# A Defeasible Logic of Policy-Based Intention*

Guido Governatori and Vineet Padmanabhan

School of Information Technology & Electrical Engineering
The University of Queensland, Brisbane, QLD, Australia
{guido,vineet}@itee.uq.edu.au

**Abstract.** Most of the theories on formalising intention interpret it as a unary modal operator in Kripkean semantics, which gives it a monotonic look. We argue that policy-based intentions [8] exhibit non-monotonic behaviour which could be captured through a non-monotonic system like defeasible logic. To this end we outline a defeasible logic of intention. The proposed technique alleviates most of the problems related to logical omniscience. The proof theory given shows how our approach helps in the maintenance of intention-consistency in agent systems like BDI.

## 1 Introduction

Formalising cognitive states like intention has received much attention in the AI community [7, 17, 18, 23]. All these theories are based on Normal Modal Logics (NMLs), where intention is formalised into a modal operator on the framework of kripkean possible world semantics. Due to this restriction, these theories suffer from the *logical-omniscience* problem [10, 22]. One of the solutions suggested to overcome this problem is to adopt a non-kripkean semantics as shown in [5]. In that work intention is interpreted in terms of its *content* and the intention consequence relation is explained based on the content of two intentions. There is also a *representationalist* theory of intention [11] that employs the minimal model semantics [4] to interpret the intention operator. Work has also been done relating intention to preferences [20] as well as commitments [6]. However none of these theories have explicitly addressed the need for a non-monotonic theory of intention and we argue that to capture the properties involved in *policy-based* intention we need such a non-monotonic setup.

Our claim is based on Bratman's [8] classification of intention as *deliberative, non-deliberative, policy-based* and we show that *policy-based* intention is non-monotonic (i.e. has a defeasible nature). Though, many of the theories mentioned above is based on Bratman's work, they fail to recognize the non-monotonic component involved in intention. In this paper we adopt a particular non-monotonic system, (*defeasible logic*), to study the properties involved in *policy-based* intention and show how one can relate it with an intentional system like BDI [17]. The reason for defeasible logic is due to its computational efficiency [13] and

* This research was partially supported by the University of Queensland under the grant UQRSF ITEE-03/2002001335.

T.D. Gedeon and L.C.C. Fung (Eds.): AI 2003, LNAI 2903, pp. 414–426, 2003.

easy implementation [15]. We are unaware of any existing work relating reasoning about intention with non-monotonic reasoning to the best of our knowledge. We believe that our approach helps in bridging the gap between non-monotonic reasoning and reasoning about intention.

The proposed method provides solutions to the problem of *logical-omniscience* which usually accompanies intention-formalisms based on normal modal logics. 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. Moreover, we outline a proof-theory whereby one can reason about ways of maintaining intention consistency in agent systems like BDI. The new approach facilitates the designer of an agent system like BDI in describing rules for constructing intentions from goals and goals from knowledge. This is important as it is in alliance with the *commitment* axioms of Rao and Georgeff [17] and also provides an explanation on the practical nature of intentional systems like BDI. In this paper we don't want to recast the whole BDI theory but focus on the intention part supplemented by the factual knowledge and its underlying theory. Moreover similar considerations can be applied to the GOAL component.

In the next section we make the case for a non-monotonic theory of intention based on Bratman's classification of intention. In the third section we outline the problem of logical omniscience and in the fourth we give an overview of defeasible logic. The fifth section argues for a defeasible logic of intention. In the final section we make a comparison between our work and the work in policy-based reasoning

## 2 The Case for Non-monotonic Reasoning

An important classification of intention that is useful in computer science is that of *intending* versus *doing intentionally*, where the former involves the true intentions or preferences of the agent whereas the latter applies to the actions or states that the agent performs or brings about but not with any prior intention to do so. Based on this division Bratman classifies intentions as *deliberative, non-deliberative and policy-based*. When an agent $i$ has an intention of the form $\mathrm{INT}_i^{t_1}\varphi, t_2$ (read as *agent $i$ intends at $t_1$ to $\varphi$ 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 have retained it from $t_0$ to $t_1$ without reconsidering it then it is called *non-deliberative*. There can be a third case when intentions can be general and concern potentially recurring circumstances in an agent's life. Such general intentions constitute *policy-based intentions*, and is defined as follows: *when the agent $i$ has a general-(policy/intention) to $\varphi$ in circumstances of type $\psi$ and $i$ notes at $t_1$ that $i$ am (will be) in a $\psi$-type circumstance at $t_2$, and thereby arrive at an intention to $\varphi$ at $t_2$*. The difference here is that there is no *present* deliberation concerning the action to be performed as the agent already

has a general intention to do a particular action (*doing intentionally*). Whether the agent is able to perform that action or not depends on the circumstances.

When dealing with such general policies/intentions (hereafter intention), we have to take into account two cases. General intentions could be either (1) *periodic* or (2) *circumstance-triggered*. They are *periodic* in the sense that their occasion for execution is guaranteed by the mere passage of a specific interval of time. For instance, the general intention of patching up and rebooting the Unix server, *hobbit* in our department on every friday at 7pm. In contrast to this, general intention could be *circumstance triggered* as in the case of being *Root if one is Super-user*. Its occasion is not guaranteed by the mere passage of time but require that certain specific circumstances obtain. In both cases one can find that the general intention has an underlying defeasible nature. The defeasible nature is explained as follows. Consider the above example for *circumstance-triggered* general intention:

$$SU(X) \Rightarrow Root(X) \tag{1}$$

which means, (super-users are typically root). Suppose, there exists an agent $i$ (a software program) that monitors tasks related to giving root permissions as and according to whether a user is a normal-user (NU) or Super-User (SU) and $i$ has a general intention like (1). This general intention has a defeasible nature in the sense that, if $i$ knows that $X$ is a SU then $i$ may conclude that $X$ is Root, unless there is other evidence suggesting that $X$ may not be root (for instance, when $X$ has only read and write permissions but not execute permission). But this does not mean that the agent $i$ should know all such conditions but, only those he considers necessary to the intended outcome and that he/she isn't confident of their being satisfied. Hence our definition of general intention boils down to:

> *An agent intends all the necessary consequences of his performing his general intention and he isn't confident of their being satisfied.*

In order to intend the necessary consequence the agent has to make sure that all the evidence to the contrary has been defeated which basically is a defeasible logic conclusion. This is different from the usual NML interpretation where the agent intends all the consequences.

The formation of such general policies helps in extending the influence of deliberation as it is a partial solution to the problems posed by our limited resources for calculation and deliberation at the time of action. General policies also facilitate co-ordination. It may sometimes be easier to appreciate expectable consequences (both good and bad) of general ways of acting in recurrent circumstances than to appreciate the expectable consequences of a single case.

## 3 Logical Omniscience and Non-monotonicity

As we mentioned before, most of the theories based on NML's interpret intention as a unary modal operator in Kripkean semantics which makes it vulnerable to the problem of *logical-omniscience*. The problem in its general form as stated in [22] is as follows: (where **X** could represent a mental state like intention (INT)

1. $\models \mathbf{X}\varphi \wedge \mathbf{X}(\varphi \rightarrow \psi) \Rightarrow \mathbf{X}\psi$ (*side-effect problem*)
2. $\models \varphi \rightarrow \psi \Rightarrow \models \mathbf{X}\varphi \rightarrow \mathbf{X}\psi$ (*side-effect problem*)
3. $\models \varphi \Leftrightarrow \psi \Rightarrow \models \mathbf{X}\varphi \Leftrightarrow \mathbf{X}\psi$ (*side-effect problem*)
4. $\models \varphi \Rightarrow \models \mathbf{X}\varphi$ (*transference-problem*)
5. $\models (\mathbf{X}\varphi \wedge \mathbf{X}\psi) \rightarrow \mathbf{X}(\varphi \wedge \psi)$ (*unrestricted combining*)
6. $\models \mathbf{X}\varphi \rightarrow \mathbf{X}(\varphi \vee \psi)$ (*unrestricted weakening*)
7. $\models \neg(\mathbf{X}\varphi \wedge \mathbf{X}\neg\varphi)$

None of these properties except for (7) is valid when we take intention into consideration. For instance, consider a situation where an agent $i$ goes to the bookstore with the intention of buying a *paper-back* and also with the intention of buying a *magazine* because he has a general intention to buy them.[1] Hence according to (5) it could be formally given as:

$$\mathrm{INT}_i(\mathit{paperback}) \wedge \mathrm{INT}_i(\mathit{magazine}) \rightarrow \mathrm{INT}_i(\mathit{paperback} \wedge \mathit{magazine})$$

But this general intention is *defeasible* in the sense that at the bookstore the agent might find that he doesn't have enough money to buy both of them and hence drops intention to buy each of them and now only intends to buy one of them. NMLs fail to account for such type of reasoning. In Sugimoto [20] an extra notion of *preference* is added and an ordering among the preferences is done to capture the desired effect. But we argue that, in general, such intentions are defeasible and hence a non-monotonic reasoning system would be more efficient for such occasions. The above example could be stated in a non-monotonic setup as

$$\begin{aligned} &(1)\ \ \mathit{paper\text{-}back}(X) \Rightarrow \mathit{buy}(X),\\ &(2)\ \ \mathit{magazine}(X) \Rightarrow \mathit{buy}(X),\\ &(3)\ \ \mathit{costly}(X) \rightsquigarrow \neg \mathit{buy}(X); \end{aligned}$$

where (1) and (2) are premises which reflects the agents general intention of buying a paper-back and magazine unless there is other evidence like (3) suggesting that he/she may not be able to buy. When intention is formalised in the background of NMLs it is often the case that the agent has to have a complete description of the environment before-hand or has to be omniscient in the sense of knowing all the consequences. Classically the logical omniscience problem amounts to say that an agent has to compute all consequences of its own theory. It is obvious that some of the consequences are not intended as shown above. Moreover in classical NML the set of consequences is infinite. Hence we need a system like DL (defeasible logic) which is easily implementable and where the set of consequences consists of the set of literals occurring in the agent theory i.e. in the knowledge base, which is finite.

## 4 Overview of Defeasible Logic

As shown in the previous section, reasoning about general intention has a defeasible nature (in the sense that it is fallible) and hence we need an efficient

[1] The example is a slightly modified one as given in [20].

and easily implementable system to capture the required defeasible instances. Defeasible logic, as developed by Nute [16] with a particular concern about computational efficiency and developed over the years by [3, 2, 1] is our choice. The reason being easy implementation [15], flexibility [1] (it has a constructively defined and easy to use proof theory) and it is efficient: It is possible to compute the complete set of consequences of a given theory in linear time [13]. We do not address any semantic issues in this paper but the *argumentation semantics* as given in [9] could be straightforwardly extended to the present case.

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

*Facts* are indisputable statements, for example, "Vineet is a System Administrator". In the logic, this might be expressed as $SA(vineet)$.

*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 "System-Administrators are Super-Users". Written formally: $SA(X) \rightarrow SU(X)$.

*Defeasible rules* are rules that can be defeated by contrary evidence. An example of such a rule is "Super-Users are typically root"; written formally: $SU(X) \Rightarrow Root(X)$. The idea is that if we know that someone is a superuser, then we may conclude that he/she is root, *unless there is other evidence suggesting that it may not be root.*

*Defeaters* are rules that cannot be used to draw any conclusions. Their only use is to prevent some conclusions. In other words, they are used to defeat some defeasible rules by producing evidence to the contrary. An example is "If a user is normal-user then he might not be a root". Formally: $NU(X) \rightsquigarrow \neg Root(X)$. The main point is that the information that a user is NU is not sufficient evidence to conclude that he/she is not root. It is only evidence that the user *may* not be able to become root. In other words, we don't wish to conclude $\neg root$ if $NU$, we simply want to prevent a conclusion $Root$.

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 $r : SU \Rightarrow Root$ and $r' : RW \Rightarrow \neg Root$ which contradict one another, no conclusive decision can be made about whether a Super-User with a read & write permission can be root. But if we introduce a superiority relation $>$ with $r' > r$, then we can indeed conclude that the Super-User cannot be root. The superiority relation is required to be acyclic. It turns out that we only need to define the superiority relation over rules with contradictory conclusions.

It is not possible in this short paper to give a complete formal description of the logic. However, we hope to give enough information about the logic to make the discussion intelligible. We refer the reader to [16, 3, 2] for more thorough treatments.

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

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

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

$+\Delta q$, meaning that $q$ is definitely provable in $D$ (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 $D$.

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

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

The definition of $\Delta$ describes just forward chaining of strict rules. For a literal $q$ to be definitely provable we need to find a strict rule with head $q$, of which all antecedents have been definitely proved previously. And to establish that $q$ cannot be proven definitely we must establish that for every strict rule with head $q$ there is at least one antecedent which has been shown to be non-provable.

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

Let us work through this condition. To show that $q$ is provable defeasibly we have two choices: (1) We show that $q$ is already definitely provable; or (2) we need to argue using the defeasible part of $D$ as well. In particular, we require that there must be a strict or defeasible rule with head $q$ which can be applied (2.1). But now we need to consider possible "attacks", that is, reasoning chains in support of $\sim q$. To be more specific: to prove $q$ defeasibly we must show that $\sim q$ is not definitely provable (2.2). Also (2.3) we must consider the set of all rules which are not known to be inapplicable and which have head $\sim q$ (note that here we consider defeaters, too, whereas they could not be used to support

the conclusion $q$; this is in line with the motivation of defeaters given earlier). Essentially each such rule $s$ attacks the conclusion $q$. For $q$ to be provable, each such rule $s$ must be counterattacked by a rule $t$ with head $q$ with the following properties: (i) $t$ must be applicable at this point, and (ii) $t$ must be stronger than $s$. Thus each attack on the conclusion $q$ must be counterattacked by a stronger rule. In other words, $r$ and the rules $t$ form a team (for $q$) that defeats the rules $s$.

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

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

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

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

## 5 Defeasible Logic for Intentions

As we have seen in section 3 NMLs have been put forward to capture the intensional nature of mental attitudes such as, for example, intention. Usually modal logics are extensions of classical propositional logic with some intensional operators. Thus any classical (normal) modal logic should account for two components: (1) the underlying logical structure of the propositional base and (2) the logic behavior 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 in so far as it requires complete, consistent and reliable information. Hence any modal logic based on classical propositional logic is doomed to suffer from the same problems.

On the other hand the logic should specify how modalities can be introduced and manipulated. Some common rules for modalities are Necessitation and RM [4]. 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. Moreover an agent need not be intending all the consequences of a particular action it does. It might be the case that it is not confident of them being successful. Thus the two rules are not appropriate for a logic of intention.

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

Accordingly a defeasible intention theory is a structure $(F, R^K, R^I, >)$ where, as usual $F$ is a set of facts, $R^K$ is a set of rules for knowledge (i.e., $\rightarrow_K$, $\Rightarrow_K$, $\rightsquigarrow_K$), $R^I$ is a set of rules for intention (i.e., $\rightarrow_I$, $\Rightarrow_I$, $\rightsquigarrow_I$), and $>$, the superiority relation, is a binary relation over the set of rules (i.e., $> \subseteq (R^K \cup R^I)^2$).

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

In what follows we provide the appropriate inference rules for intentions, and we identify strong intentions – i.e., intentions for which there are no alternatives – using $\pm\Delta_I$; goals using $\pm\Sigma_I$, and intentions using $\pm\partial_I$.

In order to correctly capture the notion of intention we extend the signature of the logic with the modal operator INT; thus if $l$ is literal then INT$l$ and $\neg$INT$l$ are modal literals. However we impose some restrictions on the form of the rules: modal literals can only occur in the antecedents of rules for intention.

Derivability for knowledge ($\pm\Delta_K$, $\pm\partial_K$) has the same conditions as those given for derivability in Section 4. It is true that the complete and accurate definition of the inference conditions is cumbersome but the intuition is natural and easy to understand. The conditions for deriving an intention are as follows:

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

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

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

To prove that a strong intention $A$ does not hold ($-\Delta_I A$), first, $A$ should not be a basic intention (1); then we have to discard all possible reasons in favour of

it. If $A$ is a definite consequence of $B$, that is $B \rightarrow_K A \in R^K$, we can disprove it if we can show that (2.1) $B$ is not the case (i.e., $-\Delta_K B$) or (2.2) $B$ is not strongly intended (i.e., $-\Delta_I B$). In case of strict policies for $A$ (3), such as, for example the strict rule for intention $\mathrm{INT}B, C \rightarrow_I A$, we have to show that either $B$ is not strongly intended (3.1), or the fact triggering the policy is not the case (3.2).

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

$+\Sigma_I$: if $P(i+1) = +\Sigma_I p$ then
(1) $\mathrm{INT}p \in F$ or
(2) $\exists r \in R^K_s[p] \forall a \in A(r) : +\Sigma_I a \in P(1..i)$ or
(3) $\exists r \in R^I_s[p]$ such that
  (3.1) $\forall \mathrm{INT}a \in A(r) : +\Sigma_I a \in P(1..i)$ and
  (3.2) $\forall a \in A(r) : +\Sigma_K a \in P(1..i)$.

$-\Sigma_I$: if $P(i+1) = -\Sigma_I p$ then
(1) $\mathrm{INT}p \notin F$ and
(2) $\forall r \in R^K_s[p]$
  (2.1) $\exists a \in A(r) : -\Sigma_K a \in P(1..i)$ or
  (2.2) $\exists a \in A(r) : -\Sigma_I a \in P(1..i)$; and
(3) $\forall r \in R^I_s[p]$ either
  (3.1) $\exists \mathrm{INT}a \in A(r) : -\Sigma_I a \in P(1..i)$ or
  (3.2) $\exists a \in A(r) : -\Sigma_K a \in P(1..i)$.

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

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

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

$-\partial_I$: if $P(i+1) = -\partial_I p$ then
1) $-\Delta_I p \in P(1..i)$ and
2.1) $+\Delta_K \sim p$ or $+\Delta_I \sim p \in P(1..i)$ or
2.2) both
  .1) $\forall r \in R^K_{sd}[p]$ $\exists a \in A(r) : -\partial_K a \in P(1..i)$, and $\exists a \in A(r) : -\partial_I a \in P(1..i)$; and
  .2) $\forall r \in R^I_{sd}[p]$ $\exists \mathrm{INT}a \in A(s) : -\partial_I a \in P(1..i)$ or $\exists a \in A(s) : -\partial_K a \in P(1..i)$; or
2.3)
  .1) $\exists s \in R^K[\sim p]$ $\forall a \in A(s) : +\partial_K a$ or $\forall a \in A(s) : +\partial_I a$, or
    $\exists s \in R^K[\sim p]$ $\forall a \in A(s) : +\partial_K a$ and $\forall \mathrm{INT}a \in A(s) : +\partial_I a$; and
  .2) $\forall t \in R[p]$ either $t \not> s$ or
    if $t \in R^K[p]$ then $\exists a \in A(t) : -\partial_K a$ and $\exists b \in A(t) : -\partial_I b$; and
    if $t \in R^I[p]$ then $\exists a \in A(t) : -\partial_K a$ or $\exists \mathrm{INT}a \in A(t) : -\partial_I a$.

The conditions for proving defeasible intentions are essentially the same as those given for defeasible derivations in Section 4. The only difference is that at each stage we have to check for two cases, namely: (1) the rule used is a rule for an intention; (2) the rule is a rule for knowledge. In the first case we have to verify that factual antecedent are defeasibly proved/disproved using knowledge ($\pm\partial_K$), and intentional antecedent are defeasibly proved/disproved using intention ($\pm\partial_I$). In the second case we have to remember that a conclusion of a factual rule can be transformed in an intention if all the literals in the antecedent are defeasibly intended. The intuition behind the definition of $-\partial_I$ is a combination of the motivation for $-\partial$ and the intuition of $-\Delta_I$.

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

$$\neg rain \Rightarrow_I cricket \qquad \mathrm{INT}cricket \Rightarrow_I outdoor$$

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

If Vineet intend to travel to Italy then he intend to travel to Europe since Italy is in Europe. This argument can be formalized by the rule $Italy \rightarrow_K Europe$ plus the basic intention $\mathrm{INT}Italy$. The conclusion $+\Delta_I Europe$ follows from clause (2) of $+\Delta_I$.

Most of the BDI systems are able to express positive and negative introspection of belief and intentions. Those notions are encoded, respectively, by the following axioms.

$$\mathrm{INT}\phi \rightarrow \mathrm{BEL}(\mathrm{INT}\phi) \qquad \neg\mathrm{INT}\phi \rightarrow \mathrm{BEL}(\neg\mathrm{INT}\phi)$$

One of the main effect of positive (resp. negative) introspection is the ability of using established (resp. rejected) intentions in epistemic contexts to derive (resp. prevent the derivation of) other intentions. But this is what is done in Clause 2 of $+\Delta_I$, Clause 2.2.1 of $+\partial_I$, for positive introspection, and Clause 2.2 of $-\Delta_I$ and Clause 2.2.1 of $-\partial_I$ for negative introspection.

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

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

**Theorem 1.** *Let $\# = \Delta_K, \partial_K, \Sigma_K, \Delta_I, \partial_I, \Sigma_I$, and $D$ be a defeasible theory. There is no literal $p$ such that $D \vdash +\#p$ and $D \vdash -\#p$.*

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

**Theorem 2.** *Let $D$ be a defeasible theory, and $M \in \{K, I\}$. $D \vdash +\partial_M p$ and $D \vdash +\partial_M{\sim}p$ iff $D \vdash +\Delta_M p$ and $D \vdash +\Delta_M{\sim}p$.*

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

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

**Theorem 3.** *For every defeasible theory $D$, and $M \in \{K, I\}$*

$$1.\ \Delta^+_M \subseteq \partial^+_M \subseteq \Sigma^+_M; \qquad 2.\ \Sigma^-_M \subseteq \partial^-_M \subseteq \Delta^-_M.$$

This theorem states that strict intentions are intentions ($\Delta^+_I \subseteq \partial^+_I$), and intentions are goals ($\partial^+_I \subseteq \Sigma^+_I$), which corresponds to the BDI principle INT$\phi \rightarrow$ GOAL$\phi$. At the same time, we have that $\Delta^+_K \subseteq \partial^+_K$. Thus if we assume that $\Delta_K$ corresponds to knowledge and $\partial_K$ corresponds to belief we obtain KNOW$\phi \rightarrow$ BEL$\phi$, the standard BDI axiom relating the two epistemic notions.

The proposed theory of intention satisfies many of the properties outlined by Bratman in [8]. The role of intention as a *conduct-controlling* pro-attitude rather than *conduct-influencing* is clearly illustrated in the elaborate proof-theory outlined for the types of intention. The proposed theory supports the fact that the rationality of an agent for his intention depends on the rationality of the relevant processes leading to that intention where the relevant processes includes using superiority relations to resolve conflicts as well as satisfying the rules of inclusion as shown in Theorem 3. The new approach provides a good formalisation as to the relation between *guiding intention* and *intentional action* termed as *historical principle of policy-based rationality* in [8]. The problem in general is to account for the rationality of an agent in performing a particular policy-based intention from a general policy. In our approach the defeasibility of general policies makes it possible to block/not block the application of the policy to the particular case without abandoning the policy.

## 6 Conclusion and Discussion

Based on Bratman's classification of intention, we have outlined a *policy-based* theory of intention which differs from the usual NML-based approaches in the sense of having a non-monotonic nature. To capture the properties involved in such intentions we adopted *defeasible logic* as the non-monotonic reasoning mechanism due to its efficiency and easy implementation as well as the defeasible nature of policy-based intentions. The new approach alleviates most of the problems related to logical-omniscience. We pointed out that some of the problems related to intention re-consideration could be easily understood through such an approach.

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

The first is in alliance with the work done in [19, 12]. Here they outline a policy description language called *PDL* and use logic programs to reason about the policies. The main concern in that work is in tracing the *event* history that gives rise to an *action* history based on stable model semantics. In a similar manner our approach could be developed using the appropriate semantics (Kunen [14] or argumentation [9]) and developed from a logic programming point of view. The advantage in our approach is the use of the superiority relation ($>$) whereby we can mention a hierarchy between the rules and this is absent in other works.

The second direction in which our work could be extended is to define various rules required for constructing goals from beliefs, intentions from goals, intentions from beliefs etc. and giving a superiority relation among these rules. The recent work on BDI [21] seems to take this direction. On the other hand many new applications in emerging information technologies have advanced needs for managing relations such as authorization, trust and control among interacting agents (humans or artificial). This necessitates new models and mechanisms for structuring and flexible management of those relations. The issues of automated management of organisations in terms of policies and trust relations in highly dynamic and decentralised environments has become the focus in recent years.

Finally, as we have alluded to many semantics have been devised for defeasible logic and can be adapted straightforwardly to the extension proposed here. The method developed in [14] gives a set-theoretic fixed-point construction for $\Delta^+, \partial^+, \ldots$, which leads to a logic programming characterisation of defeasible logic. Programs corresponding to defeasible theories are sound and complete wrt Kunen semantics. The same technique is applicable in the present case with the obvious adjustments; however, it does not offer further insights on defeasible logic for BDI, because of the almost one-to-one correspondence between the inference conditions and the steps of the fixed-point construction. However semantics for defeasible BDI logic remains an interesting technical problem.

## References


[1] G. Antoniou, D. Billington, G. Governatori, and M. Maher. A flexible framework for defeasible logics. In *AAAI'2000*, pages 401–405. AAAI/MIT Press, 2000. 418

[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, April 2001. 418

[3] D. Billington. Defeasible logic is stable. *Journal of Logic and Computation*, 3:370–400, 1993. 418

[4] B. F. Chellas. *Modal Logic, An Introduction*. Cambridge University Press, Cambridge, 1980. 414, 420

[5] X. Chen and G. Liu. A logic of intention. In *ICJAI'99*, 1999. 414

[6] P. R. Cohen and H. J. Levesque. Persistence, intention and commitment. In *In proceedings Timberline workshop on Reasoning about plans and actions*, pages 297–338, 1986. 414

[7] P. R. Cohen and H. J. Levesque. Intention is choice with commitment. *Artificial Intelligence*, 42(3), 1990. 414

[8] M. E. Bratman. *Intentions, Plans and Practical Reason*. Harvard University Press, Cambridge, MA, 1987. 414, 424

[9] G. Governatori and M. J. Maher. An argumentation-theoretic characterisation of defeasible logic. In *ECAI-2000*, pages 469–473, 2000. 418, 424

[10] J. Hintikka. *Knowledge and Belief*. Cornell University Press, 1962. 414

[11] M. E. Pollock and K. Konolige. A representationalist theory of intention. In *IJCAI-93*, pages 390–395, 1993. 414

[12] J. Lobo, R. Bhatia, and S. Naqvi. A policy description language. In *AAAI-99*. AAAI/MIT Press, 1999. 424

[13] M. J. Maher. Propositional defeasible logic has linear complexity. *Theory and Practice of Logic Programming*, 1(6):691–711, 2001. 414, 418
[14] M. J. Maher and G. Governatori. A semantic decomposition of defeasible logic. In *AAAI-99*, 1999. 424, 425
[15] M. J. Maher, A. Rock, G. Antoniou, D. Billignton, and T. Miller. Efficient defeasible reasoning systems. *International Journal of Artificial Intelligence Tools*, 10(4), 2001. 415, 418
[16] D. Nute. Defeasible logic. In *Handbook of Logic in Artificial Intelligence and Logic Programming*, volume 3, pages 353–395. Oxford University Press, 1987. 418
[17] A. S. Rao and M. P.Georgeff M. P. Modelling rational agents within a BDI-architecture. In *KR'91*, pages 473–484. Morgan Kaufmann, 1991. 414, 415
[18] M. P. Singh. Semantical considerations on intention dynamics for BDI agents. *Journal of Experimental and Theoretical Artificial Intelligence*, 1998. 414
[19] T. C. Son and J. Lobo. Reasoning about policies using logic programa. AAAI-spring symposium on answer set programming, March 26-28 2001. 424
[20] T. Sugimoto. A preference-based theory of intention. In *PRICAI-2000*, Springer-Verlag, 2000. 414, 417
[21] J. Thanagrajah, L. Padgham and J. Harland. Representation and reasoning for goals in BDI agents. In *Australasian Conference on Computer Science*, 2002. 425
[22] B. Van Linder. *Modal Logic for Rational Agents*. PhD thesis, Department of Computer Science, Utrecht University, 19th June 1996. 414, 416
[23] R. Zamparelli. Intentions are plans plus wishes (and more). In *AAAI Spring symposium-93*, 1993. 414