th International Conference on Autonomous Agents and Multi-Agent Systems*, pages 513–520. ACM Press, 10–12 May 2006.

[37] Guido Governatori, Antonino Rotolo, and Giovanni Sartor. Temporalised normative positions in defeasible logic. In *Proceedings of ICAIL '05*. ACM Press, 2005.

[38] Benjamin N. Grosof. Representing e-commerce rules via situated courteous logic programs in RuleML. *Electronic Commerce Research and Applications*, 3(1):2–20, 2004.

[39] Jaakko Hintikka. *Knowledge and Belief.* Cornell University Press, 1962.

[40] Mark Jago. Epistemic logic for rule-based agents. *Journal of Logic, Language and Information*, 18(1):131–158, 2009.

[41] Kurt Konolige and Martha E. Pollock. A representationalist theory of intention. In *Proceedings of the Thirteenth International Joint Conference on Artificial Intelligence (IJCAI-93)*, pages 390–395, 1993.

[42] Sarit Kraus, Daniel Lehmann, and Menachem Magidor. Nonmonotonic reasoning, preferential models and cumulative logics. *Artificial Intelligence*, 44:167–207, 1990.

[43] Jorge Lobo, Randeep Bhatia, and Shamim Naqvi. A policy description language. In *Proceedings of the Sixteenth National Conference on Artificial Intelligence (AAAI-99)*, pages 291–298. AAAI/MIT Press, 1999.

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

[45] Michael J. Maher and Guido Governatori. A semantic decomposition of defeasible logic. In *Proceedings of the American National Conference on Artificial Intelligence (AAAI-99)*, pages 299–305, 1999.

[46] Michael J. Maher, Andrew Rock, Grigoris Antoniou, David Billignton, and Timothy Miller. Efficient defeasible reasoning systems. *International Journal of Artificial Intelligence Tools*, 10(4), 2001.

[47] David Makinson. On a fundamental problem of deontic logic. In Paul McNamara and Henry Prakken, editors, *Norms, Logics and Information Systems*, pages 29–53. IOS Press, Amsterdam, 1998.

[48] David Makinson and Leendert van der Torre. Permission from an input/output perspective. *Journal of Philosophical Logic*, 32:691–711, 2003.

[49] John-Jules Ch. Meyer, Roel Wieringa, and Frank Dignum. The role of deontic logic in the specification of information systems. In Jan Chomicki and Gunter Saake, editors, *Logics for Databases and Information Systems*, chapter 4, pages 71–115. Kluwer Academic Publishers, Norwell, MA, 1998.

[50] Donald Nute. Defeasible logic. In *Handbook of Logic in Artificial Intelligence and Logic Programming*, volume 3, pages 353–395. Oxford University Press, 1987.

[51] Donald Nute. Defeasible reasoning. In *Proceedings of 20th Hawaii International Conference on System Science*, pages 470–477. IEEE press, 1987.

[52] Donald Nute, editor. *Defeasible Deontic Logic*. Kluwer, Dordrecht, 1997.

[53] 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.

[54] Jeremy Pitt, editor. *Open Agent Societies: Normative Specifications in Multi-Agent Systems*. John Wiley and Sons, Chichester, 2003.

[55] 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.

[56] Giovanni Sartor. *Legal Reasoning: A Cognitive Approach to the Law*. Springer, Dordrecht, 2005.

[57] Munindar P. Singh. Semantical considerations on intention dynamics for BDI agents. *Journal of Experimental and Theoretical Artificial Intelligence*, 10(4):551–564, 1998.

[58] Tran Cao Son and Jorge Lobo. Reasoning about policies using logic programs. AAAI-spring symposium on answer set programming, March 26-28 2001.

[59] Toru Sugimoto. A preference-based theory of intention. In *Sixth Pacific Rim International Conference on Artificial Intelligence (PRICAI-2000)*, Lecture notes in AI, pages 308–317. Springer-Verlag, 2000.

[60] John Thanagrajah, Lin Padgham, and James Harland. Representation and reasoning for goals in BDI agents. In *Twenty-Fifth Australasian Computer Science conference (ACSC-2002)*, volume 4 of *CRPIT*, pages 259–265, 2002.

[61] Richmond H. Thomason. Desires and defaults: A framework for planning with inferred goals. In Anthony G. Cohn, Fausto Giunchiglia, and Bart Selman, editors, *KR2000: Principles of Knowledge Representation and Reasoning*, pages 702–713, San Francisco, 2000. Morgan Kaufmann.

[62] Leon van der Torre and Tan Y. Contrary-to-duty reasoning with preference based dyadic obligations. *Annals of Mathematics and Artificial Intelligence*, 27:49–78, 1999.

[63] Bernd Van Linder. *Modal Logic for Rational Agents.* PhD thesis, Department of Computer Science, Utrecht University, 19th June 1996.

[64] Moshe Y. Vardi. On epistemic logic and logical omniscience. In *Proceedings of the 1st Conference on Theoretical Aspects of Reasoning about Knowledge (TARK)*, pages 293–305, 1986.

[65] Moshe Y. Vardi. On the complexity of epistemic reasoning. In *Proceedings of the Fourth Symposium on Logic in Computer Science (LICS)*, pages 243–252, 1989.