

# Defeasible Logic: Agency, Intention and Obligation

Guido Governatori<sup>1</sup> and Antonino Rotolo<sup>2</sup>

<sup>1</sup> School of ITEE, The University of Queensland, Australia

<sup>2</sup> CIRSFID, University of Bologna, Italy

**Abstract.** We propose a computationally oriented non-monotonic multi-modal logic arising from the combination of agency, intention and obligation. We argue about the defeasible nature of these notions and then we show how to represent and reason with them in the setting of defeasible logic.

## 1 Introduction

This paper combine two perspectives: (a) a cognitive account of agents that specifies motivational attitudes; (b) modelling societies of agents by means of normative concepts [4]. For the first approach, our background is the belief-desire-intention (BDI) architecture, where mental attitudes are taken as primitives to give rise to a set of Intentional Agent Systems [23,2]. This view has been proved to be interesting especially when the behaviour of agents is the outcome of a rational balance among their (possibly conflicting) mental states [3,24]. The normative aspect is based on some intuitions about agents and their societies, in which it is assumed that normative concepts play a decisive role, allowing for the co-ordination of autonomous agents [22,10,12].

Our approach has in general several points of contact with the BOID architecture [4,5,8,6], where a number of strategies are provided for solving conflicts among informational and motivational attitudes. BOID provides logical criteria (i) to retract agent's attitudes with the changing environment, and so (ii) to settle conflicts by stating different general policies corresponding to the agent type considered. A realistic agent thus corresponds to a conflict-resolution type in which beliefs override all other factors, while other agent types, such as simple-minded, selfish or social ones adopt different orders of overruling. As in the BOID architecture, our system is rule-based. In particular, it is developed in the setting of Defeasible Logic. All components are represented as defeasible conditionals. A rule such as  $p \Rightarrow_K q$  means that, given  $p$ , this implies defeasibly agent's belief that  $q$ . Our approach adopts a slightly different perspective. Our claim is to We develop a constructive account of BDI multi-modal logics where the rules are meant to devise suitable logical conditions for introducing modalities. If so, rules may also contain modalised literals, as for example in  $I p \Rightarrow_K q$ , where  $I$  is a BDI operator of intention. In the same spirit, possible conversions of a modality into another can be accepted, as when the applicability of  $I p \Rightarrow_K q$  may permit to obtain  $I q$ . Based on this intuitions, our focus will be on Bratman's [3] concept of *policy-based* intention [11]. The relation between mental attitudes and non-monotonicity should not sound surprising. Recent works by Thomason [27] and on BOID confirm this trend. Such a connection, with regard to epistemic logics, has already received much attention in the AI community [19]. However, the notion of defeasibility may play a new role

within a constructive theory of (modal) operators. As we said, our aim is to show how to introduce modalities in a (computationally oriented) non-monotonic formalism. In this way, the notion of defeasible derivability is crucial since rules for mental states and conditions for derivation involving them allow to introduce modal operators. This approach is motivated by the inherent computational complexity of multimodal logics [13] and, often, the notion of modality adopted for agents systems is by its own nature non-monotonic and so does not lend itself to necessitation [11]. The use of non-monotonic logics in intention reasoning allows the agent to reason with partial knowledge without having a complete knowledge of the environment. This also helps the agent in avoiding a complete knowledge of the consequences. We outline a proof theory whereby one can reason about ways of maintaining intention consistency in BDI like agent systems. The new approach facilitates the designer of an agent system like BDI in describing rules for constructing intentions from goals and goals from knowledge.

BOID system incorporates also obligations. This is crucial in characterising the interplay between internal and external factors. Such intuition is also adopted here and is framed as well within a non-monotonic setting. Even for this component, the logical aim is to devise suitable conditions for introducing modalities. Two questions may be decisive in this regard. First, it would be important to recast the logical nature of obligations and to investigate how defeasible logic, as described in the following sections, might capture the well-known defeasible character of deontic reasoning. A full analysis of the above issue is outside the scope of the paper. However, it is at least worth mentioning that our framework avoids a difficulty that is recognised in the deontic literature [7]. The source of this difficulty is the closure, classically accepted in Standard Deontic Logic, of the obligation operator under logical consequence. We simply point out that these difficulties are avoided by developing a suitable notion of logical derivation of obligations. In general, with the adoption of this strategy we preserve at least some basic properties of obligations such as the closure under logical equivalence and consistency. An important issue concerns the relation between obligations and mental states. As it is pointed out [5], a number of possible approaches are available. Here we focus shortly on some minimal principles that emerge from the agent specification approach considered in [6]. In particular, as argued there, we may adopt, for example, the schema  $Op \rightarrow Ip$ , or analogous versions for the other mental attitudes. This axiom is the strong version of intentional *norm regimentation* as it does not simply prescribe the consistency between obligations and intentions but states the inclusion of the former in the latter ones. This of course means that what is not intended is also not obligatory. Other principles, such as  $Op \rightarrow \neg I\neg p$ , correspond to weak forms of norm regimentation with regard to agent's mental states. In this sense, they also express hard constraints on agent systems. A different principle that regulate the interaction between obligations and desires may be  $(Op \wedge \text{GOAL}\neg p) \rightarrow \neg I\neg p$ . This avoids that the output of a conflict between an obligation and a desire is that of adopting a plan for obtaining what is desired. These principles can be easily encoded in our framework.

Last but not least, our framework is enriched by the notion of modal agency [9]. This aspect differentiates this system if compared, for example, to BOID architecture. The same logical strategy—a rule-based approach to introduce modalities—is also applied to this case. In particular, we will devise a set of rules to encode the action transitions occurring, under certain circumstances, as the results of actions.

We will focus on the idea of personal and direct action to realise a state of affairs. This concept is usually formalised by the well-known modal operator  $E$ , such that a formula like  $E_i p$  means that the agent  $i$  brings it about that  $p$ . Different axiomatisations have been provided for it but almost all include  $E_i p \rightarrow p$  (T, i.e., successfulness),  $\neg E_i \top$  (No),  $(E_i p \wedge E_i q) \rightarrow E_i(p \wedge q)$  (C), and are closed under logical equivalence [25,9]. This analysis, however, is here integrated by focusing on the intentional character of actions. This is done for two reasons. First, in the light of the logical framework we have defined so far it is interesting to devise criteria for handling the specific interaction between actions, intentions and the other mental states. Second, the aim is to make more precise the logical meaning of the notion of direct action. In fact, as found in the literature [26], it is not possible to capture with  $E$  the difference between the modal qualifications “sees to it” and “brings it about”. Both are usually represented by this modal operator, despite the fact that the former expression exhibits a clear intentional character, whereas the latter may refer as well to unintentional actions [14]. Thus we introduce the operator  $Z$  to express intentional actions. It is characterised by all basic properties of  $E$  plus the schema  $Zp \rightarrow Ip$ , which cannot be in general valid for  $E$ .

The interest of adding agency to a framework that includes cognitive states and obligations is evident. First, the simple combination of agency and deontic operators makes possible a more accurate representation of obligations directed to agents’ behaviour, such as in the case of  $OZp$ . In addition, it allows to express the creation of obligations, as in  $ZOp$ . As regards handling conflicts between rules, new possible types of agents can be defined, according to the order of overruling we want to adopt. In this perspective, forms of regimentation may be introduced especially for the operator  $Z$ . Finally, it is possible to embed in the system a number of interesting properties, such as  $ZOp \rightarrow Ip$ , which completes what is stated by a reasonable and analogous schema without the operator of agency, namely,  $IOp \rightarrow Ip$ .

Finally, a few notes on the meaning of rules for obligation, which emerge from focusing on their interplay with the other components we have described so far. If rules define the conditions for the introduction of modal operators, when we deal with obligations defeated by other components we may in fact adopt two different views. Suppose we have two rules like  $r_1 : p \Rightarrow_Z q$  (a rule for action) and  $r_2 : s \Rightarrow_O \neg q$  (a rule for obligation). Both are applicable and  $r_1$  defeats  $r_2$ . If so we cannot derive  $\neg q$  via  $r_2$  and so  $O\neg q$ . In a first interpretation (applicability-based obligation), that a rule for action (but the same applies to other components such as beliefs) collides with a rule for obligation means that a normative violation has occurred [4]. But if  $r_1$  prevails, in our setting we cannot argue in favour of the occurrence of  $O\neg q$ . On the other hand, a violation of an obligation does not imply the cancellation of such an obligation [28]. The obligation is still in force. This means that the existence of the actual obligation  $O\neg q$  depends on the applicability of  $r_2$ , independently of the effective derivation of its consequent. In a second interpretation (pure-derivability-based obligation) the existence of actual obligations depends on the effective derivation of the consequent of a rule. In this case we can argue as follows. On the one hand, the non-derivation of  $O\neg q$  means that, as soon as a violation occurs,  $r_2$  is nothing but a special kind of *prima facie* obligation: when violated, it does not make sense to deduce its consequent as a real obligation. On the other, and more radically, since the obligations that count in the system are those which

are derivable, we may say that, in the event the action of the agent blocks the inference of  $O\neg q$ , the agent is a sort of legislator within the system; similar considerations apply to when intentions override obligations.

## 2 Basic Defeasible Logic of Agency, Intention and Obligation

Usually modal logics are extensions of classical propositional logic with some intensional operators. Thus any classical (normal) modal logic should account for two components: (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 necessitation and 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. Such a formalisation might suit for the knowledge an agent has but definitely not for the intention part and, consequently, not for a logic of intentional agency. Furthermore, many authors have expressed concerns about the meaningfulness of  $O\top$ . 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 deontic agency. A logic of deontic agency should take care of the underlying principles governing the intention and the action of an agent. 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. Finally it should offer facilities to describe obligations and the relationships between the various modalities.

These are in short the main guidelines we will follow in this and the subsequent sections to develop a suitable framework to deal with agency, intention and obligation components. As we have argued so far, reasoning about intentions and other mental attitudes has a defeasible nature, and defeasibility is one of the proper characteristic of normative reasoning. Thus any system that aims at the integration of intentions and obligations, for example a multi-agent system, should cater for defeasibility. The two phenomena (mental attitudes and deontic notions) are both subject to defeasibility, but they might obey different and sometimes incompatible intuitions; thus we need a non-monotonic formalism that is able to deal with them in a flexible, efficient and modular way and should offers itself to a seamless integration of the relevant modal operators. Moreover we need an efficient and easily implementable system to capture the required defeasible instances.

Defeasible logic, as developed by Nute [20] with a particular concern about computational efficiency and developed over the years by [17,1], is our choice. The reason

being ease of implementation [18], flexibility [1] (it has a constructively defined and easy to use proof theory which allows us to capture a number of different intuitions of non-monotonicity) and it is efficient: it is possible to compute the complete set of consequences of a given theory in linear time [16].

A defeasible theory contains five different kinds of knowledge: facts, strict rules, defeasible rules, defeaters, and a superiority relation. In this section 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, “John is a minor”. In the logic, this might be expressed as  $minor(John)$ .

*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 minor is a person”. Written formally:  $minor(X) \rightarrow person(X)$ .

*Defeasible rules* are rules that can be defeated by contrary evidence. An example of such a rule is “every person has the capacity to perform legal acts to the extent that the law does not provide otherwise”; written formally:  $person(X) \Rightarrow hasLegalCapacity(X)$ . The idea is that if we know that someone is a person, then we may conclude that he/she has legal capacity *unless there is other evidence suggesting that h/she may not have*.

*Defeaters* are a special kind of rules. They are used to prevent conclusions not to support them. For example:  $WeakEvidence \rightsquigarrow \neg guilty$ . This rule states that if pieces of evidence are assessed as weak, then they can prevent the derivation of a “guilty” verdict; on the other hand they cannot be used to support a “not guilty” conclusion.

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

$$\begin{aligned} r : person(X) &\Rightarrow hasLegalCapacity(X) \\ r' : minor(X) &\Rightarrow \neg hasLegalCapacity(X) \end{aligned}$$

which contradict one another, no conclusive decision can be made about whether a minor has legal capacity. But if we introduce a superiority relation  $>$  with  $r' > r$ , then we can indeed conclude that the minor does not have legal capacity.

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 structure  $(F, R^K, R^I, R^Z, R^O, >)$  where  $F$  is a finite set of facts;  $R^K$ ,  $R^I$ ,  $R^Z$  and  $R^O$  are, respectively, finite set of rules (strict, defeasible rules and defeaters) for knowledge, intentions, agency, and obligations; and  $>$ , the superiority relation, is a binary relation over the set of rules (i.e.,  $> \subseteq (R^K \cup R^I \cup R^Z \cup R^O)^2$ ).

Intuitively, given an agent,  $F$  consists of the information the agent has about the world, its immediate intentions, its actions and the absolute obligations;  $R^K$  corresponds to the agent’s theory of the world, while  $R^Z$ ,  $R^I$  and  $R^O$  encode its actions, policy, and normative system;  $>$  captures the strategy of the agent (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.

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

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

Over the years a number of formulations of the proof theory of defeasible logic have been proposed (sometimes for variants of defeasible logic); here we will adopt the meta-program formalisation of [17].

The meta-program  $M$  assumes that the predicates, `fact(Head)`, `superior(Rule1, Rule2)`, `strict(Name, Operator, Head, Body)`, `defeasible(Name, Operator, Head, Body)`, and `defeater(Name, Operator, Head, Body)`, which are used to represent a defeasible theory, are defined. The interpretation of the basic predicates of the meta-program is as follows:

$$\begin{aligned}
 \text{fact}(p) & \text{ iff } p \in F \\
 \text{strict}(r, m, p, [a_1, \dots, a_n]) & \text{ iff } r : a_1, \dots, a_n \rightarrow_m p \in R_s[p] \\
 \text{defeasible}(r, m, p, [a_1, \dots, a_n]) & \text{ iff } r : a_1, \dots, a_n \Rightarrow_m p \in R_d[p] \\
 \text{defeater}(r, m, p, [a_1, \dots, a_n]) & \text{ iff } r : a_1, \dots, a_n \rightsquigarrow_m p \in R_{df}[p] \\
 \text{superior}(r, s) & \text{ iff } r > s
 \end{aligned}$$

According to the above predicates we introduce the definition of a rule.

$$\begin{aligned}
 \text{rule}(R, X, P, [A_1, \dots, A_n]) & :- \text{strict}(R, X, P, [A_1, \dots, A_n]). \\
 \text{rule}(R, X, P, [A_1, \dots, A_n]) & :- \text{defeasible}(R, X, P, [A_1, \dots, A_n]). \\
 \text{rule}(R, X, P, [A_1, \dots, A_n]) & :- \text{defeater}(R, X, P, [A_1, \dots, A_n]).
 \end{aligned}$$

We are now ready for the clause defining the meta-program describing the proof-theory of defeasible logic.<sup>3</sup> If we disregard the modal operator it is immediate to see that the following meta-program has the same structure as the meta-programs given for propositional defeasible logic in [1,17]. Essentially we have four (independent) copies of the same meta-program, one for each modality.

<sup>3</sup> We have permitted ourselves some syntactic flexibility in presenting the meta-program. However, there is no technical difficulty in using conventional logic programming syntax to represent this program. As usual with logic programming capital letters stand for variables, however we reserve  $K, O, I$  and  $Z$  for modalities, and we will use  $X, Y, W$  for variables ranging over modal operators.

```

strictly(P, K):- fact(P).
strictly(P, X):- fact(XP).
strictly(P, X):- strict(R, X, P, [A1, ..., An, Y1B1, ..., YmBm]),
    strictly(A1, K), ..., strictly(An, K),
    strictly(B1, Y1), ..., strictly(Bm, Ym).

```

The first two clauses establish that a conclusion is strictly provable if it is one of the facts, while the third corresponds to modus ponens for strict rules and strictly derivable literals. Notice that the first clause is relative to rule for knowledge; as we have argued before the rules in  $R^K$  are used to encode the description of the environment (and there is no modal operator  $K!$ ). Thus unmodalized literals can be thought of as prefixed by a virtual  $K$  modal operator.

```

defeasibly(P, X):- strictly(P, X).
defeasibly(P, X):- consistent(P, X),
    supported(R, X, P),
    not defeated(P, X, S).

consistent(P, X):- not strictly(~P, X).

defeated(P, X, S):- applicable(S, X, ~P),
    not overruled(~P, X, T, S).

overruled(P, X, T, S):- supported(T, X, P),
    superior(T, S).

applicable(R, X, P):- rule(R, X, P, [A1, ..., An, Y1B1, ..., YmBm]),
    defeasibly(A1, K), ..., defeasibly(An, K),
    defeasibly(B1, Y1), ..., defeasibly(Bm, Ym).

supported(R, X, P):- rule(R, X, P, [A1, ..., An, Y1B1, ..., YmBm]),
    defeasibly(A1, K), ..., defeasibly(An, K),
    defeasibly(B1, Y1), ..., defeasibly(Bm, Ym),
    not defeater(R, X, P, [A1, ..., An, Y1B1, ..., YmBm]).

```

The first clause allows the transformation of a strict conclusion in a defeasible conclusion. A defeasible derivation of a literal  $p$  consists of three phases. In the first phase we establish that the opposite literal is not strictly provable and then have to provide an applicable supportive rule for  $p$  (i.e., using the predicate `supported(r, p)`, where  $r$  is a supportive rule for  $p$ ), then in the second phase we build all possible counterarguments against  $p$  (i.e., `defeated(p, s)` meaning that the literal  $p$  is defeated by rule  $s$ ) and we have to verify that the conclusion is not defeated by the attacking arguments, so we try to rebut the counterarguments (i.e., `overruled(~p, t, s)`) by stronger arguments for the intended conclusion.

The relationship between proof tags on one hand and the predicates `strictly` and `defeasibly` on the other is as follows:

$$\begin{aligned}
 D \vdash +\Delta_X p \text{ iff } M \vdash \text{strictly}(p, X) & \quad D \vdash -\Delta_X p \text{ iff } M \vdash \text{not strictly}(p, X) \\
 D \vdash +\partial_X p \text{ iff } M \vdash \text{defeasibly}(p, X) & \quad D \vdash -\partial_X p \text{ iff } M \vdash \text{not defeasibly}(p, X)
 \end{aligned}$$

Let us consider a theory where  $F = \{Ia, b, Od, e\}$  and  $R = \{Ia, b \Rightarrow_Z c; e, Zc \Rightarrow_O f\}$ . Here we can prove  $+\partial_I a$ ,  $+\partial_K b$ ,  $+\partial_K e$  and  $+\partial_O d$  since they are facts. Then the first rule is applicable and we can derive  $+\partial_Z c$ , and now the second rule is applicable and we obtain  $+\partial_O f$ . If we replace the first rule with  $Ia, b \Rightarrow_K c$  we conclude  $+\partial_K c$  instead of  $+\partial_Z c$  and now the second rule is no longer applicable. We illustrate the theory with the help of a concrete example. A drunk surgeon intends to operate a patient. The surgeon is aware that operating under the influence of alcohol will result in a failure. Moreover the legal system under which the surgeon operates prescribes that people causing permanent damages as a result of negligence are responsible. Thus the two rules can be rewritten, respectively as

$$\begin{aligned} & I(\text{operate}), \text{drunk} \Rightarrow_Z \text{fail} \\ & \text{permanentDamages}, Z(\text{fail}) \Rightarrow_O \text{responsible} \end{aligned}$$

The conclusion is that the surgeon is responsible, because the damages are the result of an intentional negligence. What about when the surgeon, not on duty and he happens to be 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 permanent damages as a result of the operation, but he operates anyway. In this case we have to change the first rule in  $I(\text{operate}), \text{drunk} \Rightarrow_K \text{fail}$ . Here we derive  $+\partial_K \text{fail}$  instead of  $+\partial_Z \text{fail}$ , and thus we block the application of the second rule. Hence we cannot conclude that the surgeon is responsible.

### 3 Interaction among Agency, Intention and Obligation

The program given in the previous section does not account for the properties of the modal operators and their mutual relationships. For these we have to introduce more clauses in the meta-program.

```
strictly(P, K):- strictly(P, Z).
defeasibly(P, K):- defeasibly(P, Z).
```

These two clauses enable us to convert a conclusion in  $Z$  in a conclusion in  $K$ , and thus they mimic the successfulness of the modal operator  $Z$ .

Let us see now the relationship between the different kinds of rule we have introduced so far. Table 1 shows all possible cases and, for each kind of rule, indicates all *potential* attacks on it. Since we have defined four kinds of rules, we have to analyse twelve combinations, which are gathered in the table in six columns. Each column corresponds to a type of potential attack, such that the second rule placed in each box is nothing but the potential attack on the first one. If the potential attack fails, since the superiority does not play here any role, this means that the case at stake does not correspond to a real attack: The type of rule that wins does so in any case and independently of inspecting the strength of the rules involved (i.e., without considering superiority relation). To represent the possible attacks we have to strengthen the definitions of the predicate `consistent`.

```
consistent(P, X):- not strictly(~P, K),
                  not strictly(~P, Y1), ..., not strictly(~P, Yn).
```

$\Rightarrow_K P$	$\Rightarrow_K P$	$\Rightarrow_K P$	$\Rightarrow_I P$	$\Rightarrow_O P$	$\Rightarrow_O P$
$\Rightarrow_O \sim P$	$\Rightarrow_I \sim P$	$\Rightarrow_Z \sim P$	$\Rightarrow_Z \sim P$	$\Rightarrow_I \sim P$	$\Rightarrow_Z \sim P$
$+\partial_K P$	$+\partial_K P$	$-\partial_K P$	$-\partial_I P$	type of agent	type of agent
$\Rightarrow_O P$	$\Rightarrow_I P$	$\Rightarrow_Z P$	$\Rightarrow_Z P$	$\Rightarrow_I P$	$\Rightarrow_Z P$
$\Rightarrow_K \sim P$	$\Rightarrow_K \sim P$	$\Rightarrow_K \sim P$	$\Rightarrow_I \sim P$	$\Rightarrow_O \sim P$	$\Rightarrow_O \sim P$
$-\partial_O P$	$-\partial_I P$	$-\partial_Z P$	$-\partial_Z P$	type of agent	type of agent

**Table 1.** Basic Attacks

where  $Y_1, \dots, Y_n$  are the modalities that attack the modality  $X$ , according to Table 1. At the same time, we have to allow more types of rule in the attack phase.

$\text{applicable}(R, X, P) :- \text{rule}(R, Y, P, [A_1, \dots, A_n, W_1 B_1, \dots, W_m B_m]) ,$   
 $\text{defeasibly}(A_1, K), \dots, \text{defeasibly}(A_n, K),$   
 $\text{defeasibly}(B_1, W_1), \dots, \text{defeasibly}(B_m, W_m) .$

This clause is required for all  $Y$  that attack  $X$  in Table 1. Moreover, if  $Y = Z$  we have to include, due to the successfulness of the operator, the additional clause

$\text{applicable}(R, X, P) :- \text{rule}(R, Z \sim X P, [A_1, \dots, A_n, Y_1 B_1, \dots, Y_m B_m]) ,$   
 $\text{defeasibly}(A_1, K), \dots, \text{defeasibly}(A_n, K),$   
 $\text{defeasibly}(B_1, Y_1), \dots, \text{defeasibly}(B_m, Y_m) .$

Table 1 (and, as we shall see, Tables 2 and 3) provides some basic criteria for classifying cognitive agents [8,4]. The general assumption of Table 1 is to deal with realistic agents. In other words, we set criteria for solving conflicts in which beliefs in general override the other components. In fact, our approach considers epistemic rules as agent's basic principles of rationality about the world. The only exception to this view is that rules for action may attack rules for belief, since the former ones capture the mechanism that governs the factual results of (intentional) actions. We can speak in this case of quasi-realistic agents since, given a certain belief, a contrary evidence based on rules for action may prove that such a belief is false. Given this background, Tables 2 and 3 will consider other agent's types, such as selfish and social, plus further specifications deriving from more articulated criteria for solving conflicts. As we shall see, the double reading assigned to the rules for obligation will allow us to provide an alternative interpretation of some already established criteria for handling conflicts between deontic factors, on one hand, and mental as well as action components, on the other.

Let us focus on some examples for each type of potential attack described in Table 1. Suppose we have (first column from the left)  $r_1 : \text{forest, dry, spark} \Rightarrow_K \text{fire}$  and  $r_2 : \text{forest} \Rightarrow_O \neg \text{fire}$ . It is clear that rule  $r_2$  does not determine a real attack on  $r_1$ . Since we assume the agent is realistic, rule  $r_1$  is nothing but a principle of rationality of the agent: It says that a fire is (defeasibly) the consequence of a spark in a dry forest. Rules like  $r_1$  must prevail with regard to deontic rules, such as  $r_2$  that prohibits to light a fire in a forest. When  $r_1$  is attacked by  $r_2$ , the output that follows from  $r_1$  is not affected by this attack and the fire should be obtained since this fact is independent from any rule that forbids to light fires in the forest. Vice versa, the derivation of the obligation not to light a fire is blocked since such an obligation is meaningless when the conditions for  $r_1$  occur: Of course,  $r_2$  does not apply when *fire* is obtained according to agent's rationality.

Similar remarks apply to the case that involves rules for knowledge and intention (second column). Let us consider the rule  $r_3 : \text{cautious} \Rightarrow_I \neg \text{fire}$ . Even here it is reasonable to argue in favour of  $r_1$ . Although agent's being cautious means to intend not to light a fire, this intention does not necessarily override  $r_1$ , namely *the fact*, according to agent's knowledge, that a spark normally causes a fire in a dry forest. This means that, when  $r_3$  attacks  $r_1$ , the consequent of the latter must be obtained, while the reverse attack should prevent to get  $I\neg \text{fire}$  since such an intention is meaningless when the agent assumes rationally that the fire must spread through the forest. Different arguments may be put forward when a rule for action,  $r_4 : \text{protect\_spark} \Rightarrow_Z \neg \text{fire}$ , is considered in combination with  $r_1$ . Rule  $r_4$  states *the fact* that *fire* obtains and may be viewed as a (factual) contrary evidence with regard to  $r_1$ . In general, rules like  $p \Rightarrow_Z q$  say that a specific action performed by agent, under certain circumstances, defeasibly determines through such action the occurrence of  $q$ , and so that  $Zq$ . The applicability of these rules may thus be a factual and contrary evidence with respect to  $K$ -rules that would allow to infer  $\neg q$ . For similar (but opposite) reasons, the reverse attack ( $r_1$  on  $r_4$ ) should block the derivation of  $\neg \text{fire}$ .

Since we assume the rationality of the agent with regard to its knowledge about the world, we have set that rules for knowledge be greater in strength with regard to rules for obligation and intention. Actions may override knowledge while mutual attacks involving intentions and actions determine real attacks for the trivial reason that actions are intentional in character. It is obvious that, when we have a rule such as  $r_5 : \text{incautious} \Rightarrow_I \text{fire}$ , the attack of  $r_5$  on  $r_4$  prevents from obtaining  $\neg \text{fire}$  while the reverse attack blocks the derivation of *fire*: Actions defined by rules for  $Z$  are intentional.

On the other hand, as we have indicated in Table 1, the interplay between obligations, intentions and actions cannot be settled so easily. In the light of well-known distinctions among different kinds of agent, Table 2 summarises all combinations related to the cases indicated in Table 1, first and second columns from the right. Let

$\Rightarrow_O p / \Rightarrow_I \sim p$		$\Rightarrow_O p / \Rightarrow_Z \sim p$	
$+\partial_O p$	$+\partial_I \sim p$	Independent	$+\partial_O p$ $+\partial_Z \sim p$ Strongly independent
$+\partial_O p$	$-\partial_I \sim p$	Strongly social	$+\partial_O p$ $-\partial_Z \sim p$ Social
$-\partial_O p$	$+\partial_I \sim p$	Selfish	$-\partial_O p$ $+\partial_Z \sim p$ Strongly selfish
$-\partial_O p$	$-\partial_I \sim p$	Strongly pragmatic	$-\partial_O p$ $-\partial_Z \sim p$ Pragmatic

**Table 2.** Type of Agent: Basic Attacks

us provide some brief comments. Independent and strongly independent agents are free respectively to adopt intentions and to perform intentional actions in conflict with obligations. In particular, within a pure-derivability-based interpretation of obligations (see Section 1), strongly independent agents may correspond to true cases of normative violation, since the actual obligation is derived in presence of a contrary action. As expected, for social and strongly social agents obligations override rules for action and for intention. In addition to the standard view [4], the overruling of intentions or actions with regard to obligations may configure a case of agent legislator, when, within a pure-derivability-based interpretation, only derived obligations count as such in the system. Pragmatic and strongly pragmatic are cases where no derivation is possible and so the

agent's behaviour is open to any other course of action other than those specified in the rules considered. To illustrate the potential conflicts between obligations and intentional acts we examine 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:

$$\Rightarrow_Z \text{confess} \quad \Rightarrow_O \neg \text{confess}$$

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

Table 2 does not cover all possible types of agent. In fact, the focus is there on possible attacks that involve only two rules. Table 3 completes the scenario and provides all possible combinations when we deal with three rules. It is worth noting that we consider only the case with  $\Rightarrow_O p$ ,  $\Rightarrow_I \sim p$  and  $\Rightarrow_Z \sim p$ : The case with  $\Rightarrow_I \sim p$  and  $\Rightarrow_Z p$  is meaningless since rules for  $Z$  govern only intentional actions. For similar reasons, some combinations in Table 3 are excluded (as highlighted by adding three question marks). Some comments on Table 3. Strongly independent agents are basically as in

$\Rightarrow_O p / \Rightarrow_Z \sim p / \Rightarrow_I \sim p$			
$+\partial_O p$	$+\partial_Z \sim p$	$+\partial_I \sim p$	Strongly independent
$+\partial_O p$	$+\partial_Z \sim p$	$-\partial_I \sim p$	???
$+\partial_O p$	$-\partial_Z \sim p$	$+\partial_I \sim p$	Selfish saint
$+\partial_O p$	$-\partial_Z \sim p$	$-\partial_I \sim p$	Hypersocial
$-\partial_O p$	$+\partial_Z \sim p$	$+\partial_I \sim p$	Sinner
$-\partial_O p$	$+\partial_Z \sim p$	$-\partial_I \sim p$	???
$-\partial_O p$	$-\partial_Z \sim p$	$+\partial_I \sim p$	Social sinner
$-\partial_O p$	$-\partial_Z \sim p$	$-\partial_I \sim p$	Hyperpragmatic

**Table 3.** Type of Agent: Other Attacks

Table 2 because  $Z$  implies  $I$ . The types hypersocial and hyperpragmatic do not add conceptually anything with respect to their corresponding and weaker versions of Table 2. The new cases are the selfish saint, sinner and social sinner types. The first is given when the content of agent's intention is in conflict with an obligation, but no intentional action to realise such a content is performed. The sinner performs this action and, in parallel, the obligation is defeated. The social sinner has this intention, the derivation of the obligation is blocked but no violating action is performed. Once again, notice that sinner and social sinner may viewed, within a pure-derivability-based interpretation, as peculiar cases of legislator.

Another interesting feature that could be explained using our formalism is that of *rule conversion*. 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 let  $p, q \Rightarrow_K r$  denote that an agent knows  $r$  given  $p$  and  $q$  (or  $r$  is a consequence of  $p$  and  $q$ ). Now suppose  $I(p)$  and  $I(q)$  are given. Can we conclude  $I(r)$ ? Here we should be careful about the interpretation of the rules as  $p \rightarrow_K q$  ( $q$  is a consequence of  $p$ ),  $p \Rightarrow_O q$  (given  $p$ ,  $q$  is obligatory),  $p \Rightarrow_I q$  (given  $p$  the agent has the intention  $q$ ), and  $p \Rightarrow_Z q$  (given  $p$  the agent sees to it that  $q$ ).

The adoption of conversions should not sound strange. 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, where  $B \vdash C$  and  $A \sim B$  imply  $A \sim C$  (see, e.g., [15]). In other words, it allows the combination of non-monotonic consequence with classical consequences. While not every combination of obligations and mental attitudes or action concepts will produce meaningful results for the conversion, some of them can prove useful in the present context. For example if we want to convert rules for knowledge/belief into rules for obligations we have to determine conditions under which a rule for knowledge can be used to directly derive an obligation. The condition we have after is that all the antecedents on the rule can be shown to be obligatory. In general, when we admit conversion of rules, the situation is such that when given environmental conditions are satisfied a rule for  $X$  is transformed in a rule for  $Y$ ; accordingly we have to use the “transformed” rule both in the support and attack phases. The conditions under which a rule can be converted are that all impersonal literals are (defeasibly) provable in  $K$  and all personal literals are (defeasibly) provable in the modalities required by the conversion (see Table 4). Formally we have

```
supported(R, X, P):- rule(R, Y, P, [A1, ..., An]),
    environment(A1, W), ..., environment(An, W),
    not defeater(R, X, P, [A1, ..., An]).
```

```
applicable(R, X, P):- rule(R, Y, P, [A1, ..., An]),
    environment(A1, W), ..., environment(An, W),
```

where

```
environment(P, X):- personal(P), defeasibly(P, X).
environment(P, X):- not personal(P), defeasibly(P, K).
```

The relationships among the modalities  $X$ ,  $Y$  and  $W$  are described in Table 4<sup>4</sup>. Notice that not all cases in the Table 4 can be accepted for all types of agents: the first column

<sup>4</sup> Table 4 should be read as follows. The first and second columns indicate the modal qualifications of the antecedents of a rule. The third column specifies the type of rule while the fourth provides the possible modal qualification we may obtain in the light of the antecedents and the rule type. Fifth columns says whether the corresponding conversion holds in all cases or characterises only some particular agent types. For example (fourth row from the top), if  $Zp, Iq \Rightarrow_K r$  is applicable we may obtain  $+ \partial_I r$ . On the other hand, the derivation of  $+ \partial_I r$  from  $I p, I q \Rightarrow_O r$  is possible only if we assume a kind of norm regimentation, with which we impose that all agents intend what is prescribed by deontic rules.

from the right indicates new types of agent corresponding to each rule conversion. This is particularly evident when obligations, actions and intentions are considered. Other combinations than those here defined are possible but they are problematic. As we can

X	Y	⇒	W	
O	O	K	O	For all agents
I	I	K	I	For all agents
Z	Z	K	Z	For all agents
Z	I	K	I	For all agents
I	I	O	I	quasi-intentional-regimentation
Z	Z	O	I	quasi-intentional-regimentation
Z	Z	O	Z	quasi-behavioural-regimentation
Z	I	O	I	quasi-intentional-regimentation
I	I	Z	I	For all agents
O	O	Z	O	For all agents
I	O	Z	I	quasi-socio-intentional-regimentation

**Table 4.** Conversions

see, some of the conversions above logically characterise new types of agents. Let us focus on them. All these types correspond to weak versions of strong norm regimentation. Strong regimentation, as maintained in [6], corresponds to adopting schemata like  $Op \rightarrow BELp$ . The just mentioned conversions configure weak forms of regimentation. For instance, the conversion described in the fifth row from the top. Roughly speaking, if we want to give an intuitive reading we could conceive it as follows:  $Ip$  and  $O(p \rightarrow q)$  entail  $Iq$ . Let us see with a concrete example the meaning of some conversions. The Yale Shooting Problem can be described as follows<sup>5</sup>

$$liveAmmo, load, shoot \Rightarrow_K kill$$

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”. If the agent intends to load and shoot the gun ( $I(load), I(shoot)$ ), then, since she knows that the consequence of these actions is the death of her friend, she intends to kill him ( $+\partial_I kill$ ). Similarly if she intentionally loads the gun and intends to shoot ( $Z(load), I(shoot)$ ). To intentionally killing him she has to load and to shoot the gun intentionally ( $Z(load), Z(shoot)$ ). If she intentionally loads the gun ( $Z(load)$ ) and accidentally shoots it ( $shoot$ ), she kills the friend ( $+\partial_K kill$ ) but this is not an intentional act ( $-\partial_Z kill$ ), since not all the actions leading to this dramatic conclusion are intentional. Finally 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 ( $-\partial_K kill$ ).

So far we have only examined cases where we pass from a single modality to a different modality. However axioms  $ZOp \rightarrow Ip$  and  $IOp \rightarrow Ip$  provide modal reductions. These principles can be described in the meta-program by the following clause

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

```

defeasibly(P, I):- defeasibly(OP, Z).
defeasibly(P, I):- defeasibly(OP, I).

```

Moreover we have to add clauses for `applicable` and `supported` where we consider rules for  $Z$  and  $I$  with conclusion  $Op$  when rules for  $I$  with conclusion  $p/\sim p$  are admissible.

## 4 Related and Future Work

Let us sketch just some short conclusions, also for future research.

Nute [21] proposed a Deontic Defeasible Logic which, in some respect, is similar to the framework presented here. Beside some minor differences in the way rules are handled at the propositional level, the main difference is that he uses only one type of rule. Traditionally, in proof-theory, rules to introduce operators give the meaning of them. Thus using one and the same type of rule both for obligation and factual conclusion does not show the real meaning of the operators involved. Moreover it is not clear to us whether and how complex conversions and reductions can be dealt with in a system with only a single type of rules.

As we said, another reference of this paper is to BOID. Its calculation scheme is similar to the one proposed here. For example, as in BOID it is possible to state general orders of overruling but also local preferences involving single rules. This last job is made here by means of the superiority relation. However, our system, which also deals with agency, is designed to take care of modalised literals and modal conversions. This is due to the logical task assigned to the rules. For this reason, but in a different perspective, our logical view may be also useful to study the notion of negative permission. In fact, conditions for  $\partial Op$  may also determine the implicit introduction of a modal operator of permission in terms of non-derivability of an obligation.

As regards the complexity of the system, [16] has proved, for the propositional case, that the set of tagged literals can be derived from the theory in linear time in the number of rules in it. It is not hard to extend this result to the modal case. The distinction of different kinds of rules does not affect the complexity of the theory. The case for  $K$  is the same adopted in standard Defeasible Logic while, for the other components, we convert relevant rules into the appropriate “extended” modal literals. At this point, the inference mechanism is the same as the standard one.

Due to space limitations it was not possible to show how to model other notions of agency—such as capability (both practical [9] and deontic [12]), attempt, and so on—that have received some attention in the literature in the past few years. Here it suffices to say that those notions can be easily represented (modularly) by adopting a strategy similar to that used in [11] to derive goals from intentions in a BDI defeasible logic.

## References

1. Antoniou, G., D. Billington, G. Governatori, and M. J. Maher. A flexible framework for defeasible logics. In *AAAI-2000*. AAAI/MIT Press, 2000.
2. Bratman, M., D. Israel, and M. Pollack. Plans and resource-bounded practical reasoning. *Computational Intelligence*, 4, 1988.

3. Bratman, M. E. *Intentions, Plans and Practical Reason*. Harvard University Press, 1987.
4. Broersen, J., M. Dastani, J. Hulstijn, Z. Huang, and L. van der Torre. The BOID architecture. In *Agents-01*. 2001.
5. Broersen, J., M. Dastani, and L. van der Torre. Resolving Conflicts between Beliefs, Obligations, Intentions, and Desires. In *ECSQARU 2001*, Benferhat, S. and P. Besnard, eds. Springer, 2001.
6. Broersen, J., M. Dastani, and L. van der Torre.  $BDIO_{CTL}$ : Obligations and the specification of agent behavior. In *IJCAI-03*. 2003.
7. Carmo, J. and A. J. Jones. Deontic logic and contrary-to-duties. In *Handbook of Philosophical Logic (2nd edition)*, vol. 8, Gabbay, D. and F. Guentner, eds. Kluwer, 2000.
8. Dastani, M. and L. van der Torre. A classification of cognitive agents. In *Cogsci'02*. 2002.
9. Elgesem, D. The modal logic of agency. *Nordic Journal of Philosophical Logic*, 2, 1997.
10. Gelati, J., G. Governatori, A. Rotolo, and G. Sartor. Declarative power, representation, and mandate: A formal analysis. In *JURIX02*, Bench-Capon, T., A. Deskalopulu, and R. Winkels, eds. IOS Press, 2002.
11. Governatori, G. and V. Padmanabhan. A defeasible logic of policy-based intention. In *AI 2003: Advances in Artificial Intelligence*, Gedeon, T. D. and L. C. C. Fung, eds., vol. 2903 of *LNAI*. Springer, 2003.
12. Governatori, G. and A. Rotolo. A defeasible logic of institutional agency. In *NRAC'03*, Brewka, G. and P. Peppas, eds. 2003.
13. Halpern, J. Y. and Y. Moses. A guide to completeness and complexity for modal logic of knowledge and belief. *Artificial Intelligence*, 54, 1992.
14. Hilpinen, R. On action and agency. In *Logic, Action and Cognition: Essays in Philosophical Logic*, Ejerhed, E. and S. Lindström, eds. Kluwer, 1997.
15. Kraus, S., D. Lehmann, and M. Magidor. Nonmonotonic reasoning, preferential models and cumulative logics. *Artificial Intelligence*, 44, 1990.
16. Maher, M. Propositional defeasible logic has linear complexity. *Theory and Practice of Logic Programming*, (6), 2001.
17. Maher, M. J. and G. Governatori. A semantic decomposition of defeasible logic. In *AAAI-99*. AAAI Press, 1999.
18. Maher, M. J., A. Rock, G. Antoniou, D. Billington, and T. Miller. Efficient defeasible reasoning systems. *International Journal of Artificial Intelligence Tools*, (4), 2001.
19. Meyer, J.-J. C. and W. van der Hoek. *Epistemic Logic for AI and Computer Science*. Cambridge University Press, 1995.
20. Nute, D. Defeasible logic. In *Handbook of Logic in Artificial Intelligence and Logic Programming*, vol. 3. Oxford University Press, 1987.
21. Nute, D. Norms, priorities, and defeasible logic. In *Norms, Logics and Information Systems*, McNamara, P. and H. Prakken, eds. IOS Press, 1998.
22. Pitt, J. (ed.). *Open Agent Societies*. Wiley, 2004.
23. Rao, A. and M. Georgeff. Modelling rational agents within a BDI-architecture. In *KR'91*, Fikes, A. J. and R. E. Sandewall, eds. Morgan Kaufmann, 1991.
24. Rao, A. and M. Georgeff. BDI agents: From theory to practice. In *ICMAS'95*. 1995.
25. Santos, F. and J. Carmo. Indirect action: Influence and responsibility. In *Deontic Logic, Agency and Normative Systems*, Brown, M. and J. Carmo, eds. Springer, 1996.
26. Sergot, M. and F. Richards. On the representation of action and agency in the theory of normative positions. *Fundamenta Informaticae*, 48, 2001.
27. Thomason, R. H. Desires and defaults: A framework for planning with inferred goals. In *KR2000*, Cohn, A. G., F. Giunchiglia, and B. Selman, eds. Morgan Kaufmann, 2000.
28. van der Torre, L. and Y. Tan. The many faces of defeasibility. In *Defeasible Deontic Logic*, Nute, D., ed. Kluwer, 1997.