UNIVERSITAT POLITÈCNICA DE CATALUNYA
DEPARTAMENT DE LLENGUATGES I SISTEMES INFORMÀTICS
PROGRAMA DE DOCTORAT EN INTEL.LIGÈNCIA ARTIFICIAL

TESI DOCTORAL

Modelling rational inquiry in non-ideal agents

Memòria presentada per en Antonio Moreno i Ribas
per a optar al títol de Doctor en Informàtica per
la Universitat Politècnica de Catalunya

Directors: Dr. Ulises Cortés, Dr. Ton Sales

Tarragona, 2000

Per la meva estimada Aïda

Acknowledgements

There have been many people that have helped me in the long way that has led to this dissertation. First of all, I would like to thank my parents and brothers for their continuous confidence and support.

If there is one single person that deserves to be mentioned in this page, it is Ulises Cortés. He has been an inspiring figure for me, since we first met in 1986. He introduced me in the academic world, and he provided me with every possible means to form me as a researcher.

The main ideas underlying this dissertation were first proposed to me by Ton Sales, back in 1993. He introduced me to the logical omniscience problem, and he suggested me some ways of attacking it. Some of the ideas shown in the following pages were very influenced by his suggestions. If someone in the world may fully understand the motivation of this work, it is undoubtedly him.

I have made several stages abroad, which helped me quite a lot to become a researcher. I would like to thank Mario Furnari (Istituto de Cibernetica, Arco Felice) for providing me with a very pleasant working environment in my first research work. My two summers at the University of Bath were also important in my formation; my work there was supervised by Julian Padget.

Several people at the Human Communication Research Centre of the University of Edinburgh devoted some time to reading my half-baked ideas and provided insightful comments and suggestions. I would like to mention Andreas Schotter, Saturnino Luz, Jon Oberlander, Jean Carletta and Jasper Taylor.

The research groups on Artificial Intelligence at UPC and URV have been full of colleagues (and good friends) that have helped me in a hundred different ways during these years. My deepest thanks to all of them.

Contents

List of Figures

List of Tables

1 Introduction

1.1 Motivation

One of the main research topics in Artificial Intelligence in the last decade has been the design and construction of *rational agents*. This kind of entities may be defined in many different ways; a standard and commonly used definition, which shall be assumed throughout this dissertation, describes them as *those systems that have some kind of perception and try to act upon the environment so as to achieve their goals, given their beliefs* ([RuNo95]). Note the importance given to the agent's *beliefs* in this definition, as they are implicitly guiding its behaviour (*e.g.* by being used in order to select the most appropriate action to take between different available alternatives).

There are extensive reviews of different types of architectures for rational agents available en the literature ([WoJe95], [Müll97a], [Wool99]). Some of the most well-known architectures are shown in table 1, along with related bibliographical references.

Architecture	*References*
Reactive architectures	[Maes89], [Broo91], [RoKa96]
Logic-based architectures	[Fish94], [Lésp96], [SRG99]
Procedural Reasoning System	[GeLa87], [GeIn89], [ICAR96]
Belief-Desire-Intention architectures	[Brat87], [CoLe90], [RaGe95b]
Implicit agent architecture	[Denn78], [Denn84]
Agent-oriented programming paradigm	[Shoh90], [Shoh93], [Shoh98]
Defeasible reasoner	[Poll90], [Poll95], [Poll00]
Layered architectures	[Ferg95], [Müll97b]

Table 1: Some architectures for rational agents

This dissertation focuses on a subclass of agents, namely on the so-called *deliberative agents*. These agents must keep an internal explicit representation of their environment and of their mental state, which may be modified by some sort of syntactic inference procedure. Other kinds of architectures, such as the reactive ones, are not considered in this work. Most of the entries in the above table (except the first one) may be qualified as deliberative agents.

Their architecture is usually composed of a knowledge base, that stores relevant facts about the agent and its environment, and modules that may perform inferences from those facts, interact with the agent's environment, create and evaluate different plans, *etc.* The knowledge base constitutes a description of the world, and may be taken to explicitly represent the agent's *beliefs* about its environment (or even about itself or about other agents).

Being more specific, in this dissertation the expression *"rational agents"* is taken to refer to *those agents that, apart from complying with the previous definition, are constantly trying to make their beliefs as similar as possible to the facts that hold in the real world* (a more detailed analysis of our conception of rational agents is given in §4). They keep trying to expand their beliefs (by including facts that are true in their environment) and to get rid of wrong beliefs (those that do not reflect the actual state of the world). This process has been traditionally called *rational inquiry* ([ReBr79]). The classical philosophical tradition has considered two components in this process: a *rational* one, that consists in the application of some inference procedures to the present beliefs (resulting in the addition of new beliefs or the discovery of some incompatibility in them), and an *empirical* one, which adds or removes beliefs according to the results of the observations performed in the agent's environment ([ReBr79]). These components will find their counterparts in the *logical* and *experimental* dimensions of belief analysis performed by a especial kind of agents called *rational inquirers*, which will be defined in §4. Therefore, a rational agent's set of beliefs is constantly evolving in time, as the agent keeps updating it to take into account the results of its own internal inference procedures or the information that it may have gathered or received from the environment.

One of the main issues in Artificial Intelligence (and the main topic in this dissertation) is *how to build a formal model of the evolution of the beliefs of a rational agent*. Having such a model should be interesting and useful at least for the following reasons:

- Agent technology is developing at a tremendous rate. Agent (and multi-agent)-based systems are achieving a great complexity, and tools that may be used to describe the evolution of an agent's belief set should be useful in order to provide a description of the agent's process of inquiry at a high level of abstraction, appropriate for an adequate comprehension of the behaviour of the agent by an external user of the system.

- Being a *formal* model, it forces us to explicit every underlying assumption, resulting in a deeper and more complete understanding of the system by their designers and programmers.

- Some formal models may be even used later as the basis of specific implementations of multi-agent systems. For instance, the Belief-Desire-Intention (BDI) model has been actually used to guide the implementation of agents whose behaviour is explained in terms of these three propositional attitudes ([RaGe92], [RaGe95b], [SRG99]).

  However, it is fair to note that there is still a big gap between the formal models that describe the behaviour of multi-agent systems and the actual implementation of these systems. In this dissertation we provide a characterisation of the main activities that a rational agent may perform on its beliefs, and we suggest a particular way of implementing them in a class of agents called *rational inquirers*, which is presented in §4.

- The rationale underlying a further (and probably the most important) motivation for our work is developed in detail in the rest of this introductory chapter. Existing formal models of belief (based in the classical *possible worlds* model and its associated *Kripke semantics*, to be described in §1.2.3) model agents that must necessarily believe all classical tautologies and whose set of beliefs is necessarily closed under classical logical consequence. Thus, this model is suitable only in ideal settings, where the modelled agents are assumed to be *logically omniscient* and *perfect reasoners*. An important aim in our work is to provide a framework in which the evolution of the beliefs of a real, non-ideal, limited agent may be successfully modelled.

1.2 Formal models of belief

This section offers a brief presentation of the two main kinds of formalisms that have been traditionally used to model the reasoning processes that a rational agent may perform on its beliefs: *syntactic* (§1.2.1) and *modal* (§1.2.2). The attention is focused in the latter class of models; more specifically, in *doxastic* modal logics (*i.e. modal logics of belief*). The *possible worlds model* and the *Kripke semantics* are presented in §1.2.3; they are regarded as the

classical way of giving a natural and intuitive semantics to doxastic formulæ. However, this model has a serious drawback, explained in §1.2.4: the agents that it models must have ideal reasoning capabilities, as they must believe every classical tautology (no matter how intricate it may be) and they must also believe every logical consequence of their beliefs (regardless of the resources that may be available to the agent, *e.g.* space or time). Thus, that result motivates the need for a formal model of a rational agent's reasoning process that may be used in the case of non-ideal agents. This motivation will also be supported in §2, after providing a detailed survey of the strenghts and weaknesses of the main approaches that have been proposed in the literature in order to overcome the logical omniscience problem.

1.2.1 Syntactic treatments of belief

There are two main kinds of formalisms used to model the reasoning processes that an agent may perform on its beliefs: the *syntactic* treatments and the *modal* approaches. The so-called *syntactic* treatments of belief are not considered throughout this dissertation. In these approaches the notion of *belief* is represented in the language of predicate calculus by the predicate *bel*, where in *bel('α')*, 'α' is the name of the formula α. Thus, the language must include terms that are the names of the formulæ of the language (that is why the resulting logics are called "*reified epistemic logics*"). The main advantage of these approaches is its expressivity (*e.g.* it allows statements such as "α *believes something*" ($\exists x\ bel(\alpha, x)$) or "$\beta$ *believes everything that* α *believes*" ($\forall x\ (bel(\alpha, x) \Rightarrow bel(\beta, x))$), which may not be expressed in the modal approach. However, they also suffer from some limitations ([Kono86a], [McAr88]):

- The notation needed for a first-order meta-language is very complex, because there must exist terms that refer to the expressions in the object language, as explained above.
- Montague ([Mont63]) and Thomason ([Thom80]) showed that the (epistemic and doxastic) first-order theories that contain axioms that formalise number theory and axioms that correspond to standard modal axioms $T, 4$ and 5[1] are inconsistent. However, it may be shown that

[1] These modal axioms are defined below, in §1.2.3.

this inconsistency disappears if the first-order language is restricted to those formulæ that have a modal counterpart, see [DRLe88].

- It has also been argued that a system that uses standard theorem proving techniques over the axiomatisation of these meta-language approaches may run into severe computational problems.

The notational burden imposed on the user by the syntactic approaches has made them much less popular than the modal approaches for modelling doxastic and epistemic notions in the last years. Some of these syntactic approaches are deeply commented in [Kono86a], [McAr88], [DRLe88] and [Reic89].

1.2.2 Modal doxastic logics

This proposal is centered in those approaches in which there is a *modal* treatment of belief: *modal logics of knowledge and belief* (epistemic and doxastic logics). These modal logics are used to analyse in a formal way the reasoning about knowledge or belief performed by an agent.

In propositional modal logic two unary operators ($\Box$ and $\Diamond$) are added to propositional logic; they are called the *necessity or universal modal operator* and the *possibility or existential modal operator*, respectively. The existential modal operator may be considered as the dual of the universal operator, because it is defined in the following way: $\Diamond \mathtt{A} \equiv \neg\Box\neg\mathtt{A}$. The rules used to build formulæ in propositional modal logics are the following:

- The rules of propositional logic.
- If A is a formula, then $\Box$A and $\Diamond$A are also formulæ.

The modal operators can be interpreted in a variety of ways, which give rise to different logics. The most interesting ones are:

- *Alethic logics*: $\Box$A is interpreted as *"A is necessary"*, and $\Diamond$A as *"A is possible"*.
- *Default logics*: $\Box$A denotes the fact that *"A is normally the case"*.
- *Deontic logics*: $\Box$A is interpreted as *"A is compulsory"*, and $\Diamond$A as *"A is allowed"*.

- *Doxastic (or belief) logics*: $\Box$`A` means *"*`A` *is believed"* and $\Diamond$`A` can be interpreted as *"*`A` *is plausible"*.

- *Dynamic logics*: in this kind of logics there is a modal operator associated to each program. $\Box_\phi$`A` means *"*`A` *is the case after every execution of the program* ϕ*"*, whereas $\Diamond_\phi$`A` can be interpreted as *"There is some execution of the program* ϕ *that makes* `A` *be the case"*.

- *Epistemic (or knowledge) logics*: $\Box$`A` is interpreted as *"*`A` *is known"*, and $\Diamond$`A` as *"*`A` *is plausible"*.

- *Provability logics*: $\Box$`A` is read as *"*`A` *is provable"*, and $\Diamond$`A` means *"*`A` *is consistent"*.

- *Temporal logics*: $\Box$`A` means *"*`A` *will always be the case"*, and $\Diamond$`A` is read as *"*`A` *will (at some point in the future) be the case"*.

1.2.3 Possible worlds and Kripke semantics

In the literature of doxastic logics the universal modal operator ($\Box$) is usually called B. If several (m) agents are taken into account, a family of subscripted operators (B_1, B_2, ..., B_m) is considered (where $B_i\varphi$ is read as "$Agent_i$ *believes* φ"). The usual language of propositional doxastic logic for m agents contains a set of primitive propositions (`P`, `Q`, `R`, ...), the basic logical operators ($\neg$, $\vee$, $\wedge$ and $\Rightarrow$) and the modal belief operators B_1, B_2, ..., B_m. The formulæ of this language are the primitive propositions and the applications of the logical operators or the modal operators to other formulæ of the language.

The semantic model traditionally adopted as a basis in doxastic logics is the *possible worlds model* ([Hint62]). This model is based on the assumption that there is a set of possible states (or *possible worlds*) in which the agent can be in any moment; when the agent is in a possible world, there is a set of possible worlds which are compatible with the actual world, in the sense that the agent cannot distinguish these worlds from the actual one. The usual semantics given to the formulæ of the doxastic language described above is the *Kripke semantics* ([Krip63b]), that states that the agent believes a formula if and only if it is true in all the worlds that the agent cannot tell apart from the actual world.

Definition 1 (Kripke structures)

A normal Kripke structure is a tuple of the form $(S, \pi, R_1, ..., R_m)$, *where* S *is the set of possible worlds,* π *is a truth assignment to each primitive proposition in each world and* R_i *is the accessibility relation between worlds for* $Agent_i$ *(*$(s, t) \epsilon R_i$ *iff* s *and* t *are indistinguishable worlds for* $Agent_i$*). Given a normal Kripke structure* M*, the relation* $M, s \models \varphi$ *(read "*φ *is* true *(or* satisfied*) in state* s *of model* M*") is usually defined in the following way:*

- $M, s \models \mathtt{P}$, *being* P *a primitive proposition, if* $\pi(s, \mathtt{P}) = true$
- $M, s \models \neg\varphi$ *if* $M, s \not\models \varphi$
- $M, s \models (\varphi \vee \psi)$ *if* $M, s \models \varphi$ *or* $M, s \models \psi$
- $M, s \models (\varphi \wedge \psi)$ *if* $M, s \models \varphi$ *and* $M, s \models \psi$
- $M, s \models (\varphi \Rightarrow \psi)$ *if* $M, s \models \neg\varphi$ *or* $M, s \models \psi$
- $M, s \models B_i\varphi$ *if* $M, t \models \varphi$ $\forall t$ *such that* $(s, t) \epsilon R_i$

The last clause formalizes the conception of beliefs previously stated: an $Agent_i$ believes a proposition φ when it is true in all the worlds considered possible by the agent, *i.e.* in all the worlds that it cannot tell apart from the actual world. That means that φ has to be true in all the worlds connected to the actual world through R_i, which is the accessibility relation between worlds for $Agent_i$.

Definition 2 (Axiomatic systems and Kripke structures)

A formula ϕ *is said to be* provable *in an axiomatic system* S *(denoted as* $S\vdash\phi$, $\vdash_S\phi$ *or just* $\vdash\phi$*, if* S *is clear from the context) if it is an instance of one of the axioms of* S *or if it can be obtained by applying one of the inference rules of* S *to provable formulæ. A formula is said to be* true *in a class of Kripke structures if it is true in every state of every structure of the class. An axiomatic system* S *is* sound *with respect to a class of Kripke structures* C *if every formula provable in* S *is true in* C*.* S *is* complete *with respect to* C *if every formula which is true in* C *can be proved in* S*. An axiomatic system* characterizes *a class of Kripke structures when it is a sound and complete axiomatization of the class.*

Proposition 1 (Axiomatization of Kripke structures)

There exists a sound and complete axiomatization of the class of all normal Kripke structures for m *agents ([HaMo92]). It has two axioms and two inference rules, which are the following:*

- *A1. All the instances of tautologies of propositional calculus.*
- *A2.* $(B_i\varphi \wedge B_i(\varphi \Rightarrow \psi)) \Rightarrow B_i\psi$ *(axiom K)*
- *R1. From* $\vdash\varphi$ *and* $\vdash(\varphi \Rightarrow \psi)$ *infer* $\vdash\psi$ *(Modus Ponens)*
- *R2. From* $\vdash\varphi$ *infer* $\vdash B_i\varphi$ *(Necessitation)*

This axiomatic system is known as system K_m (or K if only one agent is considered), and it is the simplest one used to model logics of knowledge and belief. In many approaches other properties of knowledge or belief are taken into account by adding axioms to this basic system. The most popular ones are the following:

- $B_i\varphi \Rightarrow \varphi$ (Axiom of knowledge, axiom T)
- $B_i\varphi \Rightarrow B_iB_i\varphi$ (Axiom of positive introspection, axiom 4)
- $\neg B_i\varphi \Rightarrow B_i\neg B_i\varphi$ (Axiom of negative introspection, axiom 5)
- $B_i\varphi \Rightarrow \neg B_i\neg\varphi$ (Axiom of consistency, axiom D)

Axiom T states that those formulæ that are believed must also be true; this property is usually required for knowledge, but not for beliefs. Axiom 4 defends that an agent must be aware of its own beliefs, whereas axiom 5 holds in case the agent is aware of the facts that it does not believe. Axiom D holds in those situations in which the agent's set of beliefs is logically consistent (it does not contain both a formula and its negation).

System KT is defined as system K plus the axiom of knowledge. If the axiom of positive introspection is added to system KT, system $KT4$ (also known as $S4$, see [HuCr68]) is obtained; in turn $S4$ can be transformed into $S5$ ($KT45$) by adding the axiom of negative introspection. The system *weak* $S5$ ($K45$) contains the axioms of introspection but does not contain the axiom of knowledge. Adding the axiom of consistency to $K45$, the system

$KD45$ is obtained; this system is usually assumed to be the standard modal logic of (idealized) belief.

There exists a very strong relationship between these axioms and the properties of the accessibility relation between worlds R_i of the Kripke structure. This relationship is studied in a branch of modal logic known as *correspondence theory* (see *e.g.* [VBen84], [VdHo93], [Goré99]). For instance, table 2 shows the properties that the accessibility relation between worlds must have in order for the previous axioms to hold.

Axiom	*Property*
T	Reflexive $(\forall x R_i xx)$
4	Transitive $(\forall x \forall y \forall z\ (R_i xy \wedge R_i yz) \Rightarrow R_i xz)$
5	Euclidean $(\forall x \forall y \forall z\ (R_i xy \wedge R_i xz) \Rightarrow R_i yz)$
D	Serial $(\forall x \exists y\ R_i xy)$

Table 2: Examples of correspondence theory

An specially important case arises when the accessibility relations are *equivalence relations* (*i.e.* they are reflexive, symmetric and transitive). In that case it is easy to check that the Euclidean and the serial properties also hold, and the resulting logical system is $S5$, in which the four axioms shown above hold. This is usually taken to be the logic of (idealized) knowledge.

1.2.4 Logical omniscience and perfect reasoning

Regardless of the axioms that may be added to K_m, axioms (axiom schemas, in fact) A1 and A2 and rules R1 and R2 are always kept in these basic modal systems. Axiom A1 (all the instances of all propositional tautologies) and rule R1 (*Modus Ponens*) are taken directly from classical propositional logic (although some instances of the axiom schema A1 are *not* standard propositional tautologies, *e.g.* $(B_i\varphi \vee \neg B_i\varphi)$). The problems to be addressed in this work derive from axiom K and the rule of necessitation; they (seem to) commit us to model agents that are:

- *logically omniscient*, because they believe all tautologies (since all of them are true in every world), and

- *perfect reasoners*, because they also believe all logical consequences of their beliefs (*e.g.* if an agent believes `P` and (`P` $\Rightarrow$ `Q`) in a state s, it means that these two propositions are true in all the worlds compatible with s (all states R_i-accessible from s); therefore, `Q` will also be true in all of these worlds, and the agent will also believe `Q`).

These facts have very unrealistic implications; *e.g.* an agent with the basic arithmetic axioms would have to know whether the Fermat theorem is indeed a theorem or not, or an agent that knew the rules of chess would have to know whether White has a winning strategy or not ([Kono85]). The union of these problems is usually referred to in the literature as the problem of *logical omniscience*[2]. It is worth pointing out that omniscience is relative to the chosen language. Full rationality in an absolute sense would require a language isomorphic to the *real world*, but such a language is not available. Thus, language fashions and sets limits to the agents' way of seeing things ([HMP96]).

Some authors (see *e.g.* [FHMV95], [VdHM95]) make more fine-grained distinctions between different kinds of logical omniscience (because they review some approaches to this problem that solve some of these weaker forms). All these special cases of logical omniscience are relative to the notion of logical implication (and, therefore, of validity) that is considered. The most frequently mentioned forms of logical omniscience are the following:

- *Full logical omniscience*: if the agent believes all the formulæ in a set Γ, and Γ logically implies the formula ϕ, then the agent also believes ϕ (this is called *perfect reasoning* in this dissertation).

- *Belief of valid formulæ*: if ϕ is valid, then the agent believes ϕ (it can be seen as an especial case of full logical omniscience, when an initial empty belief set is considered; this is called *logical omniscience* in this dissertation).

- *Closure under logical implication*: if the agent believes ϕ and ϕ logically implies ψ, then the agent believes ψ (another especial case of full logical omniscience, when the initial set of beliefs has only one element).

[2]Some authors prefer to call it *closure under logical consequence* ([Kono85]), *consequential closure* ([Reic89]) , *tautological closure* ([Shoh93]) or the problem of *saturated belief* ([GSGF93]).

- *Closure under logical equivalence (or belief of equivalent formulæ)*: if the agent believes ϕ, and ϕ and ψ are logically equivalent, then the agent believes ψ (this is an especial case of closure under logical implication).

- *Closure under material implication (or just closure under implication)*: if the agent believes ϕ and $(\phi \Rightarrow \psi)$, then it also believes ψ (it is an especial case of full logical omniscience if ψ is a logical consequence of the set $\{\phi, (\phi \Rightarrow \psi)\}$, as it is in classical propositional logic).

- *Closure under valid implication*: if the agent believes ϕ, and the formula $(\phi \Rightarrow \psi)$ is valid, then the agent believes ψ (it is equivalent to closure under logical implication if ϕ logically implies ψ just in case $(\phi \Rightarrow \psi)$ is valid, as in classical propositional logic).

- *Closure under conjunction*: if the agent believes ϕ and ψ, it also believes $(\phi \wedge \psi)$ (again, it is an especial case of full logical omniscience in propositional calculus, where $(\phi \wedge \psi)$ is a logical consequence of the set $\{\phi, \psi\}$).

- *Weakening of belief*: if the agent believes ϕ, it also believes $(\phi \vee \psi)$, for any arbitrary formula ψ (note that, in standard propositional logic, ϕ logically implies $(\phi \vee \psi)$, for any formula ψ; thus, it is an especial case of closure under logical implication).

- *Triviality of inconsistent beliefs*: if the agent believes ϕ and $\neg\phi$ (for an arbitrary ϕ), then it believes any formula ψ (it is a consequence of closure under conjunction because in propositional logic any formula is a logical consequence of a contradiction such as $\phi \wedge \neg\phi$).

A graphical representation of the dependencies between these forms of logical omniscience is shown in figure 1 (note that the subsumptions depicted in that figure depend on the notion of logical consequence that is being considered, as pointed out above). This dissertation will focus on the (coarse grained) concepts of logical omniscience (full logical omniscience) and perfect reasoning (belief of valid formulæ), as described above.

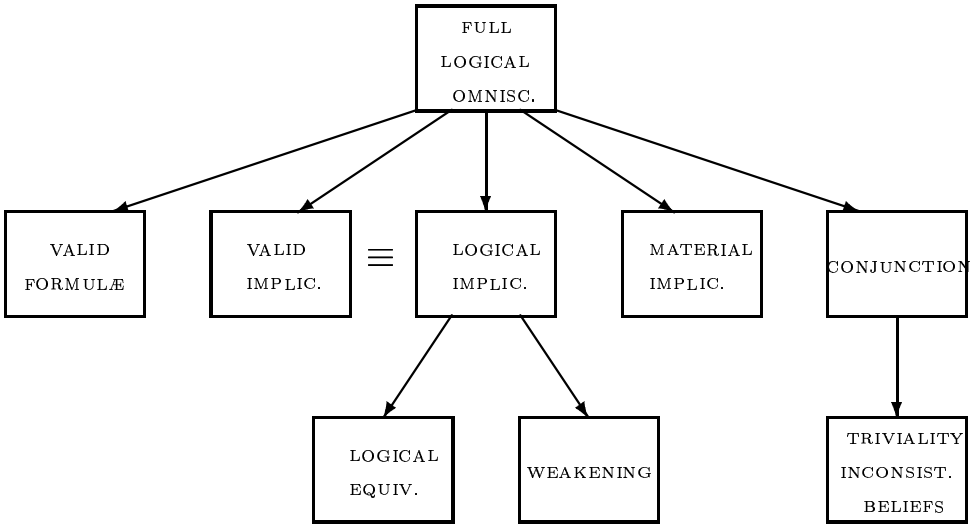

Figure 1: Dependencies between various forms of logical omniscience

1.2.5 Considering non-ideal agents

Logical omniscience and perfect reasoning can be acceptable in some circumstances and unacceptable in others, *e.g.* they are not considered as problems when epistemic logics are used for reasoning about communication protocols in distributed systems ([FHMV95])[3]. In that context there is an *external* view of knowledge. A system is described by the sets of its possibles *runs*, being a run a description of the system's behaviour over time. At each point in a run, a process is in some local state. A process at one point considers another point possible if it has the same local state in both. If points are thought of as *possible worlds*, the system's designer may ascribe knowledge to the processes by assuming that a process knows something just in case it holds in all worlds it considers possible ([FHMV95]). Processes do not compute their knowledge, and they are not required to answer questions about it. In these applications the knowledge possessed by agents is so simple that com-

[3]This dissertation deals with doxastic logics, but most of the observations are also applicable to epistemic logics, that also suffer from the logical omniscience and the perfect reasoning problems. Usually, the only difference between knowledge and belief taken into account in the literature is that one may believe false facts, but one may only know true facts. This property of knowledge is obtained when axiom T holds.

plexity of internal reasoning is usually neglected. Logical omniscience and perfect reasoning are also accepted in the classical puzzles of the literature of reasoning about knowledge, such as the *muddy children*, the *wise men* or the *unfaithful wives* ([GaSt58], [Barw81], [Gard84]). Omniscient formalisms have the major advantage of being simpler and easier to study, and can be taken as modelling ideal reasoners against which real (human or artificial) reasoners can be measured as approximations.

Some authors consider logical omniscience a sufficient reason for rejecting a model-theoretic analysis of epistemic concepts (see *e.g.* [Chom82]). Obviously there are many circumstances in which logical omniscience and perfect reasoning are unacceptable; that would be the case when the agent is supposed to be able to compute its knowledge or to take actions based on it. This would be an *internal* view of knowledge, as something that is acquired after a computation. It is clearly not a realistic model of either human agents (who are not logically omniscient) or computational agents (which have resource limitations that can prevent them from being perfect reasoners). Some of the undesirable consequences of the various forms of logical omniscience listed above are the following:

- An agent with limited resources has to be somehow capable of having access to every logical consequence of its set of beliefs (full logical omniscience).
- Every valid assertion has to be believed by the agent, regardless of its complexity (belief of valid formulæ).
- If the agent believes an assertion, it must also believe every logically equivalent assertion, no matter how complicated they may be (closure under logical equivalence).

In summary, omniscience is irreparably out of line with the needs of any real reasoning agent. A number of reasons may be given in order to justify the study of *non-ideal* agents ([Moli91]):

- Ideal agents are physically impossible, because real reasoners are not logically omniscient.
- The agents that we can construct are necessarily limited, because

  - They have only limited resources (*e.g.* a bounded finite memory).
  - They have only limited cognitive and computational capabilities.
  - They are short on time (the world does not await them).

- The set of beliefs of a real agent simply cannot be closed under logical consequence, since it would mean that the agent has a decision procedure for first-order predicate logic.
- Artificial agents have in general only incomplete and probably incorrect beliefs about reality.
- It is wrong to consider that a theory built around idealized entities will successfully apply to non-ideal ones, because
  - What is rational for an ideal agent can be judged to be irrational for a finite agent ([Cher86]). It is very different to design a system that exhibits *perfect rationality* (*i.e.* a system that acts at every instant in such a way as to maximize its expected utility, given the information it has acquired from the environment) than to build a system that has *bounded optimality* (*i.e.* a system that behaves as well as possible given its computational resources, [RuNo95]). In fact, the idea of *limited rationality* has been around AI since Newell and Simon's early work (see *e.g.* [NeSi72]).
  - By idealizing away the limitations of finite agents we resign the possibility of gaining any insights about the nature of reasoning with limited resources.

1.3 Overview of the dissertation

The main aim of this work is to develop a way to *model the process of rational inquiry* (the evolution of a rational agent's set of beliefs over time as a consequence of its interaction with the world and its internal inferential processes), *keeping the general idea of the possible worlds model and the Kripke semantics* (because, after all, they seem a very natural and intuitive semantics for modal logics of belief) *but trying to avoid the problems of logical omniscience and perfect reasoning* (in order to take into account non-ideal

agents). The steps that have been followed to reach this goal are described in this dissertation, which is structured as follows:

- An extensive survey of the state of the art is presented in §2, where the most relevant approaches that have been proposed to solve the problems of logical omniscience and perfect reasoning are presented. Some of the reviewed formalisms are the following:

  - Montague's *intensional logic of beliefs* ([Mont70]).
  - Konolige's *deduction model of belief* ([Kono86a]).
  - Fagin, Halpern and Vardi's *non-standard structures* ([FHV90b]).
  - Levesque's *logic of implicit and explicit beliefs* ([Leve84]).
  - Wooldridge's *belief models* ([Wool95]).
  - Van der Hoek and Meyer's *logic of awareness and principles* ([VdHM89]).
  - Thijsse's *hybrid sieve systems* ([Thij92]).

  This review focuses on the similarities and differences between the various solutions and shows to which degree they provide a framework in which it is possible to model non-ideal reasoners. Part of this survey was reported in [More98].

- As a result of this study, an *impossible worlds*-based framework in which logical omniscience is avoided is proposed in §3. The main idea of this proposal is that a situation may be perceived in different ways by different agents; thus, a situation may be described *subjectively* by each of them. Furthermore, in each situation an agent may have reasons to suport and/or to reject certain facts (following the suggestion made by Levesque in his *logic of implicit and explicit beliefs*, [Leve84]). The perception that an agent has of a situation will determine its (positive and negative) beliefs in that situation. These are the basic ideas underlying the *subjective situations* described in that chapter. Given a multi-agent system, the beliefs of each agent may be modelled by considering its perception of the actual situation. Within this new framework, the entities that are used to model the evolution of the beliefs of a rational agent are called *conceivable situations*; they are

descriptions of the way in which an agent perceives its current situation. This proposal is motivated by the conclusions reached in the review shown in §2. The first intuitive ideas underlying the concept of *conceivable situation* were reported in [MoSa97a] and [MoSa97b]. These ideas were refined and detailed in [MCS99a] and [MCS00a]. The framework of *subjective situations* has been described in [MCS99b] and [MCS00b].

- In order to prove the suitability of this framework to model the evolution of the beliefs of a rational agent, an analysis of the doxastic tasks which such an agent may perform is made in §4.2. After that, and with the aim of having a concrete intepretation of these tasks, a general class of non-ideal reasoners, called *rational inquirers*, is defined. These agents are constantly performing a multi-dimensional dynamic analysis of their beliefs, in order to make them as similar as possible to the facts that hold in the real world. These agents are given the following capabilities:

  - They may perform some (limited) deductive inferences on their sets of beliefs, using a modified version of the classical analytic tableaux method. This method was modified in order to allow for limited, non-ideal agents; for instance, the propositional part of the resulting tableaux method axiomatizes the logical consequence relation defined by Kleene's strong three-valued logic (as proved in appendix A).
  - They may have doubts about their beliefs, and may introduce these doubts into the analysis by adding instances of the *Axiom of the Excluded Middle* into the tableaux of the logical analysis. The introduction of these doubts permits the exploration of the two alternatives of the disjunction, and the search for examples or counter-examples needed to corroborate (or refute) the two available options.
  - They may make questions to the environment, in order to confirm or refute doubtful beliefs. The answers received from the environment are also included in the open tableaux of the logical analysis. The questions to be posed to the environment are suggested by the Skolem constants that appear in the logical analysis, linking in

a novel fashion the *rational* and *empirical* components of rational inquiry ([ReBr79]).

- They may also add to their beliefs the information that they receive directly from the environment (*e.g.* the data supplied by other agents).

In our initial works on rational inquirers (*e.g.* [MoSa97a], [MoSa97b]) the logical dimension only included the analysis of propositional disjunctions. In a later stage of our research, a full propositional analysis was considered and the other dimensions of analysis were sketched ([MCS98], [MCS99a]). Our last works already consider the use of predicate logic in the logical analysis and a more detailed account of the other dimensions of analysis ([MCS00a]).

- In §5 it is shown how the evolution of the beliefs caused by this dynamic multi-dimensional belief analysis may be formally modelled. This modelling process is made using two basic tools:

  - A new kind of semantic entities, called *conceivable situations*, is considered. This kind of *possible worlds* is closely related to the *subjective situations* described in §3, and corresponds to those scenarios that the modelled agent is capable of considering, regardless of their possible partiality or inconsistency. In our view, a two-sided analytic tableau will be seen as the representation of a class of conceivable situations.
  - The set of doxastic alternatives considered by the agent after each step of analysis will correspond to the classes of conceivable situations represented by the open tableaux. This evolution of the set of doxastic alternatives is modelled with the generation of a sequence of accessibility relations, that are used to represent which are the worlds considered possible by the agent at each point in time. This fact will allow us, through the application of a modified Kripke semantics (defined in §5.4), to have a way of finding out which are the agent's beliefs at each point in time.

We suggested the use of a dynamic accessibility relation to model the evolution of an agent's set of beliefs in [More96], [MoSa97a], [MoSa97b];

however, in these works the analysis of the beliefs was restricted to logical deduction. Conceivable situations were explicitly mentioned in [MCS98], and they have also been reported in [MCS99a]. In one of our last works we already presented how the belief change produced by all the dimensions of analysis may be modelled in our framework ([MCS00a]).

- Finally, in §6 the dissertation is summarized and some lines of future research are suggested.

1.4 Main contributions

The main contributions of the work reported in this dissertation are the following:

- We provide a very detailed review of the most interesting approaches that have tried to deal with the problems of logical omniscience and perfect reasoning. This dissertation contains the description of more than twenty different techniques that have been put forward in fields such as Computer Science, Artificial Intelligence and Philosophical Logic. To the best of my knowledge, this chapter subsumes all current published reviews of this kind of approaches.

- We provide a radically different way of tackling these problems, with the definition of a new kind of entities called *subjective situations*. The parting point of this technique is quite unorthodox and non-standard, as it proposes to deal with subjective, agent-based descriptions of possible states of affairs, instead of managing complete and consistent representations of possible worlds. We define a first-order doxastic logic on top of these situations, and show how logical omniscience and perfect reasoning are avoided.

- We have identified which are the activities (the *doxastic tasks*) that may modify the set of beliefs of a rational agent. In that way, we have abstracted from all the other actions that an agent has to carry out, in order to focus only on those that are directly related to its beliefs. We have also defined a certain type of non-ideal agents, called *rational inquirers*, that implement these doxastic tasks in a particular way. This

kind of agents is specific enough to provide us with a concrete example of agents whose beliefs we can model, and is general enough to show that the belief modelling techniques that are used in this dissertation might be used over any other kind of similar agents, irrespective of the particular way in which they carried out their doxastic activities.

- Finally, we have used the tools provided by the *subjective situations* framework in order to formally model the evolution of the beliefs of *rational inquirers*. Unlike the classical approaches to logical omniscience and perfect reasoning, we are concerned with dealing with dynamic sets of beliefs, that change in time as a consequence of the doxastic activities in which rational agents are permanently engaged. The basic idea that has been used in the modelling technique is to change the set of *conceivable situations* that are considered as doxastic alternatives by the agent after each doxastic task. This idea is not new in the literature of belief change (see *e.g.* [FHMV95]); the main contribution of our work in this respect is to provide a systematic, formal and easy-to-follow procedure for computing the doxastic alternatives after each step of belief analysis, which could be applied to any kind of non-ideal agent.

In summary, it can be argued that the main contribution of this work is to make a first step towards seriously considering the idea of real, limited, resource bounded, non-logically omniscient agents, unlike almost all previous attempts to deal formally with an agent's propositional attitudes, which provide very nice treatments of belief and knowledge but always dismiss the issues of logical omniscience and perfect reasoning.

2 Avoiding logical omniscience

There have been many authors that have tried to solve the problems of logical omniscience and perfect reasoning. This chapter contains a review of some of the more relevant approaches, classified as syntactic or semantic. In this chapter negation and conjunction are taken as the basic logical operators, and disjunction and implication are supposed to be defined from them in the usual way: $((\phi \vee \psi) \equiv \neg(\neg\phi \wedge \neg\psi)$ and $(\phi \Rightarrow \psi) \equiv \neg(\phi \wedge \neg\psi))$. Some of the proposals refer to *epistemic* logics (using the modal knowledge operator K) that, as mentioned in §1.2.5, also suffer from the problems of logical omniscience and perfect reasoning.

2.1 Syntactic approaches

2.1.1 Beliefs as sets of formulæ

The basic syntactic solution to the problems of logical omniscience and perfect reasoning is to identify the agent's beliefs as the set of formulæ contained in the agent's *belief base* ([Perl84], [Haas85]). An agent's set of beliefs is a set of formulæ of a certain language L. Intuitively, α believes ϕ, where ϕ is a sentence of L, if and only if ϕ belongs to α's set of beliefs.

This idea was already considered in the works of Eberle ([Eber74]) and Moore and Hendrix ([MoHe79]). In [FHMV95] a formalization of this approach is made using *standard syntactic assignments*. A *syntactic structure* M is a pair (S, π), where S is a set of states and π is an standard syntactic assignment, *i.e.* an assignment of truth values to all formulæ in all states that satisfies the following properties:

- $\pi(s)(\phi) = \mathit{true}$ iff $\pi(s)(\neg\phi) = \mathit{false}$.
- $\pi(s)(\phi \wedge \psi) = \mathit{true}$ iff $\pi(s)(\phi) = \mathit{true}$ and $\pi(s)(\psi) = \mathit{true}$.

The truth of a formula ϕ in an state s of a syntactic structure M is defined as follows:

$$M, s \models \phi \text{ if and only if } \pi(s)(\phi) = \mathit{true}$$

Note that syntactic assignments do not impose any constraints on the truth values given to doxastic formulæ (those with the form $B_\alpha\psi$). In fact, this model does not assume anything about the nature of beliefs. An agent can have contradictory beliefs, or it can even believe inconsistent sentences such as $(\mathtt{P} \wedge \neg\mathtt{P})$. The semantic approach analogous to this syntactic approach is Montague's *intensional logic of beliefs* ([Mont70]), which is described in §2.2.15. With this approach both logical omniscience and perfect reasoning disappear, because an agent's set of beliefs may not contain all instances of all tautologies (it may be the case that $\pi(s)(B_\alpha(\phi \vee \neg\phi)) = false$) and may also not be closed under logical consequence (it may be the case that $\pi(s)(B_\alpha\phi) = true$, $\pi(s)(B_\alpha(\phi \Rightarrow \psi)) = true$ and $\pi(s)(B_\alpha\psi) = false$). Halpern comments in [Halp86] that this option is very difficult to analyse, because using this representation of the beliefs there are no principles that can guide a knowledge-based analysis. In [FHMV95] it is argued that this option is a way of *representing* belief, rather than a way of *modelling* belief.

2.1.2 Incompleteness of deduction rules

An interesting approach is adopted by Konolige in [Kono86a]. He models the agent's beliefs with a *deduction structure*. A *deduction structure* is a tuple (B, R), where B is a base set of facts and R is a set of deduction rules (that can be logically incomplete). Using Konolige's notation, it is said that $[S_i]\phi$ if $\phi \in bel(d_i)$ (agent S_i believes a formula ϕ if it belongs to its belief set, which is defined as the *deductive* closure of the base set B using the rules R). It is a very flexible way of modelling the agent's beliefs, and it also avoids the problems of logical omniscience and perfect reasoning (*e.g.* an agent may believe $\mathtt{P}$ and $(\mathtt{P} \Rightarrow \mathtt{Q})$ and not believe $\mathtt{Q}$, if *Modus Ponens* is not included in its set of deductive rules). Deductive closure is a condition much weaker than closure under logical consequence. If the deduction rules of an agent are logically incomplete, then it will not be able to derive all logical consequences of the base set; in this way it is possible to model the reasoning about belief performed by an agent with resource limitations.

Formally, Konolige defines a *model structure* as a tuple $<\phi, v_0, U>$, where U is the universe of individuals, ϕ is a mapping from the set of constants to U, and v_0 assigns a truth value to each basic atom. A satisfiability relation $\models_m$ is defined as follows:

- $\models_m A$ iff $v_0(A^\phi) = true$, being A a basic atom. A^ϕ denotes the application of ϕ to the constants appearing in A.
- $\models_m (A \wedge B)$ iff $\models_m A$ and $\models_m B$.
- $\models_m \neg A$ iff $\not\models_m A$.
- $\models_m \exists x A$ iff $\exists k \epsilon U$ such that $\models_m A_k^x$. A_k^x denotes the substitution of x by k in A.
- $\models_m \forall x A$ iff $\forall k \epsilon U \ \models_m A_k^x$ holds.

A formula ψ is called *valid* ($\models_m \psi$) iff it is true in every model structure. A *B(L, ρ)-model* is defined as a tuple $<\phi, v_0, U, D, \gamma>$, where the first three elements define a model structure, D contains a deduction structure d_i for each agent S_i and γ contains a set of functions (γ_i) that associate an individual constant to each element of U for each agent S_i. L is the *internal* language of the agent, and ρ is the set of rules of each agent S_i (ρ_i). When these models are defined, a new clause (that deals with the belief operator for each agent, $[S_i]$) is added in the definition of $\models_m$:

- $\models_m [S_i]\phi$ iff $\phi \ \epsilon \ bel(d_i)$.

A formula is *B(L, ρ)-valid* iff it holds in every *B(L, ρ)*-model. Konolige shows how to use the *analytic tableaux* method to prove whether a formula is satisfiable. A new rule is added to the classical analytic tableaux method: a branch with the formulæ $T[S_i]\Gamma$ and $F[S_i]\Phi$ closes if $\Gamma \vdash_{\rho(i)} \Phi$. This rule is obtained from the following result, that shows the relationship between the satisfiability of sentences that contain the belief operator and derivations in the internal language of the agent: the set $\{[S_i]\Gamma, \neg[S_i]\Delta\}$ is *B(L, ρ)* unsatisfiable iff $\exists \delta \epsilon \Delta$ s.t. $\Gamma \vdash_{\rho(i)} \delta$. The method is proved to be *sound* (in every closed tableau *B(L, ρ)* headed by $F\varphi$ -the negation of φ-, φ is *B(L, ρ)*-valid) and *complete* (if φ is *B(L, ρ)*-valid, then there is a closed tableau *B(L, ρ)* for $F\varphi$). Konolige also provides a practical proof method based on *resolution*. Van der Hoek and Meyer analyse this model in [VdHM96] and show how it may be considered as a generalisation of the standard modal approach on the basis of Kripke models described in §1.2.3 (certain subclasses of deduction structures behave exactly as basic modal systems such as S4 or S5).

2.1.3 Belief models

Wooldridge describes in [Wool95] a way of modelling the belief systems of resourced-bounded reasoners. This model subsumes classical ways of belief modelling, like Konolige's *deduction model of belief* ([Kono86a], see §2.1.2) or the standard use of doxastic normal modal logics (see §1.2.3). An agent's belief system is represented using a *belief model*, that is defined as a tuple $< \Delta, BE >$, where Δ is the *base set* of (propositional) beliefs (the observations that have been made over the agent's beliefs) and BE (the *belief extension* relation) is a countable, non-empty binary relationship between sets of formulæ and formulæ satisfying the following requirements:

- If $(\Delta, \phi)\epsilon BE$, then $\forall\delta\epsilon\Delta, (\Delta, \delta)\epsilon BE$ (*reflexivity*)
- If $(\Delta, \phi)\epsilon BE$, $(\Delta', \psi)\epsilon BE$ and $\Delta \subseteq \Delta'$, then $(\Delta', \phi)\epsilon BE$ (*monotonicity*)
- If $(\Delta, \phi)\epsilon BE$ and $(\{\phi\}, \psi)\epsilon BE$, then $(\Delta, \psi)\epsilon BE$ (*transitivity*)

The belief extension relation models the agent's reasoning ability. This idea is clarified with the concept of *belief set (bel)*, that represents the agent's set of beliefs:

$$bel(< \Delta, BE >) = \{\phi \text{ such that } (\Delta, \phi)\epsilon BE\}$$

The meaning of a tuple $(\Delta, \delta)\epsilon BE$ is the following: if the agent believes the formulæ in the set Δ, it will also believe the formula δ. In fact, one of the main points of this proposal is its ability for modelling agents whose *reasoning* processes are not based on logical inference (although it may also model those that are, with an appropriate definition of the pairs included in BE).

In this model the agent does not have to necessarily believe either a formula or its negation, because it may fail to believe both (if α and $\neg\alpha$ are not included in $bel(< \Delta, BE >)$). The agent may also believe a formula and its negation (if both of them are included in $bel(< \Delta, BE >)$). These facts make it similar to Levesque's *logic of implicit and explicit belief* ([Leve84]), which is described in §2.2.4, and to our own approach, to be described in §3.

As noted in [Wool95], logical omniscience is avoided in this model because neither axiom K nor the necessitation rule of normal modal logics hold (because of the syntactic nature of belief models, and especially of belief extension relations). Wooldridge also shows how to derive the belief extension

relation for an agent, if its base set of beliefs (Δ) and its set of *legal belief states (BS)* are given. The set BS is supposed to contain all those belief states (sets of formulæ) in which the system could possibly be after some chain of events. BE would be then defined as follows:

$$BE = \{(\Delta, \delta) \text{ such that } \forall \Delta' \epsilon BS, \text{ if } \Delta \subseteq \Delta' \text{ then } \delta \epsilon \Delta'\}$$

In short, the agent would believe all those formulæ that are included in all those legal belief states that contain its base set of formulæ. Thus, this notion is similar to the *necessity* notion in normal modal logics, although Wooldridge stresses in [Wool95] that belief is not given a normal modal interpretation in his model.

2.1.4 Restrictions in deductions

Hintikka suggests another syntactic solution to logical omniscience and perfect reasoning ([Hint86a]). His proposal is to put syntactic restrictions in the deductive argument from S_1 to S_2, in order to restrict the class of logical consequences $\vdash (S_1 \supset S_2)$ for which it holds that $\{\alpha\}KS_1 \supset \{\alpha\}KS_2$ (in Hintikka's notation that formula is read *"if α knows S_1 then it also knows S_2"*). The number of individuals that is being considered in a formula is denoted by the number of free individual symbols and the number of levels of quantifiers. The basic idea is that this parameter must never be greater during the argument from S_1 to S_2 than in S_1 or S_2 ([Hint75b]). Hintikka claims that this approach yields the same results than the *urn models* semantic approach ([Rant75]), which is described in §2.2.14. He also claims that this idea is connected to many issues in Philosophy of Logic, Mathematics and the psychology of deductive reasoning ([Hint73], [Hint86b]).

2.2 Semantic approaches

2.2.1 Impossible worlds

Cresswell ([Cres72], [Cres73]) pointed out that the problems of logical omniscience and perfect reasoning could be solved by allowing *non classical worlds* in the semantics (*i.e.* worlds in which tautologies may not be true and inconsistent formulæ may be true). These worlds are called *impossible worlds* in [Hint75a] and [FHMV95], *non-designated indices* in [Scot70], *setups* in [RoRo72], *situations* in [Leve84] and *non standard worlds* in [ReBr79]. As

an example of what this kind of worlds look like, let's take Rescher and Brandom's *non standard worlds*. They are built from other worlds by using two operations: *schematization* and *superposition*. Schematization combines worlds conjunctively, whereas superposition combines them disjunctively. A formula is true in the schematization of two worlds if it is true in both of them, and it is true in the superposition of two worlds if it is true in either of them. Thus, schematized worlds may be *partial* (in the sense that it is possible that neither ϕ nor $\neg\phi$ hold in one of these worlds), and superposed worlds may be *overdetermined* (in the sense that both ϕ and $\neg\phi$ may hold in one of these worlds).

A formal account of this kind of approaches is given in [FHMV95]. An *impossible worlds structure* is defined as a tuple $(S, W, \pi, K_1, K_2, \ldots, K_n)$, where $(S, K_1, \ldots, K_n)$ is a Kripke frame, $W \subseteq S$ is the set of *possible states* or worlds, and π is a syntactic assignment that satisfies the following properties (in those states $s \in W$):

- $\pi(s)(\phi \wedge \psi) = \mathit{true}$ iff $\pi(s)(\phi) = \mathit{true}$ and $\pi(s)(\psi) = \mathit{true}$.
- $\pi(s)(\neg\phi) = \mathit{true}$ iff $\pi(s)(\phi) = \mathit{false}$.
- $\pi(s)(K_i\phi) = \mathit{true}$ iff $\pi(t)(\phi) = \mathit{true}$ for all t such that $(s,t) \in K_i$.

Furthermore, logical implication and validity are determined only with respect to possible states; thus, as agents consider impossible states when determining their knowledge, logical omniscience does not necessarily hold.

Vardi ([Vard86]) mentions some disadvantages of the impossible worlds approach:

- The intuition underlying non classical worlds is not very clear, and it is difficult to define the semantics of logical connectives in these worlds.
- Adding new worlds does not solve the perfect reasoning problem; *e.g.* the agents modelled in [Leve84] still believe all *logical consequences* of their beliefs, but not the standard logical consequences but the consequences in relevance logic ([AnBe75]).

2.2.2 Belief as possibility

Van der Hoek and Meyer made a proposal to overcome the problem of logical omniscience in [VdHM89]. They suggest to model the notion of belief using the modal possibility operator, rather than the necessity operator. Thus, they write the clause used to determine the satisfiability of a doxastic formula in a world s of a Kripke model M as follows:

$$M,\, s \models B\varphi \text{ if } \exists t \text{ such that } (s,t)\epsilon R \text{ and } M,\, t \models \varphi$$

With this definition some of the undesirable problems associated to logical omniscience disappear (*e.g.* closure under implication and closure under conjunction do not hold). However, some forms of logical omniscience are still valid and, moreover, this idea induces the addition of other constraints on the modelled sets of beliefs; for instance, all of the following properties hold in this approach:

- Closure under logical equivalence.
- Belief of valid formulæ.
- Closure under valid implication.
- Weakening of belief.
- Closure under disjunction: if an agent believes $(\phi \vee \psi)$, it must also believe ϕ and/or ψ.

As the authors note, this approach models a rather weak notion of belief: an agent believes a formula ϕ just in case it thinks it is possible to conceive a world in which ϕ holds. Thus, it may easily believe ϕ and $\neg\phi$, if it does not have any arguments that support or deny ϕ in a definitive way.

2.2.3 Non-standard structures

One way of dealing with the problems caused by the Kripkean conception of belief (as those formulæ that are *true* in every doxastic alternative) is to change the notion of truth, by giving a non-standard semantics to the logical

connectives. Fagin, Halpern and Vardi follow this approach in their *non-standard structures* ([FHV90a], [FHV90b], [FHV95]). Their main idea is to allow for a formula and its negation to have independent truth values; thus, ϕ and $\neg\phi$ may be both true or false in any state. They define non-standard structures as tuples $(S, \pi, \mathcal{K}_1, \ldots, \mathcal{K}_n, *)$, where all the components are the same as in standard Kripke structures except for *, which is a function that assigns a state to each state. This function is used to assign truth values to negated sentences in the following way:

$$(M, s) \models \neg\phi \iff (M, s^*) \not\models \phi$$

Thus, there may be (*incoherent*) worlds in which ϕ and $\neg\phi$ hold and (*incomplete*) worlds in which neither ϕ nor $\neg\phi$ hold. These situations may not arise in *standard* worlds (those worlds in which $s = s^*$). Giving this semantics to the negation operator, material implication is no longer equivalent to logical implication, and closure under material implication does not hold. Furthermore, it is shown in [FHMV95] that there are no tautologies in this logic; therefore, it is pointless to wonder whether the agent believes all valid formulæ or its set of beliefs is closed under valid implication. Other forms of logical omniscience are still valid (full logical omniscience, closure under logical implication, closure under logical equivalence and closure under conjunction). However, the presence of these properties is not as worrying as it was in classical logic, because non-standard structures define a weaker notion of logical consequence. In fact, it is also shown in [FHMV95] that non-standard agents are omniscient with respect to another implication connective (called by them *strong implication*, $\hookrightarrow$) with this semantics:

$$(M, s) \models (\phi \hookrightarrow \psi) \text{ if and only if}$$
$$(M, s) \models \psi \text{ holds whenever } (M, s) \models \phi \text{ holds.}$$

2.2.4 Explicit and implicit beliefs

One of the most well known approaches to logical omniscience and perfect reasoning is Levesque's *logic of explicit and implicit beliefs* ([Leve84]), described in this section with the notation used in [FaHa85].

Levesque uses a language with two modal operators: B for *explicit* beliefs and L for *implicit* beliefs. These operators are not allowed to be nested in the formulæ of the language. A *structure for explicit and implicit beliefs* is

defined as a tuple $M=(S, \mathcal{B}, T, F)$, where S is the set of primitive situations, $\mathcal{B}$ is a subset of S that represents the situations that could be the actual one (according to the present beliefs) and T and F are functions from Φ (the set of primitive propositions) into subsets of S. Intuitively, $T(\mathtt{P})$ contains all the situations that support the truth of P, whereas $F(\mathtt{P})$ contains the ones that support the falsehood of P. These *situations* are not classical worlds because it is not compulsory that a primitive proposition is only true or false in a given situation; it can be true, false, both of them or none of them. A situation s can be *partial*, if there is a primitive proposition P which is neither true nor false in s ($s \notin T(\mathtt{P}) \bigcup F(\mathtt{P})$) or *incoherent* if there is a proposition P which is both true and false in s ($s \in T(\mathtt{P}) \bigcap F(\mathtt{P})$).

A situation is *complete* if it is neither partial nor incoherent (it supports the truth or the falsehood of all primitive propositions, but not both of them). A complete situation s is *compatible* with a situation t if s and t agree in all the points in which t is defined. $\mathcal{B}^*$ is the set of all complete situations of S that are compatible with some situation in $\mathcal{B}$.

Now it is possible to define the relations $\models_T$ and $\models_F$ between situations and formulæ. Intuitively, $M, s \models_T \phi$ will hold when the situation s of the structure M supports the truth of ϕ, whereas $M, s \models_F \phi$ will hold when s supports the falsehood of ϕ. The definition of these relations is the following:

- $M, s \models_T \mathtt{P}$, where P is a primitive proposition, if and only if $s \in T(\mathtt{P})$
- $M, s \models_F \mathtt{P}$, where P is a primitive proposition, if and only if $s \in F(\mathtt{P})$

- $M, s \models_T \neg\varphi$ if and only if $M, s \models_F \varphi$
- $M, s \models_F \neg\varphi$ if and only if $M, s \models_T \varphi$

- $M, s \models_T (\varphi \wedge \psi)$ if and only if $M, s \models_T \varphi$ and $M, s \models_T \psi$
- $M, s \models_F (\varphi \wedge \psi)$ if and only if $M, s \models_F \varphi$ or $M, s \models_F \psi$

- $M, s \models_T B\varphi$ if and only if $M, t \models_T \varphi \; \forall t \in \mathcal{B}$
- $M, s \models_F B\varphi$ if and only if $M, s \not\models_T B\varphi$

- $M, s \models_T L\varphi$ if and only if $M, t \models_T \varphi \ \forall t \in \mathcal{B}^*$
- $M, s \models_F L\varphi$ if and only if $M, s \not\models_T L\varphi$

A formula φ is *true* in a state s if $M, s \models_T \varphi$. Levesque defines a formula φ as *valid* if it is true in all the structures $M = (S, \mathcal{B}, T, F)$ and all *complete* situations $s \in S$.

It can be proved that, with this semantics, explicit belief implies implicit belief (*i.e.* $\models (B\varphi \Rightarrow L\varphi)$ holds). It can also be proved that implicit beliefs are closed under implication and contain all propositional tautologies. Nevertheless, explicit belief does not suffer any of these problems, because:

- It is not *closed under implication* (*e.g.* $\mathrm{B}\mathtt{P} \wedge \mathrm{B}(\mathtt{P} \Rightarrow \mathtt{Q}) \wedge \neg \mathrm{B}\mathtt{Q}$ is satisfiable).
- It is not *closed under valid implication* (*e.g.* $\neg \mathrm{B}(\mathtt{P} \wedge (\mathtt{Q} \vee \neg \mathtt{Q})) \wedge \mathrm{B}\mathtt{P}$ is satisfiable).
- It does not contain all tautologies (*e.g.* $\neg \mathrm{B}(\mathtt{P} \vee \neg \mathtt{P})$ is satisfiable).
- The agent may have inconsistent beliefs without believing everything (*e.g.* $\mathrm{B}(\mathtt{P} \wedge \neg\ \mathtt{P}) \wedge \neg\ \mathrm{B}\mathtt{Q}$ is satisfiable).
- The agent may believe a formula φ and not believe a logically equivalent formula ψ (*e.g.* $\mathrm{B}(\mathtt{P} \wedge \neg\ \mathtt{P}) \wedge \neg\ \mathrm{B}(\mathtt{Q} \wedge \neg\ \mathtt{Q})$ is satisfiable).

In [FaHa85] it is pointed out that all these properties derive from the existence of incoherent situations. The following formula is valid: $\mathrm{B}\varphi \wedge \mathrm{B}(\varphi \Rightarrow \psi) \Rightarrow \mathrm{B}(\psi \vee (\varphi \wedge \neg\varphi))$. That is, it is the case that the agent's explicit beliefs are closed under implication, or otherwise there is an incoherent situation that the agent believes possible. Moreover, $B\varphi \wedge B(\neg\varphi) \equiv B(\varphi \wedge \neg\varphi)$, so it is only possible to have inconsistent beliefs if all the situations that the agent believes to be possible are incoherent, and this does not seem an appropriate idea.

It can also be proved that, even though all tautologies are not believed, the agent does believe all tautologies formed with primitive propositions which are known to the agent (those primitive propositions $\mathtt{P}$ for which $\mathrm{B}(\mathtt{P} \vee \neg\ \mathtt{P})$ holds). This fact suggests that this semantics is appropriate to model the lack of logical omniscience due to the lack of knowledge of a proposition,

but not to model the one due to lack of computational resources, because there could be tautologies formed only with known primitive propositions that could be very hard to prove.

Fagin and Halpern ([FaHa85]) add further comments to this approach:

- The relation $\models_T$ is defined for all situations, but the validity relation ($\models$) is only defined for *complete* situations. That implies that there are formulæ φ which are valid in this logic (e.g. $(\mathtt{P} \vee \neg\mathtt{P})$) whose truth may not be supported in all situations (it may exist a situation s such that $M, s \not\models_T \varphi$). All the valid formulæ of the propositional calculus are also valid in this logic (because validity is restricted to complete situations), and this idea does not seem to be very consistent with the use of situations.

- The definition of the relation $\models$ with respect to the logical operators does not seem very clear. For instance, suppose that the agent does not know a primitive proposition P. Then neither $M, s \models_T \mathtt{P}$ nor $M, s \models_F \mathtt{P}$ will hold, so $M, s \models_T (\mathtt{P} \equiv \mathtt{P})$ will not hold either. But it is easy to imagine an agent that, regardless of the knowledge of P, knows some propositional tautologies such as $\mathtt{P} \equiv \mathtt{P}$.

- This logic only deals with one agent, propositional logic and does not allow nested beliefs, so it is quite limited.

  However, similar logics that deal with many agents were presented by Halpern and Lakemeyer in [Halp93], [Lake93] or [HaLa96]. Lakemeyer extended this approach to first-order logic ([Lake91b], [Lake94]). Some partially nested beliefs were allowed in [Lake87] and [Lake91a] (as commented in the next section). Similar approaches with explicit beliefs may be found in [Delg95] and [LaLe88].

2.2.5 Implicit and explicit multi-agent nested beliefs

Lakemeyer incorporated in [Lake87] the possibility of having nested beliefs in the basic framework of implicit and explicit beliefs that has just been described. It is reported here with a multi-agent extension devised by Sim ([Sim00]). A *Lakemeyer model for nested implicit and explicit belief for n agents* is defined as a tuple $M_{LL} =< S, T, F, R_1^+, \ldots, R_n^+, R_1^-, \ldots, R_n^- >$, where S, T and F are defined as in the previous section and R_i^+ and R_i^- are

binary accessibility relations on S. These relations are used to validate beliefs and disbeliefs, respectively; thus, they only coincide in *complete* situations. Therefore, for all situations $s \epsilon B^*$ and all situations t, (sR_i^+t) if and only if (sR_i^-t).

As we are dealing now with the multi-agent case, two sets of modal belief operators $B_1, \ldots, B_n$ and $L_1, \ldots, L_n$ are considered. Modal operators may be nested, but no L_i may occur inside the scope of a B_i. The clauses that were used to define the relations $\models_T$ and $\models_F$ over modal formulæ are modified in the following way:

- M_{LL}, $s \models_T B_i\varphi$ if and only if M_{LL}, $t \models_T \varphi$ $\forall t \epsilon B$ such that (sR_i^+t)
- M_{LL}, $s \models_F B_i\varphi$ if and only if $\exists t \epsilon B$ such that (sR_i^-t) and M_{LL}, $t \not\models_T \varphi$
- M_{LL}, $s \models_T L_i\varphi$ if and only if M_{LL}, $t \models_T \varphi$ $\forall t$ ϵB^* such that (sR_i^+t)
- M_{LL}, $s \models_F L_i\varphi$ if and only if M_{LL}, $s \not\models_T L_i\varphi$

With this definition it is possible to deal with multiple agents and (restricted) nested beliefs, and some forms of logical omniscience (such as closure under material implication, triviality of inconsistent beliefs, belief of equivalent formulæ, belief of valid formulæ and closure under logical implication) are still avoided ([Sim00], [Lake87]).

2.2.6 Approximate knowledge

Schaerf and Cadoli ([ScCa92], [ScCa95]) follow a different approach (also commented in [Sim97]). Their idea is to provide a framework in which *approximate* knowledge may be explicitly represented and used.

They start by defining three kinds of *interpretations*:

- A *3-interpretation* assigns a value in the set $\{0, 1, \top\}$ to each basic proposition.

  As pointed out below, this kind of interpretations are used to represent complete (although possibly incoherent) *situations*.

- A *2-interpretation* assigns a value in the set $\{0, 1\}$ to each basic proposition (*i.e.* it is a classical interpretation).

  This kind of interpretations are used to represent *possible worlds* (*i.e.* complete and coherent situations).

- A *1-interpretation* assigns the value $\bot$ to each basic proposition.

  This kind of interpretations are used to represent coherent (although possibly incomplete) *situations*.

Schaerf and Cadoli proceed by defining more complex kinds of interpretations, in which different basic propositions may be assigned values in different sets. These new interpretations are relative to a subset S of the set of basic propositions $\mathcal{P}$:

- An *S-3 interpretation* maps every proposition in S to a value in $\{0, 1\}$ and every proposition in $\mathcal{P} - \mathcal{S}$ to a value in $\{0, 1, \top\}$.

  An S-3 interpretation has a 2-interpretation over S and a 3-interpretation over the rest of primitive propositions. Intuitively, an S-3 interpretation represents a complete (but possibly incoherent) situation.

- An *S-1 interpretation* maps every proposition in S to a value in $\{0, 1\}$ and every proposition in $\mathcal{P} - \mathcal{S}$ to the value $\bot$.

  An S-1 interpretation has a 2-interpretation over S and a 1-interpretation over the rest of primitive propositions. Intuitively, an S-1 interpretation represents a coherent (but possibly incomplete) situation.

Two families of modal operators are introduced, $\Box^1_S$ and $\Box^3_S$. The semantics of these operators is given with a variation of the classical Kripke models. A *model* M is defined as a triple (Sit, R, V), where Sit is a set of *situations*, R is a reflexive, transitive and Euclidean accessibility relation and V is a valuation that maps any situation into a S-1, S-2 or S-3 interpretation. S-1(Sit) is the set of those situations that are assigned an S-1 interpretation, $\mathcal{W}(Sit)$ is the set of possible worlds, and S-3(Sit) is the set of those situations that are assigned an S-3 interpretation. The semantics defined with these models is the following:

- $M, s \models \phi$ iff $\mathrm{V}(s)(\phi) = 1$

- $M, s \models \Box^3_S \phi$ iff $\forall t \epsilon$ S-3(Sit) (s R t) implies $M, t \models \phi$
- $M, s \models \neg\Box^3_S \phi$ iff $\exists t \epsilon$ S-3(Sit) such that (s R t) and $M, t \not\models \phi$
- $M, s \models \Box^1_S \phi$ iff $\forall t \epsilon$ S-1(Sit) (s R t) implies $M, t \models \phi$
- $M, s \models \neg\Box^1_S \phi$ iff $\exists t \epsilon$ S-1(Sit) such that (s R t) and $M, t \not\models \phi$

A formula ϕ is *valid* ($\models \phi$) if it is true in every possible world of every model. This semantics induces the following results:

- The necessitation axiom does not hold for $\Box^1_S$ ($\models \phi$ does not imply $\models \Box^1_S \phi$). However, it does hold for $\Box^3_S$ ($\models \phi$ implies $\models \Box^3_S \phi$). The same situation happens to axiom T.
- Axiom K holds for $\Box^1_S$ ($\models \Box^1_S(\phi \Rightarrow \psi) \Rightarrow (\Box^1_S \phi \Rightarrow \Box^1_S \psi)$), but it does not hold for $\Box^3_S$ ($\not\models \Box^3_S(\phi \Rightarrow \psi) \Rightarrow (\Box^3_S \phi \Rightarrow \Box^3_S \psi)$).
- Axioms 4 and 5 hold for both $\Box^1_S$ and $\Box^3_S$ (because of the properties of the accessibility relation between situations).

Sim notes the following facts ([Sim97]):

- $\Box^3_S \phi \wedge \Box^3_S(\phi \Rightarrow \psi) \wedge \neg\Box^3_S \psi$ is satisfiable (beliefs modelled with this operator are not closed under Modus Ponens).
- $\Box^3_S \phi \wedge \Box^3_S(\neg\phi) \wedge \neg\Box^3_S \psi$ is satisfiable (beliefs modelled with this operator may be inconsistent without having to believe every formula).
- $\neg\Box^3_S(\phi \wedge \neg\phi)$ and $\Box^3_S \phi \wedge \neg\Box^3_S(\phi \wedge (\psi \vee \neg\psi))$ are not satisfiable (some tautologies have to be believed).
- The converse results hold for $\Box^1_S$.

Schaerf and Cadoli argue that $\Box^1_S$ may be used to model *skeptical reasoners* (*i.e.* fully introspective agents that are capable of performing every sound inference, although they can also perform unsound inferences). On the other hand, $\Box^3_S$ may be used to model *credulous reasoners* (*i.e.* fully introspective agents that are not logically omniscient and only perform sound inferences). They also show that Levesque's modal operators (see §2.2.4) may be represented in this framework: L is equivalent to $\Box^1_S$ or $\Box^3_S$ when S=$\mathcal{P}$, and B is equivalent to $\Box^3_S$ when S=$\emptyset$.

2.2.7 Logic of general awareness

Fagin and Halpern suggest in [FaHa85] different logics that try to solve the problems of Levesque's logic of explicit and implicit beliefs (however, they keep both kinds of beliefs). Konolige comments in [Kono86b] one of them, the *logic of general awareness*.

Assume a propositional language with the usual boolean operators of negation and conjunction and an especial primitive proposition, $\perp$, which is always interpreted as false. In the case of a single agent, there are also three modal unary operators: B for *explicit* beliefs, L for *implicit* beliefs and A for *awareness*. All these operators can be nested.

A *Kripke structure of general awareness* is a tuple $M=(S, \pi, \mathcal{A}, \mathcal{B})$, where S is a set of states, $\pi(s, \mathtt{P})$ is a truth assignment for every primitive proposition $\mathtt{P}$ and every state s, and $\mathcal{B}$ is a binary relation between the elements of S (the accessibility relation between states) that is transitive, Euclidean and serial (and, therefore, this is the system KD45). In each possible world s, $\mathcal{A}(s)$ is a set of sentences of the language s.t. $\perp \epsilon\ \mathcal{A}(s)$. This is the set of formulæ that the agent is *aware* of in state s, but does not necessarily believe.

The semantics of the language is given by the relation $\models$, defined in the following way:

- $M, s \not\models \perp$
- $M, s \models \mathtt{P}$, where $\mathtt{P}$ is a primitive proposition, if $\pi(s, \mathtt{P}) = \mathit{true}$
- $M, s \models \neg\varphi$ if and only if $M, s \not\models \varphi$
- $M, s \models (\varphi \wedge \psi)$ if and only if $M, s \models \varphi$ and $M, s \models \psi$
- $M, s \models L\varphi$ if $M, t \models \varphi\ \forall t$ s.t. $(s, t)\ \epsilon\ \mathcal{B}$
- $M, s \models B\varphi$ if $\varphi\epsilon\ \mathcal{A}(s)$ and $M, t \models \varphi\ \forall t$ s.t. $(s, t)\ \epsilon\mathcal{B}$
- $M, s \models A\varphi$ if $\varphi\epsilon\ \mathcal{A}(s)$

The implicit belief operator L has the standard Kripke semantics. As the accessibility relation is transitive, Euclidean and serial, this operator will satisfy the axioms of positive and negative introspection and the axiom of consistency.

Explicit beliefs are defined as those implicit beliefs that belong to the awareness set. Therefore, $B\varphi \equiv L\varphi \wedge A\varphi$. Explicit and implicit beliefs will only be equal if the agent is aware of all implicit beliefs.

A possible interpretation of the formula $A\varphi$ could be the following: *the agent is able to prove whether φ is a consequence of a given set of premises or not in a certain time T.* That is, the agent is aware of a certain class of formulæ for which it is easy to make deductions (or to show that a certain deduction does not exist). Thus, the logic of general awareness models perfect reasoners over a restricted set of formulæ.

Konolige makes some remarks about this logic ([Kono86b]):

- With this semantics, the connection between the properties of the accessibility relation and the axioms satisfied by belief is lost.

  For instance, assume that the agent explicitly believes φ; then $B\varphi$, $L\varphi$ and $A\varphi$ are true. One can wonder whether $BB\varphi$ is true or not. $BB\varphi \equiv (LB\varphi \wedge AB\varphi) \equiv (LL\varphi \wedge LA\varphi \wedge AB\varphi)$, and (due to transitivity) $LL\varphi$ is true. The following conditions are needed:

  - $A\varphi \supset LA\varphi$
  - $B\varphi \supset AB\varphi$

  None of these conditions is affected by the accessibility relation, so the analysis of introspective properties that was possible with the standard Kripke semantics is lost (recall §1.2.3).

- The logic of general awareness can be characterized syntactically.

  Explicit beliefs are defined as the implicit beliefs of which the agent is aware: $M,\ s \models B\varphi$ if $\varphi \epsilon\ \mathcal{A}(s)$ and $M,\ t \models \varphi\ \forall t$ such that $(s,\ t)$ $\epsilon\ \mathcal{B}$. The first part of this conjunction refers to the presence in the awareness set; the second part uses the Kripke semantics, but can also be syntactically formalized. Moore ([Moor83]) proved that the system weak S5 characterizes *stable sets*. A set S is *stable* if it contains all the tautologies, is closed under *Modus Ponens* and the following conditions hold:

  - If $\varphi \in S$, then $L\varphi \in S$
  - If $\varphi \notin S$, then $\neg L\varphi \in S$

  Therefore, from a syntactic point of view, the models of the logic of general awareness are the intersection of a stable set with an arbitrary set (the awareness set).

- It does not seem very realistic to model agents which are perfect reasoners with respect to a syntactic class of formulæ.

2.2.8 Principles and implicit belief

In the logic of general awareness described in §2.2.7 the set of beliefs is restricted to those formulæ of which the agent is aware. Van der Hoek and Meyer ([VdHM89], [VdHM96]) propose in their *logic of awareness and principles* the opposite approach: to have a way to *add* arbitrary formulæ to the set of beliefs. These formulæ (called *principles*) are supposed to express propositions which the agent desires to belief, regardless of their possible inconsistency with respect to the rest of its beliefs.

Three modal operators are considered in this logic:

- $P_i\phi$: ϕ is a principle for $Agent_i$.
- $B_i\phi$: ϕ is believed by $Agent_i$.
- $B_{I,i}\phi$: ϕ is an *implicit belief* of $Agent_i$.

The modal formulæ are given a semantics with a modified version of the standard Kripke models. The ones used in this logic are defined as tuples $(S, \pi, \mathcal{P}_1, \mathcal{P}_2, \ldots, \mathcal{P}_n, R_1, R_2, \ldots, R_n)$, where S is a set of states, π is an assignment of a truth value to each basic proposition in each state, P_i is a function that returns the formulæ that are considered as *principles* by $Agent_i$ in each state, and R_i is a serial, transitive, Euclidean accessibility relation between states for $Agent_i$. The modal formulæ are given a truth value in each state of each Kripke model with these clauses:

- $(M, s) \models B_i\phi \iff \forall t\ ((s, t) \epsilon R_i \rightarrow (M, t) \models \phi)$
- $(M, s) \models P_i\phi \iff \phi \epsilon \mathcal{P}_i(s)$

- $(M, s) \models B_{I,i}\phi \Longleftrightarrow (M, s) \models B_i\phi$ or $(M, s) \models P_i\phi$

B_i behaves as the standard modal necessity operator. P_i is similar to the awareness operator in the logic of general awareness, as it only checks that a formula belongs to a set (in this case, the set of *principles* of $Agent_i$ in a given state). Implicit beliefs are similar to the explicit beliefs of the logic of general awareness, although in this logic a *disjunction* between standard beliefs and principles is used (rather than a *conjunction* between standard beliefs and the awareness set, as shown in §2.2.7). This definition provides a way of increasing the set of beliefs with a set of arbitrary formulæ in each state and, therefore, it avoids most of the forms of logical omniscience (*e.g.* closure under implication, closure under valid implication, closure under logical equivalence, closure under conjunction and weakening of belief do not hold). It is also possible to have inconsistent beliefs without believing every formula. However, agents still believe (implicitly) all valid formulæ. Moreover, Konolige's criticisms of the logic of general awareness (see §2.2.7) can still be applied to this approach.

2.2.9 Hybrid sieve systems

Thijsse ([Thij92], [Thij96]) proposes a way of using *partial* logics to deal with various forms of logical omniscience. He starts by defining a *partial model* as a tuple $(W, \mathcal{B}_1, \ldots, \mathcal{B}_n, V)$, where W is a set of worlds, $\mathcal{B}_i$ is the accessibility relation between worlds for $Agent_i$ and V is a *partial* truth assignment to the basic propositions in each world. $\top$ is a primitive proposition that is always interpreted as *true*. The truth ($\models$) and falsity ($=\!\!|$) relations are defined in the following way:

- M, $w \models \top$
- M, $w \not=\!\!| \top$
- M, $w \models \mathtt{P}$, where $\mathtt{P}$ is a primitive proposition, iff $V(\mathtt{P}, w) = 1$
- M, $w =\!\!| P$, where $\mathtt{P}$ is a primitive proposition, iff $V(\mathtt{P}, w) = 0$

- M, w $\models \neg\varphi$ iff M, w $\not\models \varphi$
- M, w $\not\models \neg\varphi$ iff M, w $\models \varphi$
- M, w $\models (\varphi \wedge \psi)$ iff M, w $\models \varphi$ and M, w $\models \psi$
- M, w $\not\models (\varphi \wedge \psi)$ iff M, w $\not\models \varphi$ or M, w $\not\models \psi$
- M, w $\models B_i\varphi$ iff M,v $\models \varphi$ $\forall v$ such that (w, v) $\epsilon \mathcal{B}_i$
- M, w $\not\models B_i\varphi$ iff $\exists v$ such that (w, v) $\epsilon \mathcal{B}_i$ and M,v $\not\models \varphi$

Validity is defined as verification: $\models \phi$ iff $\forall M, w$ $(M, w) \models \phi$. These models are similar to those in Levesque's logic of implicit and explicit belief (see §2.2.4). There are two main differences, though:

- Levesque uses two different sets in each situation: those propositions that are supported by the situation and those that are denied by the situation. Thus, a proposition may be in four different states in a situation (depending on whether it is supported and/or denied), defining a four-valued logic. In the partial models a proposition may only be in three different states with respect to a world: it may be supported, or denied, or neither supported nor denied.
- Levesque denies a modal formula in a state when there is a doxastic alternative in which the formula is not supported; Thijsse asks for a doxastic alternative in which the formula is denied.

The partiality of the valuation function causes the absence of tautologies: there are no valid formulæ in this logic. Thus, some forms of logical omniscience (belief of valid formulæ, closure under valid implication) disappear, and the following axioms (representing closure under material implication and closure under conjunction) are not valid either:

- K: $\vdash B(\phi \Rightarrow \psi) \Rightarrow (B\phi \Rightarrow B\psi)$
- C: $\vdash (B\phi \wedge B\psi) \Rightarrow B(\phi \wedge \psi)$

A notion of awareness (or rather *acquaintance*, as Thijsse puts it), may be incorporated in the logic with a new modal operator, A_i, defined as follows: $A_i\phi = \bigwedge_{\forall \mathtt{P} \ in \ \phi} B_i(\mathtt{P} \wedge \neg\mathtt{P})$. Thus, an agent is aware of a formula if each of the basic propositions that appear in the formula has a definite truth value.

In spite of these good properties, this logic also has some shortcomings:

- It eliminates too many tautologies (*e.g.* $B\mathtt{P} \vee \neg B\mathtt{P}$ seems acceptable, whereas $B(\mathtt{P} \vee \neg\mathtt{P})$ must be avoided).
- Some forms of logical omniscience still hold (albeit in a relativized way):
  - K_r: $B(\phi \Rightarrow \psi) \vdash (B\phi \Rightarrow B\psi)$
  - C_r: $B\phi \wedge B\psi \vdash B(\phi \wedge \psi)$

Thijsse proceeds by solving these problems with his *hybrid sieve* models, defined as tuples $(W, \mathcal{B}_1, \ldots, \mathcal{B}_n, \mathcal{A}_1, \ldots, \mathcal{A}_n, V)$, where W, $\mathcal{B}_i$ and V keep their previous meanings and $\mathcal{A}_i$ is the set of formulæ of which the agent is aware in each state. A new modal operator ($C_i\phi$) is introduced, with the intended meaning "$Agent_i$ consciously believes ϕ". A new satisfiability relation ($\|\!\!-$) is also introduced. The main aims of these models are:

- To provide a classical (two-valued) logic approach to the external part of the logic (so that *e.g.* $B\mathtt{P} \vee \neg B\mathtt{P}$ holds) while retaining a partial (three-valued) logic for the internal part (so that *e.g.* $B(\mathtt{P} \vee \neg\mathtt{P})$ is avoided).
- To avoid relativized forms of closure under material implication and closure under conjunction by adding the syntactic awareness filter.

The clauses that must be added to the partial models in order to give a semantic value to the new modal operator and the new consequence relation are the following:

- M, $w \|\!\!- \mathtt{P}$, where $\mathtt{P}$ is a primitive proposition, iff $V(\mathtt{P}, w) = 1$
- M, $w \|\!\!- \neg\varphi$ iff M, $w \not\|\!\!-\varphi$
- M, $w \|\!\!- (\varphi \wedge \psi)$ iff M, $w \|\!\!-\varphi$ and M, $w \|\!\!-\psi$

- $M,\ w \Vdash B_i\varphi$ iff $M,v \models \varphi\ \forall v$ such that $(w,\ v)\ \epsilon\mathcal{B}_i$
- $M,w \models C_i\varphi$ iff $M,w \Vdash C_i\varphi$ iff $M,w \models B_i\varphi$ and $\varphi\epsilon\mathcal{A}_i(w)$
- $M,w =\!| C_i\varphi$ iff $M,w =\!| B_i\varphi$ or $\varphi\ \not\epsilon\mathcal{A}_i(w)$

A formula is said to be valid ($\Vdash\phi$) just in case $M,w \Vdash\phi$ holds in each state of each model. Note that $\Vdash$ is a bivalent relation, whereas $\models$ is trivalent. These clauses impose a classical external logic in the propositional formulæ but they keep a partial internal logic in the modal formulæ (note that $M,w \models B_i\phi$ is true in exactly the same situations in which $M,w \Vdash B_i\phi$ holds). In this way some tautologies are recovered, without having the undesired property of believing valid formulæ. Note also the similarity of this approach with Fagin and Halpern's logic of general awareness (see §2.2.7), where explicit beliefs were defined as those implicit beliefs of which the agent is aware (thus, the C_i operator is very similar to the explicit belief operator (B) in that logic). Thijsse argues that with this syntactic filter the relativized forms of axioms C and K do not longer hold. Moreover, he affirms that this hybrid sieve models provide a framework in which any modal logic that extends classical propositional logic may be modelled.

2.2.10 Logic of local reasoning

Fagin and Halpern ([FaHa85], [FaHa88]) proposed a model that allows the agent to have non-trivial inconsistent beliefs (*i.e.* the agent may believe ϕ and $\neg\phi$ without believing everything). This model is called the *logic of local reasoning*, and it is based on the idea that a real (*e.g.* human) agent may often hold inconsistent beliefs, because it can easily fail to take into account all of its beliefs in every inference. Depending on the issues it is currently considering (call it its actual *context*) it may perform some inferences on a subset of its beliefs and inadvertedly deduce facts that are inconsistent with respect to other beliefs. Fagin and Halpern consider different states in which the agent may be (in their notation, different *frames of mind*, or *frames of reference* as Van der Hoek and Meyer put it in [VdHM96]). In each of these situations the agent considers a different set of doxastic alternatives and, therefore, it may held a different set of beliefs in each of them. While each *local* set of beliefs is consistent, the union of all of them (all the formulæ that

may be believed by the agent in some frame of mind) may turn out to be inconsistent.

Formally, a *local reasoning structure* is defined as a tuple $(S, \pi, \mathcal{C}_1, \ldots, \mathcal{C}_n)$, where S is a set of states, π assigns a truth value to each basic proposition in each state and $\mathcal{C}_i$ assigns to each state a non-empty set of subsets of S (where each of these subsets represents the set of doxastic alternatives considered in a frame of mind that $Agent_i$ may have in that state). The clause used to assign a truth value to doxastic formulæ is the following:

$$(M, s) \models B_i\phi \Longleftrightarrow \exists T \epsilon \mathcal{C}_i(s) \text{ such that } \forall t \epsilon T \ (M, t) \models \phi$$

Thus, the agent believes ϕ if it holds in all the doxastic alternatives, relative to a given frame of reference. In this way the agent may have inconsistent beliefs, because it may believe ϕ in one frame of mind and $\neg\phi$ in another. Note that this situation is very different from believing $(\phi \wedge \neg\phi)$. Some of the forms of logical omniscience considered in this dissertation do not hold in this logic (*e.g.* closure under implication and closure under conjunction). However, other forms of this problem are not solved by this approach (*e.g.* closure under valid formulæ, closure under valid implication, closure under logical equivalence and weakening of beliefs).

2.2.11 Logic S5P

Meyer and Van der Hoek proposed ([MvdH91], [MvdH92], [MvdH93]) the logic *S5P* as a way of representing incoherent beliefs, following the same basic idea than Fagin and Halpern in their *logic of local reasoning* (§2.2.10): an agent can focus on different *contexts* or *frames of reference* during its reasoning processes, and its beliefs may be different in each of them. This intuitive idea is formalized with an *S5P model*, which is defined as a tuple $(S, \pi, R, S_1, \ldots, S_n)$, where S is a set of worlds, π is an assignment of truth values to each basic proposition in each world, R is a universal relation on S (*i.e.* $\forall s, t \epsilon S \ (sRt)$) and each S_i is a (possibly empty) subset of S (representing a given frame of reference). There is a knowledge operator K and several *plausible belief* operators P_i (one for each frame of reference). The clauses that deal with these modal operators are defined as follows:

- $M, s \models K\phi$ if and only if $\forall t \epsilon S \ M, t \models \phi$
- $M, s \models P_i\phi$ if and only if $\forall t \epsilon S_i \ M, t \models \phi$

With these definitions, knowledge is represented by an $S5$ logic whereas each belief operator is ruled by a $KD45$ logic. Therefore, many of the forms of logical omniscience may not be avoided (for instance all tautologies must be believed in all contexts, and the beliefs within each context must be logically closed). However, it is possible to represent inconsistent beliefs, because $(P_i\phi \wedge P_j\neg\phi)$ is satisfiable (without implying $P_k\psi$, for arbitraries k and ψ). In this way it is possible to represent the inconsistency that may arise from information obtained from different sources. The main difference of the *logic S5P* with the *logic of local reasoning* described above is that in the former it is possible to select any context (using the appropriate P_i operator), whereas in the latter we can only check whether there exists a context in which a certain formula holds.

2.2.12 Non-standard belief structures

Fagin and Halpern's logic of local reasoning (see §2.2.10) may be shown to be equivalent to Vardi's *non-standard belief structures* ([Vard86]). These structures take Rescher and Brandom's *non-standard worlds* (see §2.2.1, [ReBr79]) as their basic notion. A *non-standard possible world expression* E is the smallest set that satisfies these conditions:

- $W \subseteq E$, where W is a set of worlds.
- If $w_i \epsilon E$ for all i in an index set I, then $\cap\, w_i \epsilon E$.
- If $w_i \epsilon E$ for all i in an index set I, then $\cup\, w_i \epsilon E$.

A *non-standard belief structure* M is a triple (W, N, Π), where W is a set of worlds, Π returns the *intension* of each basic proposition (the set of worlds in which it is satisfied) and N assigns to each agent in each world the non-standard world that the agent believes to be the actual one. The satisfiability relation is defined as follows:

- $M, w \models \mathtt{P}$, where $\mathtt{P} \epsilon \mathcal{P}$, if $w \epsilon \Pi(\mathtt{P})$
- $M, w \models \neg\phi$ if $M, w \not\models \phi$
- $M, w \models (\phi \wedge \psi)$ if $M, w \models \phi$ and $M, w \models \psi$
- If $N(a, w) = \cap\, w_i$, then $M, w \models B_a\phi$ if $M, w_i \models \phi$ for all $i \epsilon I$
- If $N(a, w) = \cup\, w_i$, then $M, w \models B_a\phi$ if $M, w_i \models \phi$ for some $i \epsilon I$

Sim ([Sim97]) reviews this approach and comments that schematized belief may be viewed as a semantic counterpart of Schaerf and Cadoli's $\Box_S^1$ epistemic operator ([ScCa95], see §2.2.6), whereas superposed belief is a semantic counterpart of their $\Box_S^3$ operator.

2.2.13 Fusion models

Jaspars ([Jasp91], [Jasp93]) proposed a model which is very similar to the logic of local reasoning described in §2.2.10. He also defines a kind of models (called *fusion models*) that allow the agent to deal with non-trivial inconsistent beliefs. His models are tuples of the form $(S, \pi, R_1, \ldots, R_n)$, where S is a set of states, π assigns a truth value to each basic proposition in each state and R_i is a relation between a state and a set of sets of states (as in the logic of local reasoning). These sets are seen as a kind of *superstates* in which contradictory information may hold. The definition of satisfiability of a doxastic formula in a state is given by this clause:

$$(M, s) \models B_i\phi \Longleftrightarrow \forall T \subseteq S\ (R_i(s, T) \Longrightarrow \exists t \epsilon T\ (M, t) \models \phi)$$

Note that this approach may be considered as dual to the logic of local reasoning described in §2.2.10. That logic looked for a set of doxastic alternatives in which all the members of the set supported the truth of ϕ, whereas the fusion model looks for a world that supports the truth of ϕ in each set of doxastic alternatives.

It is interesting to note that, regarding the different forms of logical omniscience, this approach has the same properties than the local reasoning one. It permits to have inconsistent beliefs without believing everything, and it also avoids closure under implication and closure under conjunction. However, other restricted forms of logical omniscience (such as closure under valid implication, closure under logical equivalence, belief of valid formulæ and weakening of belief) are not avoided in fusion models.

2.2.14 Urn models

Hintikka suggested in [Hint86a] a syntactic approach to the problem of logical omniscience (see §2.1.4). He comments that this approach is equivalent to the semantic approach taken by Rantala ([Rant75]) with his *urn models*.

This approach starts with a generalization of the concept of world, which is a variation of the notion of *urn models* in probability theory, and is called in the same way. The nested quantifiers of a formula represent successive *draws* of individuals from an *urn* (that is, the domain of the model), or successful searches of individuals of the model. The concept of urn model is obtained by allowing the set of individuals to vary between successive draws.

Rantala explains that not all urn models are appropriate for the role of impossible worlds. They are only useful if they vary so subtly between successive draws that they cannot be told apart from the (invariant) classical models with sequences of draws as long as those involved in a given sentence ([Hint75a]). It can be shown that the conditionals $(S_1 \supset S_2)$ which are true in these (almost invariant) urn models are precisely those for which the argument from $\vdash(S_1 \supset S_2)$ to $\{\alpha\}KS_1 \supset \{\alpha\}KS_2$ is allowed with the syntactic restriction mentioned in §2.1.4.

2.2.15 Intensional logic of beliefs

One of the semantic approaches to the problem of logical omniscience is Montague's *intensional logic of beliefs* ([Mont70]). Let W be a set of possible worlds, and assume that the relation that describes whether a formula φ is satisfied in a world w ($w \models \varphi$) has been defined. The *intension* of a formula φ, $I(\varphi)$, is the set of worlds in which it is satisfied. In this context, the semantics of φ is fully determined by its intension. Therefore, if two formulæ φ and ϕ have the same intension they are *semantically equivalent* (i.e. if an agent believes φ, then it also believes ϕ).

With this approach the problems of logical omniscience and perfect reasoning are partially solved, because it is no longer true that if α believes φ and φ semantically implies ϕ, then α believes ϕ. An agent does not have to believe all tautologies, and it can even believe contradictory sentences. Nevertheless, if α believes φ and φ is semantically equivalent to ϕ, then α believes ϕ. This is the price to be paid to see an agent's beliefs as a connective applied to intensions.

The semantic equivalence relation partitions the set of formulæ L into equivalence classes. Therefore, if α believes φ, one could say that α believes $[\varphi]$, where $[\varphi]$ is the equivalence class of φ. There is a natural mapping between equivalence classes and subsets of W, because all the formulæ of a certain class have the same intension. This implies that the agent believes a set of propositions or, equivalently, a set of subsets of W.

A *belief structure* is defined as a tuple $M = (W, N, \Pi)$, where W is the set of possible worlds, Π is a function that returns the intensions of the atomic propositions and N is a function that assigns to each agent the set of propositions that it believes in a certain world. A similar kind of structures are called *Montague-Scott structures* in [FHMV95]. These structures contain a set of states, a truth assignment to the primitive propositions for each state and a set of subsets of S for each state (which is the semantic way of representing the set of formulæ that is believed in each state).

Let L be the language that contains all the atomic propositions $\mathcal{P}$, is closed with respect to the boolean connectives and contains $B_a\varphi$ (a believes φ) if φ is in L and a is an agent. The semantics of this language is the following:

- $M, w \models \mathtt{P}$, where $\mathtt{P} \in \mathcal{P}$, if $w \in \Pi(\mathtt{P})$
- $M, w \models \neg\varphi$ if $M, w \not\models \varphi$
- $M, w \models (\varphi \wedge \psi)$ if $M, w \models \varphi$ and $M, w \models \psi$
- $M, w \models B_a\varphi$ if $\{u: M, u \models \varphi\} \in N(a, w)$

The last line formalizes the idea that an agent believes φ in a world w when the intension of φ is included in the propositions believed by the agent in w. Vardi ([Vard86]) comments this approach and makes some criticisms:

- The notion of possible world is left as a primitive notion, giving no intuitions about the nature of these worlds.
- It is also left open the issue of how to obtain the set W of possible worlds.

2.2.16 Belief worlds

In order to overcome these drawbacks, Vardi ([Vard86]) describes another way of modelling beliefs using propositions. He defines *belief worlds* in a constructive way, and thus he can take W as the set of *all* possible worlds. Therefore, the semantic equivalence relation is the logical equivalence relation. This kind of structures are called *knowledge structures* by Van der Hoek and Meyer in [VdHM96]; they mention that this approach has also been proposed in [FHV84] and [FHV91]. They also show how knowledge based on this kind of structures is completely axiomatized by the system $S5_n$.

Formally, a *0-order assignment* f_0 is defined as an assigment of truth values to the set of atomic propositions. $<f_0>$ is called a *1-ary world*. Assume that *k-ary worlds* ($<f_0, ..., f_{k-1}>$) are defined inductively. Let W_k be the set of all k-ary worlds. A *k-order assignment* is a function f_k that relates each agent to a set of propositions, where each proposition is a set of k-ary worlds. An infinite sequence $<f_0, f_1, f_2, ...>$, where each prefix $<f_0, f_1, ..., f_{k-1}>$ is a k-ary world, is called an *infinitary world*. W_w is the set of all infinitary worlds. Assignments are restricted in the following way: if the k-th level is removed from all k+1-level propositions, all the propositions in level k are obtained.

The notion of satisfiability of a formula in a finitary world is defined as follows:

- $<f_0, ..., f_r> \models \mathtt{P}$, where $\mathtt{P}$ is a primitive proposition, if $f_0(\mathtt{P}) = \text{true}$
- $<f_0, ..., f_r> \models \neg\varphi$ if $<f_0, ..., f_r> \not\models \varphi$
- $<f_0, ..., f_r> \models (\varphi \wedge \psi)$ if $<f_0, ..., f_r> \models \varphi$ and $<f_0, ..., f_r> \models \psi$
- $<f_0, ..., f_r> \models B_a(\varphi)$ if $(r > 0)$ and $\{w : w \epsilon W_r \text{ and } w \models \varphi\}\ \epsilon f_r(a)$

Belief worlds are always extensions of belief worlds from previous levels; therefore, to determine the satisfiability of a formula it is enough to consider a prefix long enough. $Depth(\phi)$ is the number of levels of nesting of the belief operator in ϕ. It can be proved that if a formula of depth k is satisfiable, it is satisfiable in a k+1-ary world, and thus the validity problem for formulæ in beliefs worlds is decidable.

Belief worlds can be characterized with the following axiomatic sytem:

- A1. All the instances of all propositional tautologies.
- R1. From $\varphi \equiv \psi$ infer $B_a(\varphi) \equiv B_a(\psi)$.

Therefore, in Vardi's model of belief worlds, validity is characterized by propositional reasoning plus substitution of equivalents, which somehow alleviates the logical omniscience problem but does not solve it. Vardi suggests two ways of attacking this problem:

- Add *non-classical* worlds to belief worlds.
- Accept that epistemic notions are noy purely intensional, trying to add some syntactic flavour into the semantic part, as Fagin and Halpern did in [FaHa85].

2.2.17 Dynamic epistemic logic

The basic idea underlying Duc's approach ([Duc95], [Duc97]) to the problem of logical omniscience will be strongly defended throughout this dissertation: it is possible to consider *rational* agents that are not *logically omniscient*, just by noticing and enforcing the fact that agents must follow an *explicit* reasoning process in order to obtain logical consequences from its set of beliefs (regardless of whether they have complete or incomplete deductive capabilities). Thus, if an agent has a certain set of beliefs S, and it has enough will, resources and logical capabilities, it will have the chance to deduce any logical consequence of S. Duc formalizes these ideas with the notion of *dynamic epistemic logic*. Two modal operators, $[F_i]$ and $< F_i >$, are introduced for each agent i; they have the following meaning:

- $[F_i]\phi$ means "ϕ is true after any course of thought of i".
- $< F_i > \phi$ means "ϕ is true after some course of thought of i".

Thus, the aim of dynamic epistemic logic is to allow the satisfiability of formulæ such as $K_i\phi \wedge K_i(\phi \Rightarrow \psi) \Rightarrow < F_i > K_i\psi$, that states that, if an agent knows ϕ and $(\phi \Rightarrow \psi)$, it could know ψ in the future (*e.g.* if it applies Modus Ponens to these formulæ). Note the important difference between that formula and $K_i\phi \wedge K_i(\phi \Rightarrow \psi) \Rightarrow K_i\psi$, which leads to logical omniscience. In this way, the modelled agent is both *rational* (because it has

the possibility of using its deductive capabilites to obtain new knowledge) but *not logically omniscient* (because it is not *forced* to know every logical consequence of its knowledge). These F_i operators are not allowed to appear inside the scope of a knowledge operator K_i.

The logic $DES4_n$ (Dynamic Epistemic $S4_n$) has the following axiomatic definition:

- PC1: $\phi \Rightarrow (\psi \Rightarrow \phi)$
- PC2: $(\phi \Rightarrow (\psi \Rightarrow \gamma)) \Rightarrow ((\phi \Rightarrow \psi) \Rightarrow (\phi \Rightarrow \gamma))$
- PC3: $(\neg\psi \Rightarrow \neg\phi) \Rightarrow (\phi \Rightarrow \psi)$
- TL1: $[F_i](\phi \Rightarrow \psi) \Rightarrow ([F_i]\phi \Rightarrow [F_i]\psi)$
- TL2: $[F_i]\phi \Rightarrow [F_i][F_i]\phi$
- DE1: $K_i\phi \wedge K_i(\phi \Rightarrow \psi) \Rightarrow < F_i > K_i\psi$
- DE2: $K_i\phi \Rightarrow \phi$
- DE3: $K_i\phi \Rightarrow [F_i]K_i\phi$, provided that ϕ does not contain any F_i operator
- DE4: $< F_i > K_i(\phi \Rightarrow (\psi \Rightarrow \phi))$
- DE5: $< F_i > K_i((\phi \Rightarrow (\psi \Rightarrow \gamma)) \Rightarrow ((\phi \Rightarrow \psi) \Rightarrow (\phi \Rightarrow \gamma)))$
- DE6: $< F_i > K_i((\neg\psi \Rightarrow \neg\phi) \Rightarrow (\phi \Rightarrow \psi))$
- DE7: $< F_i > K_i(K_i\phi \Rightarrow \phi)$
- DE8: $K_i\phi \Rightarrow < F_i > K_iK_i\phi$, provided that ϕ does not contain any F_i operator
- R1 (Modus Ponens): from ϕ and $(\phi \Rightarrow \psi)$, infer ψ
- R2 (Necessitation): from ϕ infer $[F_i]\phi$

The axiom schemas PC1, PC2 and PC3, along with rule R1, axiomatize propositional calculus. The axioms DE4, DE5 and DE6 are expressing the fact that agents may use this propositional axiomatization in their reasoning. The addition of TL1, TL2 and R2 permits to axiomatize the minimal temporal logic of transitive time. Axiom DE1 has been commented above (agents may use Modus Ponens). DE2 is the classical axiom of knowledge, axiom T (recall §1.2.3). DE3 states that the knowledge of an agent is persistent (*i.e.* agents do not forget what they know). DE7 forces agents to trust in their knowledge. Finally, DE8 says that agents are potentially capable of performing positive introspection.

Duc proves that the agents modelled with this axiomatic system do not suffer from omniscience problems such as knowledge of valid formulæ, knowledge of equivalent formulæ or closure of knowledge under logical implication.

2.3 Unifying frameworks

Some authors have developed frameworks that generalize some of the proposals that have just been described; we will briefly comment Wansing's *non-normal worlds* semantics, Giunchiglia's *multi-context systems* framework and Sim's *multi-valued epistemic logic*.

2.3.1 Non-normal worlds

Wansing shows in [Wans90] how different logics of knowledge and belief may be seen as particular cases of a semantics defined in a general possible worlds framework. This semantics is defined as follows. Let L be the language of multi-modal propositional logic. A *Rantala model* M is a tuple of the form $< W, W^*, R_1, \ldots, R_n, V >$, where

- W is the non-empty set of *normal* worlds,
- W^* is the set of *non-normal* worlds,
- R_i are binary relations on $W \bigcup W^*$ and
- V is a function that assigns a boolean value to each formula of L in each world, such that, $\forall w \epsilon W$, the following conditions hold:
  - $V(\phi \wedge \psi, w) = 1 \Leftrightarrow V(\phi, w) = V(\psi, w) = 1$

- $V(\neg\phi, w) = 1 \Leftrightarrow V(\phi, w) = 0$
- $V(\Box_i\phi, w) = 1 \Leftrightarrow (\forall w' \epsilon W \bigcup W^*)(R_i w w' \Rightarrow V(\phi, w') = 1)$

Note that these conditions are defined only on normal worlds; moreover, the truth or falsity of formulæ need not be recursively specified in non-normal worlds (non-normal worlds are only considered in the truth condition of modal formulæ). A formula ϕ is *true in a model M at a world w* iff $V(\phi, w) = 1$. A formula is *valid in a model* if it is true in all normal worlds of the model. It is *valid* if it is valid in all models. Wansing shows how Levesque's *logic of explicit and implicit belief* ([Leve84], see §2.2.4), Fagin and Halpern's *logics of awareness, general awareness and local reasoning* ([FaHa85], see §2.2.7 and §2.2.10) and van der Hoek and Meyer's *logic of awareness and principles* ([VdHM88], see §2.2.8) are especial cases of this unifying framework.

2.3.2 Multi-context systems

Giunchiglia *et al.* ([GSGF93], [Bene97]) have developed *multi-context systems*, that provide a framework in which it is possible to formalize the reasoning about belief that takes place in a multi-agent environment. They represent the agent's beliefs with a *context* (an axiomatic system), in a *deduction model of belief* fashion (see §2.1.2). This idea allows them to provide an exhaustive classification of all the ways in which an agent may happen to be non-ideal, to be *incomplete*:

- *Incompleteness in the signature:* an agent may not be aware of all basic propositions (this idea was also the motivation of the *logic of general awareness*, see §2.2.7).
- *Incompleteness in the formation rules:* an agent may not be aware of all the rules that may be used to construct formulæ from basic propositions, or may lack the resources (space, time) needed to build formulæ beyond a given complexity (*e.g.* formulæ with many basic propositions, or with many nested quantifiers).
- *Incompleteness in axioms:* the previous kinds of incompleteness may prevent a (real) agent from having (in the axiomatic system that implements it) all the axioms that it should have in order to be a perfect realisation of an ideal agent.

- *Incompleteness in deduction rules:* an agent may not know all the deductive rules that are necessary in order to perform a logical closure of its set of axioms (this kind of incompleteness was also commented in §2.1.2). It may also know all the relevant rules but be unable to apply them (*e.g.* when a certain resource has been exhausted).

In their multi-context systems every agent is modelled with an axiomatic system, called a *context*. The reasoning that is performed over the beliefs of these agents is also modelled with another axiomatic system. The interaction between the different systems is made possible with the use of *bridge rules*. The intuitive semantics of a bridge rule such as $((\phi, C_1) \Rightarrow (\psi, C_2))$ is the following: if ϕ is believed in context C_1, then ψ is believed in context C_2. Their basic system models logically omniscient agents, but they are also able (see [GSGF93]) to model non-ideal agents; in particular, their framework subsumes the *logic of explicit and implicit belief* (see §2.2.4), the *logic of general awareness* (see §2.2.7) and the *logic of local reasoning* (see §2.2.10).

2.3.3 Multi-valued epistemic logic

Sim has also designed a framework that subsumes some of the proposals reviewed in this chapter ([Sim95], [Sim00]). His idea is to define an epistemic logic in which atomic sentences are assigned truth values that belong to a bilattice ([Gins88]). A *bilattice* is a set with two partial orderings $\leq_t$ and $\leq_k$. Given two truth values x and y, $(x \leq_t y)$ means that y is at least as true as x, and $(x \leq_k y)$ means that the evidence underlying an assignment of the truth value x is subsumed by the evidence underlying an assignment of the truth value y. In particular, Sim considers a basic bilattice, B_4 (see figure 2). In the bilattice B4, the value $\bot$ means that a sentence is neither true nor false (*i.e.* there is no information about it), whereas the value $\top$ indicates that a sentence is both true and false (*i.e.* there is both positive and negative evidence for the sentence). Sim is interested in *interlaced bilattices* ([Fitt91]), which are lattices that have a unary operator, $-$, called *conflation*, with the following properties:

- $x \leq_t y \Rightarrow -x \leq_t -y$
- $x \leq_k y \Rightarrow -y \leq_k -x$

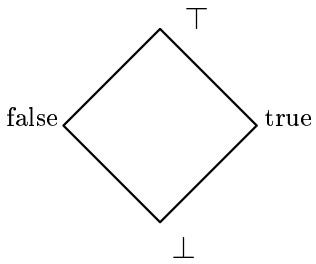

Figure 2: Bilattice B4

- $--x = x$

With this definition in mind, a *multi-valued epistemic logic* (MEL) is defined as follows: a model of MEL based on a set P of primitive propositions is a tuple $M =< S, R_1, \ldots, R_n, v, f >$, in which

- S is a set of situations.
- R_i are binary accessibility relations on S.
- v is a function that assigns a truth value in $B4$ to each primitive proposition in each situation. $v =< v^+, v^- >$, where v^+ and v^- assign an standard truth value (0 or 1) to each primitive proposition. The four possible combinations -(0,0), (0,1), (1,0) and (1,1)- represent the $B4$-values $\bot$, $false$, $true$ and $\top$, respectively.
- f is a unary function such that, for any situation $s \epsilon S$, $f(f(s)) = s$ and $v(f(s))(p) = -v(s)(p)$ for any $p \epsilon P$.

W is the set of *dual* situations (those situations s such that $s = f(s)$). The epistemic operators L_i and B_i (recall §2.2.5) are considered over *dual* (*i.e.* complete) and arbitrary situations, respectively, as shown in the following clauses:

- $M, s \models_t p$ iff $v^+(s)(p) = true$
- $M, s \models_f p$ iff $v^-(s)(p) = true$

- $M, s \models_t \neg\phi$ iff $M, s \models_f \phi$
- $M, s \models_f \neg\phi$ iff $M, s \models_t \phi$
- $M, s \models_t (\phi \wedge \psi)$ iff $M, s \models_t \phi$ and $M, s \models_t \psi$
- $M, s \models_f (\phi \wedge \psi)$ iff $M, s \models_f \phi$ or $M, s \models_f \psi$
- $M, s \models_t L_i\phi$ iff $\forall t \epsilon W\ (sR_it) \Rightarrow M, t \models_t \phi$
- $M, s \models_f L_i\phi$ iff $\exists t \epsilon W\ (sR_it)$ and $M, t \models_f \phi$
- $M, s \models_t B_i\phi$ iff $\forall t \epsilon S\ (sR_it) \Rightarrow M, t \models_t \phi$
- $M, s \models_f B_i\phi$ iff $\exists t \epsilon S\ (sR_it)$, $(M, t \models_f \phi)$ and $(sR_it) \Leftrightarrow (f(s)R_if(t))$

Sim ([Sim00]) argues that the agents modelled in this framework keep the desirable non-omniscient properties present in Levesque's *logic of explicit and implicit beliefs* (recall §2.2.4). He also claims that this approach subsumes not only the (standard and nested) *logics of explicit and implicit belief* (§2.2.4 and §2.2.5), but also the *logic of general awareness* (§2.2.7) and Schaerf and Cadoli's *approximate knowledge* framework (§2.2.6).

2.4 Summary

The *possible worlds* model ([Hint62]) and its associated ([Krip63b]) *Kripke semantics* have been used extensively in the last decades to give a formal semantics to the modal formulæ of doxastic and epistemic logics. The success attained by this model is based in facts such as the following:

- It provides a very intuitive approach to the processes of reasoning about knowledge and belief that an intelligent agent might carry out.
- The axioms that rule the behaviour of knowledge are closely linked to the properties of the accessibility relation between possible worlds ([VBen84]).

In spite of these positive facts, the commitment to the possible worlds model and the Kripke semantics makes it difficult to model the reasoning processes of non-ideal agents, because the modelled agents are *logically omniscient* (they believe every tautology) and *perfect reasoners* (they believe every logical consequence of their beliefs). This problem was detected and labelled by Hintikka in [Hint75a], and many authors (in Philosophy, Computer Science and Artificial Intelligence) have provided alternative formalisms that solve (or, at least, partially alleviate) this problem. In this chapter the following strategies have been considered:

- Revise the concept of possible world.

  In the classical view, a possible world is a logical model (for instance, in a propositional setting, a possible world is an interpretation of the basic propositions). Some authors have tried to alleviate logical omniscience by changing this notion. A possible world may be represented in different ways, for instance the following:

  - A situation in which some basic propositions are considered true and some basic propositions (not necessarily different from the previous ones) are considered false (*logic of implicit and explicit beliefs*, [Leve84], [Lake87]).
  - A *partial* assignment of truth values to the basic propositions (*hybrid sieve systems*, [Thij96]).
  - A state in which tautologies may be false and contradictions may be true (*impossible possible worlds*, [Cres72], [ReBr79], [Hint75a]).
  - A possible world may be represented just with the set of formulæ that are assumed to be true in it ([Perl84], [Haas85]).

- Restrict the agent's deductive capabilities.

  An agent may be prevented from having too many beliefs by restricting its reasoning capabilities. In the *deduction model of belief* ([Kono86a]) and the *belief model* ([Wool95]) each agent has its own set of deductive rules, that may not be logically complete. This solution has also been strongly advocated by Hintikka ([Hint86a]).

- Change the underlying consequence relationship.

  The most important approach that suggests a way in which the consequence relationship may be changed is the one described in §2.2.3, the *non-standard structures* ([FHV95]). In this framework the interpretations of a formula and its negation are made independent; thus, a new semantics is provided to the negation operator and, therefore, also to the conditional operator.

- Modify the standard Kripke semantics.

  The standard Kripke semantics may also be modified in a number of ways to avoid logical omniscience. The standard clause that provides a semantics to the doxastic formulæ may be either generalised (by allowing extra formulæ to be beliefs, as in the *logic of principles and implicit beliefs*, [VdHM96]) or specialized (by requering extra conditions in order for a formula to be a belief, as in the *logic of general awareness*, [FaHa85]). The doxastic operator may be modelled with the modal possibility operator, rather than with the necessity one ([VdHM89]). A different way of changing the Kripke semantics is to put *frames of mind* into the picture, as in the *logic of local reasoning* ([FaHa88]), the *logic S5P* ([MvdH93]) or the *fusion models* ([Jasp93]).

Detailed reviews of some of the approaches that have been surveyed in this chapter appear in [McPa87], [Hadl88], [McAr88], [Reic89], [VdHM95] [FHMV95], [Sim97] and [MvdH98].

3 Subjective situations

In this chapter we will suggest a new way of dealing with the logical omniscience and the perfect reasoning problems, based on the notion of *subjective situations*; these entities will replace the classical possible worlds. First, we will mention in §3.1 the basic ingredients to be used in our approach. *Subjective situations* will be intuitively motivated in §3.2, and formalised in a concrete first-order doxastic modal setting in §3.3 and §3.4. Having done that, in §3.5 the main properties of the modal operators of this doxastic logic will be studied; it will be seen how logical omniscience is avoided, while some interesting logical properties (such as introspection) are maintained. The chapter will finish with the comparison of our proposal with some of the frameworks that have been reviewed in the previous chapter.

3.1 Ways of avoiding logical omniscience

At the end of §2 we have summarised a number of ways in which the logical omniscience problem may be alleviated. Our approach is a blend of some of those ideas, as will be apparent in the rest of the dissertation. On the one side, the description of what a *possible world* is will be changed (§3.3); we will consider *subjective* descriptions of possible states of affairs, depending on each agent's perception of them. On the other side, the Kripke semantics will also be somewhat modified in our approach, in order to consider the vision that each agent has of the situation in which it is located (§5.4). The study presented in §2 shows clearly that it is quite difficult to find a natural way of dealing with logical omniscience without modifying any of these two aspects. Finally, in §4 a certain class of non-ideal reasoners, the *rational inquirers*, will be defined; as will be shown in that chapter, the reasoning capabilities of these agents will be restricted (with respect to classical logic) in order to prevent them from being ideal reasoners.

3.2 Motivation of subjective situations

The most popular way of dealing with the logical omniscience issue is to change the concept of what a *possible world* is, as shown in §2. Regardless of the way in which the concept of possible world is modified, there is a kernel that never changes: the formal representation of a possible world is

not related in any way with the notion of *agent*. Thus, it may be said that all the approaches in the literature present an *objective* view of what a possible world is (*i.e.* a world is the same for all the agents, it is independent of them). In an standard *Kripke structure*, the only item that depends on each agent is its own *accessibility relation* between possible worlds (recall definition 1).

The traditional meaning assigned to the accessibility relation R_i of an $Agent_i$ is that it represents the uncertainty that $Agent_i$ has about the situation in which it is located (*e.g.* $(w_0 R_5 w_1)$ means that $Agent_5$ cannot distinguish between worlds w_0 and w_1, see [FHMV95]). This situation is quite peculiar, because the formulæ that are true in two worlds that are linked by an accessibility relation are, in principle, totally unrelated (*i.e.* in a Kripke structure there is no relationship between the accessibility relation between states and the function that assigns truth values to the basic propositions in each of them). This fact is also pointed out in [Alva98], where it is argued that the accessibility relations should follow Kripke's original idea that two worlds w and w' should be related whenever every proposition that is true in w is also true in w' ([Krip59], [Krip63a]). In that way, w' would be an alternative state conceivable *on the basis of the present world* w.

Our proposal is quite different, and it may be intuitively motivated by the following scenario (the rest of the chapter will be of a more formal nature). Imagine two human agents (α and β) that are watching a football match together. In a certain play of the game, a fault is made and the referee awards a penalty kick. α thinks that the referee is right, because it has noticed that the fault was made inside the penalty area (let us represent this fact with proposition $\mathtt{P}$); at the same time, β is thinking that the referee was wrong because, in its perception of the situation, the fault was made just an inch outside the penalty area. How can this situation (and the beliefs of the two agents) be formally represented?

Following the standard approach, we could model the fact that α believes $\mathtt{P}$ and β believes $\neg\mathtt{P}$ by assuming that in all the (objectively described) worlds considered as possible in the current state by α the proposition $\mathtt{P}$ holds, whereas in all the worlds considered as possible by β (β's doxastic alternatives) $\mathtt{P}$ is false. This account of each agent's doxastic state does not seem very satisfactory to us, at least for two reasons:

- It does not tell us how each agent's perception of the situation influences in its own beliefs. Recall that we are interested in modelling the

evolution of the beliefs of agents with bounded rationality ([NeSi72]). This kind of agents have a finite set of resources available to them. The standard modal approach, in which an agent conceives a set of complete and consistent possible worlds as doxastic alternatives, does not seem to us to have any realistic interpretation. If an agent α believes $\mathtt{P}$ and $(\mathtt{P} \Rightarrow \mathtt{Q})$, what sense does it make to model those beliefs as if they had been obtained by the agent from an intersection of the formulæ holding in an infinite set of completely specified possible worlds? Furthermore, as we noted in the previous chapter, taking this stance leads to logical omniscience, as the agent is also forced to believe every logical consequence of those two formulae, such as $\mathtt{Q}$ or $(\mathtt{P} \vee \neg\mathtt{P})$. We defend a much simpler picture: the fact that α believes $\mathtt{P}$ and $(\mathtt{P} \Rightarrow \mathtt{Q})$ only means that it has perceived those facts as true in its present situation, and it has used these formulæ to build a partial description of its present state of affairs. It is a framework in which the agent may keep a partial description of the situation in which it is located, and in which it can use the facts that it keeps perceiving from the environment in order to keep increasing and refining its beliefs. It is even possible to give a concrete interpretation of the positive and negative formulæ that constitute the agent's subjective description of a situation: they could be just data structures stored in the agent's memory, that would be updated by its perception of the environment. Huang *et al.* have also noted that, in the presence of bounded rationality, the standard approach to representing incomplete information is to have *partial* descriptions (that, due to this partiality, encompass *classes* of possible worlds), rather than having complete descriptions of uniquely identified possible worlds ([HMP96]). This idea will be very important in our work, as will become apparent in §5, when classes of possible situations will be represented with two finite lists of formulæ that will provide a partial description of the facts that hold (or do not hold) in those situations.

- Assuming that the fault was indeed made inside the penalty area, most philosophers would argue that α not only *believes* $\mathtt{P}$ but also *knows* it (being $\mathtt{P}$ true in the real world), whereas β believes $\neg\mathtt{P}$ but can not possibly know $\neg\mathtt{P}$, being it actually false[4]. Thus, in a *magical* way, one

[4] It could be argued that we are somehow neglecting in this argumentation the need of

agent would have some knowledge (that would coincide with reality) whereas the other wouldn't.

In our opinion, this state of affairs (the actual situation, comprising both the football match and the agents, along with their beliefs) may not be adequately described with a simple assignment of truth values to the basic propositions. Even if we had an accurate description of the real world, does it really matter very much whether the fault was made inside the penalty area in order to model the beliefs of the two agents involved in the scene?

The situation (s) is obviously the same for the two agents α and β (they are watching the same match together). From α's point of view, the description of s should make true proposition $\mathtt{P}$; however, from β's perspective, in the present situation $\mathtt{P}$ should be considered false. Obviously, there would be many aspects of s in which α and β would agree; *e.g.* both of them would consider that the proposition representing the fact *"We are watching a football match on TV"* is true in s.

As far as beliefs are concerned, we argue that, in this situation, α should be capable of stating that $B_\alpha \mathtt{P}$ (α has seen the fault and has noticed that it was made inside the penalty area; thus, it believes so). It would not seem very acceptable a situation in which α perceived the fault to have been made inside the penalty area and defended that it did not believe that a penalty kick should have been awarded (the only possible explanation being that α is a strong supporter of the offending team). It also seems reasonable to say that α cannot fail to notice that it believes that the fault was made inside the penalty area; thus, α may also assert in s that $B_\alpha B_\alpha \mathtt{P}$. In a similar way, in this situation β cannot state that $B_\beta \mathtt{P}$ (β cannot defend that it believes that the referee is right, in a situation in which it perceived the fault to have been made outside the penalty area). Thus, it seems clear that each agent's point of view on a situation strongly influences (or we could say even *determines*) its positive and negative beliefs in that situation.

In our framework we want to include the intuition that agents are smart enough to know that other agents may not perceive reality in the same way as they do. In the previous example, without further information (*e.g.* α shouting *"Penalty!"*), β should not be capable of supporting (or rejecting)

---

a *justification* for the belief in order for it to become knowledge (as knowledge is usually defined in the philosophical literature as *true justified belief*). But, what could possibly count more as a justification that each agent's own on-site perception of the situation?

that B_αP; analogously, α could not affirm (or deny) that B_βP. That means that the communication between the agents is the main way in which an agent may attain beliefs about other agent's beliefs[5]. We could have chosen other alternatives; for instance, we could have stated that an agent believes that the other agents perceive reality in the same way as it does, provided that it does not have information that denies that fact. In that case, in the example α would assume that β also believes that P is true, as far as it does not have any reason not to think so (*e.g.* β saying *"This referee is really blind"*).

A final reflection on the meaning of the accessibility relation between situations for $Agent_i$ (R_i) is necessary. It will be assumed that an agent cannot have any doubts about its own perceptions and beliefs in a given state. *E.g.* if, in situation s, α looks at the match and thinks P, then it surely must realise this fact and believe P in s (and even believe that it believes P, were it to think about that). Thus, if R_α links s with all those situations that α cannot tell apart from s, it must be the case that α also perceives P as true in all those states as well (otherwise, those states would be clearly distinguishable by α, because in some of them it would support P whereas in some of them P would be rejected). The only uncertainty that α may have is *about the perception of s by the other agents*. In the example, α does not know whether it is in a situation in which β supports P or in a situation in which β rejects P. Therefore, α's accessibility relation must reflect this uncertainty.

Summarising, the main points that have been illustrated with the previous discussion are the following:

- A situation may be considered not as an entity that may be objectively described, but as a piece of reality that may be perceived in different ways by different agents.

  Thus, it is necessary to think of a *subjective* way of representing each situation, in which each agent's point of view is taken into account. In the previous example, the description of s should include the fact that α is willing to support P, whereas β isn't.

- An agent's beliefs in each situation also depend on its point of view.

[5]However, as will be seen in §4 and §5, in this dissertation we will only be concerned with the agent's beliefs about the world, and not with its beliefs about other agent's beliefs.

In the situation of the example, $B_\alpha \mathtt{P}$ would hold from α's perspective, whereas it would not be either supported or rejected by β. Thus, we argue that it does not make sense to ask whether $B_\alpha \mathtt{P}$ holds in s or not; that question must be referred *to a particular agent's point of view*. We have also defended that the beliefs of an agent depend more on its *perception* of reality than in what is actually true or false in its environment.

- The interpretation of the meaning of each agent's accessibility relation is slightly different from the usual one.

  Each accessibility relation R_i will keep its traditional meaning, *i.e.* it will represent the uncertainty of $Agent_i$ with respect to the situation in which it is located. However, our intuition is that an agent may only be uncertain about the other agents' perception of the present state, not about its own perception.

3.3 Formalization of subjective situations

These intuitive ideas are formalized in the *structures of subjective situations*, which are defined as follows ([MCS99b], [MCS00b]):

Definition 3 (Structure of Subjective Situations)

An structure of subjective situations for n agents E is a tuple

$$< S, R_1, ..., R_n, \mathcal{T}_1, ...\mathcal{T}_n, \mathcal{F}_1, ..., \mathcal{F}_n >, \textit{ where}$$

- *S is the set of possible situations.*
- *R_i is the binary accessibility relation between situations for $Agent_i$.*
- *$\mathcal{T}_i$ is a function that returns, for each situation s, the set of first-order formulæ that are perceived as true by $Agent_i$ in s.*
- *$\mathcal{F}_i$ is a function that returns, for each situation s, the set of first-order formulæ that are perceived as false by $Agent_i$ in s.*

$\mathcal{E}$ is the set of all structures of subjective situations.

Note that this definition does not put any constraints on the $\mathcal{T}_i$ and $\mathcal{F}_i$ functions. We may have situations such as the following:

- $\phi \epsilon \mathcal{T}_i(s)$ and $\phi \epsilon \mathcal{F}_i(s)$.

  In this case s is an *inconsistent* situation, because $Agent_i$ has reasons both to support and to reject ϕ. With this kind of situations it is possible to model for instance states of affairs in which an agent has received contradictory information from different sources.

- $\phi \not\epsilon \mathcal{T}_i(s)$ and $\phi \not\epsilon \mathcal{F}_i(s)$.

  In this case s is a *partial* situation, because $Agent_i$ does not have any reason to support or to reject ϕ. This kind of situations might be used to model an state of affairs in which the agent is unaware of the existence of a certain predicate, or simply it does not have any positive or negative evidence about its truth.

- $(\phi \vee \neg\phi) \not\epsilon \mathcal{T}_i(s)$

  Agents do not have to necessarily support all classical tautologies.

- $\phi \epsilon \mathcal{T}_i(s)$, $(\phi \Rightarrow \psi) \epsilon \mathcal{T}_i(s)$, $\psi \not\epsilon \mathcal{T}_i(s)$.

  The formulæ that are supported by an agent do not have to be closed under logical consequence.

- $\forall x(P_x \Rightarrow Q_x) \epsilon \mathcal{T}_i(s)$, $P_a \epsilon \mathcal{T}_i(s)$, $Q_a \not\epsilon \mathcal{T}_i(s)$.

  Another example of the same fact, now in the first-order case. These situations are basic to formalise non-ideal agents, which may have to devote some effort and resources to derive some logical consequences of their sets of beliefs (even if they seem so trivial to obtain as the ones we have just shown).

This kind of situations was already considered by Levesque in his *logic of explicit and implicit beliefs*, described in §2.2.4. A detailed comparison of our proposal and that of Levesque is offered in §3.6.

The accessibility relation between situations for $Agent_i$ has to reflect its uncertainty about the way in which the actual situation is perceived by the other agents. Thus, R_i has to link all those states that $Agent_i$ perceives in the same way but that may be perceived in different ways by other agents. This intuition is formalized in the following condition:

Definition 4 (Condition on Accessibility Relations)

$$\forall s,t \epsilon S, (sR_it) \textit{ if and only if } (\mathcal{T}_i(s) = \mathcal{T}_i(t)) \textit{ and } (\mathcal{F}_i(s) = \mathcal{F}_i(t))$$

It is easy to check that this condition forces the accessibility relations to be *equivalence relations* (as they are reflexive, symmetric and transitive). This result links this approach with the classical $S5$ modal system, in which this condition also holds. In $S5$ the presence of this condition makes true axiom 4 (positive introspection), axiom 5 (negative introspection) and axiom T (the axiom of knowledge); the modal operators of the system proposed in this dissertation will have similar properties, as will be shown in §3.5.

3.3.1 Managing uncertainty

As stated above, in the standard modal approach the agents' uncertainty is reduced to the accessibility relations R_i of the Kripke structures. As these relations define which are the doxastic alternatives of each agent, they determine, *via* the Kripke semantics, which will be the agents' sets of beliefs. In the framework suggested in this dissertation, there are two types of uncertainty that have to be dealt with:

- On the one hand, the agent's perception of a situation may be both *partial* (because it may not have information about every possible fact) and *inconsistent* (because it may have reasons to believe and disbelieve that a certain fact holds in a given situation). This kind of uncertainty is represented by the functions $\mathcal{T}_i$ and $\mathcal{F}_i$ of the structures of subjective situations.

- On the other hand, the agent is also uncertain about the perception of the current situation by the other agents composing the multi-agent system. The accessibility relations of the structures of subjective situations are used precisely to model this fact.

We believe that these two types of uncertainty may not be appropriately modelled within the standard modal approach, in which the agent may only be uncertain about which one is its present situation, among a set of *complete* and *consistent* possible worlds. This model is quite natural and intuitive, but it must be abandoned if the aim is to model the beliefs of rational, real, limited, non-ideal agents.

3.4 Satisfiability relations

Most of the approaches to logical omniscience reviewed in the previous chapter considered propositional modal logics of belief. In our work, we will turn to the predicate case. In the classical tradition, the formulae of a first-order doxastic language are constructed as follows:

Definition 5 (Language of first-order modal logic)

The formulæ F of the standard language of predicate doxastic logic may be built in the following way:

- $F \Rightarrow$ *Standard formulæ of the predicate calculus.*
- $F \Rightarrow (\neg F) \mid (F \vee F) \mid (F \wedge F)$
- $F \Rightarrow B_i F$

Note that this definition does not allow the presence of modal formulæ within the scope of a quantifier[6]. A simplified version of the doxastic first-order language for n agents is considered in this dissertation, as shown in the following definition:

Definition 6 (Doxastic Modal Language $\mathcal{L}$)

Consider a set of modal belief operators for n agents (B_1, ..., B_n). $\mathcal{L}$ is the language formed by all first-order formulæ (built in the standard way from a set of predicates, constants, variables, the quantifiers ($\exists$, $\forall$) and the logical operators ($\neg, \vee, \wedge, \rightarrow$)), preceded by a (possibly empty) sequence of (possibly negated) modal operators. $\mathcal{L}_{PC}$ is the subset of $\mathcal{L}$ that contains those formulæ that do not have any modal operator. The modal formulæ of $\mathcal{L}$ are called linearly nested.

Thus, $\mathcal{L}$ contains formulæ such as $\mathtt{P}$, $B_3\mathtt{Q}$, $B_1B_5(\mathtt{R} \vee \mathtt{T})$, $B_3\neg B_2\mathtt{S}$ and $\neg B_1B_1\neg\mathtt{T}$, but it is not expressive enough to represent formulæ such as $(B_2\mathtt{P} \rightarrow B_3\mathtt{Q})$ or $(\mathtt{P} \vee B_5\mathtt{Q})$. In most practical applications, an agent in a multi-agent system will only need to represent what it believes (or not) to

[6]Konolige discusses in [Kono86a] the implications of this constraint, and the severe logical difficulties that arise when it is abandoned.

be the case in the world and what it believes (or not) that the other agents believe (or not). This is just the level of complexity offered by *linearly nested* formulæ.

In an structure of subjective situations each $Agent_i$ has *positive* and *negative* information about some first-order formulæ in each situation (given by $\mathcal{T}_i$ and $\mathcal{F}_i$, respectively). This allows us to define two relations (of satisfiability, $\models_i$, and unsatisfiability, $=\!|_i$) between situations and formulæ *for each* $Agent_i$. Given an structure of subjective situations E and a situation s, the expression $E, s \models_i \phi$ should hold whenever $Agent_i$ has some reason to think that ϕ is true in situation s. Similarly, $E, s =\!|_i \phi$ should hold whenever $Agent_i$ has some reason to reject ϕ in situation s.

Notice that $E, s \not\models_i \phi$ should not imply that $E, s =\!|_i \phi$ (i.e. $Agent_i$ not having any reason to support ϕ does not mean that it must have reasons to reject it). In the same spirit, $E, s \models_i \phi$ should not imply that $E, s \neq\!|_i \phi$ ($Agent_i$ could have reasons both to support and to reject a certain formula in a given situation). These facts will indeed be true, as will be seen in the next section, due to the presence of partial and inconsistent situations commented in §3.3.

The clauses that define the behaviour of these relations are shown in the following definition:

Definition 7 (Relations $\models_i$ and $=\!|_i$)

- $\forall E \epsilon \mathcal{E}, \forall s \epsilon S, \forall Agent_i, \forall \phi \epsilon \mathcal{L}_{PC}$

$$E, s \models_i \phi \Leftrightarrow \phi \epsilon \mathcal{T}_i(s)$$
$$E, s =\!|_i \phi \Leftrightarrow \phi \epsilon \mathcal{F}_i(s)$$

- $\forall E \epsilon \mathcal{E}, \forall s \epsilon S, \forall Agent_{i,j}, \forall \phi \epsilon \mathcal{L}$

$$E, s \models_i B_j \phi \Leftrightarrow \forall t \epsilon S \ ((sR_it) \ implies \ E, t \models_j \phi)$$
$$E, s =\!|_i B_j \phi \Leftrightarrow \exists t \epsilon S \ ((sR_it) \ and \ E, t =\!|_j \phi)$$

- $\forall E \epsilon \mathcal{E}, \forall s \epsilon S, \forall Agent_{i,j}, \forall \phi \epsilon \mathcal{L}$

$$E, s \models_i \neg B_j \phi \Leftrightarrow E, s =|_i B_j \phi$$
$$E, s =|_i \neg B_j \phi \Leftrightarrow E, s \models_i B_j \phi$$

A first-order formula ϕ is supported in a given situation s by an $Agent_i$ if and only if $Agent_i$ has reasons to think that ϕ is true in s. Analogously, ϕ will be rejected if and only if there are reasons that support its falsehood (recall that a formula may be both supported and rejected in a given situation). As far as beliefs are concerned, in a given situation s $Agent_i$ supports that $Agent_j$ believes ϕ just in case $Agent_j$ supports ϕ in all the situations that are considered possible by $Agent_i$ in s ($Agent_i$'s doxastic alternatives). Similarly, $Agent_i$ may reject the fact that $Agent_j$ believes ϕ if it may think of a possible situation in which $Agent_j$ rejects ϕ. Finally, $Agent_i$ will support that $Agent_j$ does not believe ϕ if it may reject the fact that $Agent_j$ believes ϕ. We do not need more clauses to define the behaviour of the satisfiability and unsatisfiability relationships due to the restriction to linearly nested formulæ imposed in definition 6.

3.4.1 Derivability and validity

We will briefly discuss in this section the properties of the logical notions of derivability and validity that are induced from the satisfiability relationship that has just been explained. Let us consider the following definitions of these concepts:

Definition 8 (Derivability and Validity)

Being Γ a set of linearly nested formulæ, we represent with the expression $M, s \models_i \Gamma$ the fact that $\forall \gamma \in \Gamma$, the expression $M, s \models_i \gamma$ holds.

- *A linearly nested formula ψ is* i-derivable *from a set of linearly nested formulæ Γ, denoted $\Gamma \models_i \psi$, if and only if*

$$\forall \text{ structures of subjective situations } M, \forall \text{ situations } s,$$
$$(M, s \models_i \Gamma) \Longrightarrow (M, s \models_i \psi)$$

- *Two linearly nested formulæ ϕ and ψ are called* i-equivalent *if ϕ is i-derivable from $\{\psi\}$ and ψ is i-derivable from $\{\phi\}$.*
- *A linearly nested formula ψ is* i-valid, *denoted $\models_i \psi$, if and only if*

$$\forall \text{ structures of subjective situations } M, \forall \text{ situations } s, \\ M, s \models_i \psi \text{ holds}$$

This is arguably the most natural way of defining validity and derivability in the *subjective situations* framework. The analysis of validity is trivial, as shown in the following proposition:

Proposition 2 (Valid formulae)

There does not exist any i-valid *linearly nested formula.*

It is easy to check that there is not any *i-valid* formula. We only need to consider an structure of subjective situations with a single world, w, such that $\mathcal{T}_i(w)$ and $\mathcal{F}_i(w)$ are empty. Then, there would be no linearly nested formula ψ such that $M, w \models_i \psi$. As there are no valid formulae, we do not have to worry about agents believing all valid formulæ or having their beliefs closed under valid implication.

With respect to derivability, the following proposition holds:

Proposition 3 (Characterization of predicate derivability)

For all sets of predicate formulæ Γ and all predicate formulæ γ,

$$\Gamma \models_i \gamma \text{ if and only if } \gamma \in \Gamma.$$

The proof of this proposition is quite straightforward. This result is stating that, if we take all the situations and structures in which a given set of first-order formulæ hold, we can only expect those formulæ to hold, and there would be no other formula (neither classical tautologies nor classical logical consequences of those formulæ) satisfied in those structures and situations. This result is precisely stating that the agents modelled with this framework are neither logically omniscient nor perfect reasoners, as we desired.

If we turn our attention to linearly nested formulæ, we have the following result:

Proposition 4 (I-equivalence of linearly nested formulae)

For any linearly nested formula ϕ,

- *ϕ is i-equivalent to $B_i\phi$*
- *$\neg\phi$ is i-equivalent to $\neg B_i\phi$*

This result is a direct consequence of some of the propositions that will be proved in §3.5, and will be discussed with more detail there. In a nut shell, it is formally stating the intuitions that we suggested at the beginning of this chapter: $Agent_i$'s positive and negative beliefs will be determined by its perception of reality.

3.5 Properties of the belief operators

The definition of an structure of subjective situations, the fact that the accessibility relations are equivalence relations and the clauses that describe the behaviour of the satisfiability (and unsatisfiability) relations compose a framework in which the modal belief operator of each $Agent_i$ has several interesting logical properties (that, in our opinion, make it an appropriate operator to model the notion of belief for a non-ideal agent). Some of these properties are described in this section.

3.5.1 General results

Proposition 5 (Lack of Logical Omniscience)

In the framework of subjective situations, none of the following forms of logical omniscience (as defined in §1.2.4) holds:

- *Full logical omniscience.*
- *Belief of valid formulæ.*
- *Closure under logical implication.*
- *Closure under logical equivalence.*
- *Closure under material implication.*

- *Closure under conjunction.*
- *Weakening of beliefs.*
- *Triviality of inconsistent beliefs.*

Proof: Let us take a state s in which $T_i(s) = \{\mathtt{P}, (\mathtt{P} \rightarrow \mathtt{Q}), \neg\mathtt{P}\}$ and $F_i(s) = \{\mathtt{P}\}$. Consider an structure for subjective situations E that only contains the situation s.

- $E, s \models_i B_i\mathtt{P}$ and $E, s \models_i B_i(\mathtt{P} \rightarrow \mathtt{Q})$ hold, but $E, s \models_i B_i\mathtt{Q}$ does not hold. Therefore, neither full logical omniscience nor closure under material implication hold.
- $E, s \models_i B_i(\mathtt{Q} \vee \neg\mathtt{Q})$ does not hold. Therefore, there is no belief of valid formulæ.
- $E, s \models_i B_i\mathtt{P}$ holds, but $E, s \models_i B_i(\mathtt{P} \vee \mathtt{Q})$ does not hold. Therefore, closure under logical implication and weakening of belief do not hold.
- $E, s \models_i B_i(\mathtt{P} \rightarrow \mathtt{Q})$ holds, but $E, s \models_i B_i(\neg\mathtt{Q} \rightarrow \neg\mathtt{P})$ does not. Therefore, beliefs are not closed under logical equivalence or under valid implication.
- $E, s \models_i B_i\mathtt{P}$ and $E, s \models_i B_i(\mathtt{P} \rightarrow \mathtt{Q})$ hold, but the expression $E, s \models_i B_i(\mathtt{P} \wedge (\mathtt{P} \rightarrow \mathtt{Q}))$ does not hold. Therefore, there is no closure under conjunction.
- $E, s \models_i B_i\mathtt{P}$ and $E, s \models_i B_i\neg\mathtt{P}$ hold, but $E, s \models_i B_i\mathtt{Q}$ does not hold. Therefore, there is no triviality of inconsistent beliefs. □

There are two basic reasons that account for the failure of all these properties:

- $\mathcal{T}_i$ and $\mathcal{F}_i$ are defined on sets of (arbitrary) formulæ (not on basic propositions).
- $\mathcal{T}_i$ and $\mathcal{F}_i$ are unrelated. Thus, a given formula may belong to both sets, to only one of them or to none of them.

It is possible to impose any of the above properties on the belief operators by requiring these sets of formulæ to satisfy some conditions (for instance, if $(\phi \wedge \psi)\epsilon\mathcal{T}_i(s)$ implies that $\phi\epsilon\mathcal{T}_i(s)$ and $\psi\epsilon\mathcal{T}_i(s)$, then $Agent_i$'s set of beliefs would be closed under conjunction).

Proposition 6 (Relation between $\models_i$ and $=\!|_i$)

For any linearly nested formula ϕ

$E, s \not\models_i \phi$ *does not imply* $E, s =\!|_i \phi$
$E, s \models_i \phi$ *does not imply* $E, s \neq\!|_i \phi$

Proof: Take the structure of subjective situations E described in the proof of the previous proposition. It is easy to check these facts:

- $E, s \not\models_i B_i\mathtt{R}$ and $E, s \neq\!|_i B_i\mathtt{R}$. Therefore, $E, s \not\models_i \phi$ does not imply $E, s =\!|_i \phi$.
- $E, s \models_i B_i\mathtt{P}$ and $E, s =\!|_i B_i\mathtt{P}$. Therefore, $E, s \models_i \phi$ does not imply $E, s \neq\!|_i \phi$. □

3.5.2 Results on positive introspection

Proposition 7 (Characterization of positive beliefs)

For any linearly nested formula ϕ,

$$E, s \models_i \phi \text{ if and only if } E, s \models_i B_i\phi$$

Proof: The *only if* side of the formula coincides with proposition 8. The *if* side may be proven as follows:

$E, s \models_i B_i\phi \implies \forall t(sR_it), E, t \models_i \phi$. As R_i is reflexive, (sR_is); therefore, $E, s \models_i \phi$. □

This result states that $Agent_i$ believes ϕ in state s if and only if ϕ is one of the facts that is supported by $Agent_i$ in that state. In fact, the "if" side of the proposition is the classical axiom of knowledge, axiom T. Thus, in our framework the difference between *belief* and *knowledge* vanishes: both concepts have to be understood as the propositional attitude that the agents adopt towards those formulæ that they perceive to be true in the

environment. Therefore, the (rather philosophical) difference between those beliefs that are true in the real world (that constitute knowledge) and those that are not (*plain* beliefs) is not taken into account.

Proposition 8 (Belief of supported formulae)

For any linearly nested formula ϕ*,*
$E, s \models_i \phi$ *implies* $E, s \models_i B_i\phi$

Proof: There are five cases to be considered:

- $\phi \epsilon \mathcal{L}_{PC}$
  $E, s \models_i \phi$ and $\phi \epsilon \mathcal{L}_{PC} \implies \phi \epsilon \mathcal{T}_i(s) \implies \forall t(sR_it), \phi \epsilon \mathcal{T}_i(t)$
  $\implies \forall t(sR_it), E, t \models_i \phi \implies E, s \models_i B_i\phi$

- If ϕ is a modal formula that starts with an affirmed belief operator B_i (*i.e.* $\phi = B_i\psi$), this fact is exactly the next proposition.

- If ϕ is a modal formula that starts with an affirmed belief operator B_j (*i.e.* $\phi = B_j\psi$), this statement coincides with proposition 10, that will be proved below.

- If ϕ is a modal formula that starts with a negated belief operator B_i (*i.e.* $\phi = \neg B_i\psi$), this fact is the one proved as proposition 14.

- If ϕ is a modal formula that starts with a negated belief operator B_j (*i.e.* $\phi = \neg B_j\psi$), this fact is the one proved as proposition 15. □

This proposition is telling us that an agent believes all formulæ that it has reasons to support, as suggested in the motivating example. However, this proposition has an added value over our intuitions, because it refers to any kind of linearly nested formulæ, and not only to non-modal formulæ.

Proposition 9 (Single-agent positive introspection)

For any linearly nested formula ϕ*,*
$E, s \models_i B_i\phi$ *implies* $E, s \models_i B_iB_i\phi$

Proof: If $E, s \models_i B_i\phi$, that means that $E, s \models_i \phi$ holds in all the situations R_i-related to s. Being R_i an equivalence relation, these situations are exactly the ones included in the equivalence class of s induced by R_i. This class is also the set of situations that may be accessed from s in two steps (in fact, in any number of steps) *via* R_i, and ϕ is supported by $Agent_i$ in all of them. Thus, $\forall s'(sR_is')\forall s''(s'R_is'')E, s'' \models_i \phi$, and $E, s \models_i B_iB_i\phi$ also holds. □

This proposition states that axiom *4* (the classical axiom of positive introspection) holds for each belief operator B_i (*i.e.* every agent has introspective capabilities on its own positive beliefs).

Proposition 10 (Generation of positive beliefs)

$$E, s \models_i B_j\phi \textit{ implies } E, s \models_i B_iB_j\phi$$

Proof: $E, s \models_i B_j\phi \implies \forall t(sR_it), E, t \models_j \phi$. Thus, $E, t \models_j \phi$ holds in all the worlds t that belong to the same equivalence class as s (considering the partition defined by R_i). Therefore, in all the worlds accessible from s *via* R_i in any number n of steps, $E, t \models_j \phi$. Taking the case $n = 2$, we obtain that $E, s \models_i B_iB_j\phi$. □

If an agent has reasons to support a certain belief of another agent, then that belief will be included in its set of beliefs.

Proposition 11 (Inter-agent positive introspection)

$$E, s \models_i B_j\phi \textit{ implies } E, s \models_i B_jB_j\phi$$

Proof: $E, s \models_i B_j\phi \implies \forall t(sR_it), E, t \models_j \phi$. Using the result given in proposition 4, that formula implies that $\forall t(sR_it), E, t \models_j B_j\phi$; thus, $E, s \models_i B_jB_j\phi$. □

This result is more general (proposition 9 reflected the case $i = j$). It states that each agent is aware of the fact that the other agents also have introspective capabilities.

Proposition 12 (Multi-agent positive introspection)

In general, it does not hold (for three different agents $Agent_i$, $Agent_j$ and $Agent_k$ and a linearly nested formula ϕ) that

$$E, s \models_i B_j\phi \textit{ implies } E, s \models_i B_kB_j\phi$$

Proof: We will show a counterexample. Take an structure for subjective situations E with two situations, s and t, such that (sR_kt) holds, but (sR_it) and (sR_jt) do not. Take a formula ϕ such that $\phi \epsilon \mathcal{T}_j(s)$ and $\phi \not\epsilon \mathcal{T}_j(t)$. In this state of affairs, $E, s \models_i B_j\phi$ holds but $E, s \models_i B_kB_j\phi$ does not hold. □

This proposition states a negative result. It is telling that even if $Agent_i$ has reasons to support that $Agent_j$ believes something, that is not enough for $Agent_i$ to think that any other $Agent_k$ will have that belief. This proposition is essentially expressing the uncertainty of $Agent_i$ about the beliefs of a different $Agent_k$.

3.5.3 Results on negative introspection

Proposition 13 (Characterization of negative beliefs)

For any linearly nested formula ϕ,

$$E, s =\!|_i \phi \text{ if and only if } E, s \models_i \neg B_i\phi$$

Proof: The *only if* side of the proposition may be proven as follows. As we know that $E, s =\!|_i \phi$ and (sR_is), it may be said that $\exists t(sR_it), E, t =\!|_i \phi$. Therefore, $E, s =\!|_i B_i\phi$, which is equivalent to $E, s \models_i \neg B_i\phi$.

The *if* side of the proposition (*i.e.* $E, s \models_i \neg B_i\phi$ implies $E, s =\!|_i \phi$) will be proved considering five different cases (as we did in the proof of proposition 8):

- $\phi \epsilon \mathcal{L}_{PC}$

  $E, s \models_i \neg B_i\phi \Longrightarrow E, s =\!|_i B_i\phi \Longrightarrow \exists t(sR_it), E, t =\!|_i \phi$. As $\phi \epsilon \mathcal{L}_{PC}$, $E, t =\!|_i \phi$ implies that $\phi \epsilon \mathcal{F}_i(t)$; as (sR_it), $\phi \epsilon \mathcal{F}_i(s)$. Therefore, $E, s =\!|_i \phi$.

- ϕ is a modal formula that starts with an affirmed belief operator B_i (*i.e.* $\phi = B_i\psi$).

  $$E, s \models_i \neg B_i\phi \Longrightarrow E, s \models_i \neg B_iB_i\psi \Longrightarrow E, s =\!|_i B_iB_i\psi \Longrightarrow$$
  $$\exists t(sR_it), E, t =\!|_i B_i\psi \Longrightarrow \exists t, u(sR_it), (tR_iu), E, u =\!|_i \psi.$$

  As R_i is transitive, (sR_it) and (tR_iu) imply that (sR_iu). Thus, we may state that $\exists u(sR_iu), E, u =\!|_i \psi$. Therefore, $E, s =\!|_i B_i\psi$, which is equal to $E, s =\!|_i \phi$.

- ϕ is a modal formula that starts with an affirmed belief operator B_j (*i.e.* $\phi = B_j\psi$).

$$E, s \models_i \neg B_i\phi \Longrightarrow E, s \models_i \neg B_iB_j\psi \Longrightarrow E, s =\!|_i B_iB_j\psi \Longrightarrow$$
$$\exists t(sR_it), E, t =\!|_i B_j\psi \Longrightarrow \exists t, u(sR_it), (tR_iu), E, u =\!|_j \psi.$$

  As R_i is transitive, (sR_it) and (tR_iu) imply that (sR_iu). Thus, we may state that $\exists u(sR_iu), E, u =\!|_j \psi$. Therefore, $E, s =\!|_i B_j\psi$, which is equal to $E, s =\!|_i \phi$.

- ϕ is a modal formula that starts with a negated belief operator B_i (*i.e.* $\phi = \neg B_i\psi$).

$$E, s \models_i \neg B_i\phi \Longrightarrow E, s \models_i \neg B_i\neg B_i\psi \Longrightarrow E, s =\!|_i B_i\neg B_i\psi \Longrightarrow$$
$$\exists t(sR_it), E, t =\!|_i \neg B_i\psi \Longrightarrow \exists t(sR_it), E, t \models_i B_i\psi \Longrightarrow$$
$$\exists t(sR_it)\forall u(tR_iu), E, u \models_i \psi.$$

  In this expression, t is a world that belongs to the same class of equivalence than s (according to the partition defined by R_i), and u represents all the worlds that belong to t's class of equivalence; thus, u ranges over all the worlds belonging to s's class of equivalence (all the worlds that are accessible from s via R_i in any number n of steps). If we take $n = 1$, we get that $\forall t(sR_it), E, t \models_i \psi$. Thus, $E, s \models_i B_i\psi$, which is equivalent to $E, s =\!|_i \neg B_i\psi$. Therefore, $E, s =\!|_i \phi$.

- ϕ is a modal formula that starts with a negated belief operator B_j (*i.e.* $\phi = \neg B_j\psi$).

$$E, s \models_i \neg B_i\phi \Longrightarrow E, s \models_i \neg B_i\neg B_j\psi \Longrightarrow$$
$$E, s =\!|_i B_i\neg B_j\psi \Longrightarrow \exists t(sR_it), E, t =\!|_i \neg B_j\psi \Longrightarrow$$
$$\exists t(sR_it), E, t \models_i B_j\psi \Longrightarrow \exists t(sR_it)\forall u(tR_iu), E, u \models_j \psi.$$

  In this expression, t is a world that belongs to the same class of equivalence as s (according to the partition defined by R_i), and u represents all the worlds that belong to t's class of equivalence; thus, u ranges over all the worlds belonging to s's class of equivalence (all the worlds that are accessible from s via R_i in any number n of steps). If we take $n = 1$, we get that $\forall t(sR_it), E, t \models_j \psi$. Thus, $E, s \models_i B_j\psi$, which is equivalent to $E, s =\!|_i \neg B_j\psi$. Therefore, $E, s =\!|_i \phi$. □

$Agent_i$ does not believe ϕ at s if and only if ϕ is one the facts that is rejected by $Agent_i$ at s. Again, this proposition agrees with the intuitions that we had in the example that was used to motivate the need for the framework of subjective situations.

Proposition 14 (Single-agent negative introspection)

$$E, s \models_i \neg B_i\phi \textit{ implies } E, s \models_i B_i \neg B_i \phi$$

Proof: $E, s \models_i \neg B_i\phi \Longrightarrow E, s \not\models_i B_i\phi \Longrightarrow \exists t(sR_it), (E, t \not\models_i \phi)$. Thus, there exists at least one world (say w) such that (sR_iw) and $E, w \not\models_i \phi$. In order to prove the proposition, we have to notice that R_i is Euclidean (*i.e.* whenever (sR_it) and (sR_iu), (tR_iu) also holds)[7]. Therefore, w is R_i accessible from all worlds that are R_i accessible from s, and we may state that $\forall t(sR_it), (tR_iw)$ and $E, w \not\models_i \phi$. Thus, $\forall t(sR_it)\ \exists u(tR_iu) E, u \not\models_i \phi$. Thus, $\forall t(sR_it)\ E, t \not\models_i B_i\phi$, which is equivalent to $\forall t(sR_it)\ E, t \models_i \neg B_i\phi$. Therefore, we have shown that $E, s \models_i B_i \neg B_i \phi$. □

This proposition states that axiom 5 (the classical axiom of negative introspection) holds for each belief operator B_i (*i.e.* every agent has introspective capabilities on its own negative beliefs).

Proposition 15 (Generation of negative beliefs)

$$E, s \models_i \neg B_j\phi \textit{ implies } E, s \models_i B_i \neg B_j \phi$$

Proof: $E, s \models_i \neg B_j\phi \Longrightarrow E, s \not\models_i B_j\phi \Longrightarrow \exists t(sR_it), E, t \not\models_j \phi$. Let us call w to any of the worlds referred to by this existential quantifier. Being R_i Euclidean, we know that $\forall t(sR_it), (tR_iw)$; therefore, we may say that $\forall t(sR_it) \exists u(tR_iu), E, u \not\models_j \phi$. Thus, $\forall t(sR_it), E, t \not\models_i B_j\phi$, which is equivalent to $\forall t(sR_it), E, t \models_i \neg B_j\phi$. Therefore, $E, s \models_i B_i \neg B_j \phi$. □

This proposition is expressing the fact that $Agent_i$ can make positive introspection on negated beliefs of other agents.

[7] It is easy to prove that any relation that is symmetric and transitive is also Euclidean.

Proposition 16 (Inter-agent negative introspection)

In general, for two different agents ($Agent_i$, $Agent_j$) it does not hold that

$$E, s \models_i \neg B_j \phi \text{ implies } E, s \models_i B_j \neg B_j \phi$$

Proof: Consider the following counterexample. Imagine an structure for subjective situations E with three situations s, t and u, such that (sR_it) and (tR_ju). Suppose that $\mathtt{P} \epsilon \mathcal{F}_j(s)$ but $\mathtt{P} \not\epsilon \mathcal{F}_j(t)$ (and note that $\mathcal{F}_j(t) = \mathcal{F}_j(u)$). It is easy to check that $E, s \models_i \neg B_j \mathtt{P}$ holds, whereas $E, s \models_i B_j \neg B_j \mathtt{P}$ does not. □

This result states that each $Agent_i$ is aware of the fact that, even if it has reasons to think that $Agent_j$ does not believe ϕ, it may just be the case that $Agent_j$ believes ϕ indeed (and, therefore, $Agent_j$ would believe that it believed ϕ). Thus, it is another expression of the uncertainty that any agent has about the beliefs of the other agents.

3.5.4 Summary of the main properties

Summarising the main results shown in this section:

- All forms of logical omniscience are avoided.

  None of the restricted forms of logical omniscience usually considered in the literature holds in the framework of subjective situations. This result is due to the presence of partial and inconsistent situations and to the fact that the description of a situation is formed with positive and negative information about formulæ (and not about basic propositions).

- Each agent is aware of its positive and negative beliefs, and is also aware of the fact that the other agents enjoy this introspective capability.

  However, an agent is uncertain about the way the present situation is perceived by other agents and, therefore, it is unable to know anything about the other agent's beliefs.

- The positive and negative beliefs of an agent in an state reflect, as our intuitions suggested, the facts that are taken as true or false by the agent in that state.

  Thus, an agent's perception determines its beliefs in a given situation, as it might be expected.

3.6 Comparison with previous proposals

The main difference of our proposal with previous works ([More98]) is the idea of considering *subjective* situations, that may be perceived in different ways by different agents. Technically, this fact implies two differences of our approach with respect to others:

- A situation is described with two functions ($\mathcal{T}_i$ and $\mathcal{F}_i$) *for each* $Agent_i$.

  Thus, we take into account each agent's perception of the actual situation, considering a *subjective* description of each state.

- Two satisfiability and unsatisfiability relations between situations and formulæ ($\models_i$ and $=\!|_i$) are also defined *for each* $Agent_i$.

  Having a *subjective* description of each state, it makes sense to consider satisfiability relations that depend on each agent.

These differences make it impossible to embed the *subjective situations* framework in the standard doxastic modal setting or in any of the unifying proposals that were reported in §2.3, as none of these options allow the agent modeller to conceive a world from the point of view of each agent. Furthermore, we may point out other general differences between our approach and some of the ones that were described in the previous chapter:

- It is worth stressing that most of the proposals reviewed in §2 are concerned with propositional doxastic logics, whereas *subjective situations* are described with predicate formulæ. This detail will be important in the rest of the dissertation, because the use of first-order formulæ is instrumental in the way in which *rational inquirers*, to be defined in §4.3, analyse their beliefs (and *subjective situations* will be used to model the dynamic evolution of the beliefs of this kind of agents).

- The notion of *awareness* has been used in different ways to avoid some of the forms of logical omniscience, as we saw in the *logic of general awareness* ([FaHa85], §2.2.7), the *logic of awareness and principles* ([VdHM96], §2.2.8) and the *hybrid sieve systems* ([Thij96], §2.2.9). In our approach it is being implicitly assumed that all agents are aware of all the predicates and constants of the language and, therefore, this concept does not play any esential role in the *subjective situations* framework.

- Duc's approach (*dynamic epistemic logic*, [Duc97], §2.2.17) is indeed quite interesting, and seems to be inspired in intuitions rather similar to the ones that have motivated this dissertation. However, our presentation differs from Duc's in two basic points. On the one hand, our proposal to solve the logical omniscience problem, as we have seen in this chapter, is not related in any way with dynamic modal logics. On the other hand, we consider that *rational* agents must not only be capable of performing deductive inferences over their beliefs, but they also must perform other doxastic activities, that may influence their beliefs. This point will be thoroughly discussed in §4.2.

The rest of the section is devoted to a more detailed comparison of our proposal with the two approaches to the problem of logical omniscience with which it shares more similarities: Levesque's *logic of explicit and implicit beliefs* ([Leve84]) and Thijsse's *hybrid sieve systems* ([Thij96]). In both cases we begin by recalling the basic points of these frameworks and then compare them with our own.

3.6.1 Levesque's logic of implicit and explicit beliefs

Levesque uses a language with two modal operators: B for *explicit* beliefs and L for *implicit* beliefs. These operators are not allowed to be nested in the formulæ of the language. An *structure for explicit and implicit beliefs* is defined as a tuple $M{=}(S, \mathcal{B}, T, F)$, where S is the set of primitive situations, $\mathcal{B}$ is a subset of S that represents the situations that could be the actual one and T and F are functions from the set of primitive propositions into subsets of S. Intuitively, $T(\mathtt{P})$ contains all the situations that support the truth of $\mathtt{P}$, whereas $F(\mathtt{P})$ contains those that support its falsehood. A situation s can be *partial* (if there is a primitive proposition which is neither true nor false in s) and/or *incoherent* (if there is a proposition which is both true and false in s). A situation is *complete* if it is neither partial nor incoherent. A complete situation s is *compatible* with a situation t if s and t agree in all the points in which t is defined. $\mathcal{B}^*$ is the set of all complete situations of S that are compatible with some situation in $\mathcal{B}$.

The relations $\models_T$ and $\models_F$ between situations and formulæ are defined as follows:

- $M,s \models_T \mathtt{P}$, where P is a primitive proposition, if and only if $s \in T(\mathtt{P})$
- $M,s \models_F \mathtt{P}$, where P is a primitive proposition, if and only if $s \in F(\mathtt{P})$
- $M,s \models_T \neg\varphi$ if and only if $M,s \models_F \varphi$
- $M,s \models_F \neg\varphi$ if and only if $M,s \models_T \varphi$
- $M,s \models_T (\varphi \wedge \psi)$ if and only if $M,s \models_T \varphi$ and $M,s \models_T \psi$
- $M,s \models_F (\varphi \wedge \psi)$ if and only if $M,s \models_F \varphi$ or $M,s \models_F \psi$
- $M,s \models_T B\varphi$ if and only if $M,t \models_T \varphi \ \ \forall t \in \mathcal{B}$
- $M,s \models_F B\varphi$ if and only if $M,s \not\models_T B\varphi$
- $M,s \models_T L\varphi$ if and only if $M,t \models_T \varphi \ \ \forall t \in \mathcal{B}^*$
- $M,s \models_F L\varphi$ if and only if $M,s \not\models_T L\varphi$

There are some similarities between our approach and Levesque's logic of implicit and explicit beliefs. However, they are more apparent than real, as shown in this listing:

- Levesque also considers a satisfiability and an unsatisfiability relation between situations and doxastic formulæ.

  However, these relations are not considered *for each agent* (even when Levesque's framework is extended to the multi-agent case, as described in §2.2.5).

- Levesque also describes each situation with two functions $\mathcal{T}$ and $\mathcal{F}$.

  Again, these functions are not indexed by each agent, as our functions are (Levesque considers an objective description of what is true and what is false in each situation). This comment is also applicable in the multi-agent extension of Levesque's framework shown in §2.2.5. Another important difference is that Levesque's functions deal with basic propositions, and not with predicate formulæ as our functions do.

- Both approaches allow the presence of *partial* or *inconsistent* situations.

  However note that, in our case, it is not the (objective) description of the situation that is partial or inconsistent, but the *subjective* perception that an agent may have of it. Thus, the notions of partiality and inconsistency have a much more natural interpretation in our framework.

- Both approaches avoid all the forms of logical omniscience.

  The reason is different in each case, though. In Levesque's logic of explicit and implicit beliefs, it is the presence of incoherent situations that prevents logical omniscience. In our proposal, there is no need to have inconsistent situations to avoid logical omniscience. In fact, we solve that problem by defining $\mathcal{T}_i$ and $\mathcal{F}_i$ over arbitrary sets of formulae, and not over basic propositions.

- There are accessibility relations between situations for each agent in both systems.

  Levesque's accessibility relation between situations is left implicit; our accessibility relations are explicit. Furthermore, the intuition underlying these relations is somewhat different, as explained in §3.2.

Other differences with Levesque's approach that may be mentioned are the following:

- Levesque only considers one agent, and does not allow nested beliefs. Thus, his agents do not have any introspective capabilities. This statement also holds in the multi-agent extension of Levesque's ideas (§2.2.5), because the accessibility relations that are used to verify the validity of modal formulæ do not have to be neither transitive nor Euclidean (there are no constraints imposed on them).

- Levesque defines *explicit* and *implicit* beliefs, whereas we do not make this distinction.

- Even though Levesque avoids logical omniscience, his agents must necessarily believe all those tautologies that are formed by *known* basic propositions (those propositions $\mathtt{P}$ for which the agent believes $(\mathtt{P} \vee \neg \mathtt{P})$),

regardless of their complexity. This is not the case in our approach, because we deal directly with formulæ.

- There is a different treatment of the unsatisfiability relation when applied to beliefs, because he transforms $\dashv$ into $\not\models$, whereas we do not.

3.6.2 Thijsse's hybrid sieve systems

Thijsse ([Thij96]) proposes a way of using *partial logics* to deal with various forms of logical omniscience. He defines a *partial model* as a tuple $(W, \mathcal{B}_1, \ldots, \mathcal{B}_n, V)$, where W is a set of worlds, $\mathcal{B}_i$ is the accessibility relation between worlds for $Agent_i$ and V is a *partial* truth assignment to the basic propositions in each world. $\top$ is a primitive proposition that is always interpreted as *true*. Truth ($\models$) and falsity ($\dashv$) relations are defined in the following way:

- $M,w \models \top$
- $M,w \not\dashv \top$
- $M,w \models \mathtt{P}$, where $\mathtt{P}$ is a primitive proposition, iff $V(\mathtt{P}, w) = 1$
- $M,w \dashv \mathtt{P}$, where $\mathtt{P}$ is a primitive proposition, iff $V(\mathtt{P}, w) = 0$
- $M,w \models \neg\varphi$ iff $M,w \dashv \varphi$
- $M,w \dashv \neg\varphi$ iff $M,w \models \varphi$
- $M,w \models (\varphi \wedge \psi)$ iff $M,w \models \varphi$ and $M,w \models \psi$
- $M,w \dashv (\varphi \wedge \psi)$ iff $M,w \dashv \varphi$ or $M,w \dashv \psi$
- $M,w \models B_i\varphi$ iff $M,v \models \varphi$ $\forall v$ such that (w, v) $\epsilon\mathcal{B}_i$
- $M,w \dashv B_i\varphi$ iff $\exists v$ s.t. (w, v) $\epsilon\mathcal{B}_i$ and $M,v \dashv \varphi$

The most important similarities between our approach and Thijsse's are:

- n agents and n explicit accessibility relations are considered.

  However, as in Levesque's case, there are no restrictions on these relations, and the intuitive meaning of our accessibility relations is slightly different from the standard one, as argued in §3.2.

- Two relations (of satisfiability and unsatisfiability) are defined. Moreover, a similar clause is used to provide a meaning to the unsatisfiability relation with respect to the belief operator.

  As before, the main difference is that we provide two relations *for each agent*.

- There are no tautologies in Thijsse's system; therefore, he does not have to care about some forms of logical omniscience (closure under valid implication and belief of valid formulæ).

- Closure under material implication and closure under conjunction do not hold in Thijsse's approach either.

The main difference with Thijsse's proposal is that he uses *partial* assignments of truth values *over basic propositions* for each state; thus, a proposition may be true, false or undefined in each state. We deal with formulae, not with basic propositions, and each formula may be supported *and/or* rejected by *each agent* in each state. Therefore, Thijsse's approach is three-valued, whereas ours is more of a four-valued kind, such as Levesque's.

3.7 Summary

In this chapter, it has been argued that each $Agent_i$ perceives its actual situation in a particular way, which may be different from that of other agents located in the same situation. The vision that an $Agent_i$ has of a situation determines its (positive and negative) beliefs in that situation. This intuitive idea has been formalized with the notion of *subjective situations*. These entities are the base of a doxastic logic, in which the meaning of the belief operators seems to fit with the general intuitions about how the doxastic attitude of a non-ideal agent should behave. In particular, logical omniscience is avoided while some interesting introspective properties are maintained. A detailed comparison of this approach with Levesque's *logic*

of implicit and explicit beliefs ([Leve84]) and Thijsse's *hybrid sieve systems* ([Thij96]) has also been provided.

4 Rational inquirers

4.1 Introduction

In §1 we introduced the problems of logical omniscience and perfect reasoning, that arise when doxastic logics are used to model the reasoning processes that a rational agent may perform on its set of beliefs. In §2 we reviewed the most relevant approaches that have been suggested to solve these problems. In §3 we suggested a new way of tackling them, using the concept of *subjective situations*. Our aim in the rest of the dissertation will be to show how (part of) the framework developed in §3 may be used to formally model the evolution of the beliefs of real agents (that are neither logically omniscient, because they do not believe all tautologies, nor perfect reasoners, because they may believe a set of facts without having to necessarily believe all its logical consequences). It is important to stress the fact that, in the rest of the dissertation, we will mostly concentrate on the analysis of the beliefs of a single agent about its environment; the problems that would be encountered if a whole multi-agent system were considered will be sketched in §6.

4.2 Considering rational agents

The concept of *rational agent* has been given different interpretations in the Artificial Intelligence literature (some examples of architectures for rational agents were given in table 1 in §1.1). It would be impossible for us to consider in this chapter all the different agent architectures that have been proposed in the past and to show how the evolution of the beliefs of each kind of agents may be modelled in our framework; therefore, we will follow a different approach. In order to encompass a wide variety of actual agents, a very general (and quite informal) definition of rational agents will be considered below. Later, in §4.3, a specific kind of agents, called *rational inquirers*, will be defined by giving a concrete interpretation to the general characteristics that, in our opinion, a rational agent should have. After that, in §5 it will be shown how to formally model the evolution of the beliefs of this kind of agents.

The starting point of our discussion on rational agents will be the following intuitive definition:

Definition 9 (Rational agents)

Rational agents are those agents that are permanently analysing their belief sets in order to make them as similar as possible to the facts that hold in the real world. Rational agents try to get rid of those beliefs that do not reflect accurately what is true in their environment, and they also try to keep increasing their sets of beliefs by adding logical consequences of their beliefs. This continuous process of analysis will be called rational inquiry.

It may be argued that an agent behaves in a rational fashion when it tries to reduce the gap between its beliefs and the reality that surrounds it. Obviously this is what an intelligent agent should be always doing, because beliefs will be the base to form the agent's intentions and, therefore, they will be guiding its future behaviour[8]. The further are the agent's beliefs from what is true in its environment, the least effective will be the actions that it will take to try to reach its goals. In the worst of all scenarios, if the set of beliefs does not describe the agent's environment in a faithful way, the actions that it might take could be even harmful or counterproductive.

In order to build an *abstract*, general model of any kind of non-ideal agent, it is important to focus on those tasks performed by the agent *that have a direct influence on its beliefs*; these tasks will be referred to as *doxastic tasks*. More specifically, in our model of *rational agents* the following doxastic tasks will be taken into account:

- *Deductive capabilities.*

  It is usually accepted that any rational agent must be capable of performing deductions on its beliefs and of adding the results of these deductions as new beliefs. What we will *not* do is to consider that the agent believes, right from the beginning, all tautologies and all the logical consequences of its beliefs, without any kind of effort. Agents will be able to deduce facts (possibly not all those that are classical logical consequences, if they have limited deductive capabilities), but this inferential process will have to be explicit; *e.g.* if an agent believes

[8]Throughout this chapter we will be using a BDI-oriented vocabulary, *i.e.* we will consider that the main elements that determine the behaviour of an agent are its *beliefs* (its conception of the environment), its *desires* (the agent's goals) and its *intentions* (the goals that the agent actually decides to try to accomplish). Having said that, we insist that we are not arguing that rational agents have to fit to some specific architecture.

P and (P ⇒ Q), it will have to perform an explicit deductive step before believing Q. Notice the similarity of this basic idea with the one underlying Duc's *dynamic epistemic logic*, which was reviewed in §2.2.17. It must be noted that we are not demanding that all rational agents have the same deductive capabilities or that they have perfect reasoning capabilities; in fact, in the concrete instantiation of rational agents that we are going to define in the next section (the *rational inquirers*) they will be able to deduce some logical consequences of its beliefs (using a certain analytic tableaux calculus, as will be seen in §4.3.1), but not all of them.

Thus, the possibility of having ideal agents that are (by definition) logically omniscient and perfect reasoners is forbidden right from the very beginning. Considering that a deductive inference requires a conscious and explicit step may seem intuitive, natural and almost obvious to most of the readers of the previous paragraph, but this assumption has not been made by many of the researchers in the Artificial Intelligence field of belief systems. For instance Gärdenfors, in his seminal work on belief dynamics ([Gärd88]) presents *epistemic states* as rational idealizations of psychological states and models them in a logically idealized way, by representing an epistemic state with a *belief set*, which is defined as a consistent and logically closed set of sentences (*i.e.* a logical theory). By taking this stance, he can make a very nice and detailed formal analysis of the logical properties that should govern the basic operations on belief sets: *expansion* (adding a new belief), *contraction* (removing a present belief) and *revision* (changing a belief into a disbelief, or *viceversa*). These properties are usually called the *rationality postulates*. As we will assume that an agent's set of beliefs may be any set of formulæ (even a logically inconsistent one), we will have to deal with different logical frameworks, such as the *subjective situations* presented in §3.

Performing logical deductions is a direct way of increasing one's beliefs. For instance, if an agent γ has received from agent α the information that P is true, and agent β tells it that (P ⇒ Q), it seems clear that it will be useful for γ to draw the immediate inference and, at least as long as it does not receive opposite information, believe that Q also holds in the real world. Performing logical inferences seems to be such an obvious

way of obtaining new information as receiving it from sensors or from other agents. By performing these steps of deduction, γ avoids having to waste its time (and other agents' time) wondering whether `Q` holds in the real world. If agents were not capable of deducing consequences from their beliefs, they would have very limited cognitive capabilities, and they would have to be continuously asking for information to their environment or to other agents. Thus, having certain deductive power is useful to close the gap between the agent's beliefs and those facts that hold in its environment.

- *Addition of external information.*

  Every agent performs its tasks in an environment in which there may be several sources of information. It will be considered that rational agents must be capable of receiving information from the environment (*e.g.* coming from sensors or from other agents), and that this information modifies directly their beliefs. We will make the assumption that all external inputs (be they from sensors or from other agents) are treated in the same way. Of course, this is a very strong assumption. We can make it because, as stated in §4.1, we are only going to deal in this dissertation with a single agent's beliefs about the world, and not with its beliefs about the beliefs of other agents. Thus, it will be equivalent to receive the input `P` from a sensor as receiving the message `P` from an agent. The distinction between these two different ways of receiving information from the environment is indeed made by many researchers. For instance, Parsons (see *e.g.* [PaGi98]) distinguishes between four types of propositions that an agent has to deal with: *basic facts* (the agent's initial beliefs), *observations* (things that the agent perceives in its environment), *communiqués* (messages received from other agents) and *deductions* (propositions that the agent derives from its own beliefs). However, although the distinction between observations and communiqués is made, they are treated in a very similar way: the credibility of an observation depends on the reliability of the sensor or source (based on its past behaviour) whereas the credibility of a communiqué depends on the reliability of its sender (also based on its past behaviour).

  A stronger assumption will also be made in our model of rational agents:

the received information will be directly incorporated into the agent's set of beliefs. Therefore, these agents might be qualified as *credulous*, because they will take external sources of information as correct (the effects that the introduction of an external input has on a *rational inquirer*'s set of beliefs will be described in §4.4). It must be noted that this assumption is also made in the classical approaches to *belief revision* and *belief update*, as all of them incorporate a rationality postulate that indicates that a revised (or updated) belief set always contains the formula that has caused the revision (or the update). This constraint permits the agent to revise its wrong beliefs about a static world or to adapt its beliefs to the changes that occur in a dynamic environment; however, other types of behaviour could have been considered. For instance, we could have dealt with degrees of trust associated to each new belief, depending on its source, using a formal framework for trust evolution and update based on previous experiences, such as the one suggested in [JoTr99], or the degrees of credibility proposed in [PaGi98]. We could also have considered agents of a more *persistent* flavour, that do not necessarily accept new inputs that contradict in a direct way their present beliefs (*i.e.* they would not accept $\mathtt{P}$ if they are currently believing that $\neg\mathtt{P}$ is the case); see *e.g.* [Poll98].

- *Request information from the environment.*

  Rational agents, as mentioned in the previous definition, keep trying to reflect in their beliefs what is true in their environment; in order to accomplish this goal, they are going to be continuously analysing their beliefs, trying to *confirm* or *refute* those that are uncertain, in a *Popperian* style ([Popp34]). Thus, there will be occasions in which the agent will have to *search* in its environment the data that it needs to make these corroborations or refutations, to decide which beliefs should be kept or eliminated. This search will be left in this degree of abstraction at this point; in an actual implementation of rational agents, it could take the form of making an experiment, asking questions to other agents, looking up information stored in databases, looking for an item of information through the Internet, consulting the measure given by a sensor, *etc.* In our actual instantiation of rational agents, the *rational inquirers* described in §4.3, they will be allowed to make some questions to the environment and to use the answers to these questions to update

their sets of beliefs; this process will be explained in §4.3.3.

- *Generation of doubts.*

  In many cases the agent will need to know whether a certain fact may be deduced from its beliefs; that may be necessary for instance to allow the agent to decide the action to take at a particular point in time to reach a given goal. *Rational agents* will be capable of *having doubts*, of wondering whether they (implicitly) believe or not a given fact. These doubts will have a big influence on the agent's beliefs, because they will form the basis of an analysis that will lead the agent to deduce whether the given fact was (implicitly) believed or not. The way in which our concrete instantiations of rational agents, the *rational inquirers*, will manage these doubts will be covered in §4.3.2.

A general model of non-ideal agents has to include the four doxastic tasks that have been just stated. These are the activities that have a more direct influence on the agent's beliefs and, therefore, are the ones that have to be included in any model of doxastic activity. It is indeed a very abstract model because it does not take into account many other important tasks that a rational agent should perform. We will not pay attention to them because we want to focus on the agent's doxastic activity and not in those tasks in which its set of beliefs is not modified. In particular, the following kinds of tasks will *not* be taken into consideration:

- *Inference capabilities.*

  We have argued that a rational agent must have some deductive capabilities in order to be able to draw logical consequences from its set of beliefs (*i.e.* from the information that it perceives in the environment, from the data that it receives from other agents, *etc.*). This is the only kind of inference that takes one from true premises to true conclusions, and that is why is the capability that has been assumed to hold for all rational agents. However, it is possible to think of many kinds of real agents that could have different types of inference capabilities; for instance, an agent could be able to perform induction, abduction or case based reasoning, just to name a few possibilities that are not accounted for in our framework.

- *Determination of goals.*

  The agent's behaviour will be guided by its goals. The process that the agent follows to decide which are these goals will not be taken into account; it will be assumed that they are externally given, by the agent's builder, or that they are dynamically generated, by the agent itself (*e.g.* if it decomposes an initial goal into a set of subgoals). In any case, we will consider that the determination of the goals to be achieved does not modify the agent's beliefs about the world.

- *Construction of plans.*

  When the agent knows its goals, it will design plans to try to achieve them. Beliefs will certainly play an important role in the planning module; however, this process will not modify them. Beliefs will only be considered to decide which actions may be added to a plan or which plans are feasible.

- *Decision making.*

  When the agent has studied all the plans that can lead to reach its goal, it will have to decide which one to execute. The algorithm used to make this decision would take into account its model of the world (*i.e.* the agent's beliefs) to determine which is the most appropriate plan, but it would not modify its beliefs.

- *General control strategy of the analysis.*

  In this model the rational agent will be continuously performing the tasks mentioned above (making deductions, receiving information from the environment, asking for information or having doubts). The model will be left in this level of abstraction, without determining the strategy that the agent should use in order to decide which of these tasks to perform at each point in time. In this way the model will be general enough to include *e.g.* *logical* agents (oriented to the performance of deductions) or *experimental* agents (more biased towards asking questions to the environment in order to obtain information).

- *Execution of actions.*

  When the agent executes an action of a plan, their beliefs should reflect the changes produced in its environment as a result of this action (*e.g.* it

should believe $open(d, 90\_degrees)$ after executing $open\_door(d)$). The main problem with this strategy is that the action may not have had its intended effect (*e.g.* there was a chair behind the door and it has only opened 45 degrees). To prevent the agent from having wrong beliefs, it will be assumed that it cannot be sure about the effects of an action if it does not check them (*e.g.* it can use sensors to make certain that the door has indeed opened in an angle of 90 degrees, or it can ask to another agent whether that is the case). If the agent is not sure about the result of an action, it can always generate a doubt and obtain data from the environment that helps it to confirm whether the intended effect has been really accomplished.

The reader should be aware by now that it is not our intention to develop a comprehensive architecture that may be used to actually implement rational agents. If that had been our aim, we should have had to consider the different attitudes that should guide the agent's behaviour (for instance beliefs, knowledge, desires, doubts, intentions, wants, wishes, obligations, commitments or choices), how these attitudes are related to each other, which are the agent's capabilities (*i.e.* which are the actions that the agent may perform), how actions are characterized (*e.g.* in terms of preconditions and postconditions), how the agent may acquire and decompose its goals, which is its planning process, how it can communicate with other agents or interact with its environment, *etc.* Our (much more modest) goal is to have a formal way of modelling the evolution of the set of beliefs of a rational (and not logically omniscient) agent.

4.3 Rational inquirers

We have just described the limits that we have established as to which activities are included in our model of rational agents (the *doxastic activities*) and which are not. Having done that, we need to be more precise in determining how these doxastic activities can be really implemented, to be able to show in §5 how the evolution of the beliefs of a rational agent may be formally modelled. This section describes the behaviour of the *rational inquirers*, which will be the concrete instantiation of rational agents that we will consider in the rest of the dissertation. First, the types of activities that

this kind of agents will be able to perform when they are analysing their beliefs are defined; after that, each of them is briefly described.

The definition of *rational inquirers* includes the aspects that have to be covered by any rational agent, as discussed in the previous section:

Definition 10 (Rational inquirers)

A rational inquirer (see e.g. [MCS98], [MCS99a], [MCS00a]) is an agent that will be continuously making a dynamic multi-dimensional analysis of its beliefs, trying to keep them as similar as possible to the facts that hold in the real world. The components of this analysis are the following:

- *A logical dimension of analysis, in which the agent may perform (limited) deductive inferences on its own set of beliefs (see §4.3.1).*
- *An exploratory dimension of analysis, in which the agent may have doubts, may wonder whether it (implicitly) believes or not a given fact (see §4.3.2).*
- *An experimental dimension of analysis, in which the agent may ask the environment for data that may be used to confirm or refute some of its beliefs (see §4.3.3).*

Rational inquirers may also incorporate to their beliefs the information that they receive directly from their environment (see §4.4).

4.3.1 Logical analysis

Rational agents, as argued in the previous section, need to have some way of deducing logical consequences from their sets of beliefs. Therefore, they need to be able to use a proof method such as natural deduction, resolution, sequent calculus or analytic tableaux calculus, to name a few possibilities (see *e.g.* [Kell97] for a recent presentation of these proof techniques, both in propositional and predicate logic). All of these proof systems are quite similar, in the sense that proofs made with one of them may be usually easily translated into proofs made with another one. Thus, it does not matter very much which one of them is used by our *rational inquirers*, because the logical analysis performed by the agents could be translated into a similar analysis in another framework. Having said that, we have decided to endow

the rational inquirers with the capacity of using a modified version of the classical *analytic tableaux method* in order to perform the logical analysis of their beliefs (a detailed description of the classical analytic tableaux method can be found in [Smul68]). The rules of this tableaux calculus are shown in figure 4, using Smullyan's uniform notation[9], which is depicted in figure 3. In those figures the symbols Γ and Δ represent sets of first-order formulæ, c is a Skolem constant, o is an arbitrary individual, α, β, γ and δ represent especial kinds of first-order formulæ, ϕ and ψ represent any predicate formula, x is a variable, A_x is a formula where x is free, and $A\{r/x\}$ denotes the substitution of all the free appearences of x by r in A_x.

α-formulæ

α	α_1	α_2
$\phi \wedge \psi$	ϕ	ψ
$\neg(\phi \vee \psi)$	$\neg\phi$	$\neg\psi$
$\neg(\phi \Rightarrow \psi)$	ϕ	$\neg\psi$
$\neg\neg\phi$	ϕ	ϕ

β-formulæ

β	β_1	β_2
$\neg(\phi \wedge \psi)$	$\neg\phi$	$\neg\psi$
$\phi \vee \psi$	ϕ	ψ
$\phi \Rightarrow \psi$	$\neg\phi$	ψ

γ-formulæ

γ	$\gamma(r)$
$\forall x A_x$	$A\{r/x\}$
$\neg\exists x A_x$	$\neg A\{r/x\}$

δ-formulæ

δ	$\delta(r)$
$\exists x A_x$	$A\{r/x\}$
$\neg\forall x A_x$	$\neg A\{r/x\}$

Figure 3: Uniform notation for first order formulæ

The main differences between the classical analytic tableaux method and the one shown in figure 4 may serve to motivate the election of this particular tableaux calculus. They are the following:

- Use of *two-sided* tableaux.

  The formulæ of each tableau are divided into two sets, whereas in the classical method each tableau contains a unique set of formulæ[10]. The

[9] This notation was developed in [Smul68], and has been used *e.g.* in [Fitt83], [Fitt96] and [Fitt99]. It allows us to write in a compact way many similar rules.

[10] It must be noted that this remark is referred to the modern versions of this technique, and not to the analytic tableaux method as defined originally by Beth ([Beth55], [Fitt99]), in which each tableau did indeed contain two sets of formulæ.

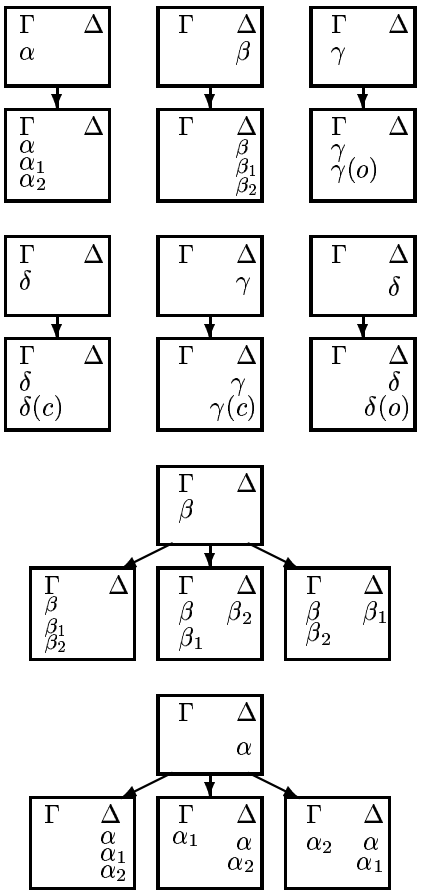

Figure 4: Rules of the logical analysis

two sides of a tableau will be called the *left column* and the *right column*. In §5 it will be argued that each of these tableaux represents a set of *potential* or *imaginable* or *conceivable* situations (those in which the formulæ in its left column hold and the formulæ in its right column do not hold). This is our main motivation for choosing this particular type of tableaux: they allow us to manage both *positive* and *negative* information that an agent may have about a given situation. Recall that, in the *structures of subjective situations* defined in §3, the description of a situation s by an $Agent_i$ was made using two functions, $\mathcal{T}_i$ and $\mathcal{F}_i$, which returned the formulæ that the agent had perceived as

true/false in s; thus, we need a way of dealing with these two types of information.

- The splitting rules have been modified.

  The *splitting rules* are those that analyse a β-formula in the left column or an α-formula in the right column, as shown in figure 4. When one of them is applied, three subtableaux are generated (whereas in the classical analytic tableaux method, only two subtableaux were generated, one for each of the components of the analysed formula). One of these subtableaux contains the two members of the analysed formula in the same column where the analysed formula appeared, whereas the other two subtableaux contain one of them in the left column and the other one in the right column. These are the possibilities of accounting for the truth/falsehood of the formula that has been analysed, namely that one (or both) of its members are true/false (depending on whether β-formulæ /α-formulæ are considered). This kind of splitting rules will allow us to make a fine-grained distinction between the class of *conceivable situations* represented by each of the resulting subtableaux. If the standard *two-sided* splitting rules were used, there would be situations represented by both subtableaux (*e.g.* those in which the two members of a disjunction hold, or those in which none of the members of a conjunction hold). This will not be the case in our way of modelling the evolution of a set of beliefs, as will be argued in §5; for instance, when we analyse a disjunction in the left column of a tableau, the three generated subtableaux represent a partition of the set of situations in which the disjunction holds in three disjoint classes of situations: in two of them one of the disjuncts holds and the other does not, and in the third one both disjuncts hold.

- Analysed formulæ are kept in the subtableaux.

  After applying a rule, the analysed proposition and all the other formulæ in the tableau are maintained in the resulting tableaux, they are not deleted. In the case of α-formulæ and β-formulæ the agent would just add a tag to the analysed proposition in other to recall that it had already been analysed. With respect to γ-formulæ and δ-formulæ the agent should take into account which instantiations (with

constants representing specific individuals or Skolem constants representing generic individuals) had already been applied to each of these formulæ in each one of the branches of the analytic tableaux tree.

This modification is not essential; however, it is quite convenient because later it will be easier to consider the expression *the set of formulæ that appear in all open tableaux* (the leaves of the tableaux tree) than the longer and more cumbersome *the set of formulæ that appear in all the branches of the tableaux tree*.

- There are two tableau closing conditions.

  A tableau may be *logically closed* by the agent when it realizes that (a) it contains either a formula and its negation in the left column or (b) it contains the same formula in both columns of the tableau. A *logically closed* tableau is permanently eliminated from the analysis. Recall that, in the classical analytic tableau method, a tableau is closed and removed from the analysis when it contains a formula and its negation (because that tableau does not represent a logical model, being so patently logically inconsistent). In our case, the tableaux that are logically closed are those that represent states of affairs in which a formula and its negation are supported or those in which a formula is both supported and rejected at the same time. Note that those situations are not *impossible*, in the sense of not being representable in our framework (recall that there are no restrictions on the functions $\mathcal{T}_i$ and $\mathcal{F}_i$ of the structures of subjective situations). By closing these tableaux, the agent is only stating that it is not interested in considering the kind of situations that they represent, not defending that they are epistemically impossible (especially in the case in which a formula is both supported and defended in a situation, which may easily arise in case the agent has several sources of information or the world is dynamic).

  Another important difference with the classical method has to be mentioned at this point: we do *not* assume that a tableau is *automatically* closed when one of the above conditions holds. It may be the case that the rational inquirer, being non-ideal, does not notice that circumstance; we demand that it has to explicitly notice this fact, and close the tableau on purpose, with a conscious step. This characteris-

tic permits the agent to consider situations in which it has reasons to support a formula and its negation or situations in which it has reasons to support and to deny a given formula, if it does not close a *closable* tableau (these types of inconsistent situations were considered in the *subjective situations* framework defined in §3).

The conditions under which a tableau is *logically closed* are not the only ones that may cause the closing of an analytic tableau. Later, in §4.3.3, it will be explained that a tableau may not only be *logically closed* but also *empirically closed*. The underlying idea is that the agent will empirically close not those tableaux that represent situations that are logically inconsistent, but those that have information that does not reflect what is true in its environment. However, the agent will be allowed to *re-open* any empirically closed tableau, if it finds out in a later stage of the analysis that the content of the tableau matches with the information provided by the reality that surrounds it. Thus, the empirical closing of tableaux will be quite different to the logical one that has just been described, both in the circumstances that may cause it and in the effects that it may have in the agent's analysis of its set of beliefs.

- Non-ideal agents.

  The agents that use the classical analytic tableaux method may be qualified as *ideal*, in the sense that they always build the *whole* tree of analysis of an initial tableau in order to discover whether the set of formulæ contained in the tableau may (or may not) be embedded in a logical model. We do *not* require a rational inquirer to make a fully fledged, complete analysis of an initial set of beliefs. In fact, as will be apparent in the rest of the dissertation, a rational inquirer will combine the different dimensions of analysis in time, intertwining the logical analysis *via* tableaux with the experimental or the exploratory analysis, and with the introduction of externally obtained information about its environment.

  It must also be noted that, even if the agent performed an exhaustive analysis of its initial set of beliefs, it would still not be a perfect reasoner in predicate calculus, due to the kind of rules that compose the tableaux calculus shown in figure 4. In fact, it may be proved that the

propositional part of this calculus is sound and complete with respect to Kleene's three-valued logic ([Klee52]). Our own proof of this statement is included in this dissertation as appendix A. This fact is also quite interesting, because it allows us to consider a type of agents that have limited reasoning capabilities, and show how they can still perform some (limited) steps of inference and keep deducing some logical consequences of their beliefs in the logical dimension of analysis. It is important to notice that this result does *not* mean at all that *rational inquirers* have to be considered logically omniscient and perfect reasoners with respect to Kleene's three-valued logic. A detailed argument on this topic is made in §4.5.

4.3.2 Exploratory analysis

As we have seen in the previous section, the classical analytic tableaux method has been modified in a number of ways (the tableaux contain two sets of formulæ, the splitting rules generate three subtableaux, there are two tableau closing conditions, *etc.*). There is yet another important difference between our analytic tableaux method and the classical one. In the standard tableaux method it is possible to add any tautology into any tableau at any point of the development of the tableaux tree, because a tableau is meant to represent a logical model and, in all standard logical models, all tautologies are true. However, in our tableaux method the agent is explicitly *not* allowed to add to an open tableau any tautology (an exception to this rule will be considered below). Some reasons that may be given to support this constraint are the following:

- If this constraint were eliminated, we would leave an open door to introducing (and later *believing*, as will become apparent in §5) all tautologies effortlessly, which is something that we argued against in §4.2, when we defended that an agent should be able of deducing facts from its set of beliefs, but always with a conscious and deliberate use of a deduction mechanism. Thus, logical omniscience (in its weaker form, *i.e.* believing all valid formulæ, as defined in §1.2.4) could hardly be avoided.

- The previous point may be strengthened by noting that a tautology may be as intrincate and complicated as we wish, so it may not be

at all readily apparent to an agent (be it human or computational) that a formula is indeed a tautology. Thus, it would not seem either reasonable or credible to assume that a non-ideal resource-bounded limited agent knows which formulæ are tautologies and which are not, so it can know which formulæ may be added *for free* in the tableaux and which formulæ may not be introduced in the tableaux.

- In the next chapter it will be shown that a tableau may be seen just as a partial representation of a set of *conceivable situations*, *i.e.* of situations that the agent may consider as real ([MoSa97a], [MoSa97b], [MCS00a]). Being only a representation of a certain set of possible states or possible worlds, but *not* a description of a logical model (as in the standard case), there seems to lack a clear justification to allow the agents to modify them arbitrarily, with no apparent reason.

- It could also be argued (see *e.g.* [Jasp94]) that the situation where no tautological information is believed could model, for instance, the initial state of information in the environmentalist *tabula rasa* theories of Locke and Rousseau, where it is assumed that children form their sets of beliefs starting from scratch, with no initial information at all.

Having made this argument against the free introduction of tautologies, we will now consider an important exception. Although it is not possible to add an arbitrary tautology into the tableaux of the rational inquirer's *logical* analysis, in the *exploratory* dimension of analysis agents *may* still be capable of adding some formulæ to the open tableaux (in fact, there will be other ways of adding information into the tableaux, as will be further explained in §4.3.3 and §4.4).

We argued in §4.2 that a rational agent must be capable of posing itself questions, of introducing *doubts* in the analysis, of wondering whether a certain formula ϕ is or not the case. Technically, this idea is implemented in the rational inquirers' exploratory dimension of analysis by allowing them to introduce instances of the *Axiom of the Excluded Middle* (*AEM*), *i.e.* formulæ with the form $(\phi \vee \neg\phi)$, in the left columns of the open tableaux of the logical analysis. As a rational inquirer deals with beliefs about the world, but not with beliefs about other agent's beliefs, we will stick to the case in which ϕ is just a first-order formula, and *not* a modal formula; therefore, a

rational inquirer α may not consider whether it is the case that β believes P, but may consider whether P is or not the case.

The use of this particular tautology seems a natural way to allow the agent to have *doubts*, to wonder whether it believes some formula (ϕ) or its negation ($\neg\phi$). This exception allows the introduction of the formula ($\phi \vee \neg\phi$) in a tableau, which will be later split (with one of the splitting rules of the logical analysis, the one that analyses disjunctions in the left column, shown in figure 4) into two subtableaux containing ϕ in one column and $\neg\phi$ in the other. Notice that the third subtableau generated in the logical analysis would be immediately (logically) closed because it would contain ϕ and $\neg\phi$ in the left column and the agent, having added the two formulæ at the same time, could hardly fail to notice the blatant inconsistency. In this way, the agent can explore both alternatives independently, and the logical analysis can guide the search of examples or counter-examples needed to give more credence to one side of the doubt than to the other. A detailed example that illustrates this issue is given in §4.6, where the logical analysis of an instance of the AEM, included in the exploratory dimension, suggests to the agent which questions may be made to the environment (in the *experimental* dimension of analysis, to be described in §4.3.3) in order to check whether one of the alternatives actually holds and, in this way, adjust the agent's beliefs to what is true in the real world.

In fact, the possibility of adding this kind of tautologies in the analytic tableaux is a well-known idea in the tradition of classical proof theory, as described for instance in [BeMa77]. It has also been suggested by Hintikka in a general theory of argumentation called the *interrogative model of inquiry* (see *e.g.* [Hint81], [Hint86a], [Hint87], [Hint88], [Hint92]). In that framework the process of scientific inquiry is modelled with plays of the *interrogative game*. In that game there are two players, named Inquirer and Nature. The Inquirer seeks to know whether a certain conclusion is true (given a certain set of premises). In order to answer this question, it can perform two kinds of actions:

- *Deductive moves*

  These moves model the deductive capacity of the Inquirer. They are controlled with Beth-style ([Beth55]) subtableaux. They consist of the application of rules of inference over the formulæ contained in an open tableau. The rules of analysis (presented in [Hint92]) are very similar

to the ones that we have considered in figure 4, because the tableaux are also two-sided.

- *Interrogative moves*

  This kind of moves permit the Inquirer to put questions to Nature. There are two types of questions that may be formulated:

  - *Disjunctive questions*

    If $(\phi \vee \psi)$ appears in the left column of a tableau, the Inquirer may question which of the two components of the disjunction is actually the case.

  - *Existential questions*

    If $\exists x S_x$ appears in the left column of a tableau, the Inquirer may ask for an specific individual o for which S_o holds.

  In these interrogative moves Hintikka allows the introduction of (some) instances of AEM in the left side of the subtableaux. The purpose of these formulæ is to serve as the presuppositions ([Hint76]) of the disjunctive questions.

The introduction of self-posed questions (by the agent, *via* Excluded Middle) and the consequent splitting of tableaux, suggest a simple explanation of a logical issue: the closing of all tableaux generated by a set of statements (representing *e.g.* one's beliefs) now means that all the conceivable situations potentially represented by the given set are impossible; the immediate consequence is that the agent, after having explored all the open possibilities, ceases to believe in the set (a process discovered with much fracas by the Pythagoreans, and usually called *reductio ad absurdum*). This point is very much related to the Artificial Intelligence area of *belief revision* ([Gärd88]), as will be discussed in §4.7.

4.3.3 Experimental analysis

An agent can acquire information from the world in many different ways, for instance the following:

- Looking up information stored in databases.

- Asking other (human or artificial) agents.
- Making some experiences or tests in the real world.
- Searching information on the Internet.
- Studying the measures obtained by a sensor.

All these ways of acquiring external information are embedded in our *rational inquirers* in their capability of obtaining information from the environment in the *experimental* dimension of analysis (recall that, in §4.2, we have argued that a rational agent needs to have a way of requesting information from its environment while analysing its beliefs). In this dimension agents are allowed to make questions to the environment, and to add the corresponding answers (externally obtained, as opposed to the internally obtained propositions of the logical analysis) to the left columns of the open tableaux of the logical analysis.

The root of this dimension can be traced (as Hintikka points out in [Hint88]) as far as Kant, who argued in his *Critique of pure reason* that Reason has to take into account observations of the environment. He argued that Reason must not approach Nature as a student, that takes everything that its teacher chooses to say for granted, but as a judge who formulates questions and compels the witnesses to answer them. This is the spirit that has inspired this dimension of analysis.

We allow the agent to make questions of the following style in the experimental dimension of analysis:

> *Does it exist an individual that has the properties* P_1, P_2, ..., P_n *and does not have the properties* P_{n+1}, ..., P_m*?*

These questions are more general than the *existential* questions suggested by Hintikka in his *interrogative model of inquiry* (briefly commented in §4.3.2), because the existence of an individual that has (or does not have) several properties is inquired. In these questions, P_i represent basic predicates. We assume that there are only two kinds of admissible answers from the environment:

- No, there are no known individuals that satisfy those properties.

- Yes, individual r satisfies those requirements.

Thus, positive answers must actually provide an specific individual that satisfies the requirements stated in the question made to the environment (*i.e.* it must be an individual o such that $P_1(o)$, $P_2(o)$, ..., $P_n(o)$ hold but $P_{n+1}(o)$, ..., $P_m(o)$ do not). Thus, this dimension of analysis has a distinctive *intuitionistic* or *constructive* flavour: it is not enough to know that there exists an individual with a given set of characteristics, but we ask for the name of one such specific individual.

This dimension is closely related to the logical dimension of belief analysis. The logical analysis can actually *guide* the experimental analysis, by suggesting which experiences or tests the agent should perform in the actual world (which questions should be put to the environment) to gain knowledge. The idea is that the agent could keep (logically) analysing formulæ until it finds some atomic formulæ that contain *Skolem constants*. If you look back at the tableaux calculus of the logical analysis shown in figure 4, you will notice that a new Skolem constant is generated in two different cases:

- Each time that a γ-formula (an existentially quantified formula or the negation of a universally quantified formula) is analysed in the left column of a tableau.
- Each time that a δ-formula (a universally quantified formula or the negation of an existentially quantified formula) is analysed in the right column of a tableau.

These constants do not refer to any specific object, but they refer to unknown individuals that must have the properties represented by the atomic predicates. Thus, the agent may increase its beliefs by eliminating the tableaux in which these Skolem constants appear, if it finds out that there are no individuals in the real world that satisfy the required properties (in §5 it will be shown how closing a tableau during the logical analysis implies a possible increase in the agent's set of positive beliefs). Therefore, the presence of Skolem constants in the atomic formulæ of the open tableaux of the logical analysis is the trigger of the experimental dimension of analysis, linking in a novel fashion the traditional *rational* and *empirical* components of rational inquiry ([ReBr79]).

Unlike other authors, we have not tried to provide in this dissertation a formal explanation of the circumstances in which a question may arise from a set of formulæ. The interested reader may consult *e.g.* [Wiśn95], in which Wiśniewski, in the context of *erotetic logic*, provides such a formal framework (and reviews many other similar frameworks, such as Hintikka's *interrogative model of inquiry*). He argues that a set of declarative sentences S raises a question q when the following conditions hold:

- S does not contain any direct answer to q.
- It is not possible to derive from S any direct answer to q.
- All presuppositions of q may be derived from S[11].
- If all the formulæ in S hold, q must have a true direct answer.

Most of these conditions would be applicable to the questions made in the experimental dimensions of analysis, except the second one (as our agents are not ideal reasoners they do not know all the logical consequences of their beliefs and, therefore, they cannot know whether an answer to the question put to the environment may be deduced from their explicit beliefs[12]). The framework developed in [Wiśn95] considers several types of questions, for instance the following:

- Given an exhaustive finite set of possible answers, the environment may be asked which one is the real answer (*e.g.* choose one from the set $\{A, \neg A\}$).
- Given a certain predicate, the environment may be asked for a set of individuals satisfying it (*e.g.* a search for one pair of individuals x, y such that $father_{x,y}$ holds).
- The latter kind of questions may be generalised to require n sets of individuals that satisfy the predicate.

[11] A formula p is a *presupposition* of a question q iff p is entailed by each direct answer to q ([Beln66]).

[12] In fact, even Wiśniewski argues in [Wiśn95] that this particular postulate is somewhat controversial.

- In a further generalisation, it is possible to ask for *all* sets of individuals that satisfy a given predicate.

A part of the example developed in §4.6 may help to illustrate our vision about how the exploratory dimension of analysis should behave, especially the implications that each kind of admissible answer may have in the development of the analysis of the agent's beliefs. This example is also useful to notice the relationships that may be established between the different dimensions of analysis, in particular between the logical and the experimental ones. Suppose that the agent doubts of the validity of a general law such as *All birds fly* (*i.e.* it is not sure whether $\forall x\ (Bird(x) \Rightarrow Flies(x))$ holds). It may introduce this doubt into the analysis by adding to its current belief set (in the exploratory dimension of analysis) the formula $\forall x(Bird(x) \Rightarrow Flies(x)) \vee \neg\forall x(Bird(x) \Rightarrow Flies(x))$, which is an instance of the AEM. When the agent analyses this disjunction (in the logical dimension of analysis, using the rule that permits the analysis of β-formulæ located in the left column of the tableaux), it will have access to two subtableaux: one of them would contain the law in the left column, whereas the second would contain $\neg\forall x(Bird(x) \Rightarrow Flies(x))$. From this latter formula, by standard first-order tableaux processing the agent would get $Bird(a)$ and $\neg Flies(a)$ (for some (undetermined) a, represented by an Skolem constant). At this point of the analysis, the agent can notice that it can increase its beliefs if it can close this tableau. In order to do that, it can make the following question in the experimental dimension of analysis: is there any individual a such that $Bird(a)$ and not $Flies(a)$? The agent could react in different ways, depending on the received answer:

- The answer is positive.

  In this case, the answer must provide an individual with those properties, and then the Skolem constant may be replaced by that specific individual before proceeding with the analysis. Note that a positive answer only modifies the content of the tableau that contains the Skolem constant that has triggered the question to the environment. The rest of the tableaux are not modified because the aim of the question was basically to corroborate (or refute) the existence of an specific individual with the properties represented by the predicates that are applied to the Skolem constant in that tableau, in order to find out whether the

situations represented by that tableau match the agent's environment. Having said that, it would not have been unreasonable to consider the possibility of adding the positive answer to the left columns of all the open tableaux of the logical analysis, as if it had been directly given by the environment as an external unsolicited input (as will be seen in §4.4).

- The answer is negative.

  Then, the agent can conclude that an individual with these properties does not exist; when this conclusion is finally reached, the agent can see the tableau that contains the formulæ with the Skolem constants as the representation of a class of *empirically impossible* situations, and so it will have grounds not to consider it any more as a conceivable, realizable alternative. Then this tableau would be *empirically* closed and dismissed from the analysis. At that moment all the open tableaux would contain $\forall x(Bird(x) \Rightarrow Flies(x))$ and the agent would believe this law (as will be argued in the formal modelling of the evolution of the beliefs shown in §5). This example (extended in §4.6) shows how the agent may combine different dimensions of analysis in order to keep refining its set of beliefs.

A few comments about *empirical closings* of tableaux are in order:

- When a tableau is *empirically closed*, it is not taken into account when the (somewhat modified) Kripke semantics is applied to compute the agent's actual set of beliefs (as will be seen in §5). In that respect, *empirical closings* behave like the *logical closings* made in the logical dimension of analysis.

- However, there is an important difference between *logical* and *empirical* closings. If a tableau is *logically closed*, it is dismissed from the analysis permanently, and it can never be considered again. However, tableaux that are only *empirically* closed do still form part of the tableaux tree. The agent is still allowed to continue with the analysis of formulæ contained in those tableaux. The spirit of this difference is that information acquired *a posteriori* may make the agent reconsider the *empirical* closing (*e.g.* it may learn about the existence of a bird that does not

fly), whereas a logical contradiction may never be retracted. Thus, this difference provides the agent with a way of having non-monotonic beliefs.

Thus, as Hintikka suggested in [Hint86a], doubt can be understood as the beginning of a *dynamic* process that, by reducing the number of conceivable situations that the agent considers, reinforces one side of the doubt over the other, sets the conditions to verify it (and falsify the other) and tendentially gives credence to it. Such a process can be made compatible, in a natural way, with *e.g.* falsification strategies in the Popperian philosophy of science ([Popp34]). It can also explain why humans finally *know* things; they simply make themselves present, not directly (through the senses, as it were) but indirectly, through their involvement in a reinforcement/disabling process which eliminates conceivable situations (that are then seen as "impossible") and so reinforces -as *belief*, now turned into *knowledge*- what active experience has indirectly but forcefully shown. This analysis not only matches some dynamic models of concept formation in Psychology or Artificial Intelligence, but it additionally suggests a simpler approach to the "justification" and "truth-tracking" concepts as the philosophers' missing ingredient for *knowledge*; it is certainly more akin to Barwise and Perry's idea that knowledge is *"succesful belief"* ([BaPe83]) than to standard epistemological traditions (although the idea of *"succesful belief"* in [BaPe83] is of a probabilistic nature).

4.4 External inputs

We also allow the agent to incorporate into the analysis the information that it receives directly from the environment. This information, unlike the one obtained in the experimental dimension of analysis, is not requested by the agent; it may be received from sensors or from other agents of the system. These new pieces of information are added to the analysis by introducing them in the left columns of the open tableaux of the logical analysis. This decision will cause the agent to belief immediately the received information, as will be shown in §5; thus, as argued in §4.2, rational inquirers are very *credulous* agents, that, in principle, believe everything they are told (as usual in the Artificial Intelligence literature of belief revision and update). We also argued in §4.2 the feasibility of treating all kinds of external inputs in the

same way, regardless of their source; therefore, a rational inquirer will not make a distinction between a message received from another agent and a piece of data measured by a sensor.

The addition of the new information may cause a tableau to be closed, if it fulfills any of the tableau closing conditions and the agent realizes that fact. Thus, a tableau may be closed by different reasons:

- As a result of the application of a rule of the tableaux calculus of the *logical* analysis, a tableau may contain a formula and its negation in the left column, or it may contain the same formula in both columns. If the agent notes that any of these conditions holds, it will close that tableau (a purely *logical closing*) and it will cease to take it into account in the rest of the belief analysis.

- As a result of information received directly from the environment (an external input), one of the two closing conditions may hold in a tableau; thus, if the agent realizes that situation, it will close it. This is also a *logical* closing (due to empirical reasons, if you wish). It will be argued in §4.7 that an agent that performed some kind of belief revision procedures could distinguish between these two different styles of logical closing in order to decide which formulæ to eliminate from its inconsistent set of beliefs in order to regain consistency.

- The agent may fail to find in the real world a set of individuals that satisfy the properties stated in a tableau (in an inquiry made in the *experimental* dimension of analysis). Thus, it may consider that the tableau represents a situation that does not match reality, and it may *empirically* close the tableau in order to (at least, temporarily) eliminate it from further analysis.

There is one important issue associated to the process of closing tableaux, namely the difference between *monotonic* and *non-monotonic* beliefs. The former are the ones obtained from the purely logical closing of tableaux, while the latter are those derived from the *empirical* closings of tableaux, that may be defeated later. For instance, the agent can, at a later stage of the belief analysis, notice that the answer from the environment to a previous question (in the experimental dimension of analysis) was not accurate, and that there

indeed exists an individual with the requested set of properties. This means that this work has some connections with the AI field of *belief revision*; some comments on this issue will be made in §4.7. An example where the agent's set of beliefs evolves in a non-monotonic way is developed in §4.6 and §5.8.

4.5 On logical omniscience and perfect reasoning in rational inquirers

After having described all the different dimensions of belief analysis that rational inquirers may perform, it is worth pausing for a moment and wondering to what extent it may be defended that this kind of agents are really an example of limited, non-ideal, non-logically omniscient agents. In particular, some of the readers of this dissertation may have thought, after reading the section devoted to the logical analysis and appendix A, that these agents are still logically omniscient and perfect reasoners (albeit not in standard predicate calculus, but in Kleene's three-valued logic). In this section we want to argue that this is certainly *not* the case.

Let us first consider the issue of logical omniscience (necessarily believing all tautologies). We can make two comments concerning this point. On the one hand, it is very easy to prove that there are no tautologies in Kleene's three-valued logic, because there does not exist any formula which is valid in the three-valued interpretation that assigns the truth value ω to all basic propositions. Therefore, the problem of believing all the tautologies in this logic vanishes. On the other hand, if we are concerned about *classical* first-order tautologies, we have provided in §4.3.2 an argument against the free introduction of tautologies in the tableaux of the belief analysis. The only classical tautologies that may be introduced in the tableaux in the exploratory dimension of analysis are the instances of the *Axiom of the Excluded Middle*. This allowance, in turn, does *not* mean that rational inquirers believe, by definition, all the possible instances of this axiom; rather, it is saying that the agent has the possibility of, at a certain point of the analysis, believing a *particular* instance of AEM. The agent has to make a conscious effort to introduce this formula in the tableaux, has to spend both time and space to do it. In summary, a limited, resource-bounded rational inquirer could never believe the infinite possible instances of the AEM; it will only believe those that it purposefully introduces in the analysis. This state of

affairs is radically different from believing, right from the very beginning, every valid formula.

As far as perfect reasoning is concerned, several comments are in order. At a given stage of the analysis, a rational inquirer will only believe, as will be seen in §5, those formulæ that appear in all open tableaux (the leaves of the tableaux tree). It should be clear by now that the sets of formulæ contained in each tableau are *not* logically closed (neither in classical first-order logic nor in Kleene's three-valued logic). In particular, it is impossible for a finite set of formulæ to be logically closed in any of these logics (and each tableau contains two finite sets of formulæ because a resource-bounded agent cannot maintain lists of infinite formulæ). Therefore, the (finite) belief set of a rational inquirer will never be logically closed.

It may also be considered whether the formulæ that are believed from the agent, from an analysis of an initial belief set Δ, are precisely the logical consequences of Δ in Kleene's three-valued logic. We can consider two different cases:

- If the agent only uses the *logical* dimension of analysis, it will indeed only believe formulæ that are logical consequences of its initial beliefs in Kleene's three-valued logic. However, it must be noted that the formulæ that are generated in the subtableaux are only *subformulæ* of the analysed formulæ; therefore, by restricting the analysis to this dimension, the agent could only obtain as new beliefs those consequences (in Kleene's three-valued logic) of the initial set of beliefs that are subformulæ of these initial formulæ. In summary, the agent would be very far away from being a perfect reasoner. If a rational inquirer wants to deduce a certain logical consequence ϕ of the initial set Δ, at least it needs to use the *exploratory* dimension of analysis to introduce the formula $(\phi \vee \neg\phi)$ in the tableaux, and then perform several steps of logical analysis until the option in which $\neg\phi$ holds is dismissed.

  In the extreme case, if the agent has enough resources to perform an exhaustive analysis of the tableaux tree, it could only discover whether the initial set of beliefs is inconsistent in Kleene's three-valued logic (if all the branches of the tableau tree are logically closed). It is quite hard to perform such an exhaustive analysis. One reason is that the number of tableaux may be exponential with respect to the number of initial beliefs; another one is that quantified formulæ may be instantiated in a

large number of ways, if there are many different objects in the domain of discourse. Therefore, it seems very unlikely that a resource bounded agent may be willing to engage in such an expensive analysis.

- If the agent uses all the dimensions of belief analysis, there are several facts that may cause it to believe formulæ that are *not* logical consequences of its initial beliefs. For instance, if the agent receives an external input, that (arbitrary) formula is introduced in all the left columns of all open tableaux; therefore, it is instantly believed. Another situation that may arise (as will be seen in the example used in §5) is that the agent poses a question, in the experimental dimension of analysis, and receives a negative answer. In that case, one of the tableaux is *empirically* closed and the agent could believe formulæ that are *not* logical consequences of its present beliefs. In the example presented in §5, the agent will believe, at a certain point of the analysis, that all birds fly, and this fact may not be deduced at all (neither in classical logic nor in Kleene's three-valued logic) from its initial beliefs.

In summary, a rational inquirer's set of beliefs does not coincide with the consequences (in Kleene's three-valued logic) of its initial beliefs:

- The agent will never believe all the logical consequences of its initial beliefs (in particular, because this is always an infinite set of formulæ).
- A rational inquirer may easily come to believe facts that are not deducible from its initial set of beliefs, due to aspects of the analysis such as the acquisition of external information, the *empirical* closing of tableaux that may happen in the *experimental* dimension of analysis, and the re-opening of empirically closed tableaux that may also occur in that dimension of analysis.

The example provided in the following section, which will be carefully analysed in §5.8, serves to illustrate these comments.

4.6 Analysing a set of beliefs

The aim of this section is to show, *via* an small example, how a *rational inquirer* can use the different dimensions of analysis over an initial set of

beliefs. In §5.8, the same example will be considered to show how the use of the *subjective situations* framework permits the formal modelling of the evolution of the beliefs of the agent during the analysis process. Let us assume that the agent's initial set of beliefs is the following:

$$\{Bird_{Tweety}, Flies_{Tweety}, Bird_{Piolin}, Flies_{Piolin}, Bird_{Woody}, \\ \forall x(Penguin_x \Rightarrow \neg Flies_x)\}$$

This set will be called Δ throughout the example. The predicates and constants will be abbreviated so that this set will be written as follows:

$$\Delta \equiv \{\mathtt{B_T}, \mathtt{F_T}, \mathtt{B_P}, \mathtt{F_P}, \mathtt{B_W}, \forall x(\mathtt{P}_x \Rightarrow \neg \mathtt{F}_x)\}$$

The initial state of the belief analysis is represented by a tableau, T_0, that contains in its left column the agent's initial set of beliefs. The agent may start the analysis by wondering whether all birds fly (note that neither this fact nor its negation may be deduced from the initial set of beliefs). It can incorporate this doubt into the analysis by using the *exploratory* dimension, in which it is allowed to add (in the left column of all open tableaux) instances of the *Axiom of the Excluded Middle*, *i.e.* formulæ of the form $(\phi \vee \neg\phi)$. Thus, the agent may add the formula $(\forall x(\mathtt{B}_x \Rightarrow \mathtt{F}_x) \vee \neg\forall x(\mathtt{B}_x \Rightarrow \mathtt{F}_x))$ in the left columns of all open tableaux (T_0). A new tableau, T_0', which is shown in figure 5, would be generated by the agent.

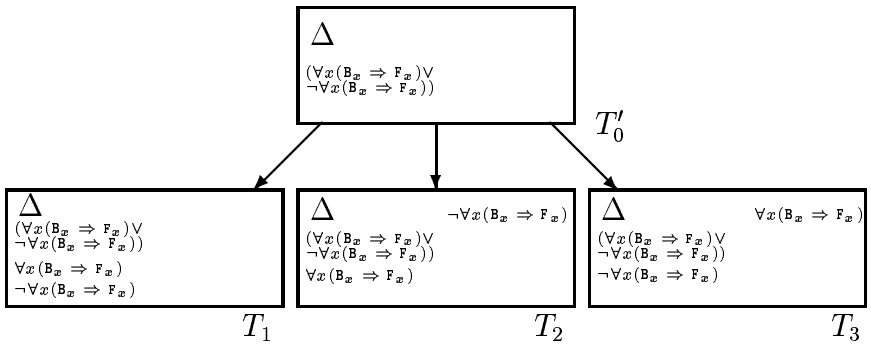

Figure 5: Logical analysis of $(\forall x(\mathtt{B}_x \Rightarrow \mathtt{F}_x) \vee \neg\forall x(\mathtt{B}_x \Rightarrow \mathtt{F}_x))$ in T_0'

The agent may proceed the analysis of its beliefs by applying one of the splitting rules of the tableaux calculus of the *logical* analysis (the one

that is used to analyse disjunctions located in the left columns of tableaux) to the formula that has just been introduced in the exploratory dimension, $(\forall x(\mathtt{B}_x \Rightarrow \mathtt{F}_x) \vee \neg\forall x(\mathtt{B}_x \Rightarrow \mathtt{F}_x))$. The result is the generation of three subtableaux, T_1, T_2 and T_3, as shown in figure 5.

The first subtableau, T_1, contains a formula and its negation in the left column. As both formulas come from the analysis of the same disjunction, the agent would surely recognise this fact and would *logically close* this tableau. Therefore, the remaining open tableaux at this point would be T_2 and T_3.

The agent may now choose to proceed its inquiry by performing the logical analysis of the formula $\neg\forall x(\mathtt{B}_x \Rightarrow \mathtt{F}_x)$ in tableau T_3. It must be noted that, in this example, we are not considering which process would be followed by the rational inquirer in order to decide which dimension of analysis should be applied or which formula should be analysed at each step of analysis. We are roughly following the strategy outlined in §4.3.3: the agent may be especially interested in analysing, in the logical dimension, those formulas that generate Skolem constants (δ-formulas on the left column or γ-formulas on the right column, recall figure 4). These constants guide the experimental dimension of analysis, which makes questions to the environment in order to adjust the agent's beliefs to what is true in its environment. The result of the logical analysis of $\neg\forall x(\mathtt{B}_x \Rightarrow \mathtt{F}_x)$ is a new subtableau, T_4, that contains (apart from all the formulæ in T_3) an instantiation of the (negated) universally quantified formula with a new *Skolem constant*, c (this tableau may be seen in figure 6).

Following the same strategy, now the formula $\neg(\mathtt{B}_c \Rightarrow \mathtt{F}_c)$ may be analysed in tableau T_4. This step causes the generation of a subtableau, T_5, that contains all the formulæ of T_4 and also $\mathtt{B}_c$ and $\neg\mathtt{F}_c$ (see figure 6).

The agent can now notice that an open tableau (T_5) contains some atomic formulæ with Skolem constants. These formulæ therefore, refer to properties held by generic individuals. It may try to find out whether an specific individual with the properties expressed in these atomic formulæ exists or not. If the agent is not capable of finding such an individual, that may be due to two reasons:

- Such an individual does in fact not exist.
- The agent's sources of information are not good enough (*e.g.* none of the databases accessible by the agent contains any information about any individual with the required properties).

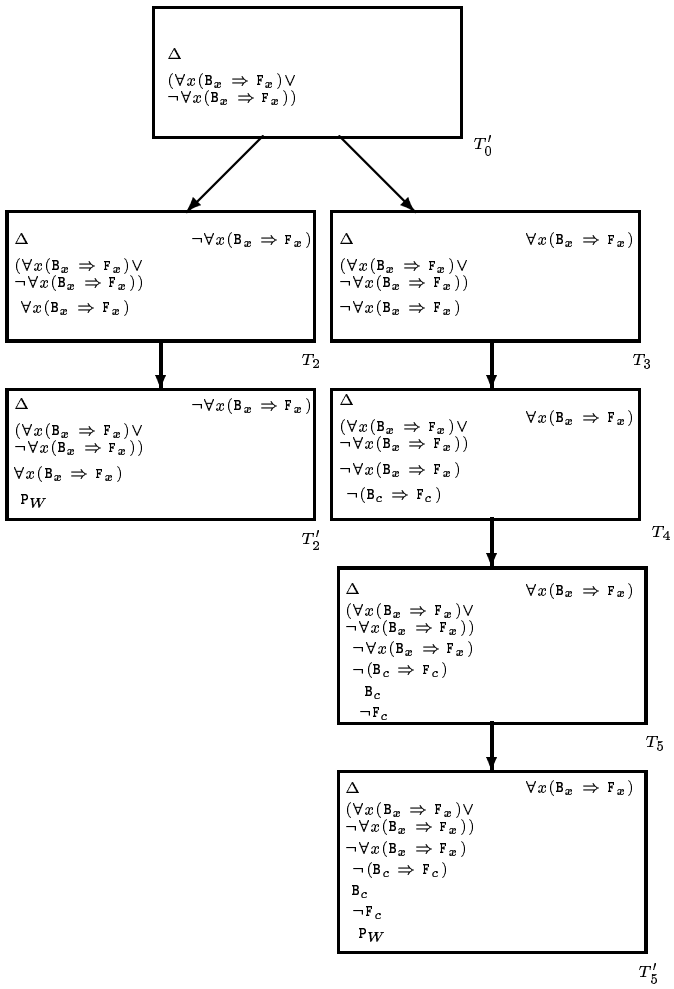

Figure 6: Incorporation of $\mathtt{P}_W$

As the agent does not know which is the case, it must leave an open door, so that, were it to find later an individual with the desired properties, it could accommodate this information into its set of beliefs. Thus, *rational inquirers* will have *non-monotonic* beliefs (*i.e.* a belief held at a certain point in time may be later dismissed, in the face of new information).

As will be apparent in §5, the main idea in the experimental dimension of analysis is that the agent may gain some beliefs by dismissing doxastic alternatives that contain atomic formulæ applied to Skolem constants, if it

is not able to find any specific individual with the properties represented by the predicates of the atomic formulæ.

In our example, the agent may start by checking whether T_5 (see figure 6) contains an individual that has property $\mathtt{B}$ and does *not* have property $\mathtt{F}$. There are some individuals that have property $\mathtt{B}$ (*Tweety*, *Piolin* and *Woody*) but none of them is known not to have property $\mathtt{F}$ (in fact, *Tweety* and *Piolin* do have it). Thus, the agent must resort to external sources of information in order to try to find other individuals that satisfy properties $\mathtt{B}$ and $\neg\mathtt{F}$.

Recall that in the experimental dimension of analysis (see §4.3.3) the agent may perform questions of this form:

> *Does it exist an individual that has the properties P_1, P_2, ..., P_n and does not have the properties P_{n+1}, ..., P_m?*

In the example developed in this section, the agent can formulate this question:

> *Does it exist an individual that has property* $\mathtt{B}$ *and does not have property* $\mathtt{F}$*?*

Notice that the experimental dimension of analysis provides the agent with a (somehow indirect) way of asking whether all birds fly. The two answers that can be received from the environment are:

- Yes, individual r satisfies those requirements (*i.e.* $\mathtt{B}_r$ and $\neg\mathtt{F}_r$ hold).

  In this case, the agent could substitute all the appearances of the Skolem constant c in T_5 by constant r and proceed with the belief analysis (*i.e.* $\neg(\mathtt{B}_c \Rightarrow \mathtt{F}_c), \mathtt{B}_c$ and $\neg\mathtt{F}_c$ would be replaced by $\neg(\mathtt{B}_r \Rightarrow \mathtt{F}_r), \mathtt{B}_r$ and $\neg\mathtt{F}_r$). As noted in §4.3.3, it could also be possible to add these formulæ in all the open tableaux.

- No, there are no known individuals that satisfy those properties (being a bird and not flying).

For the sake of proceeding with the example, let us assume that the agent makes the previous question (in the experimental dimension of analysis) and receives the second answer (*e.g.* it searches in the Internet for information

about birds that do not fly and it is not capable of finding any such individual). In this case, the agent may decide that the tableau T_5 is representing an *empirically* impossible class of worlds (because the Skolem constant c may not be given any specific value). Therefore, the agent may decide not to have this class of situations into account when computing its beliefs. The actual implementation of this decision is the *empirical closing* of T_5.

Recall the implications of the *empirical* closing of a tableau (see §4.3.3):

- An *empirically closed* tableau is no longer taken into account to calculate the agent's set of beliefs (see §5).
- An *empirically closed* tableau may still be used in the multi-dimensional belief analysis (*e.g.* its formulæ may still be logically analysed). However, all the branches that may be generated by that analysis are also considered to be *empirically closed* (some of them may even be *logically closed*, if some subtableau is considered *logically impossible*).
- An *empirically closed* tableau may be *re-opened* at a later stage of the analysis, if an individual with the required properties is finally found (*e.g.* after performing more logical analysis, or by receiving direct external information).

After having *empirically closed* the tableau T_5, the only open tableau of the logical analysis is T_2. Imagine that, at this point of the analysis, the agent receives (from some external source) the information that *Woody* is a penguin ($\mathtt{P}_W$). As mentioned in §4.4, the difference between the process of receiving external information and the experimental dimension of analysis is that, in the latter, the agent makes an specific question and waits for a concrete answer to that question, whereas in the former the agent cannot control the information that it receives unexpectedly from the environment (*e.g.* it can not prevent another agent from sending any kind of information).

As we said in §4.4 the behaviour of our *rational agents* in front of external information will be somehow *credulous*: they will take that information as reliable, and they will incorporate the formulæ representing that information in the left columns of all non-logically closed tableaux (*i.e.* all open tableaux and all *empirically closed* tableaux). This decision is motivated by the aim of showing that our framework may model both the reception of presumably defeasible information (in the experimental dimension of analysis) and the

acquisition of presumably reliable information (in the addition of external inputs). Following the example, the agent now adds the new information ($\mathtt{P}_W$) in the left columns of all non-logically closed tableaux (T_2 and T_5), generating two new subtableaux, T_2' and T_5'. This situation is depicted in figure 6.

Remember that one of the formulæ of the initial belief set (Δ) was $\forall x(\mathtt{P}_x \Rightarrow \neg\mathtt{F}_x)$. Therefore, this formula appears in the left column of both T_2' and T_5'. The agent can decide to analyse it in T_5' (recall that we allow the agent to perform analysis of formulæ that are contained in tableaux that are only *empirically closed*, as mentioned in §4.3.3). To analyse a universally quantified formula contained in the left column of a tableau, a specific instantiation of that formula has to be chosen. Assume that the agent instantiates the formula with the constant W; then the result of the analysis is the generation of a new subtableau, T_6, that contains all the formulæ in T_5' and also $(\mathtt{P}_W \Rightarrow \neg\mathtt{F}_W)$ (see figure 7). The agent could have chosen this particular instantiation due to the fact that the formula that has just received from the environment matches the antecedent of the conditional. If the agent decides to analyse the formula $(\mathtt{P}_W \Rightarrow \neg\mathtt{F}_W)$ in T_6 it will generate three new subtableaux, T_7, T_8 and T_9, as shown in figure 7.

The agent could now notice that both T_7 and T_8 contain a formula (P_W) and its negation in the left column. Having noted this fact, it could decide to *logically close* these tableaux, and to remove them from the tableaux tree. Summarising, at this point of the tableaux analysis the situation is the following:

- The tableau T_2' is the only open tableau (and, as such, is the only one that should be taken into account when computing the agent's set of beliefs).
- The only other remaining tableau in the tableaux tree is T_9. This tableau appears in the branch of tableau T_5 and, therefore, is considered to be *empirically closed*.

Recall that tableau T_5 was *empirically* closed because it contained the formulæ $\mathtt{B}_c$ and $\neg\mathtt{F}_c$ (being c an Skolem constant), and the agent did not know (and could not find in its environment, in the experimental dimension of analysis) any individual with those properties (being a bird and not flying).

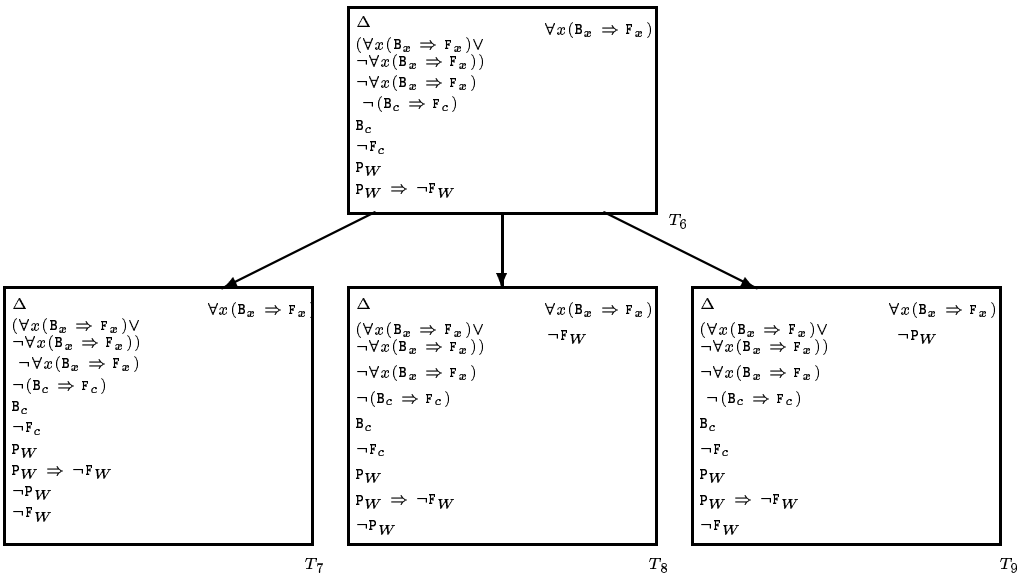

Figure 7: Logical analysis of $(\mathtt{P}_W \Rightarrow \neg\mathtt{F}_W)$ in T_6

Thus, that tableau was temporarily dismissed from the analysis. Now, after having introduced external information in the tableaux of the belief analysis ($\mathtt{P}_\mathtt{W}$) and having performed some logical analysis, the situation has changed radically. The only subtableau in the branch of T_5, T_9, contains $\mathtt{B}_W$ (that belongs to Δ) and also $\neg\mathtt{F}_W$; thus, the agent now has discovered an individual with the desired properties. Therefore, the motivation for closing that branch of analysis has disappeared, and the agent should decide to reconsider its previous closing decision by *re-opening* the tableaux of that branch of the tableaux tree (only T_9, because T_7 and T_8 were *logically* (and, therefore, permanently) closed). Moreover, the presence of the Skolem constant c, that denotes an unknown, generic individual, also seems irrelevant, now that a real individual with the desired properties is known. Therefore, the agent could decide to generate a subtableau of T_9, called T'_9, in which all the appearances of c have been replaced by constant W. This generation is shown in figure 8 (recall that $\mathtt{B}_W$ is included in Δ).

If the agent analyses the formula *"There does not exist any penguin that flies"* (that belongs to Δ) in T'_2, instantiating this formula with the constant W (as it did with the same formula in T'_5), it will obtain the tableau T_{10},

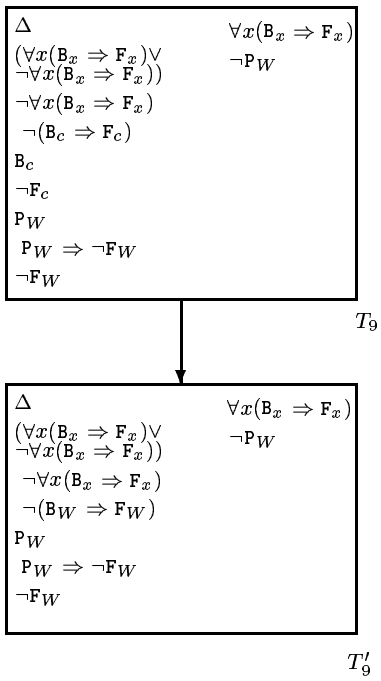

Figure 8: Replacing the Skolem constant c by W

shown in figure 9.

The agent could proceed the logical analysis with the formula that it has obtained in the previous step, $(\mathtt{P}_W \Rightarrow \neg\mathtt{F}_W)$ (as it did with the same formula in T_6). The result of the analysis of a conditional in the left column of a tableau is the generation of three subtableaux (T_{11}, T_{12} and T_{13}) as shown in figure 9.

The tableaux T_{11} and T_{12} contain P_W and $\neg P_W$ in their left columns; therefore, they would be *logically closed* and eliminated from the analysis. Thus, the only open tableaux at this point of the analysis would be T_9 and T_{13}.

The agent might now notice that it has not yet explored one of the sides of the doubt introduced in the exploratory dimension of analysis in the first step of the belief analysis. It can logically analyse the formula $\forall x(\mathtt{B}_x \Rightarrow \mathtt{F}_x)$, contained in the left column of T_{13}, instantiating it with the constant W. The result is the generation of a new subtableau, T_{14}, in which this instantiation

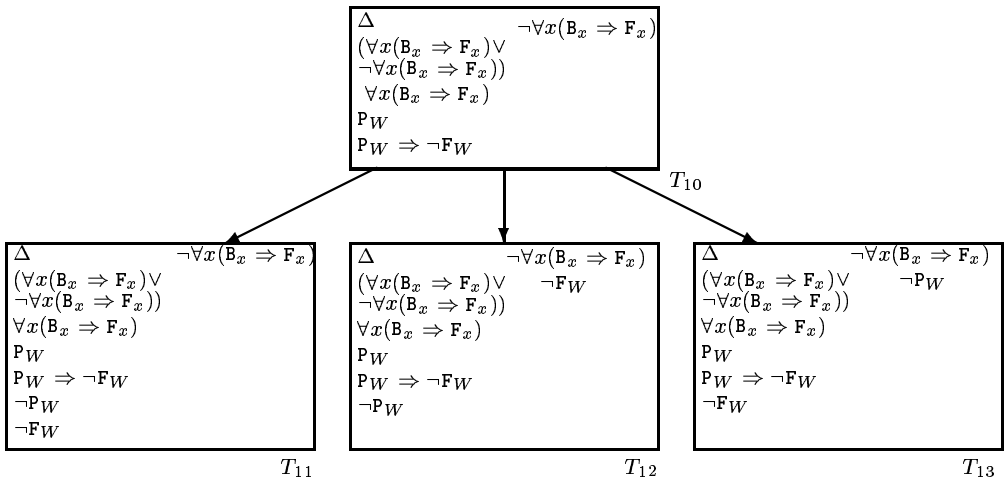

Figure 9: Logical analysis of $(\mathtt{P}_W \Rightarrow \neg\mathtt{F}_W)$ in T_{10}

of the analysed formula appears in the left column, along with all the formulæ that were already in T_{13} (see figure 10).

In the last step of this example the agent could analyse the formula $(\mathtt{B}_W \Rightarrow \mathtt{F}_W)$ in T_{14}. This logical analysis causes the generation of three new subtableaux (T_{15}, T_{16} and T_{17}) as shown in figure 10.

Notice that these three subtableaux may be all *logically closed*, because all of them contain a formula and its negation in the left column ($\mathtt{B}_\mathtt{W}$ in T_{15} and T_{16}, and $\mathtt{F}_\mathtt{W}$ in T_{15} and T_{17}). Having done this, the only remaining open tableau of the analysis would be T_9.

4.6.1 Summary of the example

The analysis performed in this example may be summarised as follows:

- The agent starts the analysis by wondering whether all birds fly. It uses the *exploratory dimension* of analysis in order to introduce this doubt into the analysis. In this way, the agent may explore the two available alternatives and determine whether any of them is logically impossible, or whether there is any question that it can make to its environment in order to confirm or refute any of the alternatives.

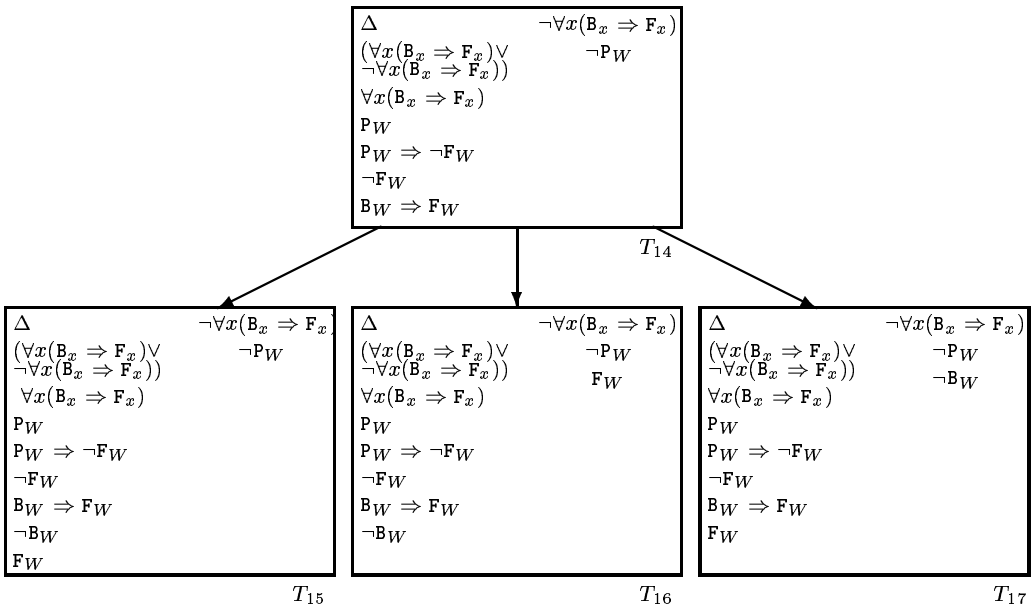

Figure 10: Logical analysis of $(\mathtt{B}_W \Rightarrow \mathtt{F}_W)$ in T_{14}

- After some *logical analysis*, the agent discovers that it can dismiss one of the options if it can check whether there is an individual that is a bird and does not fly. The *experimental* dimension of analysis is used to (unsuccessfully) search in the environment for an individual with these properties.

- As the agent does not find any individual with the desired properties, it decides to temporarily dismiss that alternative; then, it should believe that *"All birds fly"* (because it has not been able to find any counterexample).

- Afterwards, the agent receives external information that assures that *"Woody is a penguin"*. The agent incorporates this information in the analysis by adding it to the left columns of all the tableaux that have not been logically closed.

- After some logical analysis, the agent discovers that there is indeed an individual that is a bird and does not fly (*Woody*), and that it may

re-open the tableau that was empirically closed. Having made this discovery, it should cease to believe that *"All birds fly"*.

- After further logical analysis, the agent discovers that all those situations in which *"All birds fly"* are *logically impossible*. Then, it should reach the final conclusion that *"Not all birds fly"*.

The actual formal modelling of the non-monotonic evolution of the agent's beliefs in this example will be performed in §5.8; there, it will be shown how the successive agent's sets of beliefs after each step of analysis match the ones that would be intuitively expected.

4.7 Belief revision and update

There is a whole area within Artificial Intelligence devoted to deal with the issues related to *belief revision* (how to change the beliefs about an static world) and *belief update* (how to keep updating the beliefs to match an evolving world). These topics have received much attention, especially from the logical approaches taken by Alchourrón, Gärdenfors and Makinson (see *e.g.* [AGM85], [Gärd88], and some of the comments made in §4.2) and Katsuno and Mendelzon ([KaMe91]). One of the big issues in belief revision is what to do when finding out that a set of beliefs is contradictory; in our framework, that would be the case when all the analytic tableaux are closed. This is indeed an important topic, but it has not been tackled in this dissertation. We will just assume that the rational inquirers could use any of the proposals that have been made in the belief revision area in order to decide which formulæ should be abandoned when it is discovered that the present set of beliefs is logically inconsistent. Having done that, they could start a brand new analysis of its new set of beliefs. We do have a suggestion to make, though: *the agent should take into account where the formulæ of the tableaux came from when it has to decide which of them should be given up in order to regain consistency*. Notice that a formula may have several origins:

- It may be an initial belief.
- It may have been deduced from previous beliefs by deductive inference rules.

- It may have been obtained in the *experimental* dimension of analysis (see §4.3.3) as an answer to a question put to the environment by the agent.
- It may have been a measure obtained by a sensor.
- It may have been communicated by external agents.

A rational agent may have different policies (strategies) to deal with an inconsistent set of beliefs. It could decide for instance that it should try always to keep its initial belief set and its deduced beliefs, and that it may prefer to eliminate those formulæ that come from not-fully reliable sensors or from untrustworthy agents. In a different setting, in which there were perfect sensors and the good behaviour of all the agents were assumed, a rational agent could for instance choose to trust the information proceeding from these sources and to get rid of initial beliefs; that could be a way of updating the belief set in front of a dynamically evolving world. Thus, we leave the choice of the belief revision strategy to be used by the rational inquirers absolutely open, so that it may be adapted to the varying circumstances in which they may be located. These strategies could be implemented with the basic belief revision frameworks proposed in the literature. For instance, following Gärdenfors' suggestions ([Gärd88]), the agent could assign different degrees of *epistemic entrenchment* to its different beliefs, according to its procedence. In that way, in the face of contradictory information, the agent would get rid of those beliefs that had the minimal entrenchment. It should be noted, though, that most of the formalisms that have been developed in Artificial Intelligence deal with logically closed sets of beliefs; therefore, they should probably have to be somewhat modified to deal with arbitrary sets of beliefs.

There have been two recent dissertations ([Gerb99], [Lomu99]) that have considered the complex issue of belief revision policies in the context of multi-agent systems; however, both of them deal with ideal agents. Gerbrandy ([Gerb99]) assumes a $K45$ doxastic logic. She acknowledges the problem of logical omniscience, but limits herself to comment that it may be somewhat weakened if the universal modal operator is interpreted as *information*, rather than as belief or knowledge (following an idea advocated by Barwise in [Barw88]). Another difference of her approach with ours is that she deals with the change of information of an agent that learns *modal* formulæ, whereas

our rational inquirers only deal with first-order formulæ. The main tool used in her modelling technique is *dynamic logic*, which has not been used either in this dissertation. She also reviews other approaches to modelling information change in multi-agent settings, but all of them belong to the standard tradition, in which logical omniscience is simply ignored; see *e.g.* Veltman's *update semantics* ([Velt96]), in which the underlying doxastic logic is $S5$, or the approach taken in [FHMV95], where agents are also assumed to be ideal believers, under the logic $K45$. Lomuscio ([Lomu99]) studies how the agents in a multi-agent system may share knowledge through communication. He also admits that he restricts his attention to *ideal* agents, whose knowledge is ruled by the modal logic $S5$. His approach to evolving knowledge relies heavily in the framework described in [FHMV95]. Friedman and Halpern ([Frie97], [FrHa97], [FrHa99]) have also proposed a way of modelling belief in evolving systems. In their framework, belief is defined on top of the notions of knowledge and plausibility: an agent believes ϕ if it knows that the plausibility of ϕ is greater than that of its negation. They affirm that, assuming natural properties of these two basic notions, belief turns out to be axiomatised by the standard modal logic $KD45$.

All these approaches show the interest that the Artificial Intelligence community has in the topic of multi-agent belief revision, but they also reflect the fact that the logical omniscience problem has been, up to now, to the best of my knowledge, mostly ignored.

4.8 Summary

We started this chapter by arguing which were the main activities that influenced on a rational agent's set of beliefs. We identified four types of doxastic activities:

- Performing deductions on the set of beliefs.
- Incorporating doubts into the set of beliefs.
- Adding information received from the environment to the set of beliefs.
- Requesting specific items of information from the environment, in order to keep the set of beliefs as close as possible to the facts that hold in the agent's environment.

All of these activities are oriented towards the rational agents' main goal, which is to keep its beliefs as close as possible to the facts that are true in the real world. In order to present a specific way of modelling the evolution of the beliefs of a rational agent in §5, we have defined in this chapter a more concrete instantiation of this kind of agents, namely the *rational inquirers*. These agents are permanently engaged in a multi-dimensional belief analysis, which matches the doxastic activities listed above:

- They may perform some deductive inferences, using a modified version of the analytic tableaux calculus (*logical* dimension of analysis).
- They may introduce doubts into the analysis by adding instances of the Axiom of the Excluded Middle to the tableaux (*exploratory* dimension of analysis).
- They may request information from the environment, in the *experimental* dimension of analysis, in order to refine its set of beliefs. These questions are triggered by some of the formulæ obtained in the logical analysis.
- They may also incorporate to the tableaux the formulæ that they receive directly from the environment.

We have showed a small example in which the agent uses all these dimensions of analysis. In §5 the same example will be used to show how we can model the evolution of the agent's beliefs during the analysis process.

5 Modelling the evolution of beliefs

5.1 Introduction

We started this proposal by stating the problems of logical omniscience and perfect reasoning and reviewing the main ideas that have been put forward to reduce them as much as possible. In §3 we made our own proposal, based on the notion of *subjective situations*. In §4 we proposed a certain class of rational agents, called *rational inquirers*, and we described how they are capable of analysing an initial set of beliefs, with the aim of keeping it as close as possible to what is true in their environment. The underlying idea is that the agent's beliefs must surely change after each step of analysis (*e.g.* each time that the agent applies a rule of the analytic tableaux method in the logical dimension of analysis, or each time that the agent introduces a doubt into the analysis in the exploratory dimension of analysis).

In this chapter our aim is to provide a way of formally modelling the evolution of the beliefs of a rational inquirer, using some of the ideas that were shown in the framework of subjective situations. As will be apparent, we are also intending to keep the flavour of the *possible worlds model* and the *Kripke semantics* as much as possible. The main elements to be used in the modelling process are the *conceivable situations*, which are the semantic entities that correspond to the subjective perception that an agent has about its environment (see §5.2) and a *sequence of accessibility relations*, that serves to keep track of the conceivable situations that are considered as doxastic alternatives of the agent at each point of the belief analysis (see §5.3). After describing these concepts, we explain with detail the modelling process: which are its basic ingredients (§5.4), how the set of doxastic alternatives is updated after each step of belief analysis (§5.5), which are the different ways in which the set of doxastic alternatives may change (§5.6) and which is the final algorithm to be used in the modelling process (§5.7). After that, we apply it over the example developed in §4.6 to show which would be the agent's beliefs after each step of belief analysis. Even though we are going to work with rational inquirers, the reader should realise that the modelling techniques used in this chapter could be easily modified in order to be applied to any kind of rational agent, regardless of the especific way in which it carried out its doxastic activities. The reader should also bear in mind that we are only going to model the evolution of the beliefs that a single agent has

about the facts that hold in its environment; the problems that should have to be faced if a whole multi-agent setting with nested beliefs were considered will be sketched in §6.

5.2 Conceivable situations

In §2 we reviewed several proposals that have tried to solve (or, at least, partially alleviate) the problems of logical omniscience and perfect reasoning, both in Artificial Intelligence and in Philosophy. A particularly interesting suggestion was made by Hintikka in [Hint75a], where he proposed the idea of considering *[logically] impossible [epistemically] possible worlds* (see §2.2.1). Hintikka defines them as *"those worlds that are so subtly inconsistent that the inconsistency could not be expected to be perceived by an everyday logician, however competent"*. He identifies this kind of worlds with those urn models which vary so subtly as to be indistinguishable from invariant ones at a certain level of analysis ([Rant75], see §2.2.14). The concept of *impossible possible worlds* has been one of the main sources of inspiration in our work.

The main roots of the problems of logical omniscience and perfect reasoning are the assumptions of *completeness* and *consistency* underlying the possible worlds model. Recall the definition of Kripke structures given in §1.2.3: since worlds are *complete*[13] the agent is forced to have beliefs about the way that everything is in all of the accessible worlds; moreover, since worlds are *consistent*[14] as well, everything that follows from the agent's beliefs must also be believed. In short, in a propositional setting a classical *possible world* may be seen as a propositional *model* (in the logical sense of the word, *i.e.* as an interpretation of the basic propositions extended in the usual way to all the formulæ of the language). Therefore, a natural solution to these problems could be reached by dropping these assumptions. What would happen if (possibly) *incomplete* and/or (possibly) *inconsistent* possible worlds were allowed? This suggestion has been dismissed by most of the logicians and logically concerned philosophers of the Western tradition since Aristotle's time. Nevertheless, this possibility can be seriously entertained,

[13] Possible worlds are *complete* because every formula α is either satisfied or unsatisfied in each state s of every Kripke structure M.

[14] Possible worlds are *consistent* because it is not possible to satisfy both a formula α and its negation $\neg\alpha$ in a state s of a Kripke structure M.

and some philosophers have indeed argued for the feasibility of this kind of worlds, *e.g.* Rescher and Brandom in [ReBr79]. Many inconsistency-tolerant logics have been proposed in the literature; just to name a few, we can cite the logics for reasoning with inconsistent knowledge proposed by Roos or Lin ([Roos92], [Lin96]), Belnap's four-valued logic ([Beln77]), Priest's non-monotonic logic of minimal inconsistency ([Prie89]) or several paraconsistent logics ([AlNe84], [DCBB95]).

The presence of *partial* and *inconsistent* situations is the basis for the failure of logical omniscience. We may make some claims in favour of this kind of non-standard states of affairs, lest the reader think that these non-classical states are totally unacceptable. Most of these remarks are based on the study of the state of the art shown in §2. The *partiality* or *incompleteness* of possible worlds has been traditionally accepted in the Artificial Intelligence literature, the most common justifications being the following:

- The agent could be unaware of certain facts (as we have mentioned in §2.2.7, this issue was already taken into account in the *logic of general awareness*, [FaHa85]).

- The agent can have limited resources (*e.g.* the time required or the space needed to perform a given inference); therefore, it must have a bounded rationality ([NeSi72]). This is the most evident justification, if rational agents have to be implemented at all in a real computer. For instance, it is obvious that an agent cannot keep a belief database with infinite entries.

- The agent can ignore some relevant rules (*e.g.* the agent may have not been told what the rule of *Modus Tollens* is). This view was clearly considered in the *deduction model of belief* ([Kono86a]), where each agent was modelled with a base set of beliefs and a (possibly incomplete) set of inference rules (see §2.1.2). A similar idea is followed in [Bene97] or [GSGF93], where the notion of a *context* (an axiomatic formal system) is used to model the reasoning capabilities of *ideal* and *real* reasoners (see §2.3).

Inconsistency is a totally different matter. Anyway, it can be argued in its favour with a number of ideas:

- Some psychological tests show that human beings have difficulty in putting together all the information they possess, because human memory appears to be structured in *frames of mind* hardly communicating between them. Thus, an agent may be unable to take all its beliefs into account in every inference; if it focuses in a subset of them (call that a *context* ([McCa93]) or a *viewpoint* ([AtSi93])), it can draw conclusions which are consistent within the context but inconsistent if all the beliefs are considered at the same time. This argument was the main motivation behind the *logic of local reasoning*, ([FaHa85], see §2.2.10) and the *fusion model* ([Jasp94], see §2.2.13). This idea also underlies the concept of *multicontext systems* ([GSGF93], [Bene97], see §2.3). Many other researchers have also pointed out this fact (*e.g.* [Stal84], [Shoh91], [Delg95]).

- It is certainly possible to conceive the concept of inconsistent worlds and to define arguably interesting procedures of inquiry over them (as shown *e.g.* in [ReBr79]). It is even possible to depict this kind of worlds, as Escher proved so many times (see *e.g.* [Hofs80]).

- Human believers are rarely consistent, in the logical sense of the term; they will often have beliefs ϕ and ψ, where $\phi \vdash \neg\psi$, without being aware of the implicit inconsistency.

- It has been argued ([Kono86a]) that logical consistency is much too strong a property for resource bounded reasoners; being non-contradictory (not believing ϕ and $\neg\phi$ at the same time) is probably the most one can reasonably demand.

- If a theory is expressed in first-order logic, it is not even decidable in general whether it is consistent or not, so these theories would be pretty useless if they could be used only in case they were previously proved to be consistent.

- It can even be said that contradiction is the *norm* in the information that most real applications have to deal with. Most systems have databases or knowledge bases where information may be obtained from different sources, or from not fully reliable sources, and we should find ways to formalize the (potentially inconsistent or contradictory) data

> appropriately. In fact, dealing with conflicting data is part and parcel of what commonsense reasoning is all about. Moreover, when an agent detects an inconsistency in its beliefs, it may interpret that fact as a signal to take external actions, such as asking the user, invoking a truth maintenance system, activating/deactivating certain inference rules, *etc.* Thus, inconsistency may be seen as a useful tool to direct the processes of reasoning and learning, rather than as a roadblock to effective commonsense that must be avoided by any means ([GaHu91], [Perl94], [Perl97]).

Therefore, we are going to avoid the classical logicians' reluctance towards (possibly) incomplete and (possibly) inconsistent possible worlds; they will be considered as (epistemologically and even ontologically) possible as the standard complete and consistent possible worlds. Consider the positive side of this move; if worlds are incomplete and inconsistent, both logical omniscience and perfect reasoning seem to vanish. The agent can clearly fail to believe some tautologies, and it does not have to believe any logical consequence of its beliefs (recall §2.2.1).

In fact, the expression *possible world* does not convey exactly the idea that we have of what a doxastic alternative is; in our framework, we will call them *conceivable situations*, rather than *possible worlds*. A *conceivable situation*, as its name suggests, is any situation that the modelled agent may conceive, irrespective of its partiality or its consistency. It may be a situation that it has experienced, that it has been told of, or even a situation that it has just imagined as possible. The only condition for an scenario to qualify as a conceivable situation is that the agent considers it so; it does not have to be either consistent or physically realizable. The main point is that a conceivable situation is *not* tantamount to a model (in the logician's sense of the term). In the rest of the dissertation the notion of conceivable situation will be considered as primitive, and will correspond to what the modelled agent considers as *"realities"*, be they experiential or just imagined.

5.2.1 Formalizing conceivable situations

We will formalize the concept of *conceivable situation* (from now on, a *cosi*) using the formal tools of the framework of subjective situations. In §3.3 we defined the concept of *structure of subjective situations*. In each of these

structures every state was subjectively described from each agent's point of view, using the functions $\mathcal{T}_i$ and $\mathcal{F}_i$ for $Agent_i$. We will assume that each of these descriptions is the representation of a *cosi*; *i.e.* each state of an structure of subjective situations is seen as a collection of *cosis*, one for each agent of the multi-agent system. As we are dealing in this chapter only with a single agent's beliefs about the world, we will just take into account the *cosis* considered by this agent (which will usually be called i). Therefore, a *cosi is sintactically represented in our framework with two lists of first-order formulæ*: those that the agent assumes to be true and those that the agent assumes to be false. Recall that there are no constraints on the definitions of the functions $\mathcal{T}_i$ and $\mathcal{F}_i$; therefore, a given formula may belong to both of them, to only one of them or to none of them. This framework is reminiscent of the *situations* defined by Levesque in his *logic of explicit and implicit beliefs*, which was described in §2.2.4 (the main differences between our approach and Levesque's were already commented in §3.6). In some respects *cosis* are also similar to Barwise and Perry's *situations* ([BaPe83], [Barw88]), because they are not required to describe every aspect of the world, but maybe just a small portion of it; however, it must be taken into account that the *situations* in Barwise and Perry's *situation theory* are inherently consistent (*incoherent abstract situations* being the exception, as defined by Devlin in [Devl91]).

It may be said that the sets $\mathcal{T}_i(s)$ and $\mathcal{F}_i(s)$ represent the amount of positive and negative information that $Agent_i$ has in situation s. Following this conception, it is possible to define a partial order among *cosis* for each $Agent_i$ in the following way:

Definition 11 (Information order for $Agent_i$)

The information order among cosis for $Agent_i$, which will be denoted $\leq_i$, is defined as follows:

For all situations s, t, $s \leq_i t$ if and only if $\mathcal{T}_i(s) \subseteq \mathcal{T}_i(t)$ and $\mathcal{F}_i(s) \subseteq \mathcal{F}_i(t)$.

If $s \leq_i t$, it will be said that s is less i-informative *than t.*

A class of situations $\overline{w_\alpha}$ will be said to be less i-informative *than another class $\overline{w_\beta}$ if for every situation $s \epsilon \overline{w_\alpha}$ and every situation $t \epsilon \overline{w_\beta}$ it holds that $s \leq_i t$. This state of affairs will be noted $\overline{w_\alpha} \leq_i \overline{w_\beta}$.*

This definition implies the existence of a different partial order $\leq_i$ for each $Agent_i$. The least informative situations in this ordering $\leq_i$ are those *cosis* s in which $\mathcal{T}_i(s)$ and $\mathcal{F}_i(s)$ are empty, whereas the most informative situations are those in which these sets contain all the first-order formulæ. In §5.5 it will be explained how some of the modifications produced in an $Agent_i$'s set of beliefs as a consequence of its continuous process of rational inquiry may be formally modelled with a change in the set of doxastic alternatives that it considers as possible, replacing a certain set of *cosis* by another set which is more i-informative.

5.3 Dynamic accessibility relations

The goal of this research is the definition of a general model of the process of rational inquiry, so something must be added to the classical *static* possible worlds model (assuming that one indeed intends to keep the general conception of the model and the Kripke semantics) to make it suitable to model the *evolution* of the beliefs due to the analysis process. A very natural idea to model dynamic beliefs is to have some kind of variability in one of the main elements of the possible worlds model: the accessibility relation R.

Imagine that the agent's beliefs in world w have to be analysed. This world is R-connected to worlds w_1,w_2 and w_3. A certain proposition `P` is true in w_2 and w_3, but not in w_1; therefore, `P` is not believed in w. Assume also that `Q` is true in w_1 but false in w_2 and w_3 (therefore, it is not believed in w either). The agent, in the course of an inferential process performed on its beliefs, could reach the conclusion that `Q` is clearly unacceptable (*e.g.* it contradicts a large set of other actual beliefs). Therefore, it could conclude that the accessible worlds that contain `Q` are not viable alternatives to its present world, and thus they do not have to be considered accessible any more. This fact would imply that (w R w_1) would no longer hold, and that the set of doxastic alternatives to w would be reduced to $\{w_2,\ w_3\}$. But these two worlds contain `P`, and thus, *via* the standard Kripke semantics, the agent would now believe `P` in w (see figure 11). This example shows how a modification of the accessibility relation (in this case a *restriction* in the set of possible doxastic alternatives) can indeed model a modification of the beliefs caused by an internal inferential process of the agent (other sources of information could have been considered; *e.g.* the agent could have noticed the impossibility of `Q` as a result of an observation in its environment).

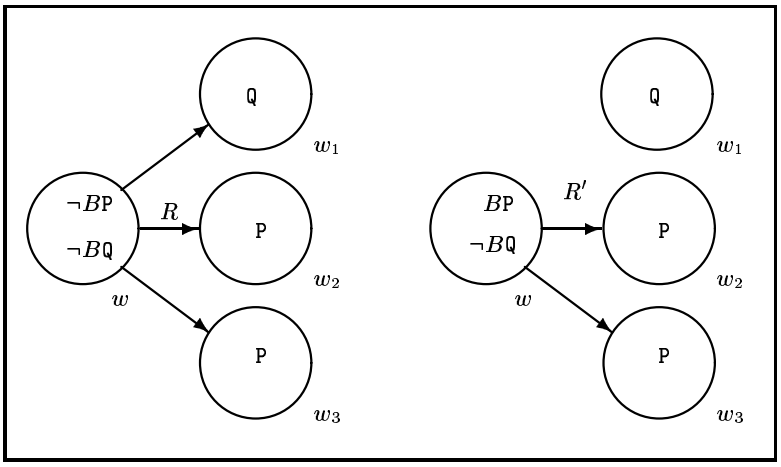

Figure 11: Belief change due to a change in R

This is another of the main ingredients of our modelling system: *we will model the evolution on time of the agent's set of beliefs as a sequence of accessibility relations*. Each of these accessibility relations will define the modelled agent's set of doxastic alternatives which, in turn, will induce (*via* a slightly modified version of the Kripke semantics, to be described in §5.4) a different set of beliefs in each point in time. In fact, the use of a change in R to model belief change is not new. Appelt describes a similar approach in [Appe85], where he argues that actions can generate knowledge by restricting the possible worlds that are consistent with the agents' knowledge after the execution of the action (following ideas from Moore, [Moor83], [Moor85]). Fagin *et al.* show in [FHMV95] how the reasoning processes followed by the (extremely idealised) *muddy children* in order to answer the question posed to them by their father (*"Does any of you know whether you have mud on your head?"*) may be modelled with a progressive restriction in the accessibility relation between the states that they consider possible[15]. As will be seen later, some of the modifications caused in a rational inquirer's

[15]However, Fagin *et al.* do not provide in [FHMV95] a formalisation of this process. Recent works ([Lomu99], [Gerb99]) have shown that it is not obvious at all how to provide an appropriate formal account of how the accessibility relations of the agents in a multi-agent system change as a consequence of their external communicative acts and their internal inference procedures.

set of beliefs by the analysis process may be modelled by using a sequence of *decreasing* accessibility relations.

Other approaches have also tried to provide a framework in which it is possible to model the evolution of a set of beliefs over time. Nirkhe *et al.* (see [NKP94]) show how *step-logics* may be used as a way to model the agent's ongoing process of reasoning; they even take into account the actual time that the agent consumes in its reasoning processes ([EMP95], [Elgo88], [ElPe90]). Kraus and Subrahmanian develop in [KrSu95] a family of temporal logics in which belief update is captured by how the agent's beliefs about the present are changing over time. A chapter of [FHMV95] is devoted to study how knowledge evolves in multi-agent systems (considering axioms that express some constraints on the semantics of temporal modal operators). The agents defined by Wooldridge in [Wool92] keep updating their beliefs, using inputs that may come from the result of previous actions of the agent or from messages received from other agents. The agents considered in [PaGi98] also modify dynamically the extent to which they trust their beliefs, taking into account the results of their inference processes or the external inputs that they may receive from sensors or other agents.

5.4 Basic ingredients of the modelling process

The following list shows the main ingredients that will be used to model the evolution of a rational inquirer's set of beliefs over time:

- Each two-columned analytic tableau (the formal object that is manipulated by the agent in the course of its inquiry) may be seen as *the representation of a class of conceivable situations: those in which all the formulæ in the left column hold and none of the formulæ of the right column holds.*

  For instance, the tableau T_3 of the example shown in §4.6 (see figure 12) represents all those *cosis* in which $(\forall x(B_x \Rightarrow F_x) \vee \neg\forall x(B_x \Rightarrow F_x))$, $\neg\forall x(B_x \Rightarrow F_x)$ and all the formulæ in Δ hold but $\forall x(B_x \Rightarrow F_x)$ does not hold. Thus, it would represent conceivable situations in which it is not true that all birds fly. In the framework of *subjective situations* outlined in §3 these situations s would be formally defined as those in which $\mathcal{T}_i(s) = \Delta \bigcup \{(\forall x(B_x \Rightarrow F_x) \vee \neg\forall x(B_x \Rightarrow F_x)), \neg\forall x(B_x \Rightarrow F_x)\}$

and $\mathcal{F}_i(s) = \{\forall x(B_x \Rightarrow F_x)\}$, being i the index of the agent whose beliefs are being modelled.

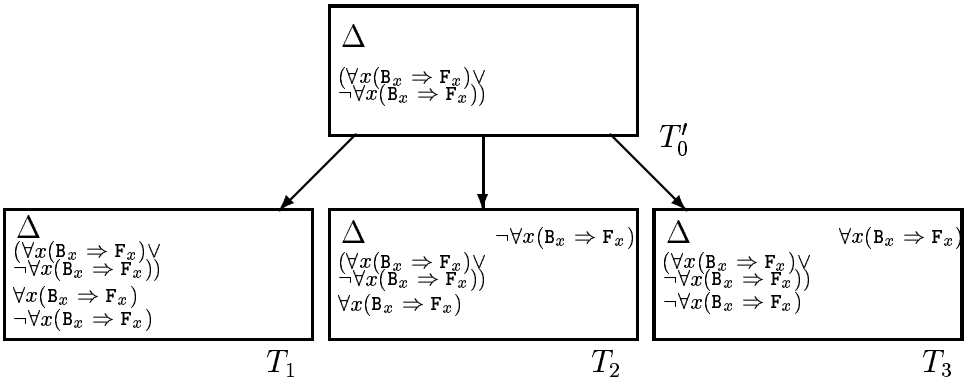

Figure 12: Logical analysis of $(\forall x(\mathtt{B}_x \Rightarrow \mathtt{F}_x) \vee \neg\forall x(\mathtt{B}_x \Rightarrow \mathtt{F}_x))$ in T'_0

In the rest of the dissertation, the reader should always bear in mind that the expression "ϕ holds in a *cosi*" is always referred *to the modelled agent's perception of the situation*; recall that the structures of *subjective situations* do not include an objective description of the possible worlds. Therefore, ϕ *"holds in* s*"* if it belongs to $\mathcal{T}_i(s)$, and it *"does not hold in* s*"* if it belongs to $\mathcal{F}_i(s)$.

- Therefore, when, at a certain stage of the analysis, the agent keeps a tableaux tree with a set of open tableaux, we may consider that *the agent's doxastic alternatives (possible states or possible worlds) are all those cosis represented by the open tableaux.*

  For instance, in the state of analysis depicted in figure 12 (and just before logically closing T_1) the agent would consider as doxastic alternatives all those *cosis* represented by the tableaux T_1, T_2 and T_3. Intuitively, the situations represented by T_1 are logically inconsistent, in the sense that they have information defending that all birds fly and that it is not true that all birds fly (these situations could easily arise, for instance when an agent receives contradictory information from different sources). The situations represented by T_2 are those in which it is assumed that all birds fly, whereas the ones in T_3 model those cases in which it is not true that all birds fly.

- At each point of the analysis, the set of doxastic alternatives will change, as the set of open tableaux changes. This continuous change will be modelled as a *sequence of accessibility relations* (a new accessibility relation will be generated after each step of belief analysis). In §5.5, we analyse the different situations that may turn out in the belief analysis performed by rational inquirers, and how the agent's set of doxastic alternatives changes in each case.

  Just to provide an example, consider the first step of logical analysis of the example shown in section 4.6, in which the agent analysed the formula $((\forall x(B_x \Rightarrow F_x) \vee \neg\forall x(B_x \Rightarrow F_x))$ in the tableau T_0' and generated three subtableaux, T_1, T_2 and T_3 (see figure 12). In the formal modelling method that we propose, the accessibility relation would change from a situation in which we consider as doxastic alternatives all those *cosis* represented by T_0' to a situation in which the doxastic alternatives would be all those *cosis* represented by any of the tableaux T_1, T_2 and T_3.

- When we have determined which are the agent's current doxastic alternatives, we can compute its present beliefs by using a modified version of the Kripke semantics, that takes into account the presence of positive and negative information about a *cosi*. The standard semantics is modified as follows:

  Definition 12 (Modified Kripke semantics)

  An agent's set of beliefs is obtained after applying the following rules to the positive and negative information of the agent's doxastic alternatives:

  - *The agent believes a formula if it holds in all its doxastic alternatives.*
  - *The agent does not believe a formula if it does not hold in at least one doxastic alternative.*

  To give an specific example, consider again the situation shown in figure 12, after the generation of T_1, T_2 and T_3. The tableau T_1 would be

logically closed and dismissed from the analysis, because it contains a formula and its negation in its left column; therefore, the only open tableaux at this stage of the analysis would be T_2 and T_3. We may obtain the agent's beliefs after this step of logical analysis by applying the modified Kripke semantics that has just been stated. The result would be the following:

- The agent would believe all the formulæ in Δ and $(\forall x(B_x \Rightarrow F_x) \vee \neg\forall x(B_x \Rightarrow F_x))$, because all these formulæ hold in all the *cosis* represented by T_2 and T_3.
- The agent would not believe $\forall x(B_x \Rightarrow F_x)$, because it does not hold in some doxastic alternatives; being more especific, it does not hold in all those *cosis* represented by T_3.
- The agent would not believe $\neg\forall x(B_x \Rightarrow F_x)$ either, because it does not hold in all the *cosis* represented by T_2.

Thus, at that point of the analysis the agent would believe a disjunction while disbelieving its two components.

5.5 Determining the doxastic alternatives

In order to prove the viability of the technique that has been presented in the previous section for the task of modelling the evolution of the beliefs of any rational inquirer, irrespective of the order in which it applies the different dimensions of analysis, it has to be carefully explained how the set of doxastic alternatives has to change after each doxastic task, *i.e.* we have to provide a systematic procedure for computing these changes. The different situations that may arise are the following:

- The rules of the analytic tableaux method of the logical dimension of analysis can be divided in two categories: the *splitting* rules (those that generate three subtableaux, *e.g.* the analysis of disjunctions in the left columns) and the *extending* rules (those that generate only one subtableau, *e.g.* the analysis of conjunctions in the left columns). The way in which the agent's set of doxastic alternatives changes after applying a logical rule of analysis is the same in both cases. First, the agent should apply the rule to a formula of a tableau, obtaining one

or three subtableaux. After that, it could devote some resources to finding out whether any of the resulting subtableaux may be *logically closed*. Having done that, there are two possibilities:

- If all the resulting subtableaux have been closed, the agent's doxastic accessibility relation is restricted by eliminating those *cosis* that were represented by the tableau that contained the analysed formula. If there was no other open tableau then the set of doxastic alternatives would be empty, and the agent would have found out that the initial set of beliefs was inconsistent; in that case, as commented in §4.7, it should apply a belief revision procedure in order to regain consistency, and start a brand new belief analysis from the revised set of beliefs.
- If there is at least one open subtableau, the agent's doxastic accessibility relation is changed from considering all the *cosis* represented by the tableau that contained the analysed formula as doxastic alternatives to just taking into account those *cosis* represented by the open subtableau(x).

All the formulæ of a tableau are always contained in the tableaux generated below it in the tableaux tree; therefore, the *cosis* represented by a tableau are always less informative (according to the agent's information order, $\leq_i$) than the *cosis* represented by its subtableaux. That implies, due to the definition of the modified Kripke semantics, that the agent's positive and negative sets of beliefs may only *grow* as a result of an step of logical analysis (assuming that at least one subtableau generated in this analysis remains open).

A logical step of analysis serves to eliminate situations such as the following:

- *Logically incomplete situations.* For instance, those in which a disjunction $(\phi \vee \psi)$ is supported but there is neither positive nor negative information about ϕ or ψ.
- *Logically inconsistent situations.* For instance, those in which there is both positive and negative evidence about a given formula.

- *Logically contradictory situations.* For instance, those in which a disjunction $(\phi \vee \psi)$ is supported but there is negative evidence about ϕ and ψ.

Having eliminated all these kinds of undesirable states of affairs, $Agent_i$ considers as doxastic alternatives situations that are more i-informative than the previous ones and, therefore, in this way it logically refines its set of beliefs.

- If the agent adds a doubt in the *exploratory* dimension of analysis, it incorporates an instance of the Axiom of the Excluded Middle (*i.e.* a disjunction of a formula and its negation) in the left column of an open tableau T. In that moment, the set of *cosis* that are considered as doxastic alternatives by the agent must change from the ones represented by T to those represented by the generated subtableau (that contains the same formulæ as T plus the new disjunction). It is theoretically possible that this tableau is logically closed by the agent, but it seems pretty unlikely. That state of affairs would arise only if T contained explicitly the negation of the introduced doubt and, in those circumstances, it seems reasonable to think that the agent would not introduce the doubt in the first place. Recall that the aim of this dimension of analysis is to generate a doubt that serves to explore two alternatives that, in principle, are not deducible (or, at least, not trivially deducible) from the present beliefs.

  As in the case of the logical analysis, note that the remaining subtableau contains the set of formulæ of its parent in the tableaux tree; therefore, the set of conceivable situations considered as doxastic alternatives by $Agent_i$ is more i-informative than the previous one and, thus, its positive set of beliefs may only *grow* (with the inclusion of the introduced disjunction, in case there are no other open tableaux in the analysis). The set of negative beliefs does not change, because the right column of the analysed tableau is not modified. In fact, the only alternatives that are dismissed with this dimension of analysis are those in which the instance of the AEM does not hold, *i.e.* non-classical situations (*intuitionistic situations*, if you wish) in which standard tautologies are not taken for granted, do not necessarily hold.

- In the *experimental* dimension of analysis, the agent delivers a question to its environment concerning the existence of an individual with certain characteristics. This question is triggered by the presence in a tableau T of a set of (affirmed and/or negated) primitive predicates applied to a given Skolem constant c. The change of the agent's set of doxastic alternatives depends on the received answer.

  If the answer is positive, the agent is provided with the name of an individual o that satisfies the required properties, and it generates a subtableau of T in which the Skolem constant c is replaced by o. Thus, the agent changes from believing that there is an individual with certain properties to believing that o has those properties. This belief change is induced by a reduction on the agent's set of doxastic alternatives, which changes from the set of those *cosis* in which there is an individual with the properties represented by the primitive predicates to the set of those *cosis* in which it is precisely o who has those properties. Obviously, this latter set of *cosis* is a subset of the former one; therefore, the agent's set of viable alternatives has reduced, and its set of positive beliefs has increased. We might say that the agent has somehow *refined* its beliefs, as it has been capable of transforming a merely existential belief into an specific belief related to a given individual. There exists the possibility that the resulting tableau has to be logically closed, as a result of the replacement of c by o; in that case, the agent's doxastic alternatives would also have to be restricted, as the set of *cosis* represented by T would be eliminated from consideration.

  If the answer is negative, the agent must *empirically* close T. The change in the agent's beliefs produced by this action is also induced by a restriction on the set of alternatives it considers possible, as it must get rid of those *cosis* which were represented by T. Thus, the way in which a *logical* closing and an *empirical* closing transform the set of accessible worlds is exactly the same: they produce the elimination of a certain set of *cosis* (those represented by the closed tableau) from the set of doxastic alternatives. There is a very important difference in the belief analysis process, though: *logical* closings are irreversible, whereas *empirical* ones may be re-considered later, in the face of new information.

- An agent may also add information that it has received directly from the environment to the left column of a tableau T. In this case, the change produced in the agent's set of beliefs may be explained by saying that it will change from considering as doxastic alternatives those *cosis* represented by T to noticing that the more $\leq_i$-informative *cosis* represented by the obtained subtableau are indeed feasible.

  As in the case of the exploratory dimension of analysis, there are two possibilities to be considered. If the new subtableau is not logically closed, the agent's set of positive beliefs will incorporate the added formula (in case there are no other open tableaux in the tableaux tree). The set of negative beliefs will not change, as the right columns of the open tableaux will not have suffered any modification. If the inclusion of the received information produces the *logical* closing of the new subtableau, there will be a restriction on the set of doxastic alternatives, as all the *cosis* represented by T will cease to be considered as potential alternatives. This restriction may cause an increase in the set of positive beliefs and a decrease in the set of negative beliefs, due to the way in which the modified Kripke semantics has been defined.

- We have just explained how there are some situations in which the evolution of the beliefs of a rational inquirer may be formally modelled with a sequence of restrictions on the set of situations that it considers possible. However, there is an important situation left to be analysed which will require a different treatment. The only occasion in which the agent will *increase* its set of doxastic alternatives is when an *empirically* closed tableau T is re-open at a later stage of the analysis. This situation may arise in three different ocasions: after an step of *logical* analysis, after receiving a positive answer in the *experimental* dimension of analysis, and after receiving a direct external input. We explain the change in the agent's set of beliefs by assuming that, at that moment, the agent will realise that the class of *cosis* represented by T is indeed feasible, and that it has to be taken into account when computing its beliefs. Thus, the agent will add the *cosis* represented by T to its present set of doxastic alternatives. This is the only way in which the agent can *reduce* its present positive beliefs, because it can consider new possible worlds in which previously held positive beliefs are not necessarily true. The set of negative beliefs may only grow,

as the new doxastic alternatives may have negative information about new formulæ.

5.6 Ways of changing the set of beliefs

In summary, an agent may modify its positive and negative beliefs in three different ways:

- It may eliminate some doxastic alternatives from consideration.

  This situation arises for instance when the agent decides to close a certain tableau; at that point, the agent ceases to consider as viable alternatives the *cosis* represented by that tableau.

  If the modified Kripke semantics is applied after eliminating some doxastic alternatives, the set of positive beliefs may only increase, as there are less alternatives to be considered. On the other side, the set of negative beliefs may only decrease, as the formulæ which were negatively supported only in the eliminated situations will cease to be negative beliefs.

- It may change some doxastic alternatives by others than are more i-informative.

  This situation arises for instance when the agent performs an step of logical analysis, and it does not close all the resulting subtableaux.

  If the $\leq_i$-informativeness of the doxastic alternatives is greater, by applying the modified Kripke semantics both the positive and the negative sets of beliefs may only grow (as all the formulæ which were already positively or negatively supported in the previous doxastic alternatives will keep the same status).

- It may consider new doxastic alternatives.

  This situation only arises when the agent decides to re-open a previously empirically closed tableau. In this case the set of positive beliefs may only decrease, as the modified Kripke semantics checks that the previously held positive beliefs are also supported in the new doxastic alternatives. The set of negative beliefs may only grow, as it will now include all the formulæ that are negatively supported in the new alternatives.

The standard possible worlds tradition (see *e.g.* [FHMV95]), in which only complete and consistent possible worlds are considered, basically includes only the first of these ways of acquiring more beliefs (restricting the set of doxastic alternatives considered by the agent). There are different approaches, such as a very interesting one developed in [Jasp94]. In that work, Jaspars proposes a way of modelling the changes that occur in the information of a group of agents as a result of their reasoning and communication processes. He uses *partial* possible worlds, in which there is a partial assignment of truth values to the basic propositions in each world. In that way, a fact may be supported, denied, or neither supported nor denied in a certain state of affairs. In his framework, information may grow along two different dimensions: an standard *eliminative* one, in which certain doxastic alternatives are eliminated, and a *constructive* one, in which the agent changes a doxastic alternative by another one which has more information, in the sense of being defined with a less partial assignment of truth values. Therefore, in some way it may be said that we share with Jaspars two of the ways of acquiring more information. However, there are many differences between the two approaches; just to name a few, he does not consider the third possibility of changing the set of beliefs (adding more doxastic alternatives, which is the only way of having non-monotonic beliefs), he uses partial logics (and, thus, his approach is more of a three-valued style), he defines a full modal calculus and he deals with multiple communicating agents. In general, his aim is much more ambitious and his framework is much more technically detailed and complex than ours.

5.7 Modelling process

Now we can give a complete account of the belief modelling process. The agent's beliefs keep changing in time because the *conceivable situations* that are considered as the agent's doxastic alternatives change after each step of analysis. The process may be summarised as follows:

- At the beginning, the agent has a given set of initial data. The agent starts the analysis with a tableau, T_0, that contains in its left column this initial information.
- In our model, we build the initial accessibility relation, R_0. The *cosis* that will be accessible through this relation (from the agent's situation,

which will be always named w_e) are the ones in which all the initial formulæ hold.

- By applying the modified Kripke semantics to R_0, we notice that the agent's initial set of beliefs corresponds to the initial set of formulæ, because these formulæ are the ones that hold in all the *cosis* represented by T_0, which are just those *cosis* that are R_0-accessible.
- While the agent decides to keep applying any of the dimensions of belief analysis, follow these steps:
  - The agent performs one step of logical, exploratory or experimental analysis, or receives some external input from the environment, and changes the tableaux tree accordingly.
  - If the agent decides so, it may logically or empirically close some of the tableaux generated in the previous step. It may also decide to re-open a previously empirically closed tableau, in the light of the information obtained in the last step.
  - In our model, we have to build a new accessibility relation, R_{i+1}. The only modification from R_i to R_{i+1} is that the agent's doxastic alternatives change from all those *cosis* represented by the open tableaux in the previous stage to all those *cosis* represented by the open tableaux after the last step of analysis.
  - The agent's set of beliefs is updated by applying the modified Kripke semantics over the *cosis* that are R_{i+1}-accessible.

5.8 Example

In §5.4 we explained the basic tools needed to model the evolution of the beliefs of a rational inquirer on time. In §5.5 and §5.6 we described how an agent's set of doxastic alternatives changes as a result of any step of analysis that a rational inquirer may carry out. This was the basic point in the modelling process shown in §5.7. The aim of this section is to illustrate all those ideas by providing an example of how the evolution of the beliefs of a rational inquirer, caused by a multi-dimensional belief analysis, may be modelled using the process that has just been explained above. We will use

the example of belief analysis developed in §4.6 to show which would be the agent's beliefs after each step of the analysis.

Before starting with the example, a notational explanation must be made. In the figures used in this section, each class of *cosis* is represented by a rectangle divided in two parts; a formula appears at the top if it holds in all the *cosis* of the class, and it appears at the bottom if it does not hold in any *cosi* of the class. The formulæ shown at the top/bottom of the square labelled w_e in the figures depicted in this section reflect the positive/negative beliefs of the agent (which are computed with the modified Kripke semantics stated above).

Let us consider now the example of §4.6. Recall that the analysis started with the following set of formulæ:

$$\Delta \equiv \{\mathtt{B_T}, \mathtt{F_T}, \mathtt{B_P}, \mathtt{F_P}, \mathtt{B_W}, \forall x(\mathtt{P}_x \Rightarrow \neg \mathtt{F}_x)\}$$

The initial state of the belief analysis was represented by a tableau, T_0, that contained in its left column the initial information of the agent. We have already commented that a two-columned tableau may be considered as a (partial) representation of a class of *cosis*: all those *cosis* that contain all the formulæ of the left column and do not contain any of the formulæ of the right column[16]. Thus, T_0 represents a class of *cosis*: all those *cosis* in which all the six formulæ of Δ hold. This class will be called $\overline{w_0}$. Therefore, $\overline{w_0}$ contains all those situations s such that $\mathcal{T}_i(s) = \Delta$ and $\mathcal{F}_i(s) = \emptyset$. The semantic modelling of the agent's doxastic state in this initial point is done by generating an initial accessibility relation, R_0, in which the agent's initial doxastic alternatives are those in class $\overline{w_0}$. This situation is depicted in figure 13.

Let us recall how the agent's beliefs are calculated in base to its doxastic alternatives:

- The agent believes a formula ϕ (*i.e.* ϕ is a *positive* belief) if it is contained in all doxastic alternatives.
- The agent does not believe a formula ϕ (*i.e.* ϕ is a *negative* belief) if there exists at least one doxastic alternative in which it is not contained.

[16] We use the expressions *"ϕ holds in the cosi"* and *"ϕ is contained in the cosi"* interchangeably.

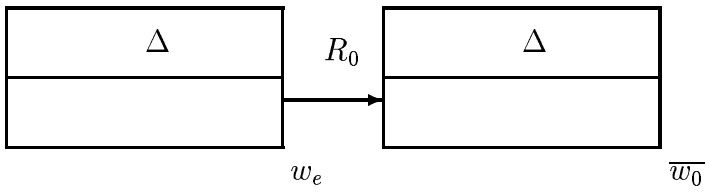

Figure 13: Initial accessibility relation

In this case, the agent believes all those formulæ contained in all *cosis* of $\overline{w_0}$, *i.e.* the six formulæ of the set Δ. Therefore, the agent's beliefs at this point (before starting the analysis) are the following:

$$\{B(\mathtt{B_T}), B(\mathtt{F_T}), B(\mathtt{B_P}), B(\mathtt{F_P}), B(\mathtt{B_W}), B(\forall x(\mathtt{P}_x \Rightarrow \neg \mathtt{F}_x))\}$$

This fact is shown at the top of w_e in figure 13, where the agent's positive beliefs are displayed.

In the first step of analysis the agent wondered whether all birds fly. It incorporated this doubt into the analysis by using the exploratory dimension, adding the formula $(\forall x(\mathtt{B}_x \Rightarrow \mathtt{F}_x) \vee \neg\forall x(\mathtt{B}_x \Rightarrow \mathtt{F}_x))$ in the left column of T_0, generating thus a new tableau, T_0', shown in figure 5.

In order to provide a formal model of the evolution of the agent's beliefs due to this step of analysis we have to consider the class of cosis $\overline{w_{0'}}$, that contains those situations s such that $\mathcal{T}_i(s) = \Delta \bigcup \{(\forall x(\mathtt{B}_x \Rightarrow \mathtt{F}_x) \vee \neg\forall x(\mathtt{B}_x \Rightarrow \mathtt{F}_x))\}$ and $\mathcal{F}_i(s) = \emptyset$.

The new tableau generated by the agent, T_0', represents class $\overline{w_{0'}}$. Therefore, the semantic counterpart of the application of the exploratory dimension of analysis is the generation of a new accessibility relation, R_1, in which the set of doxastic alternatives is changed from $\overline{w_0}$ to $\overline{w_{0'}}$. Note that $\overline{w_0} \leq_i \overline{w_{0'}}$, that is, the new doxastic alternatives are more i-informative. This change is shown in figure 14.

The agent's beliefs at this point of the analysis are obtained by applying our modified Kripke semantics. As the only accessible *cosis* are those in $\overline{w_{0'}}$, the agent's actual set of beliefs would be the following:

$$\{B(\mathtt{B_T}), B(\mathtt{F_T}), B(\mathtt{B_P}), B(\mathtt{F_P}), B(\mathtt{B_W}), B(\forall x(\mathtt{P}_x \Rightarrow \neg \mathtt{F}_x)),$$
$$B(\forall x(\mathtt{B}_x \Rightarrow \mathtt{F}_x) \vee \neg\forall x(\mathtt{B}_x \Rightarrow \mathtt{F}_x))\}$$

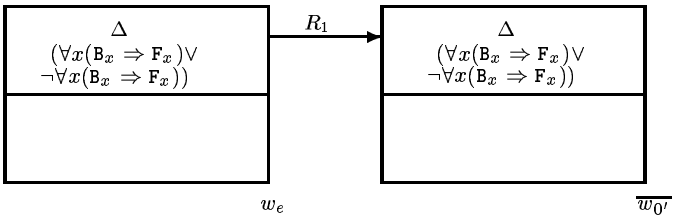

Figure 14: Generation of R_1

Thus, the change in the set of accessible *cosis* accounts for the increase in the set of positive beliefs caused by the incorporation of the agent's doubt into the analysis. The agent now would believe that either all birds fly or that it is not the case that all birds fly.

The agent proceeded with the analysis of its beliefs by logically analysing the formula that it had just introduced in the tableau, $(\forall x(\mathtt{B}_x \Rightarrow \mathtt{F}_x) \vee \neg\forall x(\mathtt{B}_x \Rightarrow \mathtt{F}_x))$. The result was the generation of three subtableaux, T_1, T_2 and T_3, as shown in figure 12.

We can now consider the following classes of *cosis*:

- $\overline{w_1}$: *cosis* in which Δ, $(\forall x(\mathtt{B}_x \Rightarrow \mathtt{F}_x) \vee \neg\forall x(\mathtt{B}_x \Rightarrow \mathtt{F}_x))$, $\forall x(\mathtt{B}_x \Rightarrow \mathtt{F}_x)$ and $\neg\forall x(\mathtt{B}_x \Rightarrow \mathtt{F}_x)$ hold (*e.g.* a situation in which a database has received contradictory information from independent sources).

- $\overline{w_2}$: *cosis* in which Δ, $(\forall x(\mathtt{B}_x \Rightarrow \mathtt{F}_x) \vee \neg\forall x(\mathtt{B}_x \Rightarrow \mathtt{F}_x))$ and $\forall x(\mathtt{B}_x \Rightarrow \mathtt{F}_x)$ hold but $\neg\forall x(\mathtt{B}_x \Rightarrow \mathtt{F}_x)$ does not hold (*i.e.* situations in which the agent has information supporting that all birds fly).

- $\overline{w_3}$: *cosis* in which Δ, $(\forall x(\mathtt{B}_x \Rightarrow \mathtt{F}_x) \vee \neg\forall x(\mathtt{B}_x \Rightarrow \mathtt{F}_x))$ and $\neg\forall x(\mathtt{B}_x \Rightarrow \mathtt{F}_x)$ hold but $\forall x(\mathtt{B}_x \Rightarrow \mathtt{F}_x)$ does not hold (*e.g.* the real world, in which it is not true that all birds fly).

Following our conception of tableaux as representations of classes of *cosis*, it may be noticed that T_1 represents $\overline{w_1}$, T_2 represents $\overline{w_2}$ and T_3 represents $\overline{w_3}$. T_1 contains a formula and its negation in its left column, so the agent may consider the class of *cosis* represented by this tableau as *logically impossible* and can dismiss these situations from the analysis by logically closing T_1. After this operation, the only open tableaux in the logical analysis are T_2

and T_3. The semantic modelling of the belief change caused by the agent's logical analysis is the generation of a new accessibility relation, R_2, that changes the set of doxastic alternatives from those in $\overline{w_{0'}}$ to those in classes $\overline{w_2}$ and $\overline{w_3}$, which are the classes represented by all open tableaux. Note that $\overline{w_{0'}} \leq_i \overline{w_2}$ and $\overline{w_{0'}} \leq_i \overline{w_3}$, *i.e.* the new doxastic alternatives are more i-informative than the previous ones; therefore, as explained above, the agent's positive and negative sets of beliefs may only grow as a result of this change of accessible situations. This state of affairs is shown in figure 15.

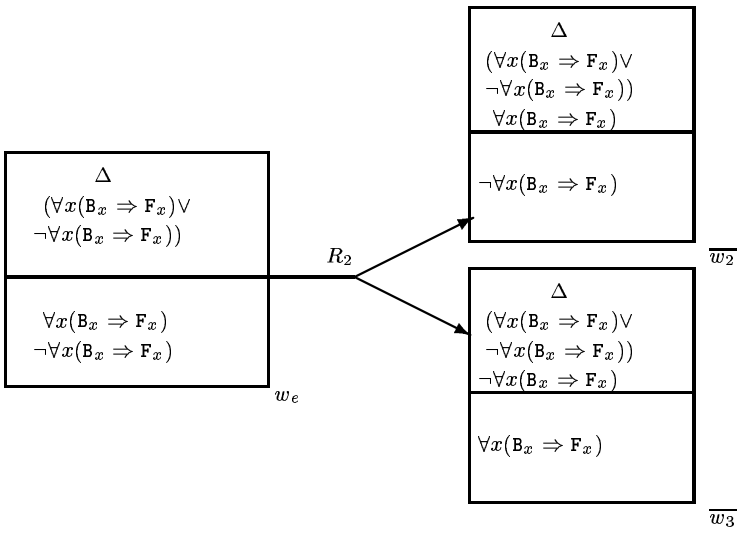

Figure 15: Generation of R_2

The agent's positive beliefs have not changed, because the only formulæ contained in all doxastic alternatives (all R_2-accessible *cosis*) are those in Δ and $(\forall x(\mathtt{B}_x \Rightarrow \mathtt{F}_x) \vee \neg\forall x(\mathtt{B}_x \Rightarrow \mathtt{F}_x))$. However, the agent has now attained two *negative* beliefs:

- The agent does not believe $\forall x(\mathtt{B}_x \Rightarrow \mathtt{F}_x)$, because the *cosis* in $\overline{w_3}$ (that are R_2-accessible) do not contain this formula.
- The agent does not believe $\neg\forall x(\mathtt{B}_x \Rightarrow \mathtt{F}_x)$, because the *cosis* in $\overline{w_2}$ (that are R_2-accessible) do not contain this formula.

Thus, the agent's beliefs at this point are:

$\{B(\mathtt{B_T}), B(\mathtt{F_T}), B(\mathtt{B_P}), B(\mathtt{F_P}), B(\mathtt{B_W}), B(\forall x(\mathtt{P}_x \Rightarrow \neg \mathtt{F}_x)),$
$B(\forall x(\mathtt{B}_x \Rightarrow \mathtt{F}_x) \vee \neg \forall x(\mathtt{B}_x \Rightarrow \mathtt{F}_x)),$
$\neg B(\forall x(\mathtt{B}_x \Rightarrow \mathtt{F}_x)), \neg B(\neg \forall x(\mathtt{B}_x \Rightarrow \mathtt{F}_x))\}$

The agent chose to proceed its inquiry by performing the logical analysis of the formula $\neg \forall x(B_x \Rightarrow F_x)$ in tableau T_3. The result of this analysis was the generation of a new subtableau, T_4, that contained a formula with a new Skolem constant c, $\neg(B_c \Rightarrow F_c)$ (recall figure 6). An Skolem constant does not refer to any specific individual; it may be considered as a *placeholder*, that occupies an space that may be later filled with an appropriate concrete value. In order to see the change produced in the agent's beliefs after this step of analysis, consider the following class of *cosis*, $\overline{w_4}$:

- $\overline{w_4}$: *cosis* in which Δ, $(\forall x(\mathtt{B}_x \Rightarrow \mathtt{F}_x) \vee \neg \forall x(\mathtt{B}_x \Rightarrow \mathtt{F}_x))$, $\neg \forall x(\mathtt{B}_x \Rightarrow \mathtt{F}_x)$ and $\neg(\mathtt{B}_o \Rightarrow \mathtt{F}_o)$ hold (for some object o), but $\forall x(\mathtt{B}_x \Rightarrow \mathtt{F}_x)$ does not hold.

The tableau generated in this step of the analysis, T_4, represents the class of *cosis* $\overline{w_4}$[17]. Note that the agent has inferred a condition that a *cosi* should satisfy in order to belong to that class. Obtaining this kind of conditions is important, because the agent can later dismiss a class of *cosis* if it can check (in the *experimental* dimension of analysis) that there are no objects in its environment that satisfy those requirements.

The evolution of the agent's set of beliefs at this point may be explained by considering that it would change from believing that all the situations in classes $\overline{w_2}$ and $\overline{w_3}$ are possible to believing that the only viable doxastic alternatives are those *cosis* in $\overline{w_2}$ and $\overline{w_4}$. Formally, this fact is represented with the generation of a new accessibility relation, R_3, in which the set of *cosis* considered as possible by the agent is changed. This situation is shown in figure 16[18]. As may be seen in this figure, the (positive and negative) beliefs of the agent have not changed, even though some of the accessible *cosis* are more i-informative than the previous ones ($\overline{w_3} \leq_i \overline{w_4}$).

[17] As a side remark, it may be noticed that the previously considered classes of *cosis* (*i.e.* $\overline{w_0}$, $\overline{w_{0'}}$, $\overline{w_1}$, $\overline{w_2}$ and $\overline{w_3}$) were equivalences classes (with respect to the partition defined by R_i). This is not the case of $\overline{w_4}$, which contains the union of several such classes (one for each object o).

[18] The term "$\exists^*$" that appears in that figure is a meta-expression, representing the fact that, in each *cosi* of the class, there must exist one object that satisfies the given condition.

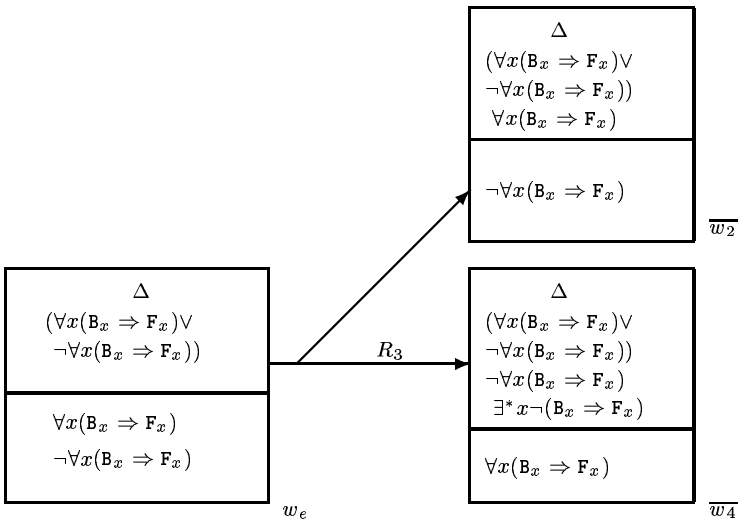

Figure 16: Generation of R_3

In the tableau generated in the following step of logical analysis (T_5, see figure 6) there are three formulæ that contain the Skolem constant c (the one that appeared in the previous step plus B_c and $\neg F_c$). Thus, this tableau postulates the existence of an object that fulfils these three conditions. The change caused in the agent's beliefs in this new step of analysis may be modelled, again, with a change in the set of scenarios considered possible by the agent. That modification may be described with the help of the following class of *cosis*, $\overline{w_5}$:

- $\overline{w_5}$: *cosis* in which Δ, $(\forall x(\mathtt{B}_x \Rightarrow \mathtt{F}_x) \vee \neg\forall x(\mathtt{B}_x \Rightarrow \mathtt{F}_x))$, $\neg\forall x(\mathtt{B}_x \Rightarrow \mathtt{F}_x)$, $\neg(\mathtt{B}_o \Rightarrow \mathtt{F}_o)$, $\mathtt{B}_o$ and $\neg\mathtt{F}_o$ hold (for some object o), but $\forall x(\mathtt{B}_x \Rightarrow \mathtt{F}_x)$ does not hold.

After this step of analysis the agent would notice that the *cosis* that are really feasible (apart from those in $\overline{w_2}$) are not those in $\overline{w_4}$, but those in $\overline{w_5}$, *i.e.* the ones represented by tableau T_5. This fact is represented in Fig. 17 with the generation of a new accessibility relation, R_4, that modifies (again) the agent's set of doxastic alternatives.

The *cosis* in $\overline{w_5}$ contain more information that those in $\overline{w_4}$; to put it more precisely, for each situation $s \epsilon \overline{w_4}$ there exists a situation $t \epsilon \overline{w_5}$ such that

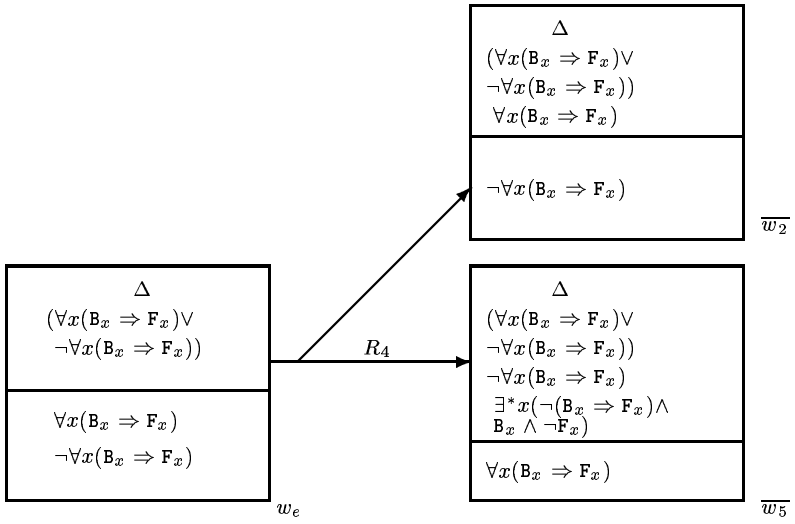

Figure 17: Generation of R_4

$s \leq_i t$. However, applying the modified Kripke semantics it may be noticed that the agent's beliefs would not change at this point of the analysis either; they would still be the following:

$$\{B(\mathtt{B}_\mathtt{T}), B(\mathtt{F}_\mathtt{T}), B(\mathtt{B}_\mathtt{P}), B(\mathtt{F}_\mathtt{P}), B(\mathtt{B}_\mathtt{W}), B(\forall x(\mathtt{P}_x \Rightarrow \neg\mathtt{F}_x)),$$
$$B(\forall x(\mathtt{B}_x \Rightarrow \mathtt{F}_x) \vee \neg\forall x(\mathtt{B}_x \Rightarrow \mathtt{F}_x)),$$
$$\neg B(\forall x(\mathtt{B}_x \Rightarrow \mathtt{F}_x)), \neg B(\neg\forall x(\mathtt{B}_x \Rightarrow \mathtt{F}_x))\}$$

The agent continued the analysis of its beliefs using the *experimental* dimension of analysis, posing the following question to the environment: *Does it exist an individual that has property B and does not have property F?* In this example, the agent received a *negative* answer, which implied the *empirical closing* of T_5. After having *empirically closed* T_5, the only open tableau of the logical analysis is T_2. The semantic counterpart of this *empirical closing* is the generation in our model of a new accessibility relation, R_5, that restricts the set of doxastic alternatives (by eliminating the class of *cosis* represented by T_5, $\overline{w_5}$). This situation is shown in figure 18.

That figure also reflects the change produced in the agent's beliefs by the *empirical* closing of T_5. That closing has reduced the doxastic alternatives

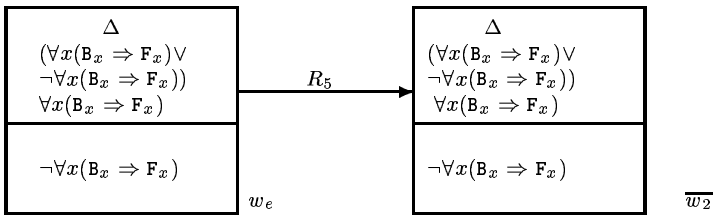

Figure 18: Generation of R_5

to those *cosis* represented by the only open tableau, T_2 (*i.e.* those conceivable situations that belong to class $\overline{w_2}$). Therefore, if our modified Kripke semantics is applied to update the agent's beliefs, the positive beliefs will be those formulæ in the left column of T_2, while the negative beliefs will be those formulæ in the right column of T_2. Thus, the agent's actual set of beliefs is the following:

$$\{B(\mathtt{B_T}), B(\mathtt{F_T}), B(\mathtt{B_P}), B(\mathtt{F_P}), B(\mathtt{B_W}), B(\forall x(\mathtt{P}_x \Rightarrow \neg \mathtt{F}_x)),$$
$$B(\forall x(\mathtt{B}_x \Rightarrow \mathtt{F}_x) \vee \neg\forall x(\mathtt{B}_x \Rightarrow \mathtt{F}_x)),$$
$$B(\forall x(\mathtt{B}_x \Rightarrow \mathtt{F}_x)), \neg B(\neg\forall x(\mathtt{B}_x \Rightarrow \mathtt{F}_x))\}$$

Notice that the agent now believes that "all birds fly" (just because it has looked for individuals that are birds and do not fly and it has not been capable of finding any such object), and it still disbelieves that "not all birds fly". This situation is quite common in the rational inquirer's analysis of belief: if it keeps reducing the number of doxastic alternatives, the set of positive beliefs may only grow. If a formula ϕ is positively believed at a certain stage of the analysis, and accessibility relations keep decreasing, ϕ will always continue to be positively believed by the agent. On the other hand, if there is an increase in the set of doxastic alternatives, the set of positive beliefs can only reduce (as some of them may not hold in the added situations).

Following the example, the agent now added a new piece of information received from its environment ($\mathtt{P_W}$) in the left columns of all non-logically closed tableaux (T_2 and T_5), generating two new subtableaux, T_2' and T_5'. This situation was depicted in figure 6.

Recall that the tableaux T_2 and T_5 represented the classes of *cosis* $\overline{w_2}$ and $\overline{w_5}$, respectively. Consider now these classes of *cosis*:

- $\overline{w_{2'}}$: *cosis* in which Δ, $(\forall x(\mathtt{B}_x \Rightarrow \mathtt{F}_x) \vee \neg\forall x(\mathtt{B}_x \Rightarrow \mathtt{F}_x))$, $\forall x(\mathtt{B}_x \Rightarrow \mathtt{F}_x)$ and $\mathtt{P}_W$ hold but $\neg\forall x(\mathtt{B}_x \Rightarrow \mathtt{F}_x)$ does not hold.

- $\overline{w_{5'}}$: *cosis* in which Δ, $(\forall x(\mathtt{B}_x \Rightarrow \mathtt{F}_x) \vee \neg\forall x(\mathtt{B}_x \Rightarrow \mathtt{F}_x))$, $\neg\forall x(\mathtt{B}_x \Rightarrow \mathtt{F}_x)$, $\mathtt{P}_W$, $\neg(\mathtt{B}_o \Rightarrow \mathtt{F}_o)$, $\mathtt{B}_o$ and $\neg\mathtt{F}_o$ hold (for some object o), but $\forall x(\mathtt{B}_x \Rightarrow \mathtt{F}_x)$ does not hold.

The new subtableaux, T_2' and T_5', represent the classes of *cosis* $\overline{w_{2'}}$ and $\overline{w_{5'}}$, respectively. Thus, our model will reflect the change of belief produced by the incorporation of the new information, $\mathtt{P}_\mathtt{W}$, by generating a new accessibility relation, R_6, that will change the set of doxastic alternatives. The *cosis* that were R_5-accessible were those in $\overline{w_2}$ (only the ones represented by T_2, because T_5 was *empirically* closed); now, the only R_6-accessible *cosis* will be those in class $\overline{w_{2'}}$ (the *cosis* represented by T_2'). The class of *cosis* represented by T_5' will not be accessible because that branch of the tableaux analysis is *empirically* closed. Thus, the resulting semantic situation is shown in figure 19.

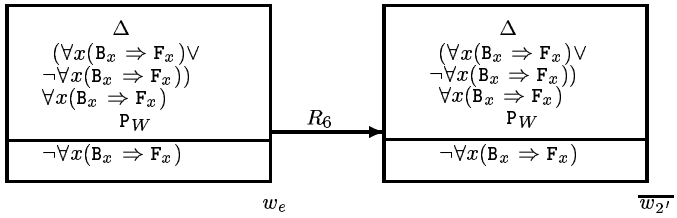

Figure 19: Generation of R_6

Let us analyse the relationship between the classes of *cosis* previously considered ($\overline{w_2}$ and $\overline{w_5}$) and those that have just been defined ($\overline{w_{2'}}$ and $\overline{w_{5'}}$). On the one hand, it is easy to see that $\overline{w_2} \leq_i \overline{w_{2'}}$; on the other hand, there is also a gain of information in the change of $\overline{w_5}$ by $\overline{w_{5'}}$ in the sense that it holds that, for any situation $s \epsilon \overline{w_5}$, there exists a situation $t \epsilon \overline{w_{5'}}$ such that $s \leq_i t$. Therefore, the change of beliefs caused by the transformation of the tableaux T_2 and T_5 into $T_{2'}$ and $T_{5'}$ is modelled as climbs in the information hierarchy defined by $\leq_i$.

As the only accessible *cosis* are those in $\overline{w_{2'}}$, the agent's positive beliefs are those formulæ known to be contained in every *cosi* of the class, whereas

the agent's negative beliefs are those formulæ known *not* to be contained in any *cosi* of the class. Therefore, the agent's beliefs at this stage of the analysis are the following:

$$\{B(\mathtt{B_T}), B(\mathtt{F_T}), B(\mathtt{B_P}), B(\mathtt{F_P}), B(\mathtt{B_W}), B(\forall x(\mathtt{P}_x \Rightarrow \neg \mathtt{F}_x)),$$
$$B(\forall x(\mathtt{B}_x \Rightarrow \mathtt{F}_x) \vee \neg \forall x(\mathtt{B}_x \Rightarrow \mathtt{F}_x)),$$
$$B(\forall x(\mathtt{B}_x \Rightarrow \mathtt{F}_x)), B(\mathtt{P}_W), \neg B(\neg \forall x(\mathtt{B}_x \Rightarrow \mathtt{F}_x))\}$$

The only difference with respect to the previous set of beliefs is that the agent has incorporated the information received from an external source, $\mathtt{P_W}$, as a new positive belief. This situation would happen with any externally obtained formula, as long as it is included in the left columns of *all* the tableaux that are still being considered in the dynamic multi-dimensional belief analysis. It would not be the case if the new formulæ were included only in *some* of the open tableaux of the open analysis.

The agent continued the belief analysis by *logically* analysing the formula $\forall x(\mathtt{P}_x \Rightarrow \neg \mathtt{F}_x)$ in T_5'. This decision caused the generation of a new tableau, T_6, that contained the formula $(\mathtt{P}_W \Rightarrow \neg \mathtt{F}_W)$, as shown in figure 7. The agent's beliefs do not change as a result of this analysis, because this branch of the tableaux tree is *empirically* closed and, therefore, the classes of *cosis* represented by the tableaux in this branch are not taken into account when applying the modified Kripke semantics to compute the agent's beliefs. Thus, in our semantic account of the evolution of the agent's beliefs, the accessibility relation between *cosis* does not change, and the only accessible *cosis* are those in class $\overline{w_{2'}}$. The only change experienced in our model is the realisation that T_6 represents the following class of *cosis*:

- $\overline{w_6}$: *cosis* in which Δ, $(\forall x(\mathtt{B}_x \Rightarrow \mathtt{F}_x) \vee \neg \forall x(\mathtt{B}_x \Rightarrow \mathtt{F}_x))$, $\neg \forall x(\mathtt{B}_x \Rightarrow \mathtt{F}_x)$, $\mathtt{P}_W$, $(\mathtt{P}_W \Rightarrow \neg \mathtt{F}_W)$, $\neg(\mathtt{B}_o \Rightarrow \mathtt{F}_o)$, $\mathtt{B}_o$ and $\neg \mathtt{F}_o$ hold (for some object o), but $\forall x(\mathtt{B}_x \Rightarrow \mathtt{F}_x)$ does not hold.

It may be seen that each situation in $\overline{w_{5'}}$ is less i-informative than its *counterpart* in $\overline{w_6}$.

The agent decided now to analyse the formula $(\mathtt{P}_W \Rightarrow \neg \mathtt{F}_W)$ in T_6, obtaining the result shown in figure 7. The classes of *cosis* represented by the new tableaux, T_7, T_8 and T_9, are the following ones:

- $\overline{w_7}$: *cosis* in which Δ, $(\forall x(\mathtt{B}_x \Rightarrow \mathtt{F}_x) \vee \neg\forall x(\mathtt{B}_x \Rightarrow \mathtt{F}_x))$, $\neg\forall x(\mathtt{B}_x \Rightarrow \mathtt{F}_x)$, $\mathtt{P}_W$, $(\mathtt{P}_W \Rightarrow \neg\mathtt{F}_W)$, $\neg\mathtt{P}_W$, $\neg\mathtt{F}_W$, $\neg(\mathtt{B}_o \Rightarrow \mathtt{F}_o)$, $\mathtt{B}_o$ and $\neg\mathtt{F}_o$ hold (for some object o), but $\forall x(\mathtt{B}_x \Rightarrow \mathtt{F}_x)$ does not hold.

- $\overline{w_8}$: *cosis* in which Δ, $(\forall x(\mathtt{B}_x \Rightarrow \mathtt{F}_x) \vee \neg\forall x(\mathtt{B}_x \Rightarrow \mathtt{F}_x))$, $\neg\forall x(\mathtt{B}_x \Rightarrow \mathtt{F}_x)$, $\mathtt{P}_W$, $\neg\mathtt{P}_W$, $(\mathtt{P}_W \Rightarrow \neg\mathtt{F}_W)$, $\neg(\mathtt{B}_o \Rightarrow \mathtt{F}_o)$, $\mathtt{B}_o$ and $\neg\mathtt{F}_o$ hold (for some object o), but $\forall x(\mathtt{B}_x \Rightarrow \mathtt{F}_x)$ and $\neg\mathtt{F}_W$ do not hold.

- $\overline{w_9}$: *cosis* in which Δ, $(\forall x(\mathtt{B}_x \Rightarrow \mathtt{F}_x) \vee \neg\forall x(\mathtt{B}_x \Rightarrow \mathtt{F}_x))$, $\neg\forall x(\mathtt{B}_x \Rightarrow \mathtt{F}_x)$, $\mathtt{P}_W$, $\neg\mathtt{F}_W$, $(\mathtt{P}_W \Rightarrow \neg\mathtt{F}_W)$, $\neg(\mathtt{B}_o \Rightarrow \mathtt{F}_o)$, $\mathtt{B}_o$ and $\neg\mathtt{F}_o$ hold (for some object o), but $\forall x(\mathtt{B}_x \Rightarrow \mathtt{F}_x)$ and $\neg\mathtt{P}_W$ do not hold.

T_7 represents $\overline{w_7}$, T_8 represents $\overline{w_8}$ and T_9 represents $\overline{w_9}$. T_7 and T_8 contain a formula ($\mathtt{P}_W$) and its negation in their left columns, so they represent classes of *logically impossible cosis*, and the agent may decide to get rid of these classes in the analysis by *logically closing* these tableaux. Thus, at this point of the analysis, the only open tableau is still T'_2, while T_9 is *empirically closed*. The agent's beliefs do not change after this logical analysis either, because the analysed formula was contained in an *empirically closed* tableau (and, therefore, the set of open tableaux, $\{T'_2\}$, remains unchanged). As happened in the previous stage of analysis, notice that each situation in $\overline{w_6}$ is less i-informative that its *counterpart* in $\overline{w_9}$.

An interesting situation arised at this point; the agent realised that T_9, that was *empirically* closed, could be now *re-open*, because it had discovered the existence of an specific individual (W) that "is a bird and does not fly". Moreover, a new tableau, T'_9, in which all the appearances of c were replaced by constant W, was generated (see figure 8).

Now there are two open tableaux in the tableaux tree, T'_2 (that represents all those *cosis* in $\overline{w_{2'}}$) and T'_9. Recall that T_9 represented the class of *cosis* $\overline{w_9}$; it is easy to describe the class of *cosis* represented by T'_9 if the following subclass of $\overline{w_9}$ is considered:

- $\overline{w_{9'}}$: *cosis* in which Δ, $(\forall x(\mathtt{B}_x \Rightarrow \mathtt{F}_x) \vee \neg\forall x(\mathtt{B}_x \Rightarrow \mathtt{F}_x))$, $\neg\forall x(\mathtt{B}_x \Rightarrow \mathtt{F}_x)$, $\mathtt{P}_W$, $\neg\mathtt{F}_W$ and $(\mathtt{P}_W \Rightarrow \neg\mathtt{F}_W)$ hold but $\forall x(\mathtt{B}_x \Rightarrow \mathtt{F}_x)$ and $\neg\mathtt{P}_W$ do not hold.

$\overline{w_{9'}}$ is a subclass of $\overline{w_9}$, as it contains the situations in this class in which the generic object (referred to as o above, in its definition) is precisely W[19].

[19] $\mathtt{B}_W$ is not missing in the definition of $\overline{w_{9'}}$, recall that it belongs to Δ.

Another interesting side-effect of this restriction is that $\overline{w_{9'}}$ is an equivalence class (according to the partition defined by R_i), whereas $\overline{w_9}$ was a collection of such classes.

The new tableau, T'_9, represents $\overline{w_{9'}}$. The change in the agent's beliefs produced by the *re-opening* decision is modelled by generating a new accessibility relation, R_7, in which the accessible *cosis* will be those of $\overline{w_{2'}}$ and $\overline{w_{9'}}$ (those *cosis* represented by the open tableaux, T'_2 and T'_9). Notice an important fact: it is the first time in this example that the set of accessible *cosis increases*, and not *decreases* (the only R_6-accessible *cosis* were those in $\overline{w_{2'}}$). Therefore, it is also the first time in which the agent's positive beliefs may be *reduced*, and not *increased*. This situation is depicted in figure 20.

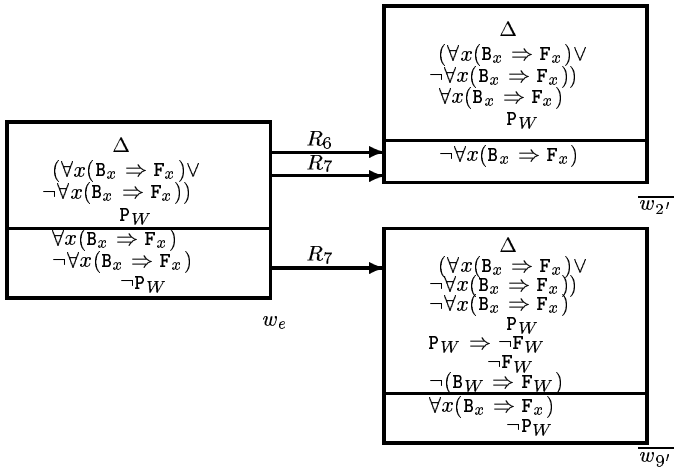

Figure 20: Generation of R_7

The only formulæ common to $\overline{w_{2'}}$ and $\overline{w_{9'}}$ (that constitute, through the application of the modified Kripke semantics, the agent's positive beliefs) are those formulæ of Δ plus $(\forall x(\mathtt{B}_x \Rightarrow \mathtt{F}_x) \vee \neg\forall x(\mathtt{B}_x \Rightarrow \mathtt{F}_x))$ and $\mathtt{P}_W$. The formulæ that are known not to appear in some doxastic alternative (that constitute, following our way of applying the Kripke semantics, the agent's negative beliefs) are $\forall x(\mathtt{B}_x \Rightarrow \mathtt{F}_x)$, $\neg\forall x(\mathtt{B}_x \Rightarrow \mathtt{F}_x)$ and $\neg\mathtt{P}_W$. Thus, the agent's

beliefs at this point of the analysis, after *re-opening* a branch of the analysis that was previously (*empirically*) closed, are:

$$\{B(\mathtt{B_T}), B(\mathtt{F_T}), B(\mathtt{B_P}), B(\mathtt{F_P}), B(\mathtt{B_W}), B(\forall x(\mathtt{P}_x \Rightarrow \neg \mathtt{F}_x)),$$
$$B(\forall x(\mathtt{B}_x \Rightarrow \mathtt{F}_x) \vee \neg \forall x(\mathtt{B}_x \Rightarrow \mathtt{F}_x)), B(\mathtt{P}_W),$$
$$\neg B(\forall x(\mathtt{B}_x \Rightarrow \mathtt{F}_x)), \neg B(\neg \forall x(\mathtt{B}_x \Rightarrow \mathtt{F}_x)), \neg B(\neg \mathtt{P}_W)\}$$

One of the effects of the addition of doxastic alternatives has been the generation of a new negative belief, $\neg \mathtt{P}_W$. A more interesting effect is the restriction in the agent's positive beliefs: the agent has stopped believing that all birds fly (which is a rational thing to do, after having discovered that there exists an accessible *conceivable situation* in which an specific bird, namely *Woody*, does not fly). Now it does not believe in this law, as it did with the previous accessibility relation; thus, our model can deal with non-monotonic beliefs (*i.e.* beliefs that hold in a particular point in time but may be retracted later, for instance in the face of new external information or new conclusions obtained in the logical analysis). Note that rational inquirers are not, however, perfect reasoners: the agent, even after having discovered a bird that does not fly, does not still believe that not all birds fly (in fact, it explicitly disbelieves this fact). The reason is that there are doxastic alternatives (those represented by the open tableau T_2') in which the law *"All birds fly"* still holds. After some steps of logical analysis in T_2', however, the agent will be able to dismiss all the *cosis* in that class and will finally reach the conclusion that not all birds fly. Those steps constitute the rest of the example.

The agent initiated those final steps with the logical analysis of the formula *"There does not exist any penguin that flies"* in T_2', instantiating it with the constant W (see figure 9, where the new subtableau, T_{10}, is shown). Recall that T_2' represented the class of *cosis* $\overline{w_{2'}}$. Consider now the following class:

- $\overline{w_{10}}$: *cosis* in which Δ, $(\forall x(\mathtt{B}_x \Rightarrow \mathtt{F}_x) \vee \neg \forall x(\mathtt{B}_x \Rightarrow \mathtt{F}_x))$, $\forall x(\mathtt{B}_x \Rightarrow \mathtt{F}_x)$, $\mathtt{P}_W$ and $(\mathtt{P}_W \Rightarrow \neg \mathtt{F}_W)$ hold but $\neg \forall x(\mathtt{B}_x \Rightarrow \mathtt{F}_x)$ does not hold.

This new class of *cosis*, which is represented by the tableau T_{10}, is more i-informative than $\overline{w_{2'}}$. The other open tableau, T_9', represents $\overline{w_{9'}}$. The change produced in the agent's beliefs by the last step of analysis is modelled with

the construction of a new accessibility relation, R_8, in which the agent has access to those *cosis* represented by the two open tableaux. This situation is reflected in figure 21.

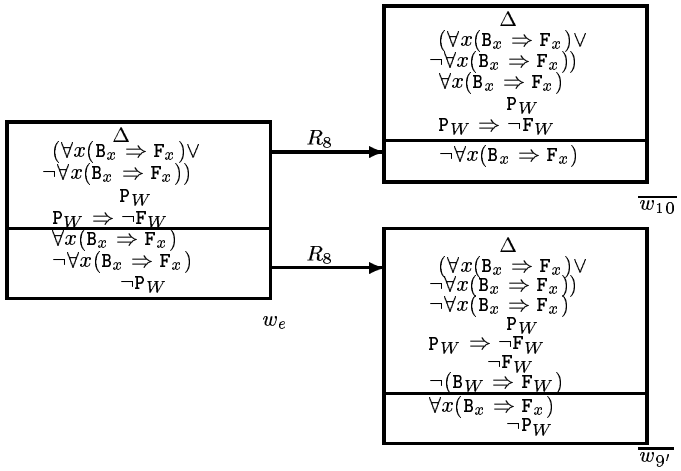

Figure 21: Generation of R_8

The agent's positive beliefs are those formulæ common to all the *cosis* in $\overline{w_{10}}$ and $\overline{w_{9'}}$ (*i.e.* the formulæ in Δ plus $(\forall x(\mathtt{B}_x \Rightarrow \mathtt{F}_x) \vee \neg\forall x(\mathtt{B}_x \Rightarrow \mathtt{F}_x))$, $\mathtt{P}_W$ and $(\mathtt{P}_W \Rightarrow \neg\mathtt{F}_W)$). The agent's negative beliefs are those formulæ that are known not to be contained in at least one doxastic alternative: in this case, the formulæ $\forall x(\mathtt{B}_x \Rightarrow \mathtt{F}_x)$ and $\neg\mathtt{P}_W$ do not appear in the *cosis* of $\overline{w_{9'}}$, whereas $\neg\forall x(\mathtt{B}_x \Rightarrow \mathtt{F}_x)$ does not appear in the *cosis* of $\overline{w_{10}}$. Summarising, the last step of analysis has produced the addition of a new positive belief, $(\mathtt{P}_W \Rightarrow \neg\mathtt{F}_W)$:

$$\{B(\mathtt{B}_\mathtt{T}), B(\mathtt{F}_\mathtt{T}), B(\mathtt{B}_\mathtt{P}), B(\mathtt{F}_\mathtt{P}), B(\mathtt{B}_\mathtt{W}), B(\forall x(\mathtt{P}_x \Rightarrow \neg\mathtt{F}_x)),$$
$$B(\forall x(\mathtt{B}_x \Rightarrow \mathtt{F}_x) \vee \neg\forall x(\mathtt{B}_x \Rightarrow \mathtt{F}_x)), B(\mathtt{P}_W), B(\mathtt{P}_W \Rightarrow \neg\mathtt{F}_W),$$
$$\neg B(\forall x(\mathtt{B}_x \Rightarrow \mathtt{F}_x)), \neg B(\neg\forall x(\mathtt{B}_x \Rightarrow \mathtt{F}_x)), \neg B(\neg\mathtt{P}_W)\}$$

The agent continued the logical analysis with the formula that it had obtained in the previous step, $(\mathtt{P}_W \Rightarrow \neg\mathtt{F}_W)$, generating three new subtableaux (T_{11}, T_{12} and T_{13}), as shown in figure 9.

These new subtableaux represent the classes of *cosis* $\overline{w_{11}}$, $\overline{w_{12}}$ and $\overline{w_{13}}$, which are more i-informative than $\overline{w_{10}}$, as shown in the following definition:

- $\overline{w_{11}}$: *cosis* in which Δ, $(\forall x(\mathtt{B}_x \Rightarrow \mathtt{F}_x) \vee \neg\forall x(\mathtt{B}_x \Rightarrow \mathtt{F}_x))$, $\forall x(\mathtt{B}_x \Rightarrow \mathtt{F}_x)$, $\mathtt{P}_W$, $(\mathtt{P}_W \Rightarrow \neg\mathtt{F}_W)$, $\neg\mathtt{P}_W$ and $\neg\mathtt{F}_W$ hold but $\neg\forall x(\mathtt{B}_x \Rightarrow \mathtt{F}_x)$ does not hold.

- $\overline{w_{12}}$: *cosis* in which Δ, $(\forall x(\mathtt{B}_x \Rightarrow \mathtt{F}_x) \vee \neg\forall x(\mathtt{B}_x \Rightarrow \mathtt{F}_x))$, $\forall x(\mathtt{B}_x \Rightarrow \mathtt{F}_x)$, $\mathtt{P}_W$, $(\mathtt{P}_W \Rightarrow \neg\mathtt{F}_W)$ and $\neg\mathtt{P}_W$ hold but $\neg\forall x(\mathtt{B}_x \Rightarrow \mathtt{F}_x)$ and $\neg\mathtt{F}_W$ do not hold.

- $\overline{w_{13}}$: *cosis* in which Δ, $(\forall x(\mathtt{B}_x \Rightarrow \mathtt{F}_x) \vee \neg\forall x(\mathtt{B}_x \Rightarrow \mathtt{F}_x))$, $\forall x(\mathtt{B}_x \Rightarrow \mathtt{F}_x)$, $\mathtt{P}_W$, $(\mathtt{P}_W \Rightarrow \neg\mathtt{F}_W)$ and $\neg\mathtt{F}_W$ hold but $\neg\forall x(\mathtt{B}_x \Rightarrow \mathtt{F}_x)$ and $\neg\mathtt{P}_W$ do not hold.

T_{11} and T_{12} contain a formula $(\mathtt{P}_\mathtt{W})$ and its negation in their left columns, so they represent classes of *logically impossible cosis*. The agent decided to eliminate these classes from the analysis by *logically closing* the tableaux that represent them. Thus, the only open tableaux at this point of the analysis were T_{13} and T'_9.

The change produced in the agent's beliefs as a consequence of the last logical step of analysis is reflected in our model in the generation of a new accessibility relation, R_9, that changes the set of doxastic alternatives (the *cosis* in $\overline{w_{13}}$ are R_9-accessible, whereas those in $\overline{w_{10}}$ cease to be considered as viable alternatives). This situation is pictured in figure 22.

The agent's belief set, built by applying the modified Kripke semantics to this state of affairs, is the following:

$$\{B(\mathtt{B}_\mathtt{T}), B(\mathtt{F}_\mathtt{T}), B(\mathtt{B}_\mathtt{P}), B(\mathtt{F}_\mathtt{P}), B(\mathtt{B}_\mathtt{W}), B(\forall x(\mathtt{P}_x \Rightarrow \neg\mathtt{F}_x)),$$
$$B(\forall x(\mathtt{B}_x \Rightarrow \mathtt{F}_x) \vee \neg\forall x(\mathtt{B}_x \Rightarrow \mathtt{F}_x)), B(\mathtt{P}_W), B(\mathtt{P}_W \Rightarrow \neg\mathtt{F}_W),$$
$$B(\neg\mathtt{F}_W), \neg B(\forall x(\mathtt{B}_x \Rightarrow \mathtt{F}_x)), \neg B(\neg\forall x(\mathtt{B}_x \Rightarrow \mathtt{F}_x)), \neg B(\neg\mathtt{P}_W)\}$$

Notice that the only change produced in the agent's beliefs (as a result of the last logical analysis) is the inclusion of a new positive belief, $\neg\mathtt{F}_W$. The agent believes that *Woody* does not fly because this information is included in the left columns of the two open tableaux of the logical analysis, T_{13} and T'_9.

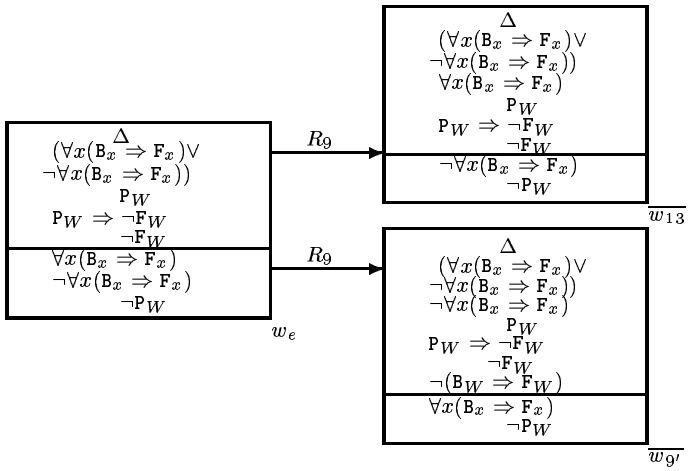

Figure 22: Generation of R_9

The agent analysed now the formula $\forall x(\mathtt{B_x} \Rightarrow \mathtt{F_x})$, contained in the left column of T_{13}, instantiating it with the constant W. The result was the generation of a new subtableau, T_{14} (see figure 10).

It was already seen that T_{13} represented the class of *cosis* $\overline{w_{13}}$. Consider now the class $\overline{w_{14}}$, which is more i-informative:

- $\overline{w_{14}}$: *cosis* in which Δ, $(\forall x(\mathtt{B}_x \Rightarrow \mathtt{F}_x) \vee \neg\forall x(\mathtt{B}_x \Rightarrow \mathtt{F}_x))$, $\forall x(\mathtt{B}_x \Rightarrow \mathtt{F}_x)$, $\mathtt{P}_W$, $(\mathtt{P}_W \Rightarrow \neg\mathtt{F}_W)$, $\neg\mathtt{F}_W$ and $(\mathtt{B}_W \Rightarrow \mathtt{F}_W)$ hold but $\neg\forall x(\mathtt{B}_x \Rightarrow \mathtt{F}_x)$ and $\neg\mathtt{P}_W$ do not hold.

The new tableau, T_{14}, represents the class of *cosis* $\overline{w_{14}}$. The only other open tableau in the tableaux tree is T'_9, that represents class $\overline{w_{9'}}$. Thus, the *cosis* in these two classes are the only doxastic alternatives considered by the agent in this stage of the analysis. This fact is reflected in our model by creating a new accessibility relation, R_{10}, that modifies again the set of accessible *cosis*; basically, it substitutes the *cosis* in $\overline{w_{13}}$ by those in $\overline{w_{14}}$, as shown in figure 23.

The agent's (positive and negative) beliefs have not changed after this logical analysis, because all the formulæ that were common to all R_9-accessible

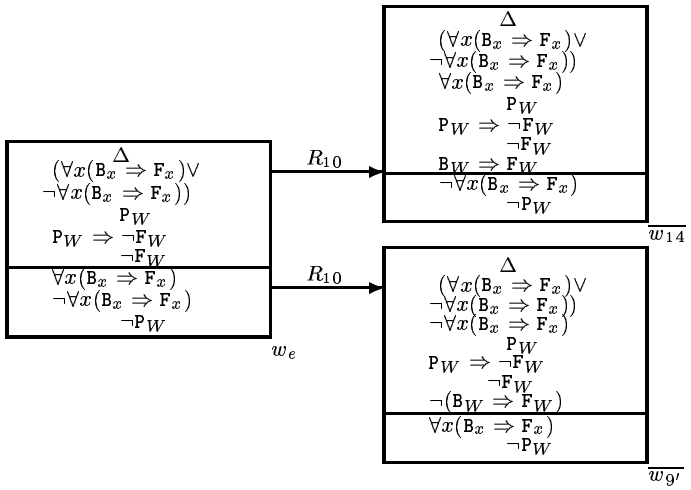

Figure 23: Generation of R_{10}

cosis also appear in those that are R_{10}-accessible, and the formulæ that were known to be missing from at least one R_9-accessible *cosi* keep the same status.

In the last step of this example, the agent analysed the formula $(\mathtt{B}_W \Rightarrow \mathtt{F}_W)$ in T_{14}. This logical analysis caused the generation of three new subtableaux (T_{15}, T_{16} and T_{17}) as shown in figure 10.

The new tableaux represent classes of *cosis* which are more i-informative than those represented by T_{14} (*i.e.* $\overline{w_{14}}$). The agent noticed that the three subtableaux represent classes of *logically impossible cosis*, because all of them contain a formula and its negation in their left columns ($\mathtt{F}_W$ in T_{15} and T_{17}, $\mathtt{B}_W$ in T_{15} and T_{16}). As it decided to *logically close* all of these tableaux, that branch of the tableaux tree was dismissed from the analysis, and the only remaining open tableau was T'_9. The change in the agent's beliefs produced by this decision of the agent is semantically reflected in our model by generating a final accessibility relation, R_{11}, that restricts the set of accessible *cosis* to those that are represented by the only open tableau, T'_9. This situation is shown in figure 24.

This reduction of doxastic alternatives has caused two changes in the agent's beliefs:

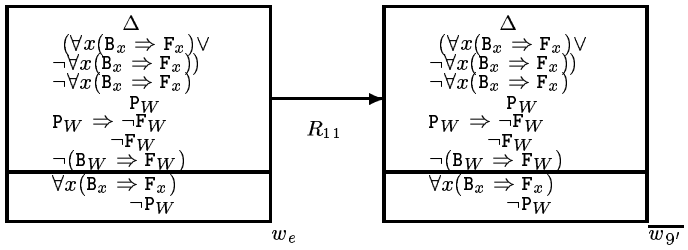

Figure 24: Generation of R_{11}

- There is a new positive belief, $\neg(\mathtt{B}_W \Rightarrow \mathtt{F}_W)$. This formula is a new positive belief because it is included in all R_{11}-accessible *cosis*, *i.e.* in all the *cosis* in class $\overline{w_{14}}$.

- The formula $\neg(\forall x(B_x \Rightarrow F_x))$ (*"Not all birds fly"*), that was *negatively* believed by the agent (when the agent's beliefs were considered after generating the accessibility relation R_{10}) is also a new *positive* belief. The agent has discovered that all the *cosis* in which the general law held were *logically impossible* and, after having dismissed them, in all its doxastic alternatives the negation of the law holds and, therefore, the agent now incorporates this formula into its positive set of beliefs.

5.8.1 Summary of the example

A summary of the example (more detailed than the one given in §4.6.1, because now we have established which are the agent's beliefs at each step) may be now given:

- The agent starts the analysis by wondering whether "all birds fly". It uses the *exploratory* dimension of analysis in order to introduce this doubt into the analysis. In this way, the agent may explore the two available alternatives and determine whether any of them is logically impossible, or whether there is any question that it can make to its environment in order to confirm or refute any of the two alternatives.

- After some *logical analysis*, the agent discovers that it can dismiss one of the options if it can check whether there is an individual that is a bird and does not fly. The *experimental* dimension of analysis is used to search in the environment for an individual with these properties.

- As the agent does not find any individual with the desired properties, it decides to temporarily dismiss that alternative, and then it believes that *"All birds fly"* (because it has not been able to find any counterexample).

- Afterwards, the agent receives external information, that assures that *"Woody is a penguin"*. The agent incorporates this information in the analysis by adding it to the left columns of all the open tableaux.

- After some logical analysis, the agent discovers that there is indeed an individual that is a bird and does not fly ($Woody$). Having (logically) discovered this fact, it ceases to believe that *"All birds fly"*, although it does not believe (yet) that *"Not all birds fly"*.

- After further logical analysis, the agent discovers that all those situations in which *"All birds fly"* are *logically impossible*. Then, it reaches the final conclusion that *"Not all birds fly"*.

Figure 25 represents the evolution of the set of doxastic alternatives considered by the agent in the course of its inquiry. Each circle represents a set of *cosis* which is an equivalence class (under the partition induced by R_i). A link between two classes denotes that the upper one is more i-informative than the lower one. Note that there are eight times in which the agent has climbed in the information hierarchy (in the generation of R_1, R_2, R_3, R_4, R_6, R_8, R_9 and R_{10}), two times in which the set of doxastic alternatives has been reduced (R_5 and R_{11}) and one case in which the set of viable alternatives considered by the agent has grown (R_7).

5.9 Summary

In this chapter we have shown how the evolution of the beliefs of a rational inquirer, caused by its continuous dynamic multi-dimensional belief analysis, may be formally modelled. The basic semantic entities that are used in this

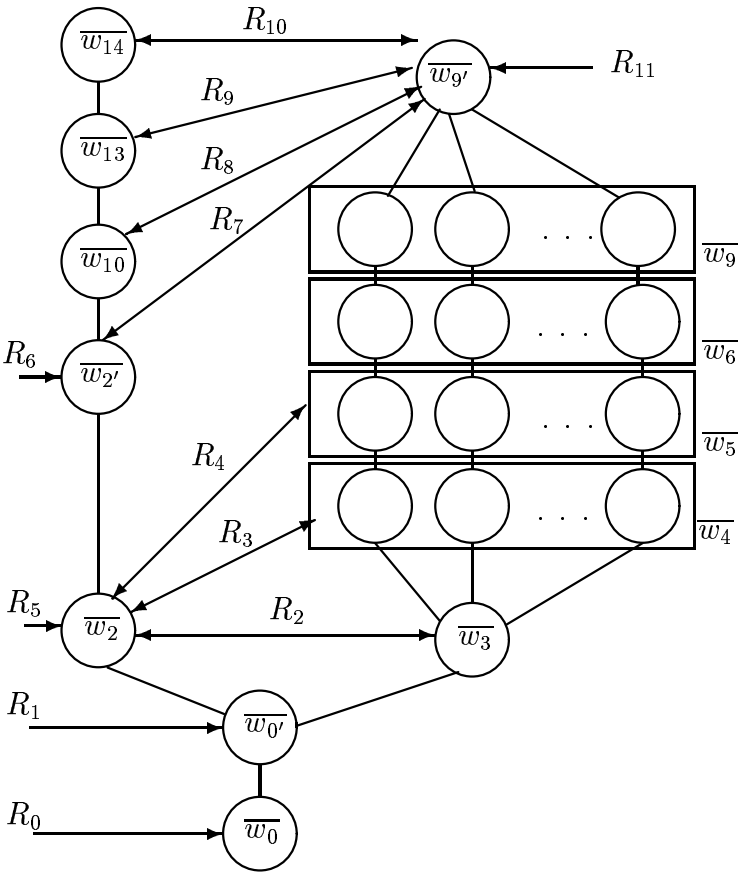

Figure 25: Sequence of accessibility relations

modelling are the *conceivable situations*, which correspond to the kind of situations described in §3. We construct a sequence of *accessibility relations* that determine, at each step of the analysis, which are the situations that are conceived as possible by the agent. The analysis of the formulæ that hold (or do not hold) in these doxastic alternatives, *via* our modified Kripke semantics, serves to establish the agent's set of beliefs at each point in time.

6 Summary and future work

In this final chapter we provide a brief summary of the main ideas that have been described in this dissertation and we suggest some lines of future research.

6.1 Summary of the proposal

The main aim of this work has been to develop a way to *model the process of rational inquiry* (the evolution of a rational agent's set of beliefs over time as a consequence of its interaction with the world and its internal inferential processes), *keeping the general idea of the possible worlds model and the Kripke semantics*. It is well known that the use of this classical framework leads to the problems of logical omniscience and perfect reasoning. As real agents do not have these properties, one of our objectives has been to avoid them. Thus, we started this proposal by making a thorough review of the main formalisms that have been put forward to solve these problems (§2, [More98]). Having done that, we suggested another alternative, based on the new concept of *subjective situations* (§3, [MCS99b], [MCS00b]). This proposal is based on the idea that a given situation may be perceived in different ways by different agents, and that this perception influences (or even determines) the agent's beliefs. These states of affairs are described by each agent with two sets of formulæ, that represent the positive and negative information that the agent has about them.

As we want to model the evolution of a *rational* agent's set of beliefs, we identified in §4 which are the main doxastic tasks that may modify the set of beliefs of a rational agent. After that, we defined a particular class of rational agents, called *rational inquirers*, that have specific ways of making these doxastic tasks ([MCS99a], [MCS00a]). These agents are constantly performing a multi-dimensional dynamic analysis of their beliefs, in order to make them as similar as possible to the facts that hold in the real world. These agents are given the following capabilities:

- They may perform some (limited) deductive inferences on their sets of beliefs, using a modified version of the classical analytic tableaux method.

- They may have doubts about their beliefs, and may introduce these doubts into the analysis by adding instances of the *Axiom of the Excluded Middle* into the tableaux of the logical analysis.

- They may make questions to the environment, in order to confirm or refute doubtful beliefs. The answers received from the environment are also included in the open tableaux of the logical analysis. The questions to be posed to the environment are suggested by the Skolem constants that appear in the logical analysis, linking in a novel fashion the *rational* and *empirical* components of rational inquiry ([ReBr79]).

- They may also add to their beliefs the information that they receive directly from the environment (*e.g.* the data supplied by other agents or the measures made by external sensors).

In §5 it is shown how the evolution of the beliefs caused by this dynamic multi-dimensional belief analysis may be formally modelled. This modelling process is made using two basic tools:

- We consider *conceivable situations* (*cosis*) as the primitive semantic entities; they include all the situations that the modelled agent is capable of imagining or considering, regardless of their partiality or inconsistency. A *cosi* is partially represented with two sets of first-order formulæ.

- The accessibility relation between *cosis* is not constant, fixed, but variable. This variability accounts for the evolution of the agent's beliefs over time (through a modified version of the usual Kripke semantics).

The agent performs steps of analysis of its beliefs over an analytic tableaux tree. Each tableau represents a class of conceivable situations, and the set of open tableaux defines the situations that are considered as doxastic alternatives by the agent at each point in time. After each step of belief analysis, the set of open tableaux changes; therefore, the set of doxastic alternatives is modified. This fact is represented in our model with the generation of an accessibility relation, that limits which are the worlds considered as possible by the agent. By applying a slightly modified version of the Kripke semantics on this set of alternatives, we obtain which are the agent's sets of positive and negative beliefs in each point in time.

In summary, the contributions of this work are:

- An extensive review of more than twenty ways in which the logical omniscience problem has been tackled.
- The definition of a radically new approach to avoid this problem, based on the new concept of *subjective situations*.
- The definition of an abstract model of rational agents, identifying the tasks that they perform on their sets of beliefs.
- The definition of a particular class of rational agents, called *rational inquirers*, in which the doxastic tasks are implemented in specific ways.
- A detailed explanation of the way in which the evolution of the beliefs of this class of agents may be formally modelled in the subjective situations framework.

6.2 Future work

The first questions to be addressed in our future work will probably be among the following:

- There are several issues related to the *logical dimension of analysis* that could be studied:
  - It could be interesting to increase the expressivity of the modal language presented in §3.4, eliminating the constraint of dealing with *linearly nested* formulæ; in that way, an agent could analyse formulæ such as $B_5\mathtt{P} \Rightarrow B_7\mathtt{Q}$ (if $Agent_5$ believes P, then $Agent_7$ believes Q). In that case each agent should have a modal calculus, and not a predicate calculus such as the one used by *rational inquirers*. This would be a change of strategy in our research because, in this dissertation, we have used modal logic as a meta-language to talk about the beliefs of an agent, but actual agents perform their internal reasonings in first-order logic.

  - It is arguable whether it is appropriate to allow an agent to keep open some tableaux that could be closed, just because it could have not noticed that the tableau contained a contradiction. We defended in §4.3.1 that this property seems appropriate to model limited reasoners, that may have not noticed that they believe a formula and its negation or that they have both positive and negative evidence of a certain fact; however, it could also be argued that, regardless of the complexity of the information contained in a tableau, the agent could not fail to observe a fact as obvious as the presence of one of the tableau closing conditions.
  - In appendix A the relationship between the propositional part of our tableaux calculus and Kleene's strong three-valued logic is shown. We wonder whether some small changes in the tableaux calculus could lead to other interesting kinds of logics, such as linear, intuitionistic or relevance logic. That possibility is quite clear, especially taking into account that the agents modelled by Levesque's logic of explicit and implicit belief are perfect reasoners in relevance logic (see [Vard86]). If that were the case, the feasibility of modelling the evolution of the beliefs of an agent in these frameworks could be also studied.

- The *experimental analysis* also raises some questions:
  - Which type of experiences should be allowed in the experimental dimension of analysis? How does the restriction on the allowed answers from the environment change the potential results of the process of inquiry? Those facts are important, as Hintikka notes in [Hint88] and [Hint92]. There is a whole hierarchy of possibilities: we could allow only atomic questions with boolean answers, or we could permit the agent to make any question and receive any answer. Some intermediate points could be also considered, *e.g.* to allow disjunctive (`P` or `Q`) or existential (give me an x such that $\mathtt{S}_x$ holds) questions, such as the ones mentioned in §4.3.2.
  - It is also important to consider which conditions must be satisfied before the agent may make a question (*e.g.* we could require the agent to have the atom `P` in an open tableau before being allowed to ask whether `P` is or not the case).

- We may also suggest some future research concerning the *exploratory dimension of analysis*:
  - Following Hintikka's ideas ([Hint86a]), it must be carefully considered whether we allow the introduction of any doubt or we only allow (in a very natural restriction) that only doubts referred to concepts known by the agent may be considered (*e.g.* instances of the Axiom of the Excluded Middle in which the used formula appears as a subformula of a formula of the tableau in which the doubt is to be added).
  - It could be considered whether it is interesting to allow the agent to introduce into the open tableaux instances of other tautologies, different from the Axiom of the Excluded Middle (that may upset the intuitionistic readers). Another possibility is to allow the agent to perform *hypothetical reasoning*, to allow it to analyse how their beliefs would evolve if some formula ϕ were true (it could add this formula to all the open tableaux and study which conclusions may be reached, always keeping in mind that it is inside a hypothetical mode of reasoning).
  - It could be thought whether it is necessary to give the agent the capability of being able to forget, at some point of the analysis, a doubt that it had considered in a previous stage of the analysis (*e.g.* it could have used a lot of resources trying to solve the doubt and it might not have come up with an answer; thus, it could decide that it is not worth wasting more efforts in that direction).
- Regarding the incorporation of *external inputs* into the tableaux, we should study specific ways in which a rational inquirer may implement a *belief revision* strategy; we should look for ways in which the agent may use the information that it has about the origin of each belief in order to determine which formulæ should be withdrawn in the presence of contradictory information, if that were necessary. For instance, some of these policies could be implemented:
  - If some of the formulæ involved in the contradiction were received as external inputs or were obtained in the experimental dimension,

they could be eliminated to avoid the contradiction (assuming that the source of the formula was not fully reliable).

- The opposite direction could be taken: the agent could maintain the last information that it has received, discarding previously held beliefs that were in contradiction with the new data. This strategy could help the agent to update its set of beliefs when the facts that hold in its environment change over time.
- If a formula comes from the exploratory dimension, it could be a sign that the introduced doubt may be solved in favour of the other alternative.
- If all the used formulæ came from the logical analysis, the agent could make some experiments in order to try to refute some of them, and thus avoid the contradiction.

The bottom line is that the presence of contradictions would not be a negative fact, to be avoided, but a positive one, that would guide the agent towards a progressive refinement of its beliefs (by eliminating alternatives, solving doubts, pointing out information that may be wrong, *etc.*).

- We should also study the different alternatives that an agent has when it analyses its set of beliefs, *i.e.* what strategies it could use to combine the different available dimensions of analysis. For instance, we could consider alternatives such as the following:
  - Logic: it only performs logical analysis, making it as exhaustive as its resources permit; it could also consider some restricted use of the exploratory dimension, posing doubts that could be logically analysed.
  - Physicist: it could make continuous questions to the environment (through experiences or tests) to incorporate new information to its beliefs. It could also use, in especial occasions, the logical and exploratory dimensions.
  - Robot: it could be constantly receiving information through its sensors, and adding this information to its beliefs. It could also use sometimes the logical dimension in order to deduce new facts.

  - Human: it combines in a rational way all the dimensions of analysis. It receives information from the environment, and takes this information into account in its beliefs. It also performs a limited logical analysis of its beliefs. Sometimes it would pose itself some questions, and it would also perform experiences in the environment in order to increase its set of beliefs by eliminating impossible alternatives.

- In this dissertation we have considered the different doxastic activities in which a single agent (more specifically, a *rational inquirer*) may engage, and we have shown a formal way of modelling the evolution of the beliefs of this kind of agents. However, it would be much harder to consider the evolution of the beliefs of all the agents composing a *multi-agent system*, because a lot of new issues that are not being considered in this proposal would arise. Some of these future topics of research could be the following:

  - A detailed study of each agent's introspective properties mentioned in §3 should be made. We think that they can be given a natural interpretation in the case of beliefs about the own agent's beliefs: it could just be said that an agent is aware of the fact that it has an internal structure where it keeps the formulæ that it has reasons to support and the formulæ that it has reasons to deny, and that these formulæ constitute its positive and negative beliefs about the state of the world.
  - It is worth pointing out that, in the modelling of beliefs made in §5, we have not used the full strength of the *subjective situations* framework defined in §3. As we have been dealing with just one agent, we have not considered a modal belief operator for each agent and the possibility of an agent having *linearly nested* beliefs (recall definition 5) about other agents (*e.g.* α believing that β believes that γ believes $\mathtt{P}$). Thus, the framework developed in §3 could probably be very useful when modelling multi-agent systems composed by non-ideal agents.
  - We should differentiate between those formulæ that are received from external sensors and those that are sent by other agents of

the system. We could assign a degree of credibility to each of these formulæ, using some degree of trust associated to each sensor and to the other agents (see *e.g.* [PaGi98], [JoTr99]).

7 References

[AGHP99] M.D'Agostino, D.Gabbay, R.Hähnle, J.Posegga (Eds.), "*Handbook of Tableau Methods*", Kluwer Academic Publishers, 1999.

[AGM85] Alchourrón, C., Gärdenfors, P., Makinson, D., "*On the logic of theory change: partial meet functions for contraction and revision*", *Journal of Symbolic Logic* 50, pp. 510-530, 1985.

[AlNe84] Almukdad, A., Nelson, D., "*Constructive falsity and inexact predicates*", *Journal of Symbolic Logic* 49, pp. 231-233, 1984.

[Alva98] Alvarado, M., "*An approach to knowledge and belief in Strong Kleene Logic: change and automatization using analytic tableaux*", PhD Thesis, Software Department, Technical University of Catalonia (UPC), Barcelona, 1998.

[AnBe75] Anderson, A., Belnap, N., "*Entailment: the logic of relevance and necessity*", Princeton University Press, Princeton, NJ, 1975.

[Appe85] Appelt, D., "*Planning English sentences*", *Studies in Natural Language Processing*, Cambridge University Press, 1985.

[AtSi93] Attardi, G., Simi, M., "*A formalization of viewpoints*", Technical Report TR-93-062, Dipartimento di Informatica, Università di Pisa, 1993.

[BaPe83] Barwise, J., Perry, J., "*Situations and attitudes*", Bradford Books, Cambridge, MA, 1983.

[Barw81] Barwise, J., "*Scenes and other situations*", *Journal of Philosophy* 78 (7), pp. 369-397, 1981.

[Barw88] Barwise, J., "*The situation in logic*", CSLI Lecture Notes, No. 17, 1988.

[Beln66] Belnap, N., "*Questions, answers and presuppositions*", *The Journal of Philosophy* 63, pp. 609-611, 1966.

[Beln77] Belnap, N., "*A useful four-valued logic*", in *Modern uses of multiple-valued logic*, Epstein,G., Dunn, J. (Eds.), pp. 5-37, Reidel, 1977.

[BeMa77] Bell, J., Machover, M., "*A course in Mathematical Logic*", North Holland, 1977.

[Bene97] Benerecetti, M., Cimatti, A., Giunchiglia, E., Giunchiglia, F., Serafini, L., "*Formal specification of beliefs in multi-agent systems*", in [MWJ97], pp. 117-130.

[Beth55] Beth, E., "*Semantic entailment and formal derivability*", from *Mededelingen van de Koninklijke Nederlandse Akademie van Wetenschappen, Afdeling Letterkunde*, Vol. 18, no. 13, pp. 309-342, 1955.

[Brat87] Bratman, M., "*Intentions, plans and practical reason*", Harvard University Press, 1987.

[Broo91] Brooks, R., "*Intelligence without representation*", *Artificial Intelligence* 47, pp. 139-159, 1991.

[Busc96] Busch, D., "*Sequent formalization of three-valued logic*", included in [Dohe96], pp. 45-76.

[Cher86] Cherniak, C., "*Minimal rationality*", Bradford Books, MIT Press, 1986.

[Chom82] Chomsky, N., "*The generative enterprise*", Foris, Dordrecht, 1982.

[CoLe90] Cohen, P., Levesque, H., "*Intention is choice with commitment*", *Artificial Intelligence* 42, pp. 213-261, 1990.

[Cres72] Cresswell, M., "*Intensional logics and logical truth*", *Journal of Philosophical Logic* 1, pp. 2-15, 1972.

[Cres73] Cresswell, M., "*Logics and languages*", Methuen, 1973.

[CRW97] Cavedon, L., Rao, A., Wobcke, W. (Eds.), "*Intelligent Agent Systems: Theoretical and Practical Issues*", Lecture Notes in Artificial Intelligence 1209, Springer Verlag 1997.

[CuPo95] Cummins, Pollock, J. (Eds.), "*Philosophy and AI: essays at the interface*", MIT Press, 1995.

[DCBB95] Da Costa, N., Béziau, J., Bueno, O., "*Aspects of paraconsistent logic*", *Bulletin of the IGPL*, Vol. 3, No. 4, pp. 597-614, 1995.

[Delg95] Delgrande, J., "*A framework for logics of explicit belief*", *Computational Intelligence* 11, pp. 47-86, 1995.

[Denn78] Dennett, D., "*Brainstorms*", Bradford Books, 1978.

[Denn84] Dennett, D., "*Elbow room*", Oxford University Press, 1984.

[Devl91] Devlin, K., "*Logic and information*", Cambridge University Press, 1991.

[Dohe96] Doherty, P. (Ed.), "*Partiality, modality, and nonmonotonicity*", Studies in Logic, Language and Information, Center for the Study of Language and Information Publications, 1996.

[DRLe88] Des Rivières, J., Levesque, H., "*The consistency of syntactical treatments of knowledge (how to compile quantificational modal logics into classical FOL)*", *Computational Intelligence* 4, pp. 31-41, 1988.

[Duc95] Duc, Ho Ngoc, "*Logical omniscience vs. logical ingorance. On a dilemma of epistemic logic*". In C.P.Pereira, N.Mamede (Eds.), *Progress in Artificial Intelligence*, Proceedings of EPIA'95, Lecture Notes in Artificial Intelligence 990, pp. 237-248, Springer, 1995.

[Duc97] Duc, Ho Ngoc, "*Reasoning about rational, but not logically omniscient agents*", *Journal of Logic and Computation* 7 (5), pp. 633-648, 1997.

[Eber74] Eberle, R., "*A logic of believing, knowing and inferring*", *Synthese* 26, pp. 356-382, 1974.

[Elgo88] Elgot-Drapkin, J., "*Step-logic: reasoning situated in time*", Ph.D. thesis, University of Maryland, 1988.

[ElPe90] Elgot-Drapkin, J., Perlis, D., "*Reasoning situated in time I: Basic concepts*", *Journal of Experimental and Theoretical Artificial Intelligence* 2, pp. 75-98, 1990.

[EMP95] Elgot-Drapkin, J., Miller, M., Perlis, D., "*Memory, reason, and time: the step-logic approach*", appears in [CuPo95], pp. 79-103.

[FaHa85] Fagin, R., Halpern, J., "*Belief, awareness and limited reasoning*", *Procs. of the Ninth International Joint Conference on Artificial Intelligence, IJCAI-85*, pp. 491-501, 1985.

[FaHa88] Fagin, R., Halpern, J., "*Belief, awareness and limited reasoning*", *Artificial Intelligence* 34, pp. 39-76, 1988.

[Ferg95] Ferguson, I., "*Integrated control and coordinated behaviour: A case for agent models*", in M.Wooldridge and N.Jennings (Eds.), *Intelligent Agents: Theories, Architectures, and Languages*, Lecture Notes in Artificial Intelligence 890, pp. 203-218, Springer Verlag, 1995.

[FHV84] Fagin, R., Halpern, J., Vardi, M., "*A model theoretic analysis of knowledge*", *Proceedings of the 25th IEEE Symposium on Foundations of Computer Science*, pp. 268-278, 1984.

[FHV90a] Fagin, R., Halpern, J., Vardi, M., "*A non-standard approach to the logical omniscience problem*", Research report RJ7234, IBM Research Division, Almaden Research Center, December 1990.

[FHV90b] Fagin, R., Halpern, J., Vardi, M., "*A non-standard approach to the logical omniscience problem*", *Proceedings of the Third Conference on Theoretical Aspects of Reasoning About Knowledge, TARK-90*, pp. 41-55, 1990.

[FHV91] Fagin, R., Halpern, J., Vardi, M., "*A model theoretic analysis of knowledge*", *Journal of the ACM* 38, pp. 382-428, 1991.

[FHV95] Fagin, R., Halpern, J., Vardi, M., "*A non-standard approach to the logical omniscience problem*", *Artificial Intelligence* 79, pp. 203-240, 1995.

[FHMV95] Fagin, R., Halpern, J., Moses, Y., Vardi, M., "*Reasoning about knowledge*", MIT Press, 1995.

[Fish94] Fisher, M., "*A survey of Concurrent Metatem - the language and its applications*", in D.Gabbay, H.Ohlbach (Eds.), *Temporal Logic - Proceedings of the First International Conference*, Lecture Notes in Artificial Intelligence 827, pp. 480-505, Springer Verlag, 1994.

[Fitt83] Fitting, M., "*Proof methods for modal and intuitionistic logics*", D. Reidel Publishing Co., 1983.

[Fitt91] Fitting, M., "*Bilattices and the semantics of logic programming*", *Journal of Logic Programming* 11, pp. 91-116, 1991.

[Fitt96] Fitting, M., "*First-order logic and automated theorem proving*", Springer Verlag, 1996.

[Fitt99] Fitting, M., "*Introduction*", in [AGHP99], pp. 1-43.

[FrHa97] Friedman, N., Halpern, J., "*Modeling beliefs in dynamic systems, part I: foundations*", *Artificial Intelligence* 95, pp. 257-316, 1997.

[FrHa99] Friedman, N., Halpern, J., "*Modeling beliefs in dynamic systems, part II: revision and update*", *Journal of Artificial Intelligence Research* 10, pp. 117-167, 1999.

[Frie97] Friedman, N., "*Modeling beliefs in dynamic systems*", PhD dissertation, Department of Computer Science, Stanford University, 1997.

[GaBo99] Garijo, F., Boman, M. (Eds.), "*Multi-Agent System Engineering*", Lecture Notes in Artificial Intelligence 1647, Springer Verlag, 1999.

[GaHu91] Gabbay, D., Hunter, A., "*Making inconsistency respectable: a logical framework for inconsistency in reasoning*", *Proceedings of the International Workshop on Fundamentals of Artificial Intelligence Research, FAIR-91*, 1991.

[Gärd88] Gärdenfors, P., "*Knowledge in flux*", Cambridge University Press, 1988.

[Gard84] Gardner, M., "*Puzzles from other worlds*", Viking Press, 1984.

[GaSt58] Gamow, G., Stern, M., "*Puzzle Math*", Vintage Press, 1958.

[GeIn89] Georgeff, M., Ingrand, F., "*Decision-making in an embedded reasoning system*", *Proceedings of the Eleventh Joint Conference on Artificial Intelligence, IJCAI-89*, 1989.

[GeLa87] Georgeff, M., Lansky, A., "*Reactive reasoning and planning*", *Proceedings of the Sixth National Conference on Artificial Intelligence, AAAI-87*, pp. 677-682, 1987.

[Gerb99] Gerbrandy, J., "*Bisimulations on planet Kripke*", PhD dissertation, Institute of Language, Logic and Information, University of Amsterdam, ILLC series DS-1999-01, 1999.

[Gins88] Ginsberg, M., "*Multivalued logics: a uniform approach to reasoning in Artificial Intelligence*", *Computational Intelligence* 4, pp. 265-316, 1988.

[Goré99] Goré, R., "*Tableau methods for modal and temporal logic*", in [AGHP99], pp. 297-396.

[GSGF93] Giunchiglia, F., Serafini, L., Giunchiglia, E., Frixione, M, "*Non-omniscient belief as context based-reasoning*", *Proceedings of the Thirteenth International Joint Conference on Artificial Intelligence, IJCAI-93*, pp. 548-554, 1993.

[Haas85] Haas, A., "*Possible events, actual events, and robots*", *Computational Intelligence* 1, pp. 59-70, 1985.

[Hadl88] Hadley, R., "*Logical omniscience, semantics, and models of belief*", *Computational Intelligence* 4, pp. 17-30, 1988.

[HaLa96] Halpern, J., Lakemeyer, G., "*Multi-agent only knowing*", in *Proceedings of the Sixth Conference on Theoretical Aspects of Rationality and Knowledge, TARK-96*, pp. 251-265, 1996.

[Halp86] Halpern, J., "*Reasoning about knowledge: an overview*", *Proceedings of the First Conference on Theoretical Aspects of Reasoning about Knowledge, TARK-86*, Ed. Halpern, J., pp. 1-17, 1986.

[Halp93] Halpern, J., "*Reasoning about only knowing with many agents*", in *Proceedings of the Conference of the American Association for Artificial Intelligence, AAAI-93*, pp. 655-661, 1993.

[HaMo92] Halpern, J., Moses Y., "*A guide to completeness and complexity for modal logics of knowledge and belief*", *Artificial Intelligence* 54, pp. 319-379, 1992.

[Hint62] Hintikka, J., "*Knowledge and belief*", Cornell University Press, Ithaca, N.Y., 1962.

[Hint73] Hintikka, J., "*Logic, language-games and information*", Clarendon Press, Oxford, 1973.

[Hint75a] Hintikka, J., "*Impossible possible worlds vindicated*", *Journal of Philosophical Logic* 4, pp. 475-484, 1975.

[Hint75b] Hintikka, J., "*Knowledge, belief, and logical consequence*", in Hintikka, J. (Ed.), *The intentions of intentionality*, D. Reidel, Dordrecht, 1975.

[Hint76] Hintikka, J., "*The semantics of questions and the questions of semantics*", *Acta Philosophica Fennica*, Vol. 28, N. 4, Helsinki, 1976.

[Hint81] Hintikka, J., "*On the logic of an interrogative model of scientific inquiry*", *Synthese* 47, pp. 69-83, 1981.

[Hint86a] Hintikka, J., "*Reasoning about knowledge in philosophy: the paradigm of epistemic logic*", *Proceedings of the First Conference on Theoretical Aspects of Reasoning about Knowledge, TARK-86*, Ed. Halpern, J., pp. 63-80, 1986.

[Hint86b] Hintikka, J., "*Mental models, semantical games, and varieties of intelligence*", in Lucia Vaina (Ed.), *Varieties of intelligence*, D. Reidel, Dordrecht, 1986.

[Hint87] Hintikka, J., "*Knowledge representation and the interrogative model of inquiry*", International Philosophy Congress, pp. 1077-1084, 1987.

[Hint88] Hintikka, J., "*What is the logic of experimental inquiry?*", *Synthese* 74, pp. 173-190, 1988.

[Hint92] Hintikka, J., "*The interrogative model of inquiry as a general theory of argumentation*", *Communication and Cognition*, Vol. 25, Nos. 2-3, pp. 221-242, 1992.

[HMP96] Huang, Z., Masuch, M., Pólos, L., "*ALX, an action logic for agents with bounded rationality*", *Artificial Intelligence* 82, pp. 75-127, 1996.

[Hofs80] Hofstadter, D., "*Gödel, Escher, Bach: an Eternal Golden Braid*", Basic Books Inc. Publishers, 1980.

[HoMo85] Hobbs, J.R., Moore, R.C. (Eds.), *Formal theories of the commonsense world*, Intellect Books, 1985.

[HuCr68] Hughes, G.E., Cresswell, M.J., "*An introduction to modal logic*", Methuen and Co. Ltd. Eds., 1968.

[ICAR96] Ingrand, F., Chatila, R., Alami, R., Robert, F., "*PRS: a high level supervision and control language for autonomous mobile robots*", *Proceedings of IEEE ICRA'96*, Minneapolis, 1996.

[Jasp91] Jaspars, J., "*Fused modal logic and inconsistent belief*", *Proceedings of the First World Conference on the Fundamentals of Artificial Intelligence*, pp. 267-275, 1991.

[Jasp93] Jaspars, J., "*Logical omniscience and inconsistent beliefs*", in *Diamonds and defaults*, M. de Rijke (Ed.), Kluwer Academic Publishers, pp. 129-146, 1993.

[Jasp94] Jaspars, J., "*Calculi for Constructive Communication*", PhD dissertation, Institute for Logic, Language and Computation, ILLC Dissertation Series 1994-4, 1994.

[JoTr99] Jonker, C., Treur, J., "*Formal analysis of models for the dynamics of trust based on experiences*", in [GaBo99], pp. 221-231.

[KaMe91] Katsuno, H., Mendelzon, A., "*On the difference between updating a knowledge base and revising it*", *Proceedings of the Second International Conference on Principles of Knowledge Representation and Reasoning, KRR'91*, pp. 387-394, 1991.

[Kell97] Kelly, J., "*The essence of Logic*", The Essence of Computing Series, Prentice Hall, 1997.

[Klee52] Kleene, S., "*Introduction to metamathematics*", North Holland, 1952.

[Kono85] Konolige, K., "*Belief and incompleteness*", in [HoMo85], pp. 359-403.

[Kono86a] Konolige, K., "*A Deduction Model of Belief*", Morgan Kaufmann, San Mateo, CA, 1986.

[Kono86b] Konolige, K., "*What awareness isn't: a sentential view of implicit and explicit belief*", *Proceedings of the First Conference on Theoretical Aspects of Reasoning about Knowledge, TARK-86*, pp. 241-250, 1986.

[Krip59] Kripke, S., "*A completeness theorem in modal logic*", *Journal of Symbolic Logic* 24, pp. 1-14, 1959.

[Krip63a] Kripke, S., "*Semantical considerations on modal logic*", *Acta Philosophica Fennica* 16, pp. 83-94, 1963.

[Krip63b] Kripke, S., "*A semantical analysis of modal logic I: normal modal propositional calculi*", *Zeitschrift für Mathematische Logik und Grundlagen Mathematik* 9, pp. 67-96, 1963.

[KrSu95] Kraus, S., Subrahmanian, V., "*Multiagent reasoning with probability, time, and beliefs*", *International Journal of Intelligent Systems* 10, pp. 459-499, 1995.

[Lake87] Lakemeyer, G., "*Tractable meta-reasoning in propositional logics of belief*", *Proceedings of the Tenth International Joint Conference on Artifical Intelligence, IJCAI-87*, pp. 402-408, 1987.

[Lake91a] Lakemeyer, G., "*On the relation between explicit and implicit belief*", *Proceedings of the International Conference on Principles of Knowledge Representation and Reasoning, KRR-91*, pp. 368-375, 1991.

[Lake91b] Lakemeyer, G., "*A computationally attractive first-order logic of belief*", in [VEij91], pp. 333-347.

[Lake93] Lakemeyer, G., "*All they know: a study in multi-agent epistemic reasoning*", in *Proceedings of the Thirteenth International Joint Conference on Artificial Intelligence, IJCAI-93*, pp. 376-381, 1993.

[Lake94] Lakemeyer, G., "*Limited reasoning in first order knowledge bases*", *Artificial Intelligence* 71, pp. 213-255, 1994.

[LaLe88] Lakemeyer, G., Levesque, H., "*A tractable knowledge representation service with full introspection*", *Proceedings of the Second Conference on Theoretical Aspects of Reasoning about Knowledge, TARK-88*, pp. 145-159, 1988.

[Lésp96] Lésperance, Y., Levesque, H., Lin, F., Marcu, D., Reiter, R., Scherl, R., *Foundations of a logical approach to agent programming*, in M.Wooldridge, J.Müller, M.Tambe (Eds.), *Intelligent Agents II*, Lecture Notes in Artificial Intelligence 1037, pp. 331-346, Springer Verlag, 1996.

[Leve84] Levesque, H.J., "*A logic of implicit and explicit belief*", *Proceedings of the Conference of the American Association for Artificial Intelligence, AAAI-84*, pp. 198-202, 1984.

[Lin96] Lin, J., "*A semantics for reasoning consistently in the presence of inconsistency*", *Artificial Intelligence* 86, pp. 75-95, 1996.

[Lomu99] Lomuscio, A., "*Knowledge sharing among ideal agents*", PhD dissertation, School of Computer Science, University of Birmingham, 1999.

[Maes89] Maes, P., "*The dynamics of action selection*", *Proceedings of the Eleventh Joint Conference on Artificial Intelligence, IJCAI-89*, pp. 991-997, 1989.

[McAr88] McArthur, G., "*Reasoning about knowledge and belief: a survey*", *Computational Intelligence* 4, pp. 223-243, 1988.

[McCa93] McCarthy, J., "*Notes on formalizing context*", *Proceedings of the Thirteenth International Joint Conference on Artificial Intelligence*, 1993.

[McPa87] McPartlin, M., "*Theories of formal epistemology*", M.Sc. Thesis, Centre for Cognitive Science, University of Edinburgh, 1987.

[MCS98] Moreno, A., Cortés, U., Sales, T., "*Modelling non-ideal inquirers*", *Proceedings of the VI Iberoamerican Conference on Artificial Intelligence, IBERAMIA-98*, Lisbon, Portugal, pp. 135-146, October 1998.

[MCS99a] Moreno, A., Cortés, U., Sales, T., "*Modelling rational inquiry in non-ideal agents*", included in *Collaboration between human and artificial societies*. Ed: J.Padget. Lecture Notes in Artificial Intelligence 1624, pp. 164-186, Springer Verlag.

[MCS99b] Moreno, A., Cortés, U., Sales, T.,"*Subjective situations*", in [GaBo99], pp. 210-220, *Proceedings of the IX European Workshop on Modelling Autonomous Agents in a Multi-Agent World, MAAMAW-99*, Valencia, Spain, June 1999.

[MCS00a] Moreno, A., Cortés, U., Sales, T., "*Inquirers: a general model of non-ideal agents*", *International Journal of Intelligent Systems* 15 (3), pp. 197-215, 2000.

[MCS00b] Moreno, A., Cortés, U., Sales, T., "*Avoiding logical omniscience by using subjective situations*", *Seventh European Workshop on Logics in Artificial Intelligence, JELIA-00*, Málaga, Spain. To be published in a special issue of Lecture Notes on Artificial Intelligence, Springer Verlag, 2000.

[MoHe79] Moore, R., Hendrix, G., "*Computational models of beliefs and the semantics of belief sentences*", Technical Note 187, SRI International, 1979.

[Moli91] Molin, T., "*Tractatus de ratione circumscripta: rationality for finite agents*", Lund University Cognitive Studies 1, 1991.

[Mont63] Montague, R., "*Syntactical treatments of modality, with corolaries on reflexion principles and finite axiomatizability*", *Acta Philosophica Fennica* 16, pp. 153-167, 1963.

[Mont70] Montague, R., "*Universal grammar*", *Theoria* 36, pp. 373-398, 1970.

[Moor83] Moore, R., "*Semantical considerations on nonmonotonic logic*", *Artificial Intelligence Center Technical Note 284*, SRI International, Menlo Park, California, 1983.

[Moor85] Moore, R., "*A formal theory of knowledge and action*", included in [HoMo85], pp. 319-358.

[More96] Moreno, A., "*Limited logical belief analysis*", *Proceedings of the Iberoamerican Conference on Artificial Intelligence, IBERAMIA-96*, Cholula, México, pp. 250-259, 1996.

[More98] Moreno, A., "*Avoiding logical omniscience and perfect reasoning: a survey*", *AI Communications* 11 (2), pp. 101-122, 1998.

[MoSa97a] Moreno, A., Sales, T., "*Dynamic belief analysis*", included in [MWJ97], pp. 87-102, *Proceedings of the Workshop on Agent Theories, Architectures and Languages, ATAL-96*, at the *European Conference on Artificial Intelligence, ECAI-96*, Budapest, 1997.

[MoSa97b] Moreno, A., Sales, T., "*Limited logical belief analysis*", included in [CRW97], pp. 104-118, *Proceedings of the Workshop on Theoretical and Practical Foundations of Intelligent Agents*, at the *Pacific Rim International Conference on Artificial Intelligence, PRICAI-96*, Cairns, Australia, 1997.

[Müll97a] Müller, J., "*Control architectures for autonomous and interacting agents: a survey*", in [CRW97], pp. 1-26, 1997.

[Müll97b] Müller, J., "*A cooperation model for autonomous agents*", in [MWJ97], pp. 245-260.

[MvdH91] Meyer, J.J., Van der Hoek, W., "*Non-monotonic reasoning by monotonic means*", in [VEij91], pp. 399-411.

[MvdH92] Meyer, J.J., Van der Hoek, W., "*A modal logic for non-monotonic reasoning*", in W. van der Hoek, J.J. Meyer, Y. Tan, C. Witteveen (Eds.), *Non-monotonic reasoning and partial semantics*, pp. 37-77, Ellis Horwood, 1992.

[MvdH93] Meyer, J.J., Van der Hoek, W., "*A cumulative default logic based on epistemic states*", in M.Clarke, R.Kruse, S.Moral (Eds.), *Symbolic and quantitative approaches to reasoning and uncertainty* (Proceedings of ECSQARU'93), pp. 265-273, Springer Verlag, 1993.

[MvdH98] Meyer, J.J., Van der Hoek, W., "*Modal logics for representing incoherent knowledge*", in D.Gabbay, P.Smets (Eds.), *Handbook of defeasible reasoning and uncertainty management systems*, Vol. 2, pp. 37-75, Kluwer Academic Publishers, 1998.

[MWJ97] Müller, J., Wooldridge, M., Jennings, N. (Eds.), "*Intelligent Agents III: Agent theories, architectures and languages*", Lecture Notes in Artificial Intelligence 1193, Springer Verlag 1997.

[NeSi72] Newell, A., Simon, H., "*Human Problem Solving*", Prentice Hall, 1972.

[NKP94] Nirkhe, M., Kraus, S., Perlis, D., "*Thinking takes time: a modal active-logic for reasoning in time*", CS-TR-3249, Computer Science Dept., University of Maryland, 1994.

[PaGi98] Parsons, S., Giorgini, P., "*On using degrees of belief in BDI agents*", *International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems*, Paris, July 1998.

[Perl84] Perlis, D., "*Non-monotonicity and real-time reasoning*", *Proceedings of the AAAI Workshop on Non-monotonic Reasoning*, 1984.

[Perl94] Perlis, D., "*Logic for a lifetime*", CS-TR-3278, Computer Science Dept., University of Maryland, 1994.

[Perl97] Perlis, D., "*Sources of, and exploiting, inconsistency: preliminary report*", *Journal of Applied Non-Classical Logics* 7, special double issue on Handling inconsistency in knowledge systems, 1997.

[Poll90] Pollock, J., "*OSCAR: A general theory of rationality*", included in the book [CuPo95], pp. 189-213, 1990.

[Poll95] Pollock, J., "*Cognitive carpentry*", Bradford Books, MIT Press, 1995.

[Poll98] Pollock, J., "*Perceiving and reasoning about a changing world*", *Computational Intelligence*, Vol. 14, No. 4, pp. 498-562, 1998.

[Poll00] Pollock, J., "*Rational cognition in OSCAR*", in *Intelligent Agents VI*, N.Jennings and Y.Lesperance (Eds.), Lecture Notes in Artificial Intelligence 1757, Springer Verlag, 2000.

[Popp34] Popper, K., "*The logic of scientific discovery*", Ed. Tecnos, 1977 printing.

[Prie89] Priest, G., "*Reasoning about truth*", *Artificial Intelligence* 39, pp. 231-244, 1989.

[RaGe92] Rao, A., Georgeff, M., "*An abstract architecture for rational agents*". *Proceedings of the 3rd International Conference on Principles of Knowledge Representation and Reasoning, KRR-92*, pp. 439-449, 1992.

[RaGe95b] Rao, A., Georgeff, M., "*BDI agents: from theory to practice*", *Proceedings of the International Conference on Multi-Agent Systems, ICMAS-95*, 1995.

[Rant75] Rantala, V., "*Urn models: a new kind of non-standard model for first-order logic*", *Journal of Philosophical Logic*, Vol. 4, pp. 455-474, 1975.

[ReBr79] Rescher, N., Brandom, R., "*The Logic of Inconsistency*", Rowman and Littlefield Eds., 1979.

[Reic89] Reichgelt, H., "*Logics for reasoning about knowledge and belief*", *Knowledge Engineering Review*, Vol. 4, No.2, pp. 119-139, 1989.

[RoKa96] Rosenschein, S., Kaelbling, L., "*A situated view of representation and control*", in P.Agre and J.Rosenschein (Eds.), *Computational Theories of Interaction and Agency*, pp. 515-540, MIT Press, 1996.

[Roos92] Roos, N., "*A logic for reasoning with inconsistent knowledge*", *Artificial Intelligence* 57, pp. 69-103, 1992.

[RoRo72] Routley, P., Routley, V., "*Semantics of first degree entailment*", *Noûs* 6, pp. 335-359, 1972.

[RuNo95] Russell, S., Norvig, P., "*Artificial Intelligence. A Modern Approach.*", Prentice Hall Series in Artificial Intelligence, 1995.

[ScCa92] Schaerf, M., Cadoli, M., "*Approximate reasoning and non-omniscient agents*", *Proceedings of the Fourth Conference on Theoretical Aspects of Reasoning about Knowledge, TARK-92*, pp. 159-183, 1992.

[ScCa95] Schaerf, M., Cadoli, M., "*Tractable reasoning via approximation*", *Artificial Intelligence* 74, pp. 249-310, 1995.

[Scot70] Scott. D., "*Advice in modal logic*", in *Philosophical problems in logic*, Ed. K. Lambert, Reidel, 1970.

[Shoh90] Shoham, Y., "*Agent-oriented programming*", Technical Report STAN-CS-90-1335, Department of Computer Science, Stanford University, October 1990.

[Shoh91] Shoham, Y., "*Varieties of context*", in *Artificial Intelligence and Mathematical Theory of Computation*, pp. 393-408, V. Lifschitz Ed., Academic Press, N.Y., 1991.

[Shoh93] Shoham, Y., "*Agent-oriented programming*", *Artificial Intelligence* 60, pp. 51-92, 1993.

[Shoh98] Shoham, Y., "*Agent Oriented Programming*", in *Readings in Agents*, M.Huhns and M.Singh (Eds.), Morgan Kaufmann, 1998.

[Sim95] Sim, K.M. "*A Multi-Valued Epistemic Logic*", PhD Dissertation, Department of Computer Science, University of Calgary, Alberta, Canada, 1995.

[Sim97] Sim, K.M., "*Epistemic logic and logical omniscience: a survey*", *International Journal of Intelligent Systems* 12, pp. 57-81, 1997.

[Sim00] Sim, K.M., "*Epistemic logic and logical omniscience II: a unifying framework*", *International Journal of Intelligent Systems* 15, pp. 129-152, 2000.

[Smul68] Smullyan, R.M., "*First-order logic*", Springer Verlag, 1968.

[SRG99] Singh, M., Rao, A., Georgeff, M., "*Formal methods in Distributed Artificial Intelligence: Logic-Based Representation and Reasoning*", in [Weis99], pp. 331-376, 1999.

[Stal84] Stalnaker, R.C., "*Inquiry*", MIT Press, 1984.

[Thij92] Thijsse, E., "*Partial logic and knowledge representation*", PhD Thesis, Tilburg University, Delft, Eburon Publishers, 1992.

[Thij96] Thijsse, E., "*Combining partial and classical semantics. A hybrid approach to belief and awareness.*", in [Dohe96], pp. 223-249, 1996.

[Thom80] Thomason, R., "*A note on syntactical treatments of modality*", *Synthese* 44, pp. 391-395, 1980.

[Vard86] Vardi, M., "*On epistemic logic and logical omniscience*", *Proceedings of the First Conference on Theoretical Aspects of Reasoning about Knowledge, TARK-86*, Ed. J.Y.Halpern, pp. 293-305, 1986.

[VBen84] Van Benthem, J., "*Correspondence theory*", in *Handbook of Philosophical Logic*, Vol. III, Eds. Gabbay, D., Guenthner, F., pp. 167-284, Kluwer Academic Publishers, 1984.

[VdHM88] Van der Hoek, W., Meyer, J.-J., "*Possible logics for belief*", Rapport IR-170, Vrije Universiteit Amsterdam, 1988.

[VdHM89] Van der Hoek, W., Meyer, J.-J., "*Possible logics for belief*", *Logique et Analyse* 127-128, pp. 177-194, 1989.

[VdHM95] Van der Hoek, W., Meyer, J.-J., "*Epistemic logic for AI and Computer Science*", Cambridge Tracts in Theoretical Computer Science 41, Cambridge University Press, 1995.

[VdHM96] Van der Hoek, W., Meyer, J.-J., "*Modalities for reasoning about knowledge and uncertainties*", in [Dohe96], pp. 77-109, 1996.

[VdHo93] Van der Hoek, W., "*Systems for knowledge and belief*", *Journal of Logic and Computation*, Vol. 3, No. 2, pp. 173-195, 1993.

[VEij91] Van Eijck, J., "*Logics in AI*" (Proceedings of the Second Workshop on Logics in AI, JELIA'90), Lecture Notes in Computer Science 478, Springer Verlag, 1991.

[Velt96] Veltman, F., "*Defaults in update semantics*", *Journal of Philosophical Logic* 25, pp. 221-261, 1996.

[Wans90] Wansing, H., "*A general possible worlds framework for reasoning about knowledge and belief*", *Studia Logica* 49 (4), pp. 523-539, 1990.

[Weis99] Weiss, G. (Ed.), "*Multiagent Systems. A Modern Approach to Distributed Artificial Intelligence*", MIT Press, 1999.

[Wiśn95] Wiśniewski, A., "*The posing of questions. Logical Foundations of Erotetic Inferences*", Synthese Library, Volume 252, Kluwer Academic Publishers, 1995.

[WoJe95] Wooldridge, M., Jennings, N., "*Intelligent agents: theory and practice*", *Knowledge Engineering Review* 10 (2), 1995.

[Wool92] Wooldridge, M., "*The logical modelling of computational multi-agent systems*", PhD Thesis, University of Manchester, 1992.

[Wool95] Wooldridge, M., "*An abstract general model and logic of resource-bounded believers*", in *Representing Mental States and Mechanims, Proceedings of the 1995 AAAI Spring Symposium*, AAAI Press, March 1995.

[Wool99] Wooldridge, M., "*Intelligent Agents*", in [Weis99], pp. 27-77, 1999.

A Soundness and completeness of the propositional tableaux calculus

Let's consider the tableaux calculus shown in figure 26:

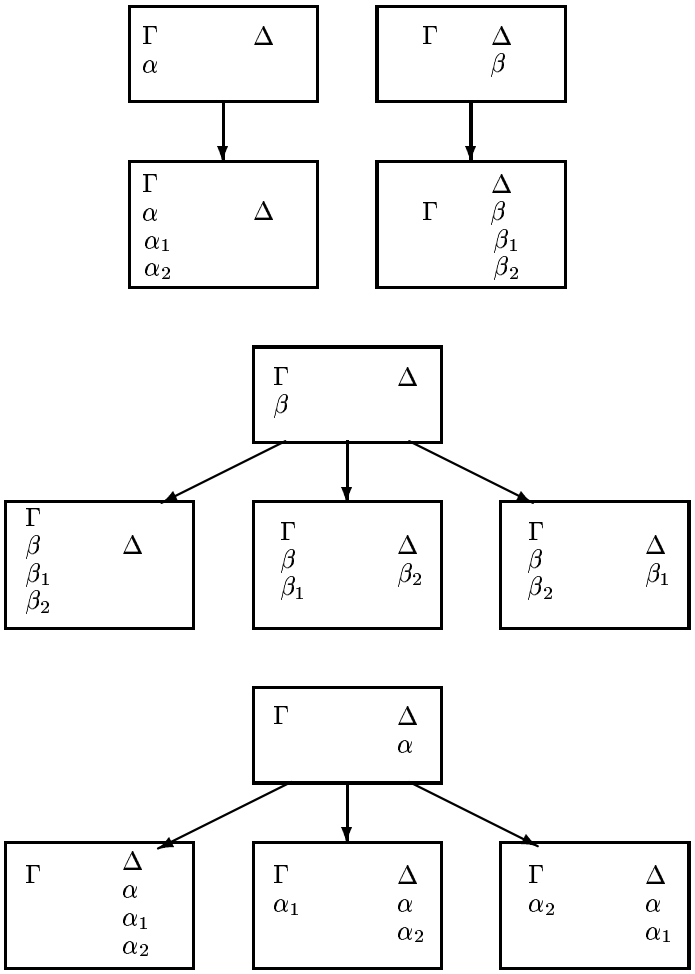

Figure 26: Rules used in the logical analysis

In this figure the symbols α and β represent especial kinds of formulæ as shown in figure 3. These rules deal with the propositional part of the tableaux calculus shown in figure 4, which is the one that may be used in

the logical analysis described in §4.3.1. In that section the conditions under which a tableau may be closed by the agent were also described:

- It contains a formula and its negation in the left column.
- It contains the same formula in both columns.

In this appendix we are going to study the power of this tableaux calculus. $\mathcal{L}$ will denote the standard language of propositional logic. We define the following consequence relationship ($\vdash_{TAB}$) between sets of propositional formulæ (Γ, Δ):

Definition 13 (Consequence relationship using tableaux)

$\forall\ \Gamma, \Delta\ \epsilon\ 2^{\mathcal{L}}$, $\Gamma \vdash_{TAB} \Delta$ *iff there exists a logical analysis of the tableau containing* Γ *in the left column and* Δ *in the right column such that all the branches of the tableaux tree are closed.*

In order to find out the sets of formulæ Γ, Δ for which the above consequence relationship holds, we need to state some preliminary definitions.

Definition 14 (Three-valued valuations)

A three-valued valuation is a total function from $\mathcal{P}$ *(the primitive propositions of* $\mathcal{L}$*) to the set* $\{0, 1, \omega\}$*. The set of all three-valued valuations over* $\mathcal{P}$ *will be called* $\mathcal{I}$*. An extended valuation is a function from the set of formulæ of* $\mathcal{L}$ *to the set* $\{0, 1, \omega\}$*. The set of all extended valuations over* $\mathcal{L}$ *will be called* $\mathcal{I}^{ext}$*. A valuation* $I \epsilon \mathcal{I}$ *may be extended to a valuation* $J \epsilon \mathcal{I}^{ext}$ *by the function* ext*, defined in the following way:* $\mathrm{ext}(I) = J$ *such that*

- $J(\mathtt{P}) = I(\mathtt{P}), \forall \mathtt{P} \epsilon \mathcal{P}$
- $J(\neg\phi) = 0$, *if* $I(\phi) = 1$
- $J(\neg\phi) = 1$, *if* $I(\phi) = 0$
- $J(\neg\phi) = \omega$, *if* $I(\phi) = \omega$

- $J(\phi \vee \psi) = 0$, *if* $I(\phi) = 0$ *and* $I(\psi) = 0$
- $J(\phi \vee \psi) = 1$, *if* $I(\phi) = 1$ *or* $I(\psi) = 1$
- $J(\phi \vee \psi) = \omega$, *otherwise*
- $J(\phi \wedge \psi) = 0$, *if* $I(\phi) = 0$ *or* $I(\psi) = 0$
- $J(\phi \wedge \psi) = 1$, *if* $I(\phi) = 1$ *and* $I(\psi) = 1$
- $J(\phi \wedge \psi) = \omega$, *otherwise*
- $J(\phi \Rightarrow \psi) = 0$, *if* $I(\phi) = 1$ *and* $I(\psi) = 0$
- $J(\phi \Rightarrow \psi) = 1$, *if* $I(\phi) = 0$ *or* $I(\psi) = 1$
- $J(\phi \Rightarrow \psi) = \omega$, *otherwise*

The previous definition of the extension function is equivalent to the use of the following three-valued truth-tables:

ϕ	$\neg\phi$
0	1
1	0
ω	ω

ϕ	ψ	$(\phi \vee \psi)$	$(\phi \wedge \psi)$	$(\phi \Rightarrow \psi)$
0	0	0	0	1
0	ω	ω	0	1
0	1	1	0	1
ω	0	ω	0	ω
ω	ω	ω	ω	ω
ω	1	1	ω	1
1	0	1	0	0
1	ω	1	ω	ω
1	1	1	1	1

In fact, all the above definitions correspond to a well-known logic, Kleene's *strong three-valued logic* ([Klee52]). Now, we define in which conditions a formula is satisfied by a valuation:

Definition 15 (Satisfiability of formulæ in a valuation)

A propositional formula ϕ will be said to be true in a valuation I ($I \models \phi$) iff the valuation of ϕ in the extension of I is 1:

$$\forall I \epsilon \mathcal{I}\ \forall \phi \epsilon \mathcal{L}\ I \models \phi \text{ iff } ext(I)(\phi)=1$$

We also define another consequence relation ($\models_p$) between sets of propositional formulæ ($\Gamma = \{\gamma_1, \gamma_2, \ldots, \gamma_n\}, \Delta = \{\delta_1, \delta_2, \ldots, \delta_m\}$) in the following way:

Definition 16 (Consequence relationship using valuations)

$$\forall\ \Gamma, \Delta\ \epsilon\ 2^{\mathcal{L}}$$
$$\Gamma \models_p \Delta \text{ iff } \forall I \epsilon \mathcal{I}^{ext}\ (I \models (\gamma_1 \wedge \gamma_2 \wedge \ldots \wedge \gamma_n) \longrightarrow I \models (\delta_1 \vee \delta_2 \vee \ldots \vee \delta_m))$$

Now we can state the following proposition, that characterizes the sets of formulæ Γ, Δ for which $\Gamma \vdash_{TAB} \Delta$ holds:

Proposition 17 (Soundness and completeness of the propositional tableaux calculus with respect to Kleene's three-valued logic)

$$\forall\ \Gamma, \Delta \epsilon 2^{\mathcal{L}}\ (\Gamma \vdash_{TAB} \Delta \longleftrightarrow \Gamma \models_p \Delta)$$

This proposition postulates that the consequence relationship induced by our modified analytic tableaux method ($\vdash_{TAB}$) is equivalent to the one defined using Kleene's three-valued logic ($\models_p$). Thus, the calculus is *sound* and *complete* with respect to three-valued logic. In the rest of this appendix we provide our own proof of this proposition[20].

[20] It is not difficult to check that a similar demostration could be made if the analysed formula were not repeated in the subtableaux obtained after the application of a rule of the tableaux calculus.

The more logically oriented readers of this dissertation may have noticed from the very beginning of this appendix that the tableau system shown in figure 26 is, in fact, the cut-free upside-down version of well-known Gentzen systems for Kleene's three valued logic (see *e.g.* [Busc96], [Jasp94], [Thij92] for works related to these proof techniques) and may therefore have skipped this proof. It has been included here because many readers may not be familiar with this kind of topics (and I may confess that, at the beginning of this research, this equivalence was not known to me, and that I actually became aware of it after constructing this proof).

A.1 Soundness

The definition of *soundness* is the following:

Definition 17 (Soundness)

The tableaux calculus shown in figure 26 is sound with respect to the consequence relationship $\models_p$ *iff the following relation holds:*

$$\forall\ \Gamma, \Delta \epsilon 2^{\mathcal{L}_{PC}}\ (\Gamma \vdash_{TAB} \Delta \rightarrow \Gamma \models_p \Delta)$$

In other words, we have to prove that, for any two sets of propositional formulæ Γ and Δ, if the tableau containing Γ in its left column and Δ in its right column may be analysed in such a way that all its branches are closed, then every three-valued interpretation that satisfies all the formulæ in Γ must satisfy at least one formula in Δ.

First we will prove that this condition holds in those tableaux that satisfy at least one of the tableaux closing conditions (*i.e.* tableaux that are leaves in the analytic tableaux tree). Then, we will complete the soundness demostration by showing that, if the condition holds in all the children of a given tableau, it must also hold in the tableau itself. Following these two steps, we will have proved that, if all the branches that appear in the analysis of a given tableau are closed, the condition related to $\models_p$ holds in every tableau of the tree (every three-valued interpretation that satisfies all the formulæ in its left column satisfies at least one formula in its right column). A tableau with the set Γ in its left column and the set Δ in its right column will be called a $\Gamma - \Delta$ *tableau*.

- *Step 1: The condition* $\Gamma \models_p \Delta$ *holds in every* $\Gamma - \Delta$ *closed tableau.*

  If T is a $\Gamma - \Delta$ closed tableau, one of the tableau closing conditions must hold; thus, there are two cases to be considered:

  - Γ *contains a formula and its negation.*

    Notice that, in this case, there cannot exist any three-valued interpretation that makes true all of the formulæ in Γ (because a three-valued interpretation cannot assign the truth value 1 to a formula and its negation). Therefore, $\Gamma \models_p \Delta$ would trivially hold.

  - *A formula* ϕ *is contained both in* Γ *and* Δ.

    If a three-valued interpretation makes true all of the formulæ in Γ, it makes true ϕ (because ϕ is contained in Γ). But ϕ also belongs to Δ; therefore, that interpretation also satisfies at least one of the formulæ of Δ, and $\Gamma \models_p \Delta$ holds.

- *Step 2: If all the children of a tableau satisfy the condition* $\Gamma \models_p \Delta$*, then it also holds in the tableau itself.*

  We must consider all the rules of the tableaux calculus and show that this proposition is true in each of them. Therefore, four cases must be studied:

  - α*-formula on the left column.*

    Assume that the condition holds in a $\{\Gamma, \alpha, \alpha_1, \alpha_2\} - \Delta$ tableau. That means that all three-valued interpretations that satisfy Γ, α, α_1 and α_2 also satisfy at least one member of Δ. The following lemma shows that the condition also holds in a $\{\Gamma, \alpha\} - \Delta$ tableau, *i.e.* its parent in the tableaux tree.

    Lemma 1 *All the interpretations that satisfy* α *also satisfy* α_1 *and* α_2.

    Proof:

    * $\alpha = (\phi \wedge \psi)$, $\alpha_1 = \phi$ and $\alpha_2 = \psi$.
      If a proposition satisfies $(\phi \wedge \psi)$, it also satisfies ϕ and ψ.

* $\alpha = \neg(\phi \vee \psi)$, $\alpha_1 = \neg\phi$ and $\alpha_2 = \neg\psi$.
  If a proposition satisfies $\neg(\phi \vee \psi)$, it assigns the truth value 0 to $(\phi \vee \psi)$. Thus, it also assigns the truth value 0 to ϕ and to ψ. Therefore, it satisfies both $\neg\phi$ and $\neg\psi$.
* $\alpha = \neg(\phi \Rightarrow \psi)$, $\alpha_1 = \phi$ and $\alpha_2 = \neg\psi$.
  If a proposition satisfies $\neg(\phi \Rightarrow \psi)$, it assigns the truth value 0 to $(\phi \Rightarrow \psi)$. Thus, it must assign the truth value 1 to ϕ and the truth value 0 to ψ. Therefore, it satisfies both ϕ and $\neg\psi$.
* $\alpha = \neg\neg\phi$, $\alpha_1 = \phi$ and $\alpha_2 = \phi$.
  If a proposition satisfies $\neg\neg\phi$, it assigns the truth value 0 to $\neg\phi$. Thus, it must assign the truth value 1 to ϕ. □

– *α-formula on the right column.*

In this case we must analyse an splitting rule. We must show that, if these statements are true:

1. The condition holds in a $\Gamma - \{\Delta, \alpha, \alpha_1, \alpha_2\}$ tableau.
2. The condition holds in a $\{\Gamma, \alpha_1\} - \{\Delta, \alpha, \alpha_2\}$ tableau.
3. The condition holds in a $\{\Gamma, \alpha_2\} - \{\Delta, \alpha, \alpha_1\}$ tableau.

then the condition also holds in a $\Gamma - \{\Delta, \alpha\}$ tableau (*i.e.* the parent of those three subtableaux in the tableaux tree).

Assume that a given interpretation I satisfies Γ. For any α formula, we must consider four different situations:

* I satisfies α_1 and α_2.
  As it satisfies α_1, statement 2 says that it satisfies α, α_2 or a member of Δ. As it satisfies α_2, statement 3 says that it satisfies α, α_1 or a member of Δ. Thus, either I satisfies α or a member of Δ (attaining the desired conclusion) or it satisfies both α_1 and α_2. In that case, the following lemma (the converse of the previous one) shows that it must also satisfy α.

  Lemma 2 *All the interpretations that satisfy α_1 and α_2 also satisfy α.*

Proof:

- $\alpha = (\phi \wedge \psi)$, $\alpha_1 = \phi$ and $\alpha_2 = \psi$.
  If a proposition satisfies ϕ and ψ, it also satisfies $(\phi \wedge \psi)$.
- $\alpha = \neg(\phi \vee \psi)$, $\alpha_1 = \neg\phi$ and $\alpha_2 = \neg\psi$.
  If a proposition satisfies $\neg\phi$ and $\neg\psi$, it assigns the truth value 0 to ϕ and to ψ. Thus, it also assigns the truth value 0 to $(\phi \vee \psi)$. Therefore, it satisfies $\neg(\phi \vee \psi)$.
- $\alpha = \neg(\phi \Rightarrow \psi)$, $\alpha_1 = \phi$ and $\alpha_2 = \neg\psi$.
  If a proposition satisfies ϕ and $\neg\psi$, it assigns the truth value 0 to ψ. Thus, it must assign the truth value 0 to $(\phi \Rightarrow \psi)$. Therefore, it satisfies $\neg(\phi \Rightarrow \psi)$.
- $\alpha = \neg\neg\phi$, $\alpha_1 = \phi$ and $\alpha_2 = \phi$.
  If a proposition satisfies ϕ, it assigns the truth value 0 to $\neg\phi$. Thus, it must assign the truth value 1 to $\neg\neg\phi$. □

* I satisfies α_1 but does not satisfy α_2.
  As it satisfies α_1, statement 2 says that it satisfies α, α_2 or a member of Δ. But we know that it does not satisfy α_2. Therefore, it must satisfy α or a member of Δ.

* I satisfies α_2 but does not satisfy α_1.
  As it satisfies α_2, statement 3 says that it satisfies α, α_1 or a member of Δ. But we know that it does not satisfy α_1. Therefore, it must satisfy α or a member of Δ.

* I does not satisfy either α_1 or α_2.
  As it satisfies Γ, statement 1 says that it satisfies α, α_1, α_2 or a member of Δ. But we know that it does not satisfy either α_1 or α_2. Therefore, it must satisfy α or a member of Δ.

– *β-formula on the left column.*

In this case we must analyse another splitting rule. We must show that, if these statements are true:

1. The condition holds in a $\{\Gamma, \beta, \beta_1, \beta_2\} - \Delta$ tableau.
2. The condition holds in a $\{\Gamma, \beta, \beta_1\} - \{\Delta, \beta_2\}$ tableau.
3. The condition holds in a $\{\Gamma, \beta, \beta_2\} - \{\Delta, \beta_1\}$ tableau.

then the condition also holds in a $\{\Gamma, \beta\} - \Delta$ tableau (*i.e.* the parent of those three subtableaux in the tableaux tree).

Assume that a given interpretation I satisfies Γ and β. The following lemma shows that it must also satisfy β_1 or β_2.

Lemma 3 *All the interpretations that satisfy β also satisfy β_1 or β_2.*

Proof:

* $\beta = (\phi \vee \psi)$, $\beta_1 = \phi$ and $\beta_2 = \psi$.
  If a proposition satisfies $(\phi \vee \psi)$, it also satisfies ϕ or ψ.
* $\beta = \neg(\phi \wedge \psi)$, $\beta_1 = \neg\phi$ and $\beta_2 = \neg\psi$.
  If a proposition satisfies $\neg(\phi \wedge \psi)$, it assigns the truth value 0 to $(\phi \wedge \psi)$. Thus, it also assigns the truth value 0 to ϕ or to ψ. Therefore, it satisfies $\neg\phi$ or $\neg\psi$.
* $\beta = (\phi \Rightarrow \psi)$, $\beta_1 = \neg\phi$ and $\beta_2 = \psi$.
  If a proposition satisfies $(\phi \Rightarrow \psi)$, it must assign the truth value 0 to ϕ or the truth value 1 to ψ. Therefore, it satisfies $\neg\phi$ or ψ. □

Thus, I satisfies β_1 or β_2. The following cases may be considered:

* I satisfies β_1 and β_2.
  As it satisfies β_1 and β_2, statement 1 says that it satisfies a member of Δ.
* I satisfies β_1 but does not satisfy β_2.
  As it satisfies β_1, statement 2 says that it satisfies β_2 or a member of Δ. But we know that it does not satisfy β_2. Therefore, it must satisfy a member of Δ.
* I satisfies β_2 but does not satisfy β_1.
  As it satisfies β_2, statement 3 says that it satisfies β_1 or a member of Δ. But we know that it does not satisfy β_1. Therefore, it must satisfy a member of Δ.

– *β-formula on the right column.*

Assume that the condition holds in a $\Gamma - \{\Delta, \beta, \beta_1, \beta_2\}$ tableau. That means that all three-valued interpretations that satisfy Γ also satisfy β, β_1, β_2 or at least one member of Δ. The following

lemma (the converse of the previous one) shows that the condition also holds in a $\Gamma - \{\Delta, \beta\}$ tableau, *i.e.* its parent in the tableaux tree.

Lemma 4 *All the interpretations that satisfy β_1 or β_2 also satisfy β.*

Proof:

* $\beta = (\phi \vee \psi)$, $\beta_1 = \phi$ and $\beta_2 = \psi$.
  If a proposition satisfies ϕ or ψ, it also satisfies $(\phi \vee \psi)$.
* $\beta = \neg(\phi \wedge \psi)$, $\beta_1 = \neg\phi$ and $\beta_2 = \neg\psi$.
  If a proposition satisfies $\neg\phi$ or $\neg\psi$, it assigns the truth value 0 to ϕ or to ψ. Thus, it also assigns the truth value 0 to $(\phi \wedge \psi)$. Therefore, it satisfies $\neg(\phi \wedge \psi)$.
* $\beta = (\phi \Rightarrow \psi)$, $\beta_1 = \neg\phi$ and $\beta_2 = \psi$.
  If a proposition satisfies $\neg\phi$ or ψ, then it assigns the truth value 0 to ϕ or to $\neg\psi$. Thus, it satisfies $(\phi \Rightarrow \psi)$. □

In this way the second step is completed and the soundness proof is over.

A.2 Completeness

The definition of *completeness* is the following:

Definition 18 (Completeness)
The tableaux calculus shown in figure 26 is complete with respect to the consequence relationship $\models_p$ iff the following relation holds:

$$\forall\ \Gamma, \Delta \epsilon 2^{\mathcal{L}_{PC}}\ (\Gamma \vdash_{TAB} \Delta \leftarrow \Gamma \models_p \Delta)$$

That expression is equivalent to the following one:

$$\forall\ \Gamma, \Delta \epsilon 2^{\mathcal{L}_{PC}}\ (\Gamma \nvdash_{TAB} \Delta \rightarrow \Gamma \not\models_p \Delta)$$

Let $\Gamma = \{\gamma_1, \gamma_2, \ldots, \gamma_n\}$ and $\Delta = \{\delta_1, \delta_2, \ldots, \delta_m\}$. Applying the definition of $\models_p$, the following expression is obtained:

$$\forall\ \Gamma, \Delta \epsilon 2^{\mathcal{L}_{PC}}$$
$$\Gamma \nvdash_{TAB} \Delta \rightarrow \exists I \epsilon \mathcal{I}^{ext} \text{ such that } I(\gamma_1 \wedge \gamma_2 \wedge \ldots \wedge \gamma_n) = 1 \text{ and}$$
$$I(\delta_1 \vee \delta_2 \vee \ldots \vee \delta_m) \neq 1.$$

We provide a constructive demostration of this proposition. Given a $\Gamma - \Delta$ tableau that cannot be closed in an exhaustive tableaux-based analysis, we build a three-valued interpretation I that satisfies all the formulæ in Γ and does not satisfy any formula in Δ.

The demostration has two steps:

- *Step 1: Selection of a leaf of the tableaux tree and construction of the desired interpretation.*

  If $\Gamma \nvdash_{TAB} \Delta$ then there is at least one open tableau in an exhaustive logical analysis of the $\Gamma - \Delta$ tableau. Let's call this (open) leaf of the tableau tree a $\Gamma^* - \Delta^*$ tableau (where $\Gamma^* = \{\gamma_1^*, \gamma_2^*, \ldots, \gamma_n^*\}$ and $\Delta^* = \{\delta_1^*, \delta_2^*, \ldots, \delta_m^*\}$). We want to build an interpretation that satisfies all the formulæ in Γ^* and does not satisfy any formula in Δ^*.

  Let's consider the following definitions:

  - Γ_+^* contains all the atoms that are affirmed in Γ^*.
  - Γ_-^* contains all the atoms that are negated in Γ^*.
  - Δ_+^* contains all the atoms that are affirmed in Δ^*.
  - Δ_-^* contains all the atoms that are negated in Δ^*.

  For instance, if $\Gamma^* = \{P, \neg Q\}$ and $\Delta^* = \{R, \neg S\}$, then $\Gamma_+^* = \{P\}$, $\Gamma_-^* = \{Q\}$, $\Delta_+^* = \{R\}$ and $\Delta_-^* = \{S\}$.

  Recall that the $\Gamma^* - \Delta^*$ tableau is open. Therefore, none of the tableau closing conditions is applicable, and the following relations hold:

  - $\Gamma_+^* \bigcap \Gamma_-^* = \emptyset$.
  - $\Gamma_+^* \bigcap \Delta_+^* = \emptyset$.
  - $\Gamma_-^* \bigcap \Delta_-^* = \emptyset$.

  We consider any three-valued interpretation I that satisfies the following requirements:

- $\forall\ p \epsilon \Gamma^*_+\ I(p) = 1$.
- $\forall\ p \epsilon \Gamma^*_-\ I(p) = 0$.
- $\forall\ p \epsilon \Delta^*_+\ I(p) = 0$ or ω.
- $\forall\ p \epsilon \Delta^*_-\ I(p) = 1$ or ω.

It is indeed possible to build at least one interpretation with those properties, because of the relationships that have been stated above. The precise rules that should be followed to construct an interpretation that meets those requirements are the following:

- If $p \epsilon \Gamma^*_+$ then $I(p) = 1$.
- If $p \epsilon \Gamma^*_-$ then $I(p) = 0$.
- If $p \epsilon \Delta^*_+$ and $p \not\epsilon \Gamma^*_-$ and $p \not\epsilon \Delta^*_-$ then $I(p) = 0$ or ω.
- If $p \epsilon \Delta^*_-$ and $p \not\epsilon \Gamma^*_+$ and $p \not\epsilon \Delta^*_+$ then $I(p) = 1$ or ω.
- If $p \epsilon \Delta^*_-$ and $p \epsilon \Delta^*_+$ then $I(p) = \omega$.

It can be easily checked that any interpretation I that fulfills these requirements has the following property:

$$I^{ext}(\gamma^*_1 \wedge \gamma^*_2 \wedge \ldots \wedge \gamma^*_n) = 1 \text{ and } I^{ext}(\delta^*_1 \vee \delta^*_2 \vee \ldots \vee \delta^*_m) \neq 1.$$

In the example shown above, we might build the following interpretation I:

- $P \epsilon \Gamma^*_+$, therefore $I(P) = 1$.
- $Q \epsilon \Gamma^*_-$, therefore $I(Q) = 0$.
- $R \epsilon \Delta^*_+$ and $R \not\epsilon \Gamma^*_-$ and $R \not\epsilon \Delta^*_-$, therefore $I(R) = 0$ or ω.
- $S \epsilon \Delta^*_-$ and $S \not\epsilon \Gamma^*_+$ and $S \not\epsilon \Delta^*_+$ then $I(S) = 1$ or ω.

Thus, $I(P \wedge \neg Q) = 1$ and $I(R \vee \neg S) = 0$ or ω, as intended.

The property stated above is trivially true for literals, and can be also shown to be true for more complex formulæ. In order to do that, we will state two preliminary definitions:

Definition 19 (Hintikka set) [Hint62]

A Hintikka set is a set S that satisfies the following requirements:

- *It does not contain a formula and its negation.*
- *If $\neg\neg\phi\epsilon S$, then $\phi\epsilon S$.*
- *If $\alpha\epsilon S$, then $\alpha_1\epsilon S$ and $\alpha_2\epsilon S$.*
- *if $\beta\epsilon S$, then $\beta_1\epsilon S$ and/or $\beta_2\epsilon S$.*

Definition 20 (Hintikka-inverse set)

We will call Hintikka-inverse sets to those sets S that satisfy these requirements:

- *If $\neg\neg\phi\epsilon S$, then $\phi\epsilon S$.*
- *If $\alpha\epsilon S$, then $\alpha_1\epsilon S$ and/or $\alpha_2\epsilon S$.*
- *if $\beta\epsilon S$, then $\beta_1\epsilon S$ and $\beta_2\epsilon S$.*

It can be easily seen that this lemma holds:

Lemma 5 (Characterization of leaves of tableaux analysis)

- *After an exhaustive logical analysis, all the formulæ in the left column of any open tableau form a Hintikka set.*
- *After an exhaustive logical analysis, all the formulæ in the right column of any open tableau form a Hintikka-inverse set.*

The proof of the previous lemma is very straightforward from the rules of the tableaux analysis and the definition of Hintikka and Hintikka-inverse sets. It is also immediate to check that the following lemma also holds:

Lemma 6 (Satisfiability in Hintikka and Hintikka-inverse sets)

- *If all the literals of a Hintikka set are satisfied by a given interpretation, all the formulæ of the set are satisfied by that interpretation as well.*

- *If none of the literals of a Hintikka-inverse set is satisfied by a given interpretation, none of the formulæ of the set may be satisfied by that interpretation.*

These lemmas complete the proof of the first step.

- *Step 2: If the interpretation of a tableau satisfies all the formulæ of its left column and does not satisfy any formula in its right column, then this property is also true in its parent in the tableaux tree.*

  Let's call $\Gamma^* - \Delta^*$ to the tableau that satisfies the property mentioned above, and $\Gamma - \Delta$ its parent in the tableaux tree. We have to show that the latter tableau also satisfies the property.

  If the $\Gamma^* - \Delta^*$ satisfies the property, that means that there exists an interpretation I that satisfies all the formulæ in Γ^* and does not satisfy any formula in Δ^*. It can be noticed that the rules of the tableaux calculus are written in such a way that $\Gamma \subseteq \Gamma^*$ and $\Delta \subseteq \Delta^*$[21]. Thus, if I satisfies all the formulæ in Γ^* it also satisfies all the formulæ in Γ, and if it does not satisfy any formula in Δ^*, it cannot satisfy any formula in Δ.

This argument closes step 2 and the completeness proof.

[21] If the rules of the tableaux calculus were modified (and the analysed formulæ were not repeated in the subtableaux) then this property would not hold, but the completeness demostration could still be easily made. For instance, in the rule of the double negation in the left column, if $I(\phi) = 1$ then $I(\neg\neg\phi) = 1$.