ARE INTENTIONAL VERBS CLOSED UNDER IMPLICATION ?

Paul GOCHET
E-mail : pgochet@ulg.ac.be

26 novembre 2009

Chapitre 1

From the B.D.I. architecture to STIT logic

Introduction

Most studies devoted to intentional logic deal with the failure of the usual rules of the *logic of identity* when applied within the scope of verbs of propositional attitude or with the analogous failure of *rules governing quantifiers* [see Gochet and Gribomont, 2006]. In this essay I shall focus on a third peculiarity of intentional logic which is shared by intentional logic and by the logic of action known as *stit*-logic : *the lack of closure under implications of various kinds*. That problem has aroused much interest in the A.I. and Computer science communities. In the first chapter I will examine some of the more influential papers in this area. In the second chapter I will present and analyze the contribution of a logician-philosopher to the same subject. I will then draw some conclusions on the philosophical significance of those contributions to the logic of intentionality.

I will start by enumerating six inferences in natural language for which the question arises whether they exemplify a failure of closure under logical consequence or not. Next I will survey the formal semantics and axiomatic systems which have been put forward to give a rigorous treatment to reasonings in which intentional terms *occur essentially* and to deal with inferences like the six inferences below :

(1) Having one's tooth filled by the dentist is believed to entail suffering pain.
Agent α aims at having its tooth filled by the dentist.
Therefore agent α aims at suffering pain. [Rao and Georgeff]

(2) If α dies painlessly, α dies.
α wants to die painlessly.
Therefore α wants to die.

(3) If the police identifies and captures the culprit, the police identifies the culprit.

The police wants to identify and capture the culprit.

Therefore the police wants to identify the culprit.

(4) There is at least one injured man who is bandaged entails that there is at least one injured man.

You see to it that there is at least one injured man who is bandaged.

Therefore you see to it that there is at least one injured man. [Belanp & Horty]

(5) If Alphonse is in Alabama and Betty buys a brick then Alphone is in Alabama.

I see to it that both Alphonse is in Alabama and Betty buys a brick.

Therefore I see to it that Alphonse is in Alabama. [adapted from Chellas]

(6) An agent intentionally defends itself.

It does not know another way of defending itself than by killing the attacker.

The agent intentionally sees to it that the attacker is killed. [Broersen]

Inferences (1), (4) and (6) are clearly invalid. Inferences (3) and (5) are clearly valid, inference (2) is dubious. It is expected that a good formal semantics will explain our logical intuitions.

1.1 The Belief-Desire-Intention architecture

In "Advice on Modal logic" Dana Scott addressed his famous warning to modal logicians : "Here is what I consider one of the biggest mistakes of all in modal logic : concentration on a system with just one modal operator. The only way to have any philosophically significant results in deontic logic or epistemic logic is to combine those operators with : tense operators (otherwise how can you formulate principles of change ?) ; the logical operators (otherwise how can you compare the relative with the absolute ?) ; operators like historical or physical necessity (otherwise how can you relate the agent to his environment ?) ; and so on and so on [Scott, 1970, 161]".

Dana Scott's warning has been taken seriously. Multi-modal logic has become an area of intensive research in several communities of logicians. In 1985. Robert Moore worked out the first integrated *Formal Theory of Knowledge and Action*. R.Moore's theory is not, however, a modal logic. It is a many-sorted first-order theory in which the concepts of the metatheory of modal logic such as "possible worlds" are introduced into the object-language. Yet Moore meets at least one of the main demands expressed by D.Scott, namely the integration of notions which were previously studied in isolation : knowledge and action [Moore, 1985, reedited 1995, see Gochet 2007].

In a very influential paper, Philip R.Cohen and Hector J.Levesque have addressed the problem raised in the introduction. They claimed that neither *goals*

nor *intentions* are closed under *logical consequence*. Considering persistent goals (P-GOALS) i.e. goals that the agent will not give up until it has been satisfied or until he thinks they will never be true, Cohen and Levesque state that P-GOALS are closed under *logical equivalence* only, hence they implictly admit that goals are not closed under the weaker relation of *logical implication* (logical consequence) and *a fortiori* under even weaker relations such as *believed implication* or *strict implication* [Cohen & Levesque, 1990, 237].

They envisage six degrees of increasing strength :

$$
\begin{array}{ll}
1 & p \supset q \\
2 & (BELx(p \supset q)) \\
3 & (BELx\Box(p \supset q)) \\
4 & \Box(BELx\Box(p \supset q)) \\
5 & \models p \supset q \\
6 & \models p = q.
\end{array}
$$

To support the claim that *intentions* are not closed under logical consequence, they provide an example which has become part of the folklore of the suject. let me quote it in full : "[. . .] an agent intended to have his teeth filled. Not knowing about anaesthetic (one could assume this took place just as they were being first used in dentistry), he believed that it was always the case that if one's teeth are filled, one will feel pain. One could even say that surely the agent *chose* to undergo pain. Nontheless, one would not like to say that he intended to undergo pain [Cohen & Levesque, 1990, 251]". It will be shown later that Cohen and Levesque's claim has to be mitigated. Intentions (and goals) do not exemplify failure under *logical consequence*. They exemplify failure under weaker forms of implications such as *believed implications* or *causal implications*.

Cohen and Levesque's formalism is first-order logic enriched with modal operators. Beliefs, goals and intentions are captured by *predicate constants*. In "Modeling Rational Agents within a BDI architecture" published in 1991, Rao and Georgeff came nearer to Dana Scott's concept of multi-modal logic. They represent beliefs, goals and intentions by *modal operators*.

Rao and Georgeff adopt the propositional branching time logic used to reason about programs (CTL) and work out a first-order extension of that logic. Besides operators such as $F\varphi$ (*sometimes in the future*) and $G\varphi$ (*always in the future*), CTL contains operators which look like quantifiers, namely A (*on all paths*) or E (*on some paths*). Events (primitive events) are represented by nodes on a tree. Time is interpreted as a binary relation which is total, transitive and backward linear. This makes it possible to enforce a single past and a branching future. Possible worlds are assumed to be time trees.

Next, as I have already said, Rao and Georgeff enrich their language by introducting modal operators designed to capture respectively beliefs, goals and intentions. For the interpretation of these operators, they fall back on possible worlds semantics [see Blackburn *et al.* 2001]. The interpretation of belief operator

(B) is standard : a formula φ is said to be believed if it is true in all the worlds reachable by the belief-accessibility relation **B**. As usual, this relation is taken to be transitive, Euclidean and serial. It is axiomatically captured by the system KD45. The goal operator (G) is interpreted in the model by an accessiblity relation **G** which is serial. This ensures that goals are consistent, as opposed to desires which can fail to be so. Seriality is captured by the axiom D. The accessibility relation **I** which interprets the operator I [for intentions] is also serial and nothing else. This does not mean, of course, that thc distinction between goals and intentions is lost. Two relations may differ (be made up of different n-tuples) even if they obey the same constraint (here seriality). All this is fairly traditional. The main contribution of the authors lies elsewhere, namely in the account they give of the *interrelations between beliefs, goals and intentions.*

In their search for a logic designed to capture the relationship between belief, goal and intention, Rao and Georgeff first spell out intuitive axioms. Next they build a semantic interpretation which makes them valid. Lastly they state an axiomatic system which they show to be sound. They did not care about completeness. A full axiomatization was provided later indirectly. In 2000, Klaus Schild filled the gap and obtained an important result : the basic BDI theory can be captured within a standard logic of concurrency by relying upon Kozen's propositional μ calculus [Schild 2000].

Among the intuitive axioms adopted by the author, let me focus on the simplest one namely (1) GOAL (α) $\supset$ BEL (α) where α is a formula of the form "Optional ψ". A formula of the form "Optional ψ" is true if and only if there is at least one *full path* of posssible worlds which makes it true. By "full path in a world w" they mean that there is an *infinite sequence* of time points such that for all $i, (t_i, t_{i+1}) \in A_w$ where A_w is the restriction of the time relation to the set of time points of the world w [Meyer & Veltman 2007, 1002].

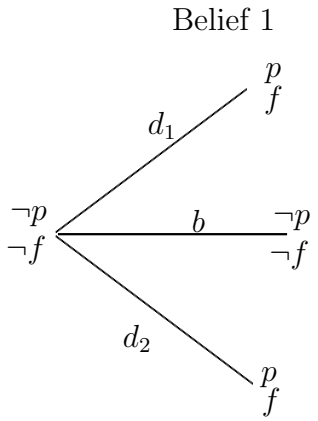
Axiom (1) states that if the agent has the goal that "Optional ψ" is true, there is at least one path in the worlds reacheable by the accessibility relation of belief (*belief-accessible worlds*) in which it is true. In other words *the goal must be compatible with the beliefs* of the agent.

To enforce this notion of compatibility, the authors require that, for each belief-accessible world w at a given moment in time t, there must be a goal-accessible world that is a *sub-world* of w at time t (or better a *subtree*), formally : $\forall w' \in B\ \exists w'' \in G$ such that w'' is a sub-world of w'. The notion of subworld can be defined in this way : "Intuitively, a world w is said to be a subworld of world w' if w has the same structure as w' but has fewer paths *and is otherwise identical.* Formally, if w, w' are worlds, then w is a subworld of w' (written $\sqsubseteq$ iff paths (w) $\subseteq$ paths (w') but w, w' agree on the interpretation of predicates and constants in common time points [Wooldridge 2000, 93]".

Suppose there is just one belief -accessible world called $b1$. Like all possible worlds, a belief accessible world is a tree. Let $b1$ be a tree with a root decorated

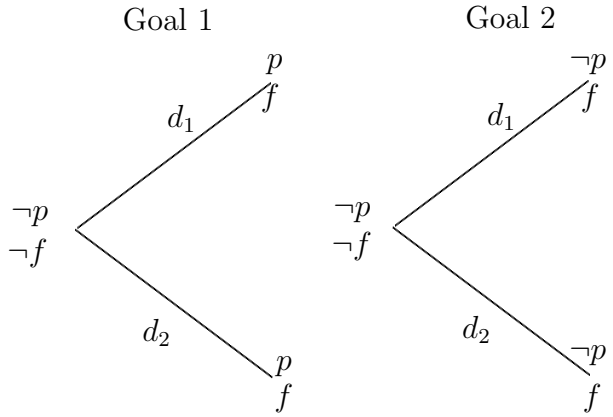
with the label $< \neg p \neg f >$ which describes the state of affairs "No pain, no tooth-filling". From the root three branches emanate, each of which leads to a new state of affairs. Each of these three edges is decorated by the name of the action which leads to the new state of affairs. The first action is "going to dentist 1". The resulting state of affairs is a node decorated with the label $< p, f >$ i.e. "pain, tooth filling". The second alternative action is "go shopping". The resulting state of affairs is a node decorated with the same label as the root since nothing changes for the patient. The third alternative action is "going to dentist 2". The result ia a node decorated with the label "pain, tooth filling" again.

See Figure 1.



Now suppose that that there are two goal-accessible worlds : $g1$ and $g2$. Goal possible world $g1$ is a tree made up of two branches : respectively the first and the third edge of the previous tree. Goal-accessible world $g1$ is a *sub*-tree of the belief–accessible world $b1$.

See Figure 2.



The converse of $\forall w' \in B \; \exists w'' \in G$ such that w'' is a sub-world of w' need not hold, there are goals which are not compatible with the beliefs of the agents. This makes it possible to construct a model in which goals are not closed under the beliefs of the agent.

To construct such a model, we have to define goal-accessible world $g2$ mentioned above as follows : $g2$ is a tree with two branches like $g1$. It has the same root as $g1$ and its edges are decorated respectively by action of going to dentist 1 and by the action of going to dentist 2. Unlike $g1$ however, the terminal nodes of $g2$ are both decorated with $< f, \neg p >$, i.e. "having tooth filled, with no pain". Hence unlike $g1$, $g2$ is not a sub-world of $b1$. In other words, goal $g2$ is a goal of the agent which is not compatible with the agent's belief.

We want to show that goals need not be closed under the beliefs of the agent. [In this case the belief is the belief that an implication holds.] For that purpose let us show that the formula below is not valid :

$$\text{GOAL } (\varphi) \wedge \text{BEL } (\text{inevitable } \varphi \supset \gamma) \supset \text{GOAL } (\gamma).$$

This amounts to showing that this formula is satisfiable :

$$\text{GOAL } (\varphi) \wedge \text{BEL } (\text{inevitable } \varphi \supset \gamma) \wedge \neg\text{GOAL } (\gamma).$$

If we substitute "have one's tooth filled" for φ and "suffer pain" for γ, the model described above shows that the second formula is satisfiable. In all belief-accessible worlds, namely in $b1$, since there is only one belief-accessible world in Rao and Georgeff's model, having one's tooth filled is believed to inevitably entail suffering pain. Clearly Rao and Georgeff do not prove that goals (or intentions) are not closed under *logical consequence*. They prove that they are not closed under *implications that are believed to be physically necessary* [relation 3 in Cohen and Levesque's hierarchy mentioned above].

Let us now consider goal-accessible worlds. In all goal-accessible worlds, namely $g1$ and $g2$, having one's tooth filled is a goal of the agent. But *in at least one goal-accessible* world ($g2$) having one's tooth filled is not associated with suffering pain. This happy situation occurs in the goal-accessible world $g2$ which is not a subworld of a belief world, i.e. in a world that the agent does not believe to be accessible.

Rao and Georgeff have solved one of the problems raised in the Introduction. They have provided a semantics which explains why inference (1) is not valid. They proceed in an analogous manner for intentions. They explain the failure of closure of intentions under *believed implication* displayed in Cohen and Levesque's famous exemple of the patient who did not know about anaesthetic.

One more step is needed to reach action. Intention *leads* to action. The authors formulate the relation between intention and action by the axiom :

(3) INTEND(does (e)) $\supset$ does (e).

The authors cautiously observe that the agant will not always succeed. Success in the execution of the action depends not only on the agent but also on the environment. If the agent intends to do something, it will at least *try* to do it. Recent researches have been made on the logic of intention and attempt [Emiliano Lorini and Andreas Herzig 2008].

Some actions are performed intentionally but not all. Hence it is worth studying a more general and more basic concept of action which leaves intention aside. The logic of action in this basic and idealized sense has been a lively field of research for several decades under the name of "*stit* logic" (the logic of "see to it that"). Very recently (in 2008), several authors have succeeded in combining BDI logic with *Stit* logic. Before describing and commenting on their achievements, it is necessary to make a brief presentation of *Stit* logic. That logic will settle the status of inference (4) of the Introduction.

1.2 Stit logic

To study the concept of action, one might start by examining action verbs as they are used in natural language. But, as Belnap, Perloff and Xu observe, natural language is not a safe guide. Some action verbs do not really denote an action. Consider the following two examples given by the authors :

(1) Ahab sailed in search of Moby Dick.

(2) The *Pequod* sailed in search of Moby Dick.

These two sentences involve the same action verb ("sailed"), but in (2) the verb is used in a metonymic way. The action of sailing cannot be ascribed to the ship but only to Captain Ahab and his crew.The sailing is not the result of a choice the ship made among alternatives open to her.

To isolate action-like verbs in which agency is *really*, as opposed to *apparently*, ascribed, a test has been invented, namely a way of paraphrasing action verbs with the expression "see to it that". Belnap and his co-authors describe this method as "...an attempt to isolate, by way of a canonical form, a particular set of English sentences in order to study more closely how they interact with each other and with other parts of language in different linguistic environments [Belnap, Perloff, Xu, 2001, 7,8]". This way of paraphrasing has nothing to do with the regimentation of a piece of natural language into a canonical notation which Quine invented in *Word and Object* [Quine 1960]. It should rather be seen as a kind of experimentation with natural language.

Compare these sentences :

(1) Ahab *stit* : Ahab searched the white whale.

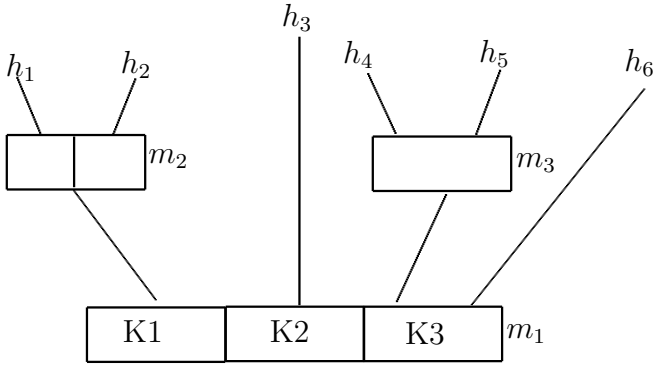
(2) Ahab *stit* :Ahab found the white whale.

The test reveals that the second is not an agentive sentence. Finding the white whale as opposed to searching it was beyond the control of Ahab.

Stit-theorists conceive action as constraining the future by making a choice among different future courses of events. Hence the conception of time which fits *stit* theory is that of moments ordered in treelike structure with forward branching representing the indeterminacy of the future and linearity of the past, its determinacy and irrevocability.

The branches are called *histories*. The truth of a sentence is evaluated relatively to a pair m/h where m denotes a moment belonging to the history h.

Only one history is realized. Not all indeterminacy is resolved by agents. There are cases where several histories remain open even after the agent has made his or her choice at a given moment. These histories are *equivalent* at that moment, *relatively to his or her choice*, and can thus be said to belong to the same *equivalence class*. This will be made clear by Figure 3.



In Figure 3, the agent in moment m_1 has to choose between three equivalence classes of histories : $K_1 = \{h_1, h_2\}, K_2 = \{h_3\}, K_3 = \{h_4, h_5 h_6\}$. Observe that in moment m_2, a choice between two histories (h_1 and h_2) is open to the agent but not in m_3 where the choice between histories h_4 and h_5 is left to Nature or to another agent. When the agent choses $K3$, three histories may occur : h_4, h_5 and h_6.

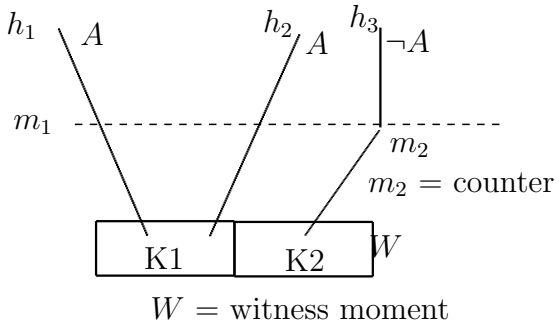
The course of history is isomorphic to the time arrow. We can draw a line across all the histories and call it "instant".

The whole frame needed to provide the operator *stit* with an interpretation is the n-tuple : $< \text{Tree}, \leq, I, C, A >$ where I denotes instants, C stands for choice function. A is the set of agents.

The choice function C represents "the constraints that an individual is able to exercise upon the course of history at a given moment, the acts or choices open to him at that moment» [Horty & Belnap 1995,388]. In Figure 4, function C associates three equivalence classes of histories $K1, K2$ and $K3$ with moment m_1 between which the agent can choose.

To complete the semantics of *stit*, it remains to turn the frame into a model by supplying an interpretation function and an evaluation rules which states the truth conditions of *stit*-formulas.

Before stating it, it is useful to look at Figure 4.



At the so-called *witness-moment*, the agent can choose between $K1$ and $K2$. After $K1$ has been chosen, two different courses of events may occur, as indicated by the fork V but this second choice is not made by the agent. In these two different histoires which are equivalent relatively to the agent's choice, the proposition A is true (h_1 describes the history which happens to be the real one). If the agent α chooses $K2$, he or she brings about history h_3 where A is false at moment m_2, called "counter". The moments m_1 and m_2 are *posterior* to the *witness moment* w. The fact that a different state of affairs occurs (A or $\neg A$), depending whether the agent chooses $K1$ or $K2$, shows that the agent's choice determines the course of events (at least to a certain extent : the agent does not determine which of histories h_1 or h_2 will become real).

The evaluation rule which provides an interpretation for the logical constant *astit* [i.e. sees to it that in the *achievement sense* of "sees to it that"] embodies two requirements which are depicted in Figure 4 : "The positive requirement is that, as a result of a prior choice by α at the witnessing moment w, things have evolved in such a way that A is now at the instant of m to be true [...]. The negative requirement is that it was not yet settled at w that A should now (at $i(m)$) be true, so that α's action at w did have some real effect in bringing about the present truth of A. [Belnap *et al.* 2001, 36-37].

Let us now examine inference (4) quoted in the Introduction which is used by Belnap and his co-authors to show that *astit* formulas are not closed under logical consequence.

Consider the statement $A \wedge B$ [There is an injured man (B) who he is bandaged (A)] and the statement which logically follows : B [There is an injured man]. The authors want to *explain* why "*You see to it that* there is at least one injured man who is bandaged" does not entail "*You see to it that* there is at least one injured man". Here for the first time, we encounter a genuine failure of closure under logical consequence. To see why the inference breaks down let us look at te model of agency for *astit*-logic built up by the authors (Figure 5) :

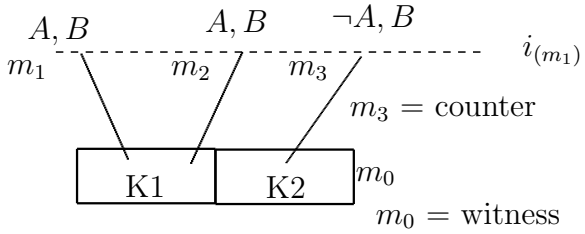


Figure 5 is designed to show that although $A \wedge B$ entails B, it is not the case that $[\alpha : \ stit\ A \wedge B]$ entail $[a : \ stit\ B]$. The reason why it does not lies in the fact that B was already settled at the beginning, i.e. at the witness moment w. Hence, as the authors stress, the negative condition for the truth of $[a : \ stit\ B]$ is violated.

The picture depicts a case where, before the action takes place there is an injured man, but after the action there is an injured man who is bandaged. The agent's choice is decisive. Had he or she chosen $K2$ instead of $K1$, the injured man would not have been bandaged. But, of course, he would have remained injured.

It is crucial to see that there are two temporal stages. At the initial stage the man is injured but, implicitly, not bandaged. At the next stage, the injured man *becomes* bandaged. The operator *astit*(*achievement stit*) precisely captures the kind of action which results in a state of affairs *becoming* true *in virtue of the choice of the agent*.

The formal semantics worked out by Belnap at *al. explains* the lack of closure under logical consequence exemplified by inference (4) in the Introduction. Let us observe that the model provided by the authors is not an *ad hoc* construction produced for the purpose of explaining our logical intuitions about (4). The model of *astit*-logic exhibits a structure which has been neglected by philosophers exclusively concerned with ontology : *the structure of agency*. Belnap and his co-authors bring that out in this passage : "There is not the slightest paradox in saying, there is neither "funny logic" nor grammatical subtlety required in calculating that from the fact that you see to it that there is at least one injured man who is bandaged, it does not follow that you see to it that there is at least one injured man, even though that there is at least one injured man who is bandaged implies that there is at least one injured man. To the contrary, it is deeply built in the real-choice-based idea of agency that such cases should be typical [Belnap et *al. Ibid.* 40]».

1.3 A solution of the problems raised by inferenes (2), (3), (5)

Inferences (3) and (5) do not raise any problem. They are no exceptions to clossure under logical consequence. Inference (2) however might seem to be invalid. From "If A dies painlessly, A dies", and "A wants to die painlessly", some are reluctant to conclude "A wants to die". They should not.

Let us insert "now" after all the occurrences of "dies" and "dies painlessly". The initial intuition of invalidity vanishes completly. This shows that in the absence of the adverb "now", readers skeptical about the validity of (2) fall prey to a subtle equivocation. They implictly assign different meanings to "dies" in "wants to die painlessly" and "wants to die". They interpret the first sentence

as «A wants to die painlessly when the time of death has come" and the second sentence as "A wants to die now".

At this stage, it should be stressed that it is of crucial importance to distinguish*logical implication, believed implication* and what, for lack of a better word, I shall call *causal implication*. Closure under implication may hold for one of these implications without holding for the others.

The following example due to Frank Veltman raises an interesting problem for the formal semantics sketched above (I thank. Joost Joosten for bringing Veltman's example to bear on the problem of lack of closure under logical consequence). The formula below is a tautology, a formula true for all uniform interpretations of the propositional variables.

$$(p \supset r) \supset ((p \wedge q) \supset r).$$

Yet the following instance is plainly false :

If I am given coffee, I am pleased, then if I am given coffee and oil in it I am pleased.

Confronted with this question Hans van Ditmarsch replied that the sentence in natural language is not a substitution instance of the formula. The *prepositional constructions* "coffee and oil in it" or "coffee with oil" cannot be rendered by the truth-functional connective $\wedge$.

That reply settles the question. It is worth observing however that there are apparently admissible substitution instances of the formula which make it false although the antecedent of the last conditional is interpreted as a truth-functional conjunction. Consider this sentence :

If I swallow poison, I die, then if I swallow poison and I swallow counterpoison I die.

The reason why this sentence is false is again to be found in its failing to be a proper substitution of the tautological formula. The latter sentence in natural language is not an admissible substitution of the formula, the contrary appearance notwithstanding. The source of the problem is this : the first and the third conditionals which occur in it are not *material implications* (truth functional implications) but *causal implications*. But causal implications cannot be rendered by the truth-functional connective $\supset$. Causal implication violates a property of material implication : causal implication is non monotonic.

Let us take stock. Over the six problematic or seemingly problematic inferences listed in the Introduction, five have been dealt with. One inference remains to be accounted for, namely inference (6). This will be done in Section 5. In Section 4 I shall examine Semmling and Wansing's work on the combination of the BDI logic with *dstit* logic.

1.4 Semmling and Wansing's combination of BDI and STIT

Before we examine Semmling and Wansing's contribution, we need to say what *dstit*-logic is. The name "dstit logic" is a contraction of "deliberative see to it logic". As opposed to *astit*-formulas whose truth depends on two separate moments : the moment at which the *astit*-formula and the outcome are evaluated and the prior moment at which the choice or action which guarantees the outcome was made, the deliberative *stit* is referred to only a single moment and, as Belnap *et al.* observe, a formula of the form $[\alpha\ dstit : A]$, which means "agent deliberately sees to it that A", is evaluated at the moment of choice and action. Nothing more needs to be said here about *dstit*-logic. The *stit*-logic which the two authors combine with BDI is *dstit*-logic. [On *dstit*-logic see Wansing 2006 for a semantic tableau proof method].

In "FROM BDI AND *stit* TO $bdi - stit$ LOGIC" Caroline Semmling and Heinrich Wansing carry further than Rao and Georgeff the move from *uni-modal* to *multi-modal* logic and succeed in combining BDI-logic with *stit*-logic. A first move toward bringing together epistemic logic and *stit*-logic had been made by Andreas Herzig and Nicolas Troquard in "Knowing How to Play. Uniform Choices in the Logic of Agency" [Herzig and Troquard 2006].

Before we examine *bdi-stit* logic, it is worth mentioning an original contribution made by Semmling and Wansing in their paper to the formal semantics of belief, desire and intention. They observe an important difference between *desires* and *beliefs* on the one hand and *intentions* on the other which is ignored by the BDI architecture. Although beliefs and desires cannot be inconsistent, beliefs entertained by an agent can conflict with one another and desires also. Intentions, on the contrary are both consistent and not conflicting.

Before describing the formal tools needed to capture the notion of conflicting beliefs and conflicting desires, it is worth looking at the concrete examples offered by the authors. Consider a man who envisages donating one of his kidneys to his brother. He may have antagonstic desires about the matter. He desires to save his brother's life but he also desires to preserve his bodily integrity although the two desires are in conflict. There is no similar conflict at the level of intentions. Once the man under consideration has committed himself to fulfilling one of the two desires, i.e. when he has formed an intention, there is no room any more for conflicting intentions.

Conflicting beliefs or conflicting desires differ from inconsistent beliefs and inconsistent desires. The man who oscillates between the desire of donating and his desire of not denoting one of his kidneys to his brother both *desires* to donate and *desires not* to donate. If he is a rational agent however he does not desire the contradictory action of *donating and not donating* one kidney to his brother. How can we capture the logical difference between "desiring to do p and desiring not

to do p" and "desiring to do p and not p" ?

We need a logic (proof-theory and formal semantics) in which the following law of modal logic fails :

$$(\Box p \wedge \Box q) \supset \Box(p \wedge q)$$

where the square stands for "necessarily", "believes" or "desires". The neighbourhood semantics introduced by Montague and Scott provides the technical tools to construct a model in which the above formula is false [Chellas 1980, 210].

A model in neighbourhood semantics is a triple $< W, N, V >$ where

- W is a set of possible worlds : $\alpha, \beta, \chi \ldots$
- N is a function which assigns the set of propositions that are necessary in α to each possible world α. Given that propositions are sets of possible worlds, we have that :
$$N : W \mapsto \wp(\wp(W)).$$
- V is a valuation function which for each propositional variable q assigns the set of possible worlds in which q is true.

[I borrow this presentation of neighbourhood semantics together with a counter-model from Marina Straetmans' M.A. thesis University of Liège 1991].

Let α be a world of neighbourhood semantics.

$M, \alpha \models \Box A$ if and only if the set of worlds α which make A true is member of N_α (which itself is a set of propositions which are necessary in α).

Let us now built a model which falsifies $(\Box p \wedge \Box q) \supset \Box(p \wedge q)$.

- $W = \{\alpha, \beta\}(\alpha$ and β distinct),
- $N\alpha = \{\{\alpha\}, \{\beta\}\}$,
- $P(p) = \{\alpha\}$,
- $P(q) = \{\beta\}$,
- $\{\alpha\} \in N_\alpha$ and $\{\beta\} \in N_\alpha$.
- Hence $\Box p$ and $\Box q$ are in N_α hence. $\Box p \wedge \Box q$ is true in α.
- But $\{\alpha\} \cap \{\beta\} = \emptyset$,
- $\emptyset \notin N_\alpha$.
- Hence $p \wedge q$ is not in N_α hence $\Box(p \wedge q)$ is not true in α.

Semmling and Wansing who want to allow for the possibility of conflicing beliefs and desires and for the impossibility of conflicting intentions took up a *neighbourhood semantics* for beliefs and desires and a *relational semantics* for intentions. With this composite semantics they can account for the fact that the following formulas are satisfiable :

$$\alpha \; bel : \varphi \text{ and } \alpha \; bel : \neg\varphi,$$

$$\alpha \; des : \varphi \text{ and } \alpha \; des : \neg\varphi.$$

On the contrary, the following formula is not satisfiable :

$$\alpha \; int : \varphi \text{ and } \alpha \; int : \neg\varphi.$$

Allowing for conflicting beliefs and desires and discarding conflicting intentions enables Semmling and Wansing to account for the relationships between desires, intentions and seeing to it that which would have been impossible with the former logics available. For instance situations of everyday life can now be formally represented :

Agent α does not desire to donate one kidney but, out of the sense of duty, intends do do it.

Agent α does not desire to donate one kidney but, nevertheless, he or she sees to it that he or she donates.

The combined *bdi* and *dsit* logic is more expressive than *bdi* logic or *dstit* logic taken separately. As the authors note, "[...] we want to be able to express that an agent believes that a certain agent sees to it that something is the case [Semmling and Wansing 2008, 196]. We are now equipped with a rich and refined formal language in which complex philosophical problems can be expressed and tackled. Moreover the above-mentioned authors provide a sound and complete axiomatic system for $bdi - stit$ logic [Semmling and Wansing, 2009].

1.5 Broersen's account of intention and the treatment of inference (6)

Jan Broersen has studied the interaction between *dstit*-logic and an operator for propositional attitude, the operator K for knowledge. He set up a complete STIT logic for knowledge and action which holds for the multi-agent case. This new system leads to an improved definition of the *dstit*-operator "Agent α sees to it that φ in the next immediate state (X)", formally : $[\alpha\ xstit]\varphi$. Relying on the operator K which belongs to his language, Broersen can expresss the fact that the agent who does something deliberately is *aware* that he or she does it and is also aware that the outcome would have been different if he or she had not acted. Moreover he can also formally express the adverbial construction "*a knowingly sees to it that* φ".

In a subsequent paper ("First Steps in the stit-Logic Analysis of Intentional Action"), Broersen proposes a new account of intention which helps us take a stance on the validity of inference (6) of the Introduction and identifies a new kind of failure of closure under implication, namely *failure of closure under causal implication.*

Contrary to Cohen and Levesque who treat intention as a *mental state*, Broersen construes it as a *mode of action.* The significance of this shift of category is of major importance. If intentions are modes of actions, we should not be surprised that they fail to be closed under causal implication. It is typical of actions not to be closed under side-effects Hence Broersen's account of intention settles the question of the validity of inference (6) – it is not valid – and provides an expla-

nation for that verdict, an explanation which could not have been given if he had stuck to Cohen and Levesque's conception of intention.

Let us see in details the formal language and semantics which fit the new conception of intention advocated by Broersen. Broersen represents action by the *stit*-operator : $[\alpha\ xstit]\varphi$.

Intention, being a kind of action, is represented by an operator very similar to the *stit*-operator : $[\alpha\ xint]\varphi$. Since the formal language defined by Broersen contains the knowledge operator $K\varphi$, it is easy to express "knowingly doing" by combining the third and the first operator : $K[\alpha\ xstit]\varphi$.

To give a rigorous meaning to these new operators, a semantics has to be spelled out. Broersen constructs the following model [slightly simplified here as we only consider single agents] :

$$< S \times H, R_{\Box}, R_a, I_a, \sim, \delta >$$

in which the first item is the Cartesian product of states and histoires followed by four accessibility relations and an interpretation function.

(1) relations of historical necessity : $R_{\Box}$,

(2) 'effectivity' relations which associate with each agent the set of outcomes for which the agent can force the outcome to come into existence R_a ,

(3) intentional effectivity relations : I_a ,

(4) an equivalence relation : $\sim$ to capture knowledge.

Evaluations rules are spelled out which give the truth-conditions of the basic constructions. For instance, the evaluation rule below provides the truth conditions for the intention operator.

$M, < s, h > \models [\alpha\ xint]\varphi$ iff φ is true in all pairs $< s', h' >$ such that $< s, h > I_a < s', h' >$.

Let us note that if the fourth relation is constrained by the standard properties of transitivity, symmetry and reflexivity, the first tree relations are constrained by several new first-order properties.

Next a system of axioms sound and complete for the above semantics is proposed. One of these axioms stipulates that if agent α *intends to do*, then agent α *does* φ knowingly, but not conversely.

With that apparatus, the dentist's puzzle can finally be solved.

(1) The agent intentionally sees to it that he visits the dentist.

(2) the agent knowingly sees to it that he visit the dentist in a way that causes him pain.

(3) the agent does not know a way of visiting the dentist without having pain.

From this, Broersen observes, we can deduce that :

(4) the agent kowingly sees to it that he has pain.

But we cannot deduce :

(5) the agent intentionally sees to it that he has pain. This can be shown by constructing a counter-model within the framework of the semantics described above.

In other words, "knowingly sees to it that" is closed under *logical consequence* and/or *causal implication* but "intentionally sees to it" is closed only under *logical consequence*.

Up to now, we have surveyed various semantics designed to deal with particular intentional constants. Graham Priest has proposed a general semantics for intentionality. In the next and much shorter chapter, his contribution will be briefly described and discussed.

Conclusion of Chapter 1

The results obtained so far can be summed up in this statement : desires, goals and intentions are closed under *logical consequence* but not under *believed implication* or *causal implication*, while *astit*-sentences are not even closed under *logical consequence*. At this stage we can take a stance on the validity/non-validity of the six inferences listed in the Introduction and provide an explanation and a justification of our verdict.

Chapitre 2

Priest's Logic of Intentionality

2.1 An overview of Priest's enterprise

In *Towards Non-Being, the logic and metaphysics of intentionality*, Graham Priest undertakes the task of providing a general semantics for intentional operators expressed by verbs such as "knows", "desires", "fears", "tries to". Commenting on the present state of the subject he writes : "But despite the recent renewed interest in intentional contexts, the semantics of intentionality are in a highly unsatisfactory state. [Priest 2005, 6]".

Priest addresses the problem of closure failure under logical consequence by constructing an entirely new semantics which I shall briefly describe. To start with Priest adopts the standard structure used to interpret modal operators (alethic, temporal, epistemic, deontic etc ...), i.e. a non empty set of *possible worlds* and a bunch of accessiblity relations over those worlds. Next, Priest brings in the *logically impossible worlds* invented by the Finnish logician Veikko Rantala to deal with the problem of logical omniscience. Finally Priest introduces *open worlds* which "realize how things are conceived to be for the contents of arbitrary intentional states [Priest, Ibid. 21]". These three sets of worlds are mutually exclusive and jointly exhaustive. The union of possible and impossible worlds constitutes the set of *closed worlds (closed under logical consequence)*.

The actual world is enclosed in the set of possible worlds. It is represented by the symbol @. Validity of an inference is defined as truth preservation at @.

Before characterizing open worlds, let us come back to logically impossible worlds. The latter share some features with open worlds. Logically impossible worlds are worlds where laws of logic may be different. Most laws of logic are entailments of the form $A \supset B$. If logic changes at impossible worlds, formulas of that form [conditionals] must behave differently. How differently ? Priest replies that formulas of that form may differ in any way whatever. We can even go so far as treating conditionals as atoms when we evaluate them relatively to *logically impossible worlds*. We can now go a step further and characterize *open worlds* as

worlds which are even more ill-behaved than logically impossible worlds in so far as *all formulas* and not merely conditionals can be treated as atoms. As Priest puts it : "...just as conditionals may behave arbirarily at impossible worlds, all formulas may behave arbitrarily at open worlds [Priest, 2005, 22]".

Another innovation due to Priest should be mentioned here. Unsually we just give only the extension, we take the co-extension for granted. For instance, we give the conditions under which a formula is true and assume that if these conditions are not fulfilled then the formula is false. We assume, e.g., that the extension of a 0-place predicate and its co-extention are both exhaustive and exclusive. Priest breaks with this common practice and provides both the extension and the co-extension of predicates and operators. This forces him to provide two interpretation functions mapping constants into set-theoretical entites the model, namely δ^+ which gives the *extension* and δ^- which gives the *co-extension*.

$$\delta^+(P, w) \cap \delta^-(P, w) = \emptyset$$

$$\delta^+(P, w) \cup \delta^-(P, w) = D^n.$$

The advantage of saying explicitly that extension and co-extension are mutually exclusive and jointly exhaustive lies in the fact that it is now possible to relax either one or the other of these two constraints or both. If we relax the second, we create *gaps* (for instance truth-value gaps). If we relax the first we create *gluts* (overlapping truth-values).

As we have three sorts of worlds at out disposal we can drop these two constraints at logically impossible worlds and obtain a semantics appropriate for relevant logic without following in Hegel's footsepts who claimed that there are contradictions in the real world.

2.2 Priest's formal semantics for intentional operators

Let us now see how Priest's semantic apparatus provides a *general interpretation* for intentional operators which ensures the lack of closure under logical consequence.

The general semantics for intentional operators involves two compoenents :

(1) a model,

(2) an evaluation rule which enables us to define what I shall call the *generic intentional operator*.

(1) Priest's model is the n-tuple $< P, I, O, @, D, \delta >$ where
P is a set of possible worlds,
I is a set of impossible worlds,

O is a set of open worlds,

@ is the actual world. It is a member of the set of possible worlds,

D is the domain shared by all the worlds (constant domain),

δ is the evaluation function.

The union of P and I is denoted by C (the set of closed worlds). The union of P, I and O is denoted by W (the set of worlds).

(2) The clause for the interpretation of intentional operators of the form $t\Psi A$, where t is the name of an agent, Ψ is an intentional operator and A is a proposition.

Let us take an intentional operator Ψ. The truth conditions are given by the following clause [I skip the clause for the co-extension].

The relation R_Ψ is an accessibility relation which serves to interpret Ψ. What is new is that these accessiblity relations are allowed to access open worlds.

$$w \models -t\Psi A \text{ iff for all } w' \in W \text{ such that } wR^{(\delta t)}w', w' \models \neg A.$$

Hintikka brought out the explanatory power of clauses of that kind in this passage : "Putting the main point very briefly and somewhat crudely, by stepping from a world to its alternatives, we can reduce the truth-conditions of modal statements to truth-conditions of nonmodal statements [Hintikka 1973, 198"].

2.3 A model for lack of closure under logical consequence

Priest spelled out a model in which an intentional operator fails to satisfy closure under logical consequence.

For that purpose, he had to produce a model in which

$$\models (Pa \wedge Qa) \supset Pa \text{ but } \not\models t\Psi(Pa \wedge Qa) \supset t\Psi Pa$$

The formula $(Pa \wedge Qa) \supset Pa$ is true by assumption.

We have to find an interpretation in which $t\Psi(Pa \wedge Qa)$ is true and $t\Psi Pa$ is false.

Priest's counter-model contains one closed world only, namelly @ and one open world only, namely w. The accessibility relation leads from @ to w and nowhere else.

In virtue of the clause defining Ψ, the task of showing that $t\Psi(Pa \wedge Qa)$ is true in @, boils down to showing that $(Pa \wedge Qa)$ is true in w. Analogously showing that $t\Psi Pa$ is false reduces to showing that Pa is false in w.

Since w is an open world, formula $(Pa \wedge Qa)$ can be treated as *atomic*. Let us treat it as if it had the form Rab and give it an interpreation which makes it

true. For that purpose, Priest assigns the extension $D \times D$ to $Px_1 \wedge Gx_2$ (i..e. to Rx_1x_2) and individuals which are members of D to the individual constants "a" and "b". Under this interpreation $Pa \wedge Qa$ is true in w and $Pt\Psi(Pa \wedge Qa)$ is true in @.

Finally we have to give an interpreation to Pa which makes it false in the open worlds. Priest assigns the extension $\emptyset$ to Px_1 and assigns an individual in D to "a". Clearly, under that interpretation, Pa is false in w and $t\Psi Pa$ is false in @.

One might be tempted to object that once the formula $(Pa \wedge Qb)$ is treated as atomic, it ceases to qualify as the antecedent of $\models (Pa \wedge Qa) \supset Pa$ and to provide the counter-model we are searching for. This objection however is misguided. Priest does not change the syntactic form of the formulas under consideration *but assigns them a non compositional semantics* which makes it possible to construct the counter-model he needs. [*I* owe the clarification of this crucial point to Shahid Rahman].

One should not conclude however that as far as open worlds are concerned, anything goes. As Priest observes, inferences involving quantifiers are preserved.

2.4 The significance of Priest's counter-model

Several counter-models have been set up in this paper. Most ot them exemplified lack of closure under believed implication or causal implication. Only the bandaged man case examplified genuine lack of closure under logical consequence. However interesting the last case could be, it required a sophisticated temporal and causal setting. On the contrary, Priest's model of lack of closure under logical consequence is purely abstract. No concrete example of closure failure is given. No particular intentional verb is mentioned. This is done on purpose. Priest strives towards a characterization of closure failure as a feature of *pure logic*. He wants to capture *generic closure failure*. The absence of constraints on the accessibility he puts to use confirms his concern *for pure* as opposed to *applied logic*. The lack of consideration for BDI and STIT logics does not betray Priest's lack of interest in these logics but a clear awareness that although the two kinds of logic belong to the same science, they are nevertheless different enterprises. It is surprising however that no typical intentional verb satisfy the truth-conditions for the abstract intentional operator Ψ.

Priest considers exceptions to the failure of closure of intentional verbs under logical consequence. A construction such as "agent a (α) is rationally commited to φ" is a case in point. With that construction, closure is restored. Yet, if no specification is made about a given intentional verb, Priest seems to assume by default that if fails to be closed under logical consequence. Does Priest's general logic of intentionality clash with the findings of the local approaches surveyed in the first chapter ? Not necessarily. Priest's rule of evaluation for the generic Ψ operator should be seen as a presentation of an ideal type, in Max Weber's sense,

of intentionality. Ideal types are not the types that are more often instanciated. They may be outrun by hybrid cases. If we bear this in mind, we can reconcile Priest's general logic of intentionality with the local logics developped by the proponents of the BDI and the STIT logic

General conclusion

In his epoch making book *Word and Object*, Quine showed how the implicit ontology of a science could be brought to light by regimenting the language of science in a canonical notation and applying his criterion of ontologicaol commitment. He showed that the minimal ontology we need is made up of physical objects and sets. His ontological investigations led to startling results. For instance he showed that contrary to the received view mathematical induction can be proved without admitting infinite sets [Quine1963].

Ontology tells us about what there is. But what there is does not exhaust reality. Besides *what there is* there is *what people do* [Antoniol 1998]. The logicians who contributed to *stit*-logic should be praised for having put emphasis on a neglected but important feature of reality : agency. The study of agency forces upon us the consideration of choice between open scenarios. Hence Quine's distrust for possible worlds ceases to be unquestionable.

Once we are ready to accommodate actions in our picture of reality, we should also accommodate *intentions* construed as modes of actions. But *knowingly doing* is also a mode of action. Hence we cannot discard it. We are engaged on a slippery slope which will compel us to acknowledge many propositional attitudes such *knowing how, knowing that, knowing who.*

It is difficult to suscribe to Quine's following statement : "If we are limning the true and ultimate structure of reality, the canonical scheme for us is the austere scheme that knows [....] no propositional attitudes but only the physical constitution and behavior of organisms [Quine 1960, 221]". The developments of intentional logic seem to support a wider scheme which accommodates new categorial distinctions such as the distinction between the internal and external perspective [see Aucher 2008]. This widening of the scheme was forshadowed by Quine himself in his latest works where he stressed the role of empathy in learning a language [Quine 1992].

Bibliographie

[1] Antoniol Lucie, 1998, *Things People Do*, PhD. thesis, University of Stirling.

[2] Aucher Guillaume,2008, *Perspectives on Belief and Change*, PhD. thesis, Otago University and Toulouse University.

[3] Belnap Nuel, Perloff Michael, Xu Ming,2001, *Facing the Future*, O.U.P,Oxford.

[4] Balckburn Patrick, de Rijke Maarten and Yde Venema, *Modal Logic*, C.U.P.Cambridge.

[5] Broersen Jan 2009 "A Complete STIT Logic for Knowledge and Action, and Some of Its Applications" in M.Baldoni *et al.* (Eds.), DALT 2008, LNAI 5397, 47-59.

[6] Broersen Jan forthcoming "First Steps in the stit-Logic Analysis of Intentional Action"

[7] Chellas Brian 1980. *Modal Logic, an Introduction* C.U.P., Cambridge.

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

[9] Gochet Paul and Pascal Gribomont 2006 "Epistemic Logic" in Dov Gabbay and John Woods *Handbook of the History of Logic*, Vol.7, Elsevier, Amsterdam, 99-195.

[10] Gochet Paul 2007, "Un Problème ouvert en Epistémologie : la formalisation du savoir faire" suivi de commentaires de M.Cozic, P.Egré et G.Sandu in Paul Gochet et Philippe de Rouilhan, *Logique épistémique & Philosophie des Mathématiques*, edited by Thierry Martin and Philippe Mongin, Vuibert Sciences.Paris, 3-37.

[11] Herzig Andreas and Nicolas Troquard, 2006, "Knowing How to Play. Uniform Choices in Logics of Agency" in Gerhard Weiss and Peter Stone, editors, *Fifth Conference on International Joint Autonomous Agents & Multi Agent Systems* (AAMAS- 06). Hakodate Japan ACM, Press 8-12.

[12] Hintikka Jaakko 1973 "Grammar and Logic. Some Borderline Problems" in K.J.Hintikka, J.M.F.Moravcsik and P.Suppes, *Approaches to Natural Language*, Dordrecht, D. Reidel, 197-214.

[13] Horty John F. and Nuel Belnap 1995, "The deliberative Stit. A Study of Action, Omission, Ability and Obligation" *Journal of Philosophical Logic*, 24, 583-644.

[14] Horty John F. 2001, *Agency and Deontic Logic*, O.U.P., Oxford.

[15] Lorini Emiliano and Andreas Herzig 2008, "A logic of Intention and Attempt", Synthese 163,45-77

[16] Meyer John-Jules and Veltman Frank 2007, "Intelligent Agents and Common Sense Reasoning" in P.Blackburn, J.van Benthem and F.Wolter, *Handbook of Modal Logic*, Elsevier, Amsterdam, 991-1029.

[17] Moore Robert C 1985, 1995, "A Formal Theory of Knowledge and Action" reprinted in Moore Robert C. *Logic and Representation*, CSLI Lecture Notes Stanford. Stanford 27-70.

[18] Priest Graham 2005, *Towards Non-Being. The logic and metaphysics of intentionality* O.U.P. Oxford.

[19] Quine Willard Van Orman 1960, Word and Object, MIT Press., Cambridge MA.

[20] Quine Willard Van Orman 1963, *Set-Theory and Its Logic*, The Belknap Press of Harvard U.P., Harvard.Cambridge MA.

[21] Rao Anand S. & Georgeff Michael P.,1991, "Modeling Rational Agents within a B.D.I. Architecture" in James Allen, Richard Fikes, Erik Sandewall Eds. *Principles of Knowledge Representation and Reasoning* San Mateo Morgan Kaufmann, 473-484.

[22] Schild Klaus, 2000 On the Relationship Between BDI Logics and Standard Logics of Concurrency in *Autonomous Agents and Multi-Agent Systems*, 3 Kluwer, Amsterdam 259-283.

[23] Scott Dana 1970 "Advice on Modal Logic", in Karel Lambert, *Philosoophical Problems in Logic*, D.Reidel Dordrecht, 143-173.

[24] Semmling Caroline and Wansing Heinrich, "From *BDI* AND *stit* TO *bdi-stit* LOGIC" *Logic and Logical Philosophy*, 17, 189-211.

[25] Semmling Caroline and Wansing Heinrich, 2009, "A sound and complete system of bdi-stit logic" in M.Pelis (ed.) *Logica Yearbook* 2008, College Publications,London, 193-210.

[26] Straetmans Marina,1991, *Logiques modales non normales et logiques modales non monotones*, M.A. thesis,. University of Liège, Dept. of Mathematics, Liège.

[27] Wansing Heinrich, 2006 "Tableaux for multi-agent deliberative stit logic" in Guido Governatori, Ian Hodkinson and Yde Venema (Eds.) *Advances in Modal Logic*, Volume 6, College Publications, London.

[28] Wooldridge, Michael 2000, *Reasoning about Rational Agents*, MIT Press Cambridge MA.

Acknowledgment

The paper is a revised version of the first two lectures delivered at the Seminar of the Department of Philosohy, Logic and Philosoophy of Science, Universidad de Sevilla (Grupó de Logica, Lenguage e Información) as a guest of Prof. Angel Nepomuceno. I thank very much my hosts and the members of his team of researchers. I owe many improvements to him and to Hans van Ditmarsch, Senior Researcher, to Dr Joost J.Joosten and to several other participants who took part in the discussion. Prof.Shahid Rahman saved me from misinterpreting a crucial passage of Graham Priest's book. I had access to many sources of information thanks to Prof. Natasha Alechina, Prof. Pascal Gribomont, Prof. Jacques Riche, Prof. Heinrich Wansing, Senior Researchers Jan Broersen, Andreas Herzig and Emiliano Lorini to whom I express my gratitude. I also thank Dr Hassan Bougrine and Mrs Palermini who facilitated access to local bibliographical resources.

Comments and criticisms welcome Paul Gochet (pgochet@ulg.ac.be)