The Cost of Social Agents*

Guido Governatori
School of ITEE
The University of Queensland
Brisbane, Australia
guido@itee.uq.edu.au

Antonino Rotolo
CIRSFID
University of Bologna
Bologna, Italy
rotolo@cirsfid.unibo.it

Vineet Padmanabhan
School of ITEE
The University of Queensland
Brisbane, Australia
vnair@itee.uq.edu.au

ABSTRACT

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

Categories and Subject Descriptors

I.2.4 [**Artificial Intelligence**]: Knowledge Representation Formalisms and Methods

General Terms

Theory, Legal Aspects, Algorithms, Performance

Keywords

Defeasible logic, Intention, Obligation, Social Agents, Computational Complexity

1. INTRODUCTION

Recent works on cognitive agents combine two apparently independent perspectives [6, 14, 8, 7, 9, 10]: (a) a classical cognitive account of agents that specifies their mental attitudes; (b) modelling agents' behaviour by means of normative concepts. For the first approach, the background is the belief-desire-intention (BDI) architecture, where mental attitudes are taken as primitives to give rise to a set of Intentional Agent Systems [20, 4]. This view is interesting especially when the behaviour of agents is the outcome of a rational balance among their (possibly conflicting) mental states. The normative aspect is rather based on the assumption that normative

---

concepts play a role to characterise the idea of social co-ordination of autonomous agents [19]. The nice result of this combination of perspectives is that of leading to an account of agents' deliberation and behaviour in terms of the interplay between mental attitudes and normative (external) factors such as obligations.

Following [14, 8, 7] we discuss how this combination can be framed in Defeasible Logic (DL). As is well-known, DL is based on a logic programming-like language and it is a simple, efficient but flexible non-monotonic formalism able to deal with many different intuitions of non-monotonic reasoning and recently applied in many fields. In addition, several efficient implementations have been developed [17, 3]. Here we discuss and extend some aspects of a non-monotonic logic of agency, based on the framework of [1], developed in [8, 7]. Indeed, DL is one of the most expressive languages that allows for the definition of large sets of patterns called agent types. Here, the focus will be in particular on philosophical and computational aspects of the notion of social agent, by which we mean a norm-complying agent.

The layout of the paper is as follows. Section 2 provides the theoretical background of our system. In particular, since the concept of social agent focuses on the interplay between obligations and intentions, we will discuss which kind of intentions have to be considered in this regard. Section 3 will present our logical framework, based on DL, which will embed our intuitions and permit to deal with social agency. Some first conclusions will be provided with reference to the problematic nature of social agents. Section 4 deals with the computational complexity of social agency.

2. SOCIAL AGENTS: OBLIGATION AND INTENTION

This section provides some theoretical background for the rest of the paper. Our focus is on the so-called policy-based attitudes. The term was coined by Bratman [5] with specific reference to the idea of intention. The intuition behind policy-based intention is based on Bratman's view regarding *future directed intention* and *general intention*. Bratman terms general intentions as *general/personal policies*. Along with general policies go policy-based intentions. For example, I have a general policy to patch up and reboot the Unix server in the department once every month. This morning, on the basis of this policy, I form the intention to reboot the machine at 7.00 pm in the evening. My intention this morning to reboot the machine this evening is a *policy-based* intention. This specific intention will play a major part in my planning process for the day as it will pose problems about means and constrain my other options. Based on this distinction Bratman makes the following classification of intentions: *deliberative, non-deliberative* and *policy-based*.

The difference between the three is the following: When an agent i has an intention of the form $INT_i^{t_1} \varphi, t_2$ (read as *agent i intends at*

---

t_1 to φ at t_2) as a process of *present* deliberation, then it is called *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 it is called *non-deliberative*. The third case arises when intentions are general and concern potentially recurring circumstances in an agent's life. Such general intentions constitute *policy-based intentions*. 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 a full-blown deliberation where an attempt is made to weigh pros and cons for and against conflicting options. It also differs from an intention in favour of necessary means, i.e., intention in favour of a specific end, in the sense that the defeasibility of general policies makes it possible to *block* the application of the policy to the particular case without *abandoning* the policy. Otherwise one could abandon the intention in favour of the end. The 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.

As argued in detail elsewhere [13], it may happen that a policy-based intention needs to be *re-considered* if not *blocked* for the application to particular cases. But this does not mean that the agent should know all such conditions in a scenario, but only those she considers necessary for the intended outcome and that she is not confident of their being satisfied. To intend the necessary consequence the agent has to make sure that all the evidence to the contrary has been defeated, which is basically a defeasible conclusion.

The starting point of this paper is to extend the policy-based approach to other attitudes and motivational factors such as beliefs and obligations. In this way, all motivational factors are represented within a rule-based system: intentions and beliefs are viewed as constituting the internal constraints (based on policies) of an agent while obligations are her external constraints (based on rules). As constraints they are defeasible. Notice, in particular, that such an extension to obligations can capture the well-known defeasible character of deontic reasoning. In this last case, a policy-based obligation –conceived of as an external motivational attitude– turns out to be simply a conditional obligation, namely, a rule that allows for the inference of an obligation whenever the antecedent of this rule holds [18, 21].

2.1 Intentions and Normativity

According to Bratman, rational agents can be basically modelled as follows:

- agents are goal-directed without being necessarily aware of their activity;

- intentions are used to choose partial plans for the realisation of a goal;

- not all consequences are intended but only some initial intentions and the goal as a result of the plan; if some side-effects occur, they are never intended.

It is worthy discussing here the notion of side-effect. This well-known problem has to deal with several variants of logical omniscience: the problem arises when the agent is required to know all the truths defined by her logic, or when the logic that depicts the agent automatically includes all the logical truths of classical logic, or, finally, if the agent knows all the logical consequences of the

known propositions [12]. Indeed, the problem is usually referred to as the *expected side-effects* problem [5], a problem which 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. 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 $\Gamma \mathrel{\vdash\mkern-6mu\sim}_{\mathbf{X}} \alpha$, where $\mathbf{X}$ is a modal operator, means that if we can prove all the formulas in Γ then we can deduce $\mathbf{X}\alpha$. In what follows we will develop a system for mental states and motivational attitudes based on this idea. However, we will allow the consequence relation for intentions and obligations to interact with the propositional module and we will also consider possible interactions between the modal operators. To this end we have to show that the expected side-effects phenomenon is not a drawback for policy-based agents: such a kind of agents must accept the expected-side effects unless they have some reasons to reject the consequences corresponding to them.

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

$$\text{INT}\textit{Smoke, Smoke} \rightarrow \textit{Cancer} \mathrel{\vdash\mkern-6mu\sim} \text{INT}\textit{Cancer} \qquad (1)$$

$$\text{INT}\textit{GoToRome, GoToRome} \rightarrow \textit{GoToItaly} \mathrel{\vdash\mkern-6mu\sim} \text{INT}\textit{GoToItaly} \qquad (2)$$

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

Accordingly, some cases of the side-effects problem are not necessarily a weakness of a theory. This holds in particular if we assume that our agents are *aware* of their activities. In our view, modelling rational agents corresponds to the following assumptions:

- agents are aware of their activities, of their policies;

- some cases of the side-effects problem can be accepted;

- if a case has to be rejected this means that its unpleasant consequences should not be intended;

- when unpleasant consequences are not intended, this only means that they are blocked by conflicting attitudes or facts.

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

intended unless it is removed in the meantime from deliberation. But this case is indeed considered within our analysis because we may have concrete contexts in which some policy-based intentions, as soon as they are applicable, turn out to be overridden by other policies: we may have reasons to argue that, if an agent intends *A* and believes that *B* is a consequence of *A*, this is not a reason for necessarily intending *B*; in fact, the derivation of *B* as an intention may be blocked, in our view, by competing attitudes or made non-applicable by concrete facts.

According to the previous discussion it should be clear that, though inspired by Bratman's [5] analysis, the notion of intention we study in this paper is slightly different, as it focuses on the idea of *intentionality*. In Bratman's view intentions are used to choose partial plans for the realisation of a goal; in this way they have a close relation to means-ends. In our view intentions should be related not only to means-ends but also to the their consequences.

This concept of intention is particularly relevant in conjunction with deontic and normative notions, for example if we want to say that an agent is legally 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 some of their consequences. Our intuition is compatible with von Wright's [22] classical theory of normative actions. Von Wright's problem is to identify what should be the content of norms. He argues that norms should deal with actions. Roughly, actions can be described in terms of state transitions and as the sets of all changes of world that follow from them. It is not our purpose discussing here von Wright's theory of action. It should be noted, however, that he considers the related problem of intentions. On the one hand, von Wright is clear when he says that any action may have an arbitrary number of consequences and not all of them are intended. On the other hand, he provides a very broad concept of action, according to which all actions in norms, strictly speaking, are intentional. If so, what are the boundaries of intentions to be considered when they interplay with obligations?

Let us see how to recast Bratman's Strategic Bomber scenario [5] 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 loses the war, and that there is a process for war crimes against him. Civil casualties are a sad but almost unavoidable consequence of war, but usually the killing of civilians does not constitute a war crime if there was no intention to kill. According to Bratman, Strategic Bomber did not commit a war crime since he did not have such an intention. 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.

Given this interpretation of intentions, we will see in the rest of this paper that some standard accounts of agent types, and of social agents in particular, are not satisfactory. We provide now some brief comments on the notion of agent type and social agent.

Classically, agent types are characterised by stating conflict resolution types in terms of orders of overruling between rules [6, 14, 7, 8]. In this perspective, agent types are meaningful within a non-monotonic setting and are nothing but general strategies to detect and solve conflicts between the different components of the cognitive profiles of agent's deliberation. In [6] 24 possible types are identified while, in [8], based on a different framework, 20 combinations are proposed. Typically, rational agents are assumed to be at least *realistic*: a realistic agent, in fact, is such that rules for beliefs override all other components, as beliefs correspond to agent's account of how the environment is. If the realistic condition is abandoned, we may have situations where intentions and desires override beliefs, thus leading to various forms of wishful thinking.

Given the minimal assumption that a rational agent should be realistic, we may further constrain agent's deliberation in order not to violate obligations: a *social agent* type requires that obligations are stronger than the other motivational components with the exception of beliefs. Various forms of social agency correspond to assuming axiom schemata such as

$$\text{OBL}\phi \rightarrow \neg\text{INT}\neg\phi \qquad (3)$$
$$\text{OBL}\phi \rightarrow \neg\text{DES}\neg\phi \qquad (4)$$
$$(\text{OBL}\phi \wedge \text{DES}\neg\phi) \rightarrow \neg\text{INT}\neg\phi \qquad (5)$$

In this paper we will consider only the interaction between intentions and obligations. But, even confining the problem to these components, the question at stake is: How to deal with social agents? The simplest solution is the classical one, corresponding to adopting schema (3): when we have two rules, one leading to INTϕ and the other to OBL$\neg\phi$, the former is blocked. Are we sure that this classical view is sufficient, given the account of policy-based attitudes we previously discussed?

3. BIO AGENTS IN DEFEASIBLE LOGIC

We focus on how beliefs, intentions and obligations jointly interplay in modelling agent's deliberation and behaviour. In particular, the system is meant to infer the goals the agent has to achieve.

The formal language contains modal literals and preferences, and is defined as follows:

DEFINITION 1. *Let* $M = \{\text{BEL}, \text{INT}, \text{OBL}\}$ *be a set of modal operators, and P a set of propositional atoms. The set of literals is defined as* $L = P \cup \{\neg p | p \in P\}$. *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).*

- *The language* $\mathscr{L}$ *is the smallest set including L and containing modal literals Xl and* $\neg Xl$ *when* $l \in L$ *is a literal and* $X \in M$ *is a modal operator.*

For $X \in \{\text{BEL}, \text{INT}, \text{OBL}\}$, we have that $\phi_1, \ldots, \phi_n \rightarrow_X \psi$ is a *strict rule* such that whenever the premises $\phi_1, \ldots, \phi_n$ are indisputable so is the conclusion ψ. $\phi_1, \ldots, \phi_n \Rightarrow_X \psi$ is a *defeasible rule* that can be defeated by contrary evidence. A rule $\phi_1, \ldots, \phi_n \rightsquigarrow_X \psi$ is a *defeater* that is used to defeat some defeasible rules by supporting evidence to the contrary.

DEFINITION 2. *A rule r consists of its* antecedent *(or* body*)* $A(r)$ *($A(r)$ may be omitted if it is the empty set), an arrow ($\rightarrow$ for a strict rule, $\Rightarrow$ for a defeasible rule, and $\rightsquigarrow$ for a defeater), and its* consequent *$C(r)$ (or* head*).*

- *The arrow is labelled with a modal operator* $X \in \{\text{BEL}, \text{INT}, \text{OBL}\}$. *If the arrow is labelled with* BEL *the rule is for belief, and similarly for the other modal operators.*

- *Given a rule r,* $A(r)$ *is a set of literals or modal literals, and* $C(r)$ *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 any R is labelled with X, that is R^X, all expressions just defined will refer to rules for X.

The purpose of the system is to derive modalised literals (goals), with the exception of rules for beliefs. As we shall see, provability for beliefs will not generate goals, as in our view beliefs concern the knowledge an agent has about the world: they may contribute to derive goals (here, intentions and obligations), but they are not in themselves motivations for action. For example, the application of $p \Rightarrow_{\text{OBL}} q$ permits to infer OBLq. Accordingly, modalities will not occur in the consequents of rules to keep the system manageable.

DEFINITION 3. *A defeasible agent theory is a structure $D = (F, R^{\text{BEL}}, R^{\text{INT}}, R^{\text{OBL}}, >)$ where F is a finite set of facts, R^{BEL} is a finite set of rules for belief, R^{INT} is a finite set of rules for intention, R^{OBL} is a finite set of rules for obligation, and >, the superiority relation, is a binary relation over the set of rules.*

The *superiority relation* > says when one rule may override the conclusion of another rule. *Facts* are indisputable statements.
   The following example illustrates the agent theory.

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

$$F = \{\text{INT}GoToShire, EntrustedByElrond\}$$
$$R = \{r_1 : EntrustedByElrond \Rightarrow_{\text{BEL}} RingBearer$$
$$r_2 : RingBearer \Rightarrow_{\text{OBL}} DestroyRing$$
$$r_3 : \text{INT}GoToShire \Rightarrow_{\text{INT}} \neg GoToMordor$$
$$r_4 : \neg GoToMordor \Rightarrow_{\text{BEL}} \neg DestroyRing\}$$
$$>= \{r_4 > r_2\}$$

3.1 Inferences with Social Agents

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

The meaning of the proof tags $+\Delta_X$, $-\Delta_X$, $+\partial_X$ and $-\partial_X$ is as follows: $+\Delta_X q$ means that q is definitly provable using only facts and strict rules for X, $-\Delta_X q$ means that it has been proved that q is not definitely provable, $+\partial_X q$ that q is defeasibly provable in D and $-\partial_X q$ that q is not defeasibly provable.
   We start with some terminology. As was explained, the following definition states the special status of belief rules, and that an introduction of a modal operator corresponds to being able to derive the associated literal using the rules for the modal operator.

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

   1. q is a literal and $P(m) = +\#_{\text{BEL}}q$ or

   2. q is a modal literal Xp and $P(m) = +\#_X p$ or

   3. q is a modal literal $\neg Xp$ and $P(m) = -\#_X p$.

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

   1. q is a literal and $P(m) = -\#_{\text{BEL}}q$ or

   2. q is a modal literal Xp and $P(m) = -\#_X p$ or

   3. q is a modal literal $\neg Xp$ and $P(m) = +\#_X p$.

The definition of Δ_X describes just forward chaining of strict rules:

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

For a literal q to be definitely provable we need to find a strict rule with head q, whose antecedents have all been definitely proved previously. And to establish that q cannot be definitely proven we must establish that for every strict rule with head q there is at least one antecedent which has been shown to be non-provable. Condition (3) says that a belief rule can be used as a rule for a different modal operator in case all literals in the body of the rule are modalised with the modal operator we want to prove. Thus given the rule $p, q \rightarrow_{\text{BEL}} s$, we can derive $+\Delta_Y s$ if we have $+\Delta_Y p$ and $+\Delta_Y q$.
   Conditions for ∂_X are more complicated. We define when a rule is applicable or discarded. A rule for a belief is applicable if all the literals in the antecedent of the rule are provable with the appropriate modalities, while the rule is discarded if at least one of the literals in the antecedent is not provable. For the other types of rules we have to take complex derivations into account called conversions [14]. In this paper we say there is a conversion from X to Y if a rule fior X can also be used as a rule for Y. We have thus to determine conditions under which a rule for X can be used to directly derive a literal q modalised by Y. Roughly, the condition is that all the antecedents a of the rule are such that $+\partial_Y a$.
   We represent all allowed conversions by a conversion relation c.

DEFINITION 6. *Let a conversion relation c be a binary relation over $\{\text{BEL}, \text{INT}, \text{OBL}\}$, such that (1) $c(X, Y)$ stands for the conversion of X rules into Y rules, (2) $Y \neq \text{BEL}$. Given a derivation P, $P(1..n)$ denotes the initial part of the derivation of length n.*

   - *A rule r in R^{BEL} is applicable iff $\forall a \in A(r)$, $+\partial_{\text{BEL}}a \in P(1..n)$ and $\forall Za \in A(r)$, where Z is a modal operator, $+\partial_Z a \in P(1..n)$.*

   - *A rule r in R^{BEL} is discarded iff $\exists a \in A(r)$ such that $-\partial_{\text{BEL}}a \in P(1..n)$ or $\exists Za \in A(r)$ such that $-\partial_Z a \in P(1..n)$.*

   - *A rule $r \in R_{sd}$ is applicable in the condition for $\pm\partial_Y$ iff*

      1. $r \in R^Y$ and $\forall a \in A(r)$, $+\partial_{\text{BEL}}a \in P(1..n)$ and $\forall Za \in A(r)$ $+\partial_Z a \in P(1..n)$, or

      2. $r \in R^X$ and $\forall a \in A(r)$, $+\partial_Y a \in P(1..n)$.

   - *A rule r is discarded in the condition for $\pm\partial_Y$ iff we prove either $-\partial_{\text{BEL}}a$ or $-\partial_X a$ for some $a \in A(r)$.*

      1. $r \in R^Y$ and $\exists a \in A(r)$ such that $-\partial_{\text{BEL}}a \in P(1..n)$ or $\exists Za \in A(r)$ such that $-\partial_Z a \in P(1..n)$; or

      2. $r \in R^X$ and $\exists a \in A(r)$ such that $-\partial_Y a \in P(1..n)$.

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

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

As a corollary of the definition of applicability, we can establish when a literal is supported (see Section 4 for the use of this notion):

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

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

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

To show that q is defeasibly provable we have two choices: (1) We show that q is already definitely provable; or (2) we need to argue using the defeasible part of a theory D. For this second case, three (sub)conditions must be satisfied. First, we require that there must be a strict or defeasible rule for q which can be applied (2.1). Second, we need to consider possible reasoning chains in support of $\sim q$, and show that $\sim q$ is not definitely provable (2.2). Third, we must consider the set of all rules which are not known to be inapplicable and which permit to get $\sim q$ (2.3). Essentially, each such a rule s attacks the conclusion q. For q to be provable, s must be counterattacked by a rule t for q with the following properties: (i) t must be applicable, and (ii) t must be stronger than s. Thus each attack on the conclusion q must be counterattacked by a stronger rule. In other words, r and the rules t form a team (for q) that defeats the rules s. $-\partial_X q$ is defined in an analogous manner.

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

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

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

3.2 The Problem of Social Agents

As suggested in [7, 8], agent types can be characterised in DL as follows:

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

While realistic agents are such that $X = \text{BEL}$ and $Y \in \{\text{INT}, \text{OBL}\}$, social agents are such that $X = \text{OBL}$ and $Y = \text{INT}$.

Unfortunately, this definition –adopted also in [6]– does not guarantee that agent's deliberation is oriented to fully complying with obligations. This drawback is mainly due to the introduction of conversions. Indeed, the notion of conversion should not sound strange. In many formalisms we can convert from one type of conclusion into a different one. Take for example the right weakening rule of non-monotonic consequence relations, where it is possible to combine non-monotonic with classical consequences: $B \vdash C$ and $A \mathrel{\vdash\!\!\!\sim} B$ imply $A \mathrel{\vdash\!\!\!\sim} C$ [15]. Here, conversions simply allow to obtain conclusions modalised by a certain X through the application of rules which are not modalised by X. In particular, they are fundamental in order to capture the fact that some side-effects should be accepted insofar as they are consequences of policies of which the agent is aware. Finally, some conversions seem useful to integrate the basic idea of social agency. For example, we may have agent types for which, given $p \Rightarrow_{\text{OBL}} q$ and $+\partial_{\text{INT}}p$, we can obtain $+\partial_{\text{INT}}q$. Of course, this is possible only if we assume a kind of norm regimentation, by which we impose that all agents intend what is prescribed by deontic rules.

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

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

$$a \Rightarrow_{\text{OBL}} q \qquad b \Rightarrow_{\text{BEL}} \sim q$$
$$+\partial_{\text{BEL}}a \qquad\qquad +\partial_{\text{INT}}b$$

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

4. THE COST OF SOCIAL AGENTS

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

DEFINITION 9. *Let # be one of the proof tags. Given a theory D, $D \vdash \pm\#p$ iff there is a derivation P in D such that for some n $P(n) = \pm\#p$.*

DEFINITION 10. *Given a theory D, the* universe *of D (U^D) is the set of all the atoms occurring in D; the* extension *of D (E^D), is defined as follows:*

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

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

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

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

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

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

PROOF. The proof is based on a modification of the algorithm given by Maher [16] to show that propositional defeasible logic has linear complexity.

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

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

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

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

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

The algorithm has three phases. (1) A pre-processing phase where we use the transformations given in [2] to transform a theory into an equivalent theory without superiority relation and defeaters; the transformation is linear. (2) A *rule loader* that parses the theory obtained in the first phase and generates the data structure that encodes the theory. (3) The *inference engine* applies transformations to the data structure, where at every step it reduces the complexity of the data structure.

The rule loader builds a data structure as follows: for every atom $\alpha \in U^D$ we create three entries α, INTα and OBLα. Each entry has associated to it a list of hash tables:

For α we have

- $+h$ is a list of (pointers to) rules in R^{BEL} where α appears in the head;

- $-h$ is the list of rules in R^{BEL} where $\sim\alpha$ appears in the head;

- $+b$ is the list of rules in R where α occurs in the body;

- $-b$ is the list of rules in R where $\sim\alpha$ occurs in the body.

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

- $+h$ is a list of rules in R^X where α appears in the head;

- $-h$ is the list of rules in R^X where $\sim\alpha$ appears in the head;

- $+h^B$ is a list of rules in R^{BEL} where α appears in the head;

- $-h^B$ is a list of rules in R^{BEL} where $\sim\alpha$ appears in the head;

- $+b$ is the list of rules in R where $X\alpha$ occurs in the body;

- $-b$ is the list of rules in R where $X\sim\alpha$ occurs in the body.

- $+b_\sim$ is the list of rules in R where $\sim X\alpha$ occurs in the body;

- $-b_\sim$ is the list of rules in R where $\sim X\sim\alpha$ occurs in the body.

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

The Inference Engine is based on an extension of the *Delores* algorithm/implementation proposed in [17] as a computational model of Basic Defeasible Logic. In turn

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

2. It scans the list of active rules for rules where the body is empty. It takes head and searches for rule (of the appropriate type) where the head is the negation of the atom or a modal literal incompatible with it. If there are no such rules then, the atom is appended to the list of facts, and removed from the rules

3. It repeats the first step.

4. The algorithm terminates when one of the two steps fails. On termination the algorithm outputs the set of conclusions.[1]

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

Given the above result it might seem that social agents are computationally feasible. However, as we have seen in Section 3.1 there are situations (let us call them deviant situations) where social agents do not behave as expected. First of all, we have to identify when we have a deviant situation and what are the reasons why we have them, and what kind of control an agent has over them. Here we assume that a deviant situation depends on some primitive intentions of an agent (i.e., intentions given as facts). Since these intentions are independent of the policy the theory describe the only alternative a social agent has is to give up some of them. In the rest of the section we study whether this is possible and what price an agent has to pay to be social. The answer is negative; we will provide a theory that is essentially deviant, and we will show that social agents are (computationally) expensive.

First of all we have to give a precise definition of the problem.

---

[1]This algorithm outputs ∂^+; ∂^- can be computed by an algorithm similar to this with the "dual actions". For Δ^+ we have just to consider similar constructions where we examine only the first parts of step 1 and 2. Δ^- follows from Δ^+ by taking the dual actions.

Restoring Sociality Problem

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

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

The specification of the problem is meant to formalise the situation we have described in the previous sections. The combination of the proof tags in the specification of the instance is only possible in case there is an applicable deontic rule for p $(+\Sigma_{\mathrm{OBL}}p)$ which would be otherwise unchallenged, i.e., there are no deontic rules to support $\sim p$ $(-\Sigma_{\mathrm{OBL}}\sim p)$ and there are no reasons to believe the opposite, is defeated, against the sociality of the agent, by the intentionality of $\sim p$ obtained as a consequence of an intention of the agent (this means it has been obtained by converting a belief rule into an intention rule). In other terms a potentially valid obligation is blocked by a consequence of an intentional behaviour.

EXAMPLE 5. Let us the consider the theory consisting of

$$F = \{\mathrm{INT}p, \mathrm{INT}s\}$$
$$R = \{r_1 : p, s \Rightarrow_{\mathrm{BEL}} q \quad r_2 : \Rightarrow_{\mathrm{OBL}} \sim q \quad r_3 : \Rightarrow_{\mathrm{BEL}} s\}$$
$$> = \{r_1 > r_2\}$$

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

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

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

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

EXAMPLE 6. Let the theory D be

$$F = \{\mathrm{INT}p\}$$
$$R = \{r_1 : \Rightarrow_{\mathrm{INT}} p \quad r_2 : p \Rightarrow_{\mathrm{BEL}} q \quad r_3 : \Rightarrow_{\mathrm{OBL}} \sim q\}$$
$$> = \{r_2 > r_3\}$$

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

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

EXAMPLE 7. Given a theory D as follows

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

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

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

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

THEOREM 2. *The Restoring Sociality Problem is NP-complete.*

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

For the NP-hard part we have to map a known NP-complete problem to the Restoring Sociality Problem. Here we use the *knapsack problem* [11, Problem MP9].

Knapsack Problem

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

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

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

- INT$load(u) \Rightarrow_{\mathrm{BEL}} load(u)$ for each $u \in U$.

- $\sum_{s(u):D\vdash+\partial_{\mathrm{BEL}}load(u)} s(u) > B \Rightarrow_{\mathrm{INT}} overload$

- $\sum_{s(u):D\vdash+\partial_{\mathrm{BEL}}load(u)} v(u) < K \Rightarrow_{\mathrm{INT}} undervalue$

- $overload \Rightarrow_{\mathrm{BEL}} \neg good$

- $undervalue \Rightarrow_{\mathrm{BEL}} \neg good$

- $\Rightarrow_{\mathrm{OBL}} good$

F is given by the relationship INT$load(u) \in F$ iff $u \in U'$.

The theory of the above construction has several interesting properties. First of all $D \vdash +\partial_{\mathrm{BEL}}load(u)$ iff INT$load(u) \in F$, which means $u \in U'$; then $D \vdash +\partial_{\mathrm{OBL}}good$ iff either of the two conditions of the knapsack problem are satisfied; notice that since there are no literals for $\neg load(u)$, the computation of the rule INT$load(u) \Rightarrow_{\mathrm{BEL}} load(u)$ can be computed independently of the rest of the theory thanks to the modularity of DL [2], thus the sums in the antecedent of the second and third rule can be considered as "facts" in the theory. In case one of the condition of the knapsack problem is not satisfied we have exactly a deviant situation as in the restoring sociality problem. The encoding of the knapsack problem in DL is clearly linear, thus any algorithm that solves the restoring sociality problem in polynomial time will solve the knapsack problem in polynomial time. Therefore the restoring sociality problem is NP-complete. $\square$

5. REVISING SOCIAL AGENTS

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

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

$$r_1 : a \Rightarrow_{\mathrm{BEL}} q \qquad r_2 : b \Rightarrow_{\mathrm{BEL}} \sim q$$
$$+\partial_{\mathrm{INT}}a \qquad\qquad +\partial_{\mathrm{OBL}}b$$
$$r_1 > r_2$$

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

6. SUMMARY

The contribution of this paper is manifold. We extend the analysis of policy-based cognitive agents with the notion of obligation and we argue that in such case side-effects do not endanger the logical analysis but on the contrary are beneficial to explain notions, e.g., intentionality, of paramount importance for agents situated in legal contexts.

Policy based agents are represented in defeasible logic extended with the modalities of belief, intention and obligation. This choice was motivated by the computational feasibility of the logic. We have demonstrated that the logic has linear complexity. As far as we know this is the first result of this kind for cognitive agents.

Finally we have studied the notion of social agent and we have proved that a proper and philosophically sound treatment of this notion leads to an increase of the computational complexity of the problem. Again this is the first result of this kind we are aware of.

7. REFERENCES

[1] G. Antoniou, D. Billington, G. Governatori, and M.J. Maher. A flexible framework for defeasible logics. In *Proc. AAAI-2000*, pages 401–405. AAAI/MIT Press, 2000.

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

[3] N. Bassiliades, G. Antoniou, and I. Vlahavas. DR-DEVICE: A defeasible logic system for the Semantic Web. In H.J. Ohlbach and S. Schaffert, editors, *Proc 2nd PPSWR*, LNCS 3208, pages 134–148. Springer, 2004.

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

[5] M.E. Bratman. *Intentions, Plans and Practical Reason*. Harvard University Press, 1987.

[6] J. Broersen, M. Dastani, J. Hulstijn, and L. van der Torre. Goal generation in the BOID architecture. *Cognitive Science Quarterly*, 2(3-4):428–447, 2002.

[7] M. Dastani, G. Governatori, A. Rotolo, and L. van der Torre. Preferences of agents in defeasible logic. In S. Zhang and R. Jarvis, editors, *Proc. Australian AI05*, LNAI 3809, pages 695–704. Springer, 2005.

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

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

[10] F. Dignum, D. Morley, L. Sonenberg, and L. Cavedon. Towards socially sophisticated BDI agents. In *Proc 4th ICMAS*, pages 111–118, 2000.

[11] M. Garey and D. Johnson. *Computers and Intractability: A Guide to the Theory of NP-Completeness*. W. H. Freeman and Company, 1979.

[12] R. Girle. *Modal Logic and Philosophy*. Acumen, 2000.

[13] G. Governatori, V. Padmanabhan, A. Rotolo, and A. Sattar. A defeasible logic for modelling policy-based intentions and motivational attitudes. *submitted*, 2005.

[14] G. Governatori and A. Rotolo. Defeasible logic: Agency, intention and obligation. In A. Lomuscio and D. Nute, editors, *Deontic Logic in Computer Science*, LNAI 3065, pages 114–128. Springer, 2004.

[15] S. Kraus, D. Lehmann, and M. Magidor. Nonmonotonic reasoning, preferential models and cumulative logics. *Artificial Intelligence*, 44:167–207, 1990.

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

[17] M.J. Maher, A. Rock, G. Antoniou, D. Billignton, and T. Miller. Efficient defeasible reasoning systems. *International Journal of Artificial Intelligence Tools*, 10(4):483–501, 2001.

[18] D. Nute, editor. *Defeasible Deontic Logic*. Kluwer, 1997.

[19] J. Pitt, editor. *Open Agent Societies*. Wiley, 2005.

[20] A.S. Rao and M.P. Georgeff. Modelling rational agents within a BDI-architecture. In *Proc. KR'91*, pages 473–484. Morgan Kaufmann, 1991.

[21] G. Sartor. *Legal Reasoning: A Cognitive Approach to the Law*. Springer, 2005.

[22] G.H. von Wright. *Norm and Action*. Routledge, 1963.