Dynamic belief analysis

Antonio Moreno
Departament de Llenguatges i Sistemes Informàtics
Universitat Politècnica de Catalunya
Pau Gargallo, 5
08028-Barcelona
amoreno@lsi.upc.es, toni@cogsci.ed.ac.uk

Abstract

The process of *rational inquiry* can be defined as the evolution of the beliefs of a rational agent as a consequence of its internal inference procedures or its interaction with the environment. These beliefs can be modelled in a formal way using *doxastic logics*. The *possible worlds model* and its associated *Kripkean semantics* provide an intuitive semantics for these logics, but they seem to model agents which are *logically omniscient* and *perfect reasoners*. These problems can be avoided with a syntactic view of possible worlds, defining them as arbitrary sets of sentences in a propositional doxastic logic. This approach does not account for any kind of evolution in the set of beliefs, though. In this work this syntactic view of possible worlds is taken, and a three-dimensional dynamic analysis of the beliefs of the agent is suggested in order to model the process of *rational inquiry* in which the agent is permanently engaged. The agent can make a *logical* analysis of the beliefs (using a modified version of the *analytic tableaux* method); the results of this analysis guide the *experimental* analysis, that allows the agent to perform certain tests in the environment, that corroborate or refute the results of the logical analysis. In a third (*axiomatic*) kind of analysis, the agent can transform its set of beliefs into an axiomatic system (a set of axioms and a set of inference rules).

Keywords: rational inquiry, logics of knowledge and belief, logical omniscience, perfect reasoning, possible worlds, Kripke semantics, dynamic accessibility relation, analytic tableaux

Most of the research described in this report was developed during the stay of the author at HCRC, University of Edinburgh, from March to August 1995. Partially funded by project VIM ERBCHRXCT930401.

1 Introduction

Some computer scientists argue that the aim of Artificial Intelligence (AI) is the construction of *autonomous rational agents*. The architecture of such agents (typically) includes a knowledge base that stores relevant facts about the agent or its[1] environment, a module capable of making (deductive, inductive, abductive, ...) inferences from those facts, modules that interact with the agent's environment, modules that create and evaluate different plans, *etc.* (see [WOOL95] for an extensive review of different approaches, including *reactive* architectures, that do not adhere to this classical picture). If the agent is to display a rational behaviour, it probably needs to have a module capable of handling its beliefs (see *e.g.* the Belief-Desire-Intention architectures considered in [BRAT88], [RAOG91]). These beliefs can be used by the agent in different tasks, *e.g.* to choose between alternative courses of action. Another subset of the AI scientific community could use this beliefs module to test different psychological theories as to how the human system of beliefs is built, maintained and used in rational behaviour.

This concern in AI is closely linked to a more general concern in philosophy, namely how to model the process of *rational inquiry* (defined by Brandom in [RESC79] as the [rationally controlled] transformation of beliefs over time). This will also be the main aim of this work. The classical tradition has considered two components in this process: a *rational* one, that consists in the application of some inference procedures to the actual beliefs (resulting in the addition of inferred beliefs or the discovery of some incompatibility in them), and an *empirical* one, which adds or removes beliefs according to the result of observations performed in the agent's environment. These two components will find its counterparts in the analysis suggested later, namely in the *logical* and *experimental* analysis of the beliefs.

In the field of mathematical logic this concern led to the development of the so-called *logics of knowledge and belief* (epistemic and doxastic logics, see *e.g.* [HALP92]). These logics are built on the language of standard propositional logic, adding the modal operators of necessity and possibility. The first operator is considered as belief/knowledge in doxastic/epistemic logics. The formulas of these languages are given a semantic interpretation in the *possible worlds model* ([HINT62]), with its associated *Kripke semantics* ([KRIP63]). Basically, this semantics postulates (in the one agent case) that there is a fixed *set of possible worlds* W (which are considered as complete and consistent descriptions of the ways the real world could be) and a fixed *binary accessibility relation* between possible worlds, R (a subset of $W * W$), that connects each world w with all the worlds that could be the real one according to the agent's beliefs in w. In the traditional Kripke semantics, the agent is said to believe a proposition p in a world w if p is true in all the worlds R-accessible from w (*i.e.* those worlds w' *s.t.* $(w\ R\ w')$).

These logics are useful in some circumstances (*e.g.* modelling states of knowledge in communication protocols in distributed environments, or solving classical puzzles (the wise men, the muddy children, the cheating wives), as described in [FAGI95]), but they are not appropriate for the purpose of this work (modelling rational inquiry) for two main reasons:

- The set of possible worlds W and the accessibility relation R are fixed, so the set of beliefs is also fixed. That means that the classical[2] approaches try to model the beliefs of the agent

[1] In this report the pronoun *it* is used to refer to the agent.

[2] Some researchers have considered time *explicitly* in epistemic logics, adding new modal operators to deal with it, in order to reason about the change of an agent's knowledge over time in distributed systems

in a particular point in time, but not its rational evolution over time as a consequence of the agent's process of inquiry.

- Even if you want to model the agent's beliefs in a certain time, there are some consequences of the commitment to the possible worlds model and the Kripke semantics which make this semantics an unsuitable one to use. The problem is the assumption that possible worlds are *complete* and *consistent* descriptions of the different ways the actual world could be. This assumption implies that all classical tautologies are true in all possible worlds, and that the set of formulas which are true in a world is closed under deductive closure. These properties are inherited by the beliefs modelled with the Kripke semantics, and thus the modelled agents are *logically omniscient* and *perfect reasoners*. These properties could be appropriate if one wished to model fairly ideal agents (*e.g.* the children in the muddy children puzzle, see [FAGI95]), but they are clearly inadequate if you consider more realistic agents (be they human or computational) which have resource limitations that do not allow them to even approach such incredible states of knowledge.

Summarizing, the aim of this work is to develop a way *to model the process of rational inquiry* (the evolution of the beliefs of a rational agent over time as a consequence of its interaction with the world and its internal inferential processes), *keeping 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*.

2 Main theoretical ideas

2.1 Logical omniscience and perfect reasoning

Let's consider in the first place the problems of logical omniscience and perfect reasoning. Several proposals that have tried to solve (or, at least, partially alleviate) them have been put forward in the AI and the philosophical literature (*e.g.* [McAR88], [REIC89] or [FAGI95] contain detailed reviews of the most relevant approaches). Hintikka made an interesting suggestion in [HINT75], where he proposed the idea of considering *[logically] impossible [epistemically] possible worlds*, but later he seems not to have pursued this idea. This is the path followed in this work, as will be apparent in the rest of this report.

The main roots of these problems, as has been stated above, are the assumptions of *completeness* and *consistency* that underlie the possible worlds model. As McArthur notes in [McAR88], since worlds are complete the agent is forced to have beliefs about the way that *everything* would be in all of the accessible worlds, and since worlds as consistent as well, everything that follows from the agent's beliefs must also be believed. 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, let's face this possibility anyway (some philosophers have indeed argued for the feasibility of this kind of worlds, *e.g.* Rescher and Brandom in [RESC79], and it is

(see [FAGI95]). Others have taken into account in their logical approaches the time that the agent must spend in the course of its reasoning (*step logic*, see [ELGO88], [NIRK94]).

also implicit in the development of the so-called *paraconsistent logics* (Belnap [BELN77], Fitting [FITT90])). It is obvious that *e.g.* humans are both incomplete and inconsistent. Humans often have to say "I don't know" when asked if they believe a certain proposition or its negation, showing their incompleteness (they do not believe or disbelief every imaginable proposition). When asked the same question in different circumstances, they can frantically support an idea the first time and disagree wholeheartedly with it the second time (*e.g.* this fact is intelligently exploited by marketing experts, who can readily construct a questionnaire which will obtain a good support for an idea and another one that will show the abhorrence of the same idea). In summary, it seems easily acceptable that mental states are indeed both partial and inconsistent. Notice that inconsistency can be interpreted as being able to agree, in a standard conversation, both with p and with $\neg p$ (which is quite different to agree with $(p \wedge \neg p)$).

The partiality or incompleteness of possible worlds has been traditionally accepted in the AI literature, the most common justifications being the following:

- The agent can be unaware of certain facts (Fagin and Halpern tried to take this fact into account in their *logic of general awareness*, [FAGI85]).
- The agent can have limited resources (*e.g.* the time required or the space needed to perform a given inference). This is the most evident justification, if rational agents have to be implemented at all in a real computer.
- 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 by Konolige in his Ph.D. thesis ([KONO86]), where each agent was modelled with a base set of beliefs and a (possibly incomplete) set of inference rules.

Inconsistency is a totally different matter. Anyway, it can be argued in its favour with a number of ideas (adding to the questionnaires example suggested above):

- The agent can be unable to take all its beliefs into account in every inference; if it focuses in a subset of them (call that a *context*), it can draw conclusions which are consistent within the context but inconsistent if all the beliefs are considered (as remarked by Shoham in [SHOH91]; Fagin and Halpern modelled this situation in their *logic of local reasoning*, [FAGI85]; Delgrande also considered this idea in [DELG95]).
- There is no theoretical problem in defining interesting procedures of inquiry over inconsistent sets of beliefs (as shown *e.g.* in [RESC79]).
- These kinds of world are perfectly conceivable (and even depictable in pictures, as Escher (see *e.g.* [HOFS80]) proved so many times).
- Some theories have been considered acceptable for a certain period of time (*e.g.* Frege's set theory) for many people, but they have been proved to be inconsistent later (as Russell did with Frege's theory). This inconsistency does not mean that the theory has been useless and that all the work made on it has to been thrown away (*e.g.* many interesting insights can be gained with the inconsistency proof).

- 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.

- A modified version of the analytic tableaux method is going to be used to perform the logical analysis of the beliefs of the agent, as will be explained later, and the conception of possible worlds as inconsistent sets induces a nice relationship between a tableau and the set of possible worlds that it represents (note that an open tableau can contain an inconsistent set of formulas, if the logical development of the tableau has not (yet) shown the contradiction that may be hidden in the set, waiting for further analysis to appear).

Therefore, *the first proposal in this work is to avoid the classical logicians' reluctance towards (possibly) incomplete and (possibly) inconsistent possible worlds and to consider them as (epistemologically and even ontologically) possible as the good old 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 before our very eyes. The agent can clearly fail to believe some tautologies, and it does not have to believe any logical consequence of its beliefs. These facts will be easier to grasp later, when the practical implementation of these ideas is shown.

2.2 Rational inquiry

Recall that the goal of this research is the definition of a general process of rational inquiry, so something must be added to the classical possible worlds model (assuming that one indeed intends to keep the general conception of the model and the Kripke semantics). A very natural idea to model dynamic beliefs is to have some kind of *variability* in the main ingredients of the possible worlds model: the (assumed *fixed*) set of possible worlds W and the (also assumed *fixed*) accessibility relation R.

Imagine that you want to study the beliefs of the agent in world w. 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. If the set of worlds could be variable and not fixed, the following situation could arise: the agent could, e.g. as a consequence of an observation performed in the real world, notice that p holds indeed. So, this observation could be included in all the accessible worlds (w_1, w_2 and w_3), producing a modification of w_1 (which, in its turn, would cause a modification in the agent's beliefs in w, which now would include p). This simple example shows that a variability in the contents of the possible worlds could account for experiences performed in reality, and could model the changes of beliefs produced by such observations.

Imagine that in the previous situation (p true in w_2 and w_3, not in w_1, so p is not believed in w), a certain proposition 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 don't 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 alternative worlds 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. This example shows how a modification of the accessibility relation (*e.g.* a restriction in the set of possible doxastic alternatives) can indeed lead to a modification of the beliefs as a consequence of an internal inferential process of the agent (other sources could have been considered; *e.g.* the agent could have observed the impossibility of q as a result of an observation in its environment). 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 performance of the action (following ideas from Moore, [MOOR83], [MOOR85]).

It could be argued that the evolution of the agent's beliefs can be modelled with just one kind of variability (be it in W or in R). Some justifications to keep both of them could be the following:

- If the agent observes a *new* fact in its environment, that had not been previously noticed, it can only be integrated in its set of beliefs if it is added in the set of accessible worlds (*i.e.*, a modification of W is required). This fact could not be taken into account by restricting the set of accessible worlds (which is the only way to modify R and acquire more knowledge).
- When R is somehow restricted, a lot of new beliefs can be added to the current set of beliefs, *via* the Kripke semantics, as explained above. Note that it would be quite difficult to achieve the same result with a modification of W (adding these beliefs to all the accessible worlds), because *many* beliefs could be involved in this process.

This is the second proposal: the process of *rational inquiry* (including the consideration of observations made by the agent and the results of its own inferential processes) *can be modelled within the possible worlds model, keeping the Kripke semantics, but allowing dynamic changes in the set of possible worlds W and in the accessibility relation R.*

2.3 Summary of the main ideas

The main goal of this work is to model the process of rational inquiry, *i.e.* the evolution of the beliefs of a rational agent over time as a consequence of its interaction with the environment or as a result of its own internal inferential processes. The main framework to be used is the possible worlds model and its associated Kripke semantics, but it will be modified in two important ways:

- a possible world will not necessarily be either complete or consistent; thus, the traditional problems of logical omniscience and perfect reasoning will be avoided.
- the set of possible worlds and the accessibility relation will not be fixed but variable; these changes in W and R will reflect the agents' process of inquiry, and they will produce, *via* the Kripke semantics, the gradual evolution of the agent's beliefs over time.

3 Dynamic belief analysis

3.1 Possible worlds

The language used to model the agent's beliefs is a subset of the language used in standard doxastic propositional logics. A *world* is *defined* as a set of formulas in this language. There are no conditions imposed on this set, so it can be both partial (most facts about the actual world will probably not be contained in each possible world) and inconsistent (although perhaps the inconsistency is not apparent in the set, it may be hidden in its deductive closure). This characterisation of the possible worlds implies that a world can fail to contain some (even all) tautologies, and that the set of formulas that defines a world does not have to be necessarily deductively closed. As mentioned above, this is a way to avoid the problems of logical omniscience and perfect reasoning.

In fact, many people have proposed to consider *impossible* possible worlds, in the sense of having possible worlds where the usual logical connectives do not behave in the usual way, or tautologies may not be true, or inconsistent formulas may hold (see *e.g. impossible possible worlds* in [HINT75], *non-classical worlds* in [CRES72], *non-standard worlds* in [RESC79] or *situations* in [LEVE84]). The drawback of these approaches is that, although they alleviate the problems of logical omniscience and perfect reasoning, they cause different problems. Basically, the main inconvenient of the syntactic approaches (such as the one followed in this report) is that they are very limiting, in the sense that the agent hasn't got a minimal set of inference capabilities (it can believe p and $(p \Rightarrow q)$ and not believe q, or it can believe $(p \wedge q)$ and not believe either p or q). In summary, there seems to be no way to make a knowledge-based analysis of the agent's beliefs (and also recall that the final aim is to model the evolution of beliefs over time, not a static set of beliefs in a certain point in time).

The rest of this report tries to give a partial answer to the following questions:

- How can syntactically-generated beliefs be analysed?
- How do they evolve in time (as a consequence of this analysis)?

3.2 Three-dimensional analysis

In this document a three-dimensional analysis of the agent's beliefs is proposed. These dimensions are:

- a *logical* dimension, which will model the inferential (strictly deductive) processes carried out by the agent
- an *experimental* dimension, which will model the changes of beliefs caused by the interaction of the agent with its environment
- an *axiomatic* dimension, which will generate a compact view of the agent's beliefs, transforming them into an axiomatic system

Section 3.3 describes the main ideas underlying the logical analysis, while section 4 gives a general perspective of the whole belief modeling system, including a brief description of the experimental and axiomatic dimensions. The final section summarizes the report and includes some of the questions that will have to be addressed in the course of this research.

3.3 Logical dimension

3.3.1 Possible worlds and accessibility

The beliefs of the agent will be expressed in a language $\mathcal{L}$, which is a subset of the language traditionally used in propositional modal logics of belief. This language is composed of a finite fixed set of basic propositions ($\mathcal{P}$), two logical operators[3] ($\neg$, $\vee$) and the modal operator B[4]. $\mathcal{L}$ is composed of the formulas that can be generated by the following grammar (where p is any basic proposition):

$$\mathrm{F} \Longrightarrow \mathrm{N} \mid \neg\mathrm{N}$$

$$\mathrm{N} \Longrightarrow B\mathrm{N} \mid \mathrm{G}$$

$$\mathrm{G} \Longrightarrow \neg\mathrm{G} \mid (\mathrm{G} \vee \mathrm{G}) \mid p$$

With this grammar, the language is restricted to standard propositional formulas prefixed by a (possibly negated and possibly empty) sequence of modal operators[5] (*e.g.* $\neg(p \vee (q \vee r))$, Bq, $\neg BB(p \vee (q \vee r))$).

In order to model agents which are neither logically omniscient nor perfect reasoners, you have to notice that the origin of these problems, as mentioned above, resides in the fact that possible worlds have always been assumed to be *complete* and *consistent*, describable with *models*, in the logical sense of the word (in a propositional setting, that means that a world is an interpretation, an assignment of a truth value to each basic proposition). If these conditions hold both problems are difficult to avoid, because tautologies are true in every model (which implies logical omniscience) and the set of propositions which are true in a model is closed by *Modus Ponens* and the ordinary deductive laws (which implies perfect reasoning).

A *world* is *defined* as a set of formulas of $\mathcal{L}$. Furthermore, no restriction is imposed on this set (it does not have to be consistent, it may be not deductively closed, it may fail to contain some tautologies, *etc.*). Intuitively, a world can be envisioned as *any situation that the agent may consider as real.* It can be a situation that has happened to the agent, or maybe something it has imagined. With this approach to possible worlds maybe it would be more appropriate to call them *conceivable* or *imaginable worlds*[6], bearing always in mind that they can be logically inconsistent or physically unrealizable. With this vision of worlds and the analysis of beliefs suggested in this report, the problems of logical omniscience and perfect reasoning vanish.

The set of formulas that defines a world w can be used to define three sets:

- *Belief formulas* (w): formulas that are prefixed by the modal operator in w.

[3] Conjunction and implication can be defined from negation and disjunction in the usual way.

[4] For the moment only one agent and propositional logic are considered; in the future this work will be extended to include a modal operator B_i for each agent i and predicate logic.

[5] This syntactic restriction will also be eliminated in the future.

[6] It would even be better to call them *conceivable situations*, in the spirit of Barwise and Perry ([BARW83]), because they are not required to describe every aspect of the world, but maybe just a small portion of it. It could be misleading, though, due to the inherent consistency of situations in that framework (*incoherent abstract situations* being the exception, as defined by Devlin in [DEVL91]).

$$BF(w)=\{\phi \mid B\phi \; \epsilon \; w\}$$

- *Negated belief formulas* (w): formulas that are prefixed by the negation and the modal operator in w.

$$NBF(w)=\{\phi \mid \neg B\phi \; \epsilon \; w\}$$

- *Propositional formulas* (w): standard propositional formulas of w.

$$PF(w) = w \text{ - } \{B\phi \mid \phi \; \epsilon \; BF(w)\} \text{ - } \{\neg B\phi \mid \phi \; \epsilon \; NBF(w)\}$$

Note that, for any world w, $BF(w)$, $NBF(w)$ and $PF(w)$ are unrelated (because the set of formulas that defines w is unrestricted). Furthermore, if $BF(w)$ and $PF(w)$ were equal that would be an unstable situation, as suggested by Smullyan in [SMUL86][7].

The standard Kripke semantics ([KRIP63]) for modal logics of knowledge and belief is used throughout the logical analysis: a formula ϕ is believed by the agent in world w if and only if ϕ appears in all the doxastic alternatives to w (all the worlds related to w by the accessibility relation R). In the classical approaches the accessibility relation between worlds is predetermined, so the set of beliefs is constant. This fact implies that these approaches are unsuitable to model changing beliefs. In this new framework the aim is to model the evolution of the agent's beliefs over time. This evolution can be due to internal logical analysis of the agent or to incorporation of data derived from tests performed in the environment in which the agent is located.

In the logical analysis the set of beliefs of the agent in a world changes over time, because the accessibility relation between worlds changes. The agent *generates* a sequence of accessibility relations; each of them (*via* the standard Kripke semantics) defines a different set of beliefs. In section 3.3.2 a semantic way of generating this sequence of accessibility relations is developed, using a modified version of the analytic tableaux method.

3.3.2 Generation of the sequence of accessibility relations

The evolution of the beliefs of the agent in a world will be described with a small example[8]. Assume that you want to study the beliefs of the agent in a world w_e. World w_e, as defined above, is just a set of formulas of $\mathcal{L}$ e.g.

$$w_e \equiv \{ \texttt{P}, \neg\texttt{Q}, (\texttt{R} \vee \texttt{S}), B\texttt{R}, B(\neg\texttt{R} \vee (\texttt{S} \vee \texttt{P})) \}$$

This set can be used to define $BF(w_e)$, $NBF(w_e)$ and $PF(w_e)$ as described previously:

$$BF(w_e) = \{ \texttt{R}, (\neg\texttt{R} \vee (\texttt{S} \vee \texttt{P})) \}$$
$$NBF(w_e) = \emptyset$$
$$PF(w_e) = \{ \texttt{P}, \neg\texttt{Q}, (\texttt{R} \vee \texttt{S}) \}$$

[7] *E.g.*, Smullyan shows how a perfect reasoner cannot believe that it is accurate (that it never believes a false proposition) without becoming inaccurate.

[8] Nested beliefs are not considered in this example; they would be treated in a way similar to normal beliefs.

Using the set $BF(w_e)$ the agent can generate the initial accessibility relation, R_0, by applying the standard Kripke semantics:

$$\forall w' \epsilon\, W,\ (w_e\ R_0\ w') \Leftrightarrow \{\ \mathtt{R},\ (\neg\mathtt{R} \vee (\mathtt{S} \vee \mathtt{P}))\ \} \subseteq w'$$

World w' is R_0-accessible from world w_e if and only if every proposition which is believed by the agent in w_e appears in w'. Moreover, the beliefs of the agent in a world w are restricted in the following way: they have to be exactly all the propositions which appear in all its doxastic alternatives, *i.e.* there can be no proposition common to all the accessible worlds from w which is not a belief in w. The ontology of possible worlds just defined is represented[9] in figure 1.

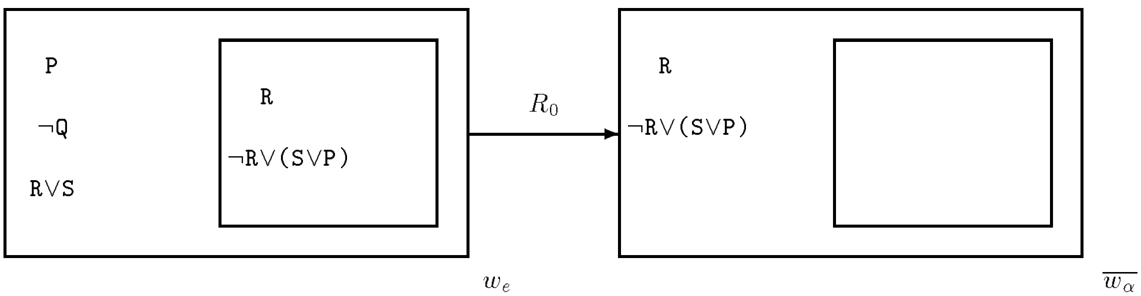

Figure 1: World w_e and its doxastic alternatives

World w_e is R_0-related to all the worlds in class $\overline{w_\alpha}$. $\overline{w_\alpha} = \{\ w_{\alpha_1}, w_{\alpha_2}, \ldots, w_{\alpha_n}\ \}$, where $w_{\alpha_i} = \{\ \mathtt{R},\ (\neg\mathtt{R} \vee (\mathtt{S} \vee \mathtt{P}))\ \} \cup \Delta_{\alpha_i}$. According to the above definitions, every proposition believed in world w_e appears in all the worlds w_{α_1}, w_{α_2}, ..., w_{α_n}; furthermore, there is no other proposition in the intersection of all these worlds, so $\bigcap \Delta_{\alpha_i} = \emptyset$.

The logical analysis of the beliefs of the agent is performed using (a modified version of) the method of the analytic tableaux (see, e.g. [SMUL68]). A tableau is divided into a left and a right column. The analysis starts with a tableau T_0 that contains in its left column the initial set of beliefs of the agent in w_e:

$$T_{0_{left}} \equiv BF(w_e) = \{\ \mathtt{R},\ (\neg\mathtt{R} \vee (\mathtt{S} \vee \mathtt{P}))\ \}$$

In the light of previous considerations, it can be argued that T_0 not only represents $BF(w_e)$; it can also be seen as a (very probably partial) description of all the worlds initially connected to w_e, because all the beliefs in w_e are true in all of these worlds (T_0 represents $\overline{w_\alpha} = \{w \mid (w_e\ R_0\ w)\}$). Thus, a tableau represents a class of worlds: all the worlds in which the propositions

[9] Each world w is represented by a rectangle that contains four columns: from left to right, they represent $PF(w)$, the formulas that cannot appear in w, $BF(w)$ and $NBF(w)$ (these last two columns are included in an inner rectangle, to reinforce its interpretation as beliefs in the world). In figure 1, the second and the fourth columns in w_e are empty. Classes of worlds ($\overline{w}$) are represented in the same way, and formulas are supposed to refer to every world in the class (*e.g.* the first column in $\overline{w_\alpha}$ shows the formulas that are included in all the worlds of that class).

contained in the tableau are true. More precisely, a tableau represents all those worlds that contain the formulas of its left column and do not contain any of the formulas of its right column.

The splitting rule of the analytic tableaux method has also been modified[10]: when it is applied to analyse a disjunction $(\phi \vee \psi)$ contained in the left column of a tableaux T, it generates three subtableaux: the first one has ϕ and ψ in its left column; the second one has ϕ in its left column and ψ in its right column, and the third one has ψ in its left column and ϕ in its right column. Thus, with this rule the agent explores all the possibilities of accounting for the truth of the disjunction, namely that one (or both) of its members are true. The three subtableaux also keep all the formulas contained in T.

After the application of this rule to analyse $(\neg\mathtt{R} \vee (\mathtt{S} \vee \mathtt{P}))$ (the only disjunction in T_0), the situation shown in figure 2 is reached.

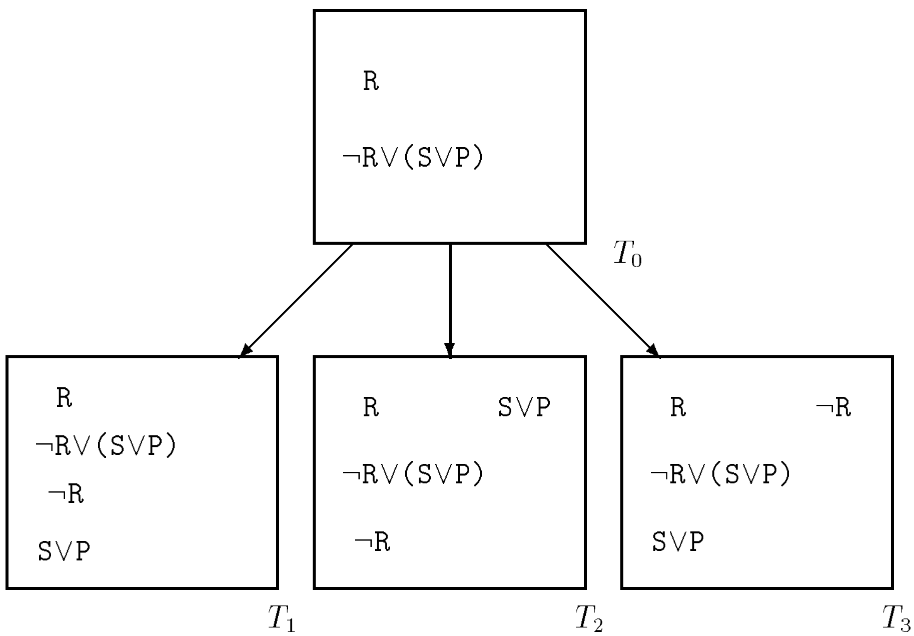

Figure 2: Splitting rule applied to $(\neg\mathtt{R} \vee (\mathtt{S} \vee \mathtt{P}))$

One can wonder now which class of worlds is represented by T_1, T_2 or T_3. Recall that T_0 represents the class of worlds $\overline{w_\alpha}$. Consider the following partition of class $\overline{w_\alpha}$:

- $\overline{w_{\alpha_1}}$: worlds of $\overline{w_\alpha}$ that contain $\neg\mathtt{R}$ and $(\mathtt{S} \vee \mathtt{P})$.
- $\overline{w_{\alpha_2}}$: worlds of $\overline{w_\alpha}$ that contain $\neg\mathtt{R}$ but do not contain $(\mathtt{S} \vee \mathtt{P})$.
- $\overline{w_{\alpha_3}}$: worlds of $\overline{w_\alpha}$ that contain $(\mathtt{S} \vee \mathtt{P})$ but do not contain $\neg\mathtt{R}$.
- $\overline{w_{\alpha_4}}$: worlds of $\overline{w_\alpha}$ that do not contain either $\neg\mathtt{R}$ or $(\mathtt{S} \vee \mathtt{P})$.

[10]Recall that, in the classical method, the analysis of a disjunction $(\phi \vee \psi)$ generates two subtableaux, one containing ϕ and the other containing ψ.

It is clear that the tableau T_1 represents the class of worlds $\overline{w_{\alpha_1}}$, T_2 represents $\overline{w_{\alpha_2}}$ and T_3 represents $\overline{w_{\alpha_3}}$. The worlds in class $\overline{w_{\alpha_4}}$ are not even considered by the analytic tableaux method, because it looks for models of the initial set of formulas, and it is not possible to have a model of the set { R, (¬R ∨ (S ∨ P)) } in which both ¬R and (S ∨ P) fail to be true.

In the classical analytic tableaux method, those tableaux that contain an atom and its negation are *closed*, and thus dismissed from the logical analysis (because these tableaux cannot represent a model of the initial set of formulas, which is the aim of the method). In the logical analysis, the agent can consider these worlds as *logically impossible*, not physically realizable, and can eliminate them from the analysis by *closing* the tableau that represents this kind of worlds. In the tableaux analysis that is being considered, a tableau will be closed if any of the following circumstances holds:

- The tableau contains a formula and its negation in its left column; it is representing a world that contains an explicit contradiction, so it can be considered logically impossible and eliminated from the analysis.

- The tableau contains the same formula both in the left and the right column[11]; it represents the worlds that contain and do not contain the given formula, so it represents the empty set of worlds and can be dismissed from the analysis.

In the example, T_1 and T_2 contain both R and ¬R in their left columns, so the agent can consider the classes of worlds represented by these tableaux ($\overline{w_{\alpha_1}}$, $\overline{w_{\alpha_2}}$) as logically impossible[12] and it can get rid of these worlds by closing T_1 and T_2 and dismissing the branches that contain them from the logical analysis. The remaining tableaux tree contains only T_0 and T_3. Thus, the effect of the analysis of the formula (¬R ∨ (S ∨ P)) in the analytic tableaux is the restriction of the doxastic alternatives of w_e that shrink from $\overline{w_\alpha}$ to $\overline{w_{\alpha_3}}$. The application of the splitting rule is modelling the generation of a new accessibility relation, R_1, that restricts the set of R_0-accessible worlds. This restriction is shown in figure 3.

w_e was R_0-related to all the worlds in class $\overline{w_\alpha}$, but is R_1-related only to the worlds in class $\overline{w_{\alpha_3}}$. The change of the accessibility relation implies a (possible) change in the beliefs of the agent. The beliefs in w_e at this point would be { R, (¬R ∨ (S ∨ P)), (S ∨ P) }. Notice that the restriction of the doxastic alternatives has caused the addition of a new belief, (S ∨ P). This formula is a new belief because it was not included in all the worlds in $\overline{w_\alpha}$ but it appears in all the worlds in $\overline{w_{\alpha_3}}$. Furthermore, as $\neg R$ is *not* included in any of the worlds in $\overline{w_{\alpha_3}}$, the analysis also shows that it is *not* a belief of the agent in w_e, and this fact is also reflected in figure 3, with the addition of $\neg R$ in the fourth column of the representation of w_e.

If the agent wants to pursue the analysis of its beliefs it can generate the next accessibility relation, R_2, by analysing the disjunction (S ∨ P) in T_3, obtaining the result shown in figure 4.

A more intuitive interpretation of the restriction from R_1 to R_2 results if you consider the following partition of $\overline{w_{\alpha_3}}$ in four subclasses:

- $\overline{w_{\alpha_{3_1}}}$: worlds of $\overline{w_{\alpha_3}}$ that contain S and P.

[11]The example used in this section does not contain any tableau with this property; this situation would appear for instance if $(\phi \vee \psi)$ is analysed in a tableau with the formulas $\{\phi, (\phi \vee \psi)\}$ in its left column.

[12]With the previous definition of worlds, those that contain an atom and its negation are not *impossible*, so it would be more appropriate to designate them with another expression; they are *conceivable, imaginable* but *physically unrealizable, incoherent, logically inconsistent, non-classical, non-standard, etc.*

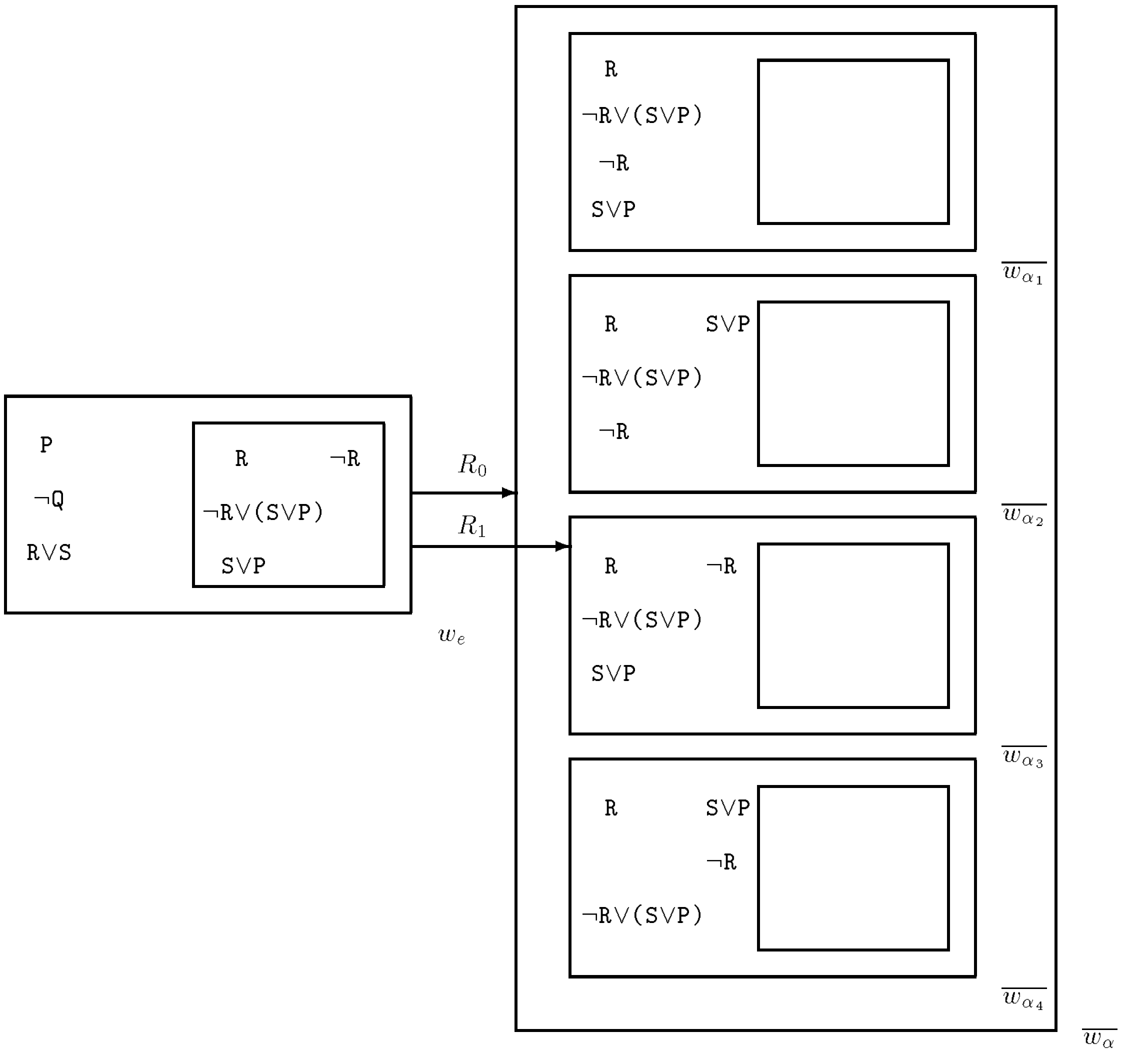

Figure 3: Reduction of doxastic alternatives

- $\overline{w_{\alpha_{3_2}}}$: worlds of $\overline{w_{\alpha_3}}$ that contain `S` but do not contain `P`.
- $\overline{w_{\alpha_{3_3}}}$: worlds of $\overline{w_{\alpha_3}}$ that contain `P` but do not contain `S`.
- $\overline{w_{\alpha_{3_4}}}$: worlds of $\overline{w_{\alpha_3}}$ that do not contain either `S` or `P`.

w_e, that was R_1-related to the worlds in $\overline{w_{\alpha_3}}$, is only R_2-related to the worlds of the subclasses $\overline{w_{\alpha_{3_1}}}$, $\overline{w_{\alpha_{3_2}}}$ and $\overline{w_{\alpha_{3_3}}}$, but not to those in $\overline{w_{\alpha_{3_4}}}$. This situation is depicted in figure 5.

Notice that the set of positive beliefs has not changed, because there is no formula common to all the R_2-accessible worlds which was not already common to all the R_1-accessible worlds from w_e. However, the set of negative beliefs has increased, because `P` is not contained in all the doxastic alternatives (because the worlds in $\overline{w_{\alpha_{3_2}}}$ do not contain it). The same situation happens to `S`, which is not contained in the worlds in class $\overline{w_{\alpha_{3_3}}}$; thus, neither `P` nor `S` are believed in w_e.

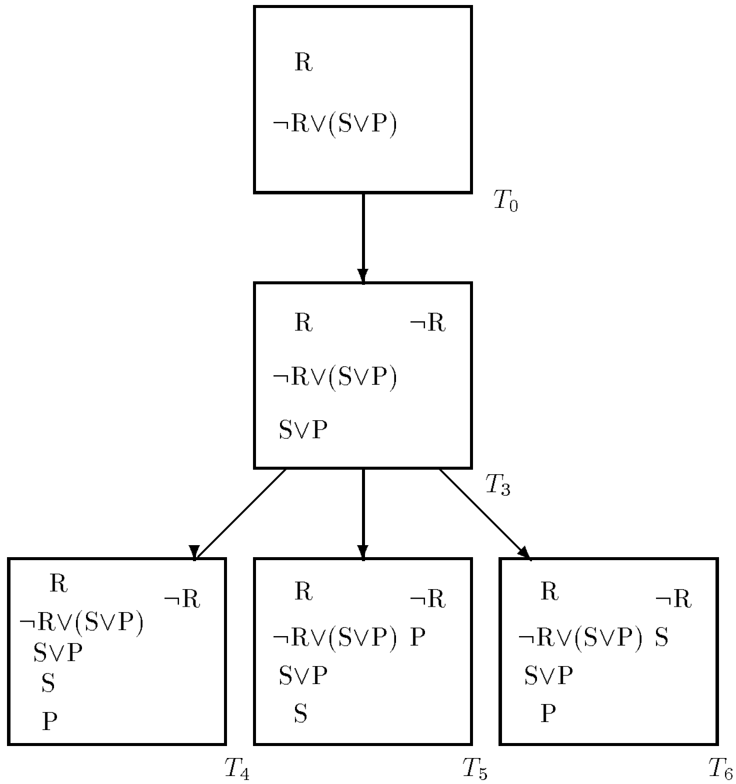

Figure 4: Analysis of (S $\vee$ P)

This fact is reflected in the fourth column of the rectangle that represents w_e in figure 5. There are no more disjunctions to be analysed in the open tableaux, so this is the end of the (purely logical) analysis that the agent can perform in w_e (without introducing instances of tautologies, topic considered in section 3.3.3).

3.3.3 Introduction of doubts in the analytic tableaux

In this logical analysis of beliefs the classical analytic tableaux method has been modified in a number of ways:

- A tableau contains two sets of formulas (the left and the right columns). It represents the set of possible worlds that contain all the formulas in its left column and do not contain any of the formulas in its right column.
- When the splitting rule is applied, three subtableaux are generated. One of them contains the two members of the analysed disjunction in its left column, while the other two 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 of the disjunction that has been analysed, namely that one (or both) of its members are true.
- After applying a splitting rule, the analysed proposition and all the other formulas in the tableau are maintained in the resulting tableaux, they are not deleted. The agent would

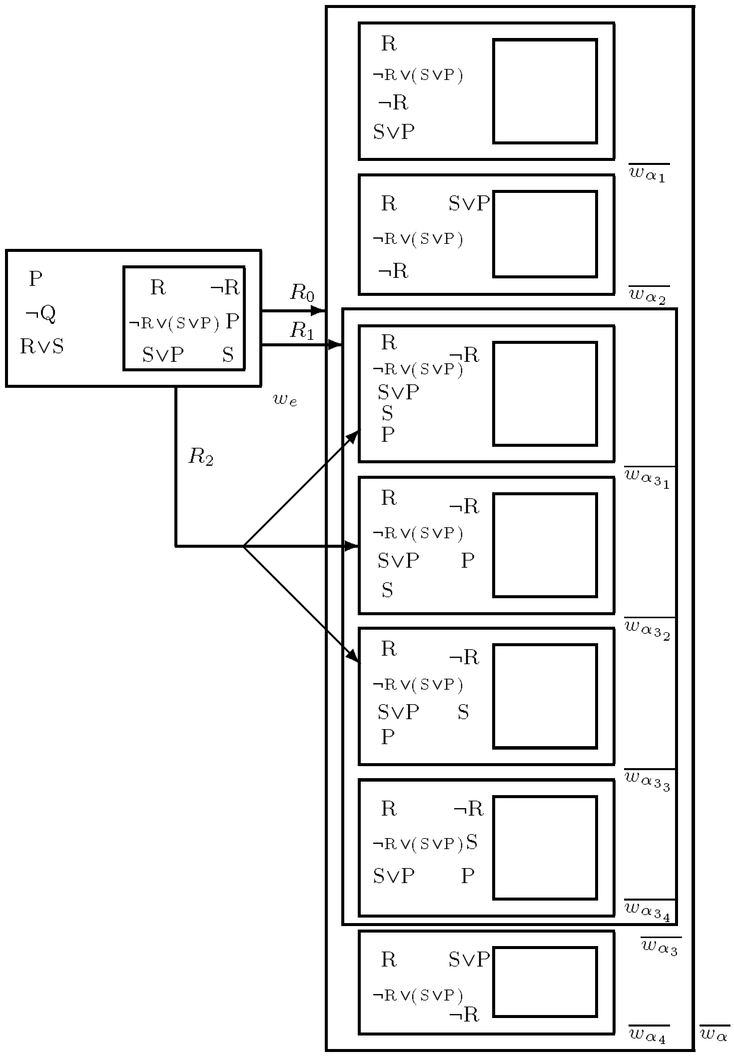

Figure 5: Further reduction of doxastic alternatives

just add a tag to the proposition in other to recall that it had already been analysed. Notice that, in the example shown in figure 2, if ($\neg$R $\vee$ (S $\vee$ P)) would not have been included in T_1, T_2 and T_3 then the set of propositions common to these tableaux would have been {R} and, therefore, the beliefs of the agent would have decreased from {R,$\neg$R $\vee$ (S $\vee$ P)} (before starting the analysis) to {R}. If analysed propositions are not eliminated during the logical analysis, then the set of beliefs of the agent cannot decrease in this process[13].

- The user is not allowed to add to an open tableau any tautology; otherwise, the problem of logical omniscience could not be avoided. This prohibition is based on the idea that a

[13]This modification is not essential; however, it is quite convenient because later it is easier to consider the set of formulas that appear in all open tableaux (the leaves of the tableaux tree) than the set of formulas that appear in all the branches of the tree.

tableau is just a mere list of the beliefs of the agent in a certain moment, and it cannot be modified with no apparent reason.

- There is only one exception to the previous rule: the agent can freely introduce (some) instances of the Axiom of the Excluded Middle, AEM, $(p \vee \neg p)$.

This exception is made for several reasons:

- It seems a very natural way to allow the agent to have *doubts*, to wonder whether it believes some formula (p) or its negation $(\neg p)$. This exception permits the introduction of the formula $(p \vee \neg p)$ in a tableau, which is later split into two subtableaux containing p in one column and $\neg p$, in the other[14]. 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 classical example is shown in section 4.1). In fact, the possibility of adding this kind of tautologies in the analytic tableaux is a classical idea in the tradition of model theory, as described *e.g.* in [BELL77].

- The introduction of instances of AEM is also a way to control the set of formulas that can be believed by the agent using the analytic tableaux method. It can be shown (see appendix A) that the set of believed formulas depends on the kind of instances of AEM that are allowed, so putting restrictions on the use of this principle is a way to constrain the set of obtainable formulas (the set of potential beliefs). In appendix B this idea is pushed even further, when the use of a different tautology is considered. As will be shown in section 4, the potential beliefs of the agent will be controlled by restricting the (instances of the) tautologies that the agent can use in the logical analysis.

- This exception also creates a strong link between this approach and the *interrogative model of inquiry* proposed by Hintikka ([HINT81], [HINT86], [HINT87], [HINT88], [HINT92]). He models the process of scientific inquiry with plays of the *interrogative game*, which is a game played by the *Inquirer* and *Nature*. The Inquirer seeks to know whether a certain conclusion C is true or not, and it can perform *deductive* moves (controlled with analytic tableaux) and *interrogative* moves, which are questions put to Nature. In these interrogative moves, Hintikka allows the introduction of (some) instances of AEM in the left side[15] of the subtableaux (those in which p is a subsentence of a formula of the subtableaux or a substitution instance of a subformula of a formula in the subtableaux with respect to names appearing in the subtableaux).

[14]The third subtableau would be immediately closed because it would contain both p and $\neg p$ in its left column.

[15]Hintikka uses the classical Beth-style analytic tableaux, as described in [BETH55].

4 General view of the belief system

This section describes briefly the different parts of the belief system, shown in figure 6.

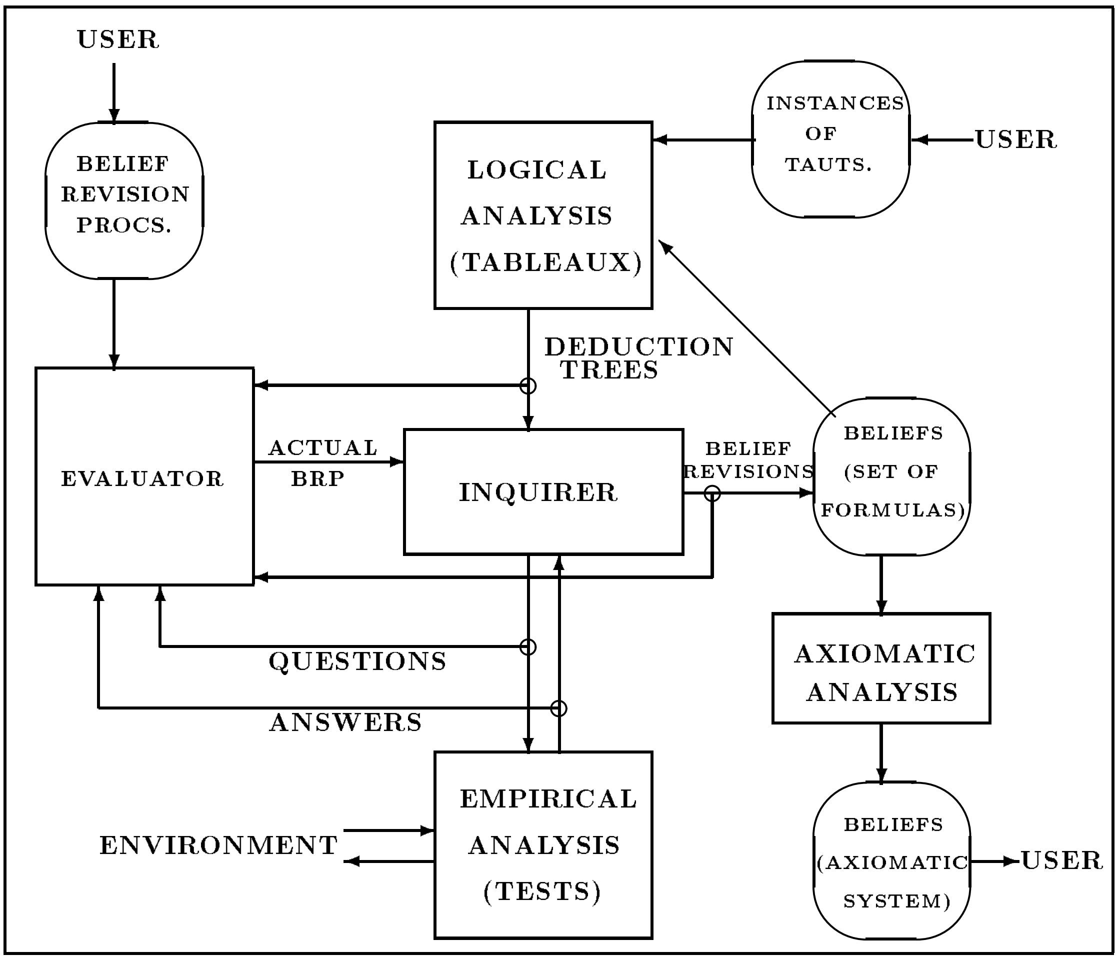

Figure 6: Modular decomposition of the belief system

The user (the designer of the agent) controls two input parameters to the belief modelling system:

- The belief revision procedures that can be used when a contradiction between the logical and the experimental analysis is found.
- The tautologies (and the kind of instances of each tautology) that can be used in the logical analysis. This is the way to control the set of potential beliefs obtainable by the agent.

The output of the system is an axiomatic system that the designer can use as a model of the beliefs of the agent. The designer would then probably use a standard theorem prover over this axiomatic system to derive the beliefs of the agent.

The process of the modelling system, briefly sketched, would be the following:

- The logical analysis receives the set of beliefs of the agent (a set of formulas) and the tautologies that it can use, and applies[16] the method of the analytic tableaux, as described in section 3.3.

- The deductions (tableaux trees) performed by the logical analysis are sent to the Inquirer, that would use the experimental analyser to refute (or corroborate) the results of the logical analysis, performing tests in the environment.

- If the results of the logical analysis and the experimental analysis contradict each other, then the Inquirer uses the actual belief revision procedure in order to modify the set of beliefs of the agent so that the result of the experimental analysis holds and the previous result of the logical analysis is no longer obtainable.

- The Evaluator module controls the whole process of inquiry. It would receive the deductions of the logical analysis, the tests performed in the experimental analysis and their results, and the belief revisions performed by the Inquirer. From time to time it would evaluate whether the agent is having a good performance (*e.g.* how many of the deductions of the logical analysis are indeed corroborated by the environment, whether the actual belief revision procedure is adequate, or maybe it is too permissive or too restrictive, *etc.*). If it considers that the actual belief revision is not appropriate, then it would choose another one from the set of belief revision procedures, and give it to the Inquirer.

The rest of this section describes very briefly the main ideas underlying the experimental and axiomatic analyses.

4.1 Experimental analysis

In the experimental dimension of analysis the agent is allowed to add propositions (externally obtained, as opposed to the internally obtained propositions of the logical analysis previously described) in all open tableaux of the logical analysis. Thus, this dimension models the agent's acquisition of data in the actual world (obtained via perception, experiences, tests, *etc.*). 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 thinks that reason must not approach Nature as a student, that takes everything that its teacher chooses to say, but as a judge who formulates questions and compels the witnesses to answer them. Nature has to be constrained to give answers to questions of reason's own determining.

This dimension is very much related to the logical dimension, and especially to the introduction of instances of AEM in the tableaux. The logical analysis can guide the experimental analysis, suggesting which experiences or tests the agent can perform in the actual world to gain knowledge.

Suppose (considering predicate logic for a second) that the agent doubts of the validity of a general law such as *All birds fly* (*i.e.* $(\forall x\ Bird(x) \Rightarrow Flies(x))$). It adds to its set of beliefs the formula $((\forall x Bird(x) \Rightarrow Flies(x)) \vee (\neg\forall x Bird(x) \Rightarrow Flies(x)))$ and, when it analyses this

[16]The way in which the logical analysis would be controlled has not been considered. The agent would possibly have a set of goals that should be achieved and they would guide this logical analysis.

disjunction, it will have access to two classes of worlds: in the first class the law will hold, while the second will contain $(\neg\forall x Bird(x) \Rightarrow Flies(x))$, from which by normal tableaux processing the agent would get $Bird(a)$ and $\neg Flies(a)$ (for some (undetermined) a). At this point, the agent can notice that it can increase its beliefs if it can dismiss this class of worlds; if the agent looks *actively* in the real world for an individual a such that $Bird(a)$ and $\neg Flies(a)$ and it cannot find it, then it can conclude that an individual with these properties does not exist; when this conclusion is finally reached, the agent can declare this class of worlds *physically impossible*, and so the agent will have grounds not to consider it any more as a conceivable, realizable alternative. Then the tableau that represents this class of worlds would be closed and dismissed from the analysis. At that moment all the worlds accessible by the agent would contain $(\forall x Bird(x) \Rightarrow Flies(x))$ so, by definition, the agent would believe this law. One important issue arises here, namely the difference between monotonic and non-monotonic beliefs. The former are the ones obtained from the purely logical analysis, while the latter are those derived from the experimental analysis, that may be defeated later (*e.g.* the agent can later be aware of the existence of an individual which is a bird and does not fly, and it will have to retract somehow its previous belief in the general law). This means that this work will have connections with the AI field of *belief revision*, which has received much attention in recent years (especially from the logical approach taken by Gärdenfors, Makinson and Alchourrón; see *e.g.* [ALCH85], [GARD88]).

Thus, as Hintikka suggested in [HINT86], doubt can be understood as the beginning of a *dynamic* process that, by reducing the number of conceivable worlds 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 ([POPP34]) philosophy of science. 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 possible worlds (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 ([BARW83]) that knowledge is "successful belief" than to standard epistemological traditions[17].

4.2 Axiomatic analysis

The idea in this third dimension is to represent the beliefs of the agent with an axiomatic system (a set of core axioms plus a set of inference rules), in a Konolige's deduction model of belief fashion ([KONO86]). The beliefs of the agent would then be all those propositions which can be obtained from the set of axioms with the application of the inference rules.

Notice that this dimension does not perform any deduction on the set of beliefs (as the logical dimension does), but merely transforms the set of beliefs into another form, namely an axiomatic system.

The main motivation of this dimension is a concept that could be called *cognitive economy*. There are several advantages of an axiomatic system *vs.* a set of beliefs, *e.g.* the following:

[17] Although the idea of "successful belief" in [BARW83] is of a probabilistic nature.

- ease of manipulation and representation
- representation of different levels of acceptance of the beliefs (*e.g.* the easier it is to obtain a proposition from the axioms, the easier it is for the agent to believe it)
- compactness of the axiomatic system
- different viewpoints on the same set of beliefs (with different choices of the axioms and inference rules)

5 Summary and future work

- The main aim of this work is to provide a way of modelling the process of rational inquiry, unlike the standard approach of modelling the beliefs of an agent in a certain fixed point in time.
- This process of rational inquiry is modelled within the possible worlds framework and its classical Kripke semantics. To avoid the problems of logical omniscience and perfect reasoning that appear in this semantics, two important changes to the classical approaches are made:
  - Possible worlds are considered as (possibly) incomplete and (possibly) inconsistent descriptions of the actual world.
  - Both the set of possible worlds W and the accessibility relation between worlds R are not constant, fixed, but variable. This variability accounts for the evolution of the agent's beliefs over time (through the usual Kripke semantics).
- In the practical implementation of these ideas, a world is defined as a set of formulas in a certain propositional doxastic language. To avoid the drawbacks of this syntactic definition, a brand new three-dimensional analysis of the beliefs is proposed; this analysis includes:
  - A logical dimension, in which the agent makes some deductive inferences in its belief set, using a modified version of the analytic tableaux method.
  - An experimental dimension, which allows the agent to take into account all the data that it can obtain from the actual world. The agent performs tests in the real world in order to increase its knowledge. The tests to be performed are suggested by the logical analysis (linking in a novel fashion the rational and empirical components of the classical notion of inquiry). The agent modifies its beliefs if it finds a contradiction between a logical derivation and experimental data.
  - An axiomatic dimension, which transforms the belief set into a more compact format, namely into an axiomatic set. This component does not appear in the classical views of inquiry in the philosophical literature.

Reading this report, it is obvious that many questions will have to be answered in the course of this inquiry. The first ones to be addressed will be the following:

- Which is the result of a logical analysis if no instance of AEM is allowed in the analytic tableaux? If some instances of AEM are indeed allowed, how does the resulting set of beliefs depend on the restrictions put on the allowed instances? Would the same results be obtained if any other tautology, different from AEM, were used?[18]

- In the experimental dimension, which type of experiences will be allowed? How does the restriction on the allowed answers from the environment change the potential results of the process of inquiry? Which will be the relationship of the addition of new information in the open tableaux with the work already developed in AI in belief revision?

- How will a set of beliefs be transformed into an axiomatic system, in the axiomatic dimension? How will the agent decide between different choices for the set of axioms and the set of rules that compose the axiomatic system? How will it even be able to find these options?

- Which will be the relationship between the three dimensions? In which conditions will the logical dimension suggest an experience to be performed in the experimental dimension? When will the set of beliefs be transformed into an axiomatic system, using the axiomatic dimension? Once an axiomatic system is built at a certain point, how will it be maintained when further logical and experimental analysis are performed? Which process will be followed by the agent to decide when to perform logical analysis with the analytic tableaux method, when to perform a certain test in its environment or when to transform its belief set into an axiomatic system?

- How would the Evaluator be able to know whether the actual belief revision procedure is appropriate or not? If it decides that the Inquirer is not having a good performance, which criteria would it use to decide which of the other belief revision procedures should be used?

References

[ALCH85] 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.

[APPE85] Appelt, D., "**Planning English sentences**", *Studies in Natural Language Processing*, Cambridge University Press, 1985.

[BARW83] Barwise, J., Perry, J., "**Situations and attitudes**", Bradford Books, Cambridge, MA, 1983.

[BELL77] Bell, J., Machover, M., "**A course in Mathematical Logic**", North Holland, 1977.

[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.

[18] These questions about the logical analysis have partly been considered, see appendixes A and B.

[BETH55] Beth, E.W., "**Semantic entailment and formal derivability**", from *Mededelingen van de Koninklijke Nederlandse Akademie van Wetenschappen, Afdeling Letterkunde*, Vol. 18, no. 13, pp. 309-342, 1955.

[BRAT88] Bratman, M., Israel, D., Pollack, M., "**Plans and resource-bounded practical reasoning**", *Computational Intelligence* 4, pp. 349-355, 1988.

[CRES72] Cresswell, M.J., "**Intensional logics and logical truth**", *Journal of Philosophical Logic* 1, pp. 2-15, 1972.

[DELG95] Delgrande, J., "**A framework for logics of explicit belief**", *Computational Intelligence* 11, pp. 47-86, 1995.

[DEVL91] Devlin, K., "**Logic and information**", Cambridge University Press, 1991.

[ELGO88] Elgot-Drapkin, J., "**Step-logic: reasoning situated in time**", Ph.D. thesis, University of Maryland, 1988.

[FAGI85] Fagin, R., Halpern, J., "**Belief, awareness and limited reasoning**", *Procs. of the Ninth IJCAI*, pp. 491-501, 1985.

[FAGI95] Fagin, R., Halpern, J., Moses, Y., Vardi, M., "**Reasoning about knowledge**", MIT Press, 1995.

[FITT90] Fitting, M., "**Modal/multiple-valued logics**", Technical Report, CUNY, 1990.

[GARD88] Gärdenfors, P., "**Knowledge in flux**", Cambridge University Press, 1988

[HALP92] 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.

[HINT75] Hintikka, J., "**Impossible possible worlds vindicated**", *Journal of Philosophical Logic* 4, pp. 475-484, 1975.

[HINT81] Hintikka, J., "**On the logic of an interrogative model of scientific inquiry**", *Synthese* 47, pp. 69-83, 1981.

[HINT86] Hintikka, J., "**Reasoning about knowledge in philosophy: the paradigm of epistemic logic**", *Proceedings of the 1986 TARK Conference*, Ed. J.Y.Halpern, pp. 63-80, 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.

[HOFS80] Hofstadter, D., "**Gödel, Escher, Bach: an Eternal Golden Braid**", Basic Books Inc. Publishers, 1980.

[KONO85] Konolige, K., "**Belief and incompleteness**", *Formal theories of the commonsense world*, pp. 359-403, Hobbs and Moore Eds., 1985.

[KONO86] Konolige, K., "**A Deduction Model of Belief**", Morgan Kaufmann, San Mateo, CA, 1986.

[KRIP63] 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.

[LEVE84] Levesque, H.J., "**A logic of implicit and explicit belief**", *Proceedings of AAAI-84*, pp. 198-202, 1984.

[McAR88] McArthur, G., "**Reasoning about knowledge and belief: a survey**", *Computational Intelligence* 4, pp. 223-243, 1988.

[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**", in *Formal theories of the commonsense world*, pp. 319-358, Hobbs and Moore Eds., 1985.

[NIRK94] 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.

[POPP34] Popper, K., "**The logic of scientific discovery**", Ed. Tecnos, 1977 printing.

[RAOG91] Rao, A., Georgeff, M., "**Modeling rational agents within a BDI-architecture**", *Proceedings of the 2nd International Conference on principles of knowledge representation and reasoning*, pp. 473-484, 1991.

[REIC89] Reichgelt, H., "**Logics for reasoning about knowledge and belief**", *Knowledge Engineering Review*, Vol. 4, No.2, pp. 119-139, 1989.

[RESC79] Rescher, N., Brandom, R., "**The Logic of Inconsistency**", Rowman and Littlefield Eds., 1979.

[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.

[SMUL68] Smullyan, R.M., "**First-order logic**", Springer Verlag, 1968.

[SMUL86] Smullyan, R.M., "**Logicians who reason about themselves**", *Proceedings of the 1986 Conference on Theoretical Aspects of Reasoning about Knowledge*, Ed. J.Halpern, pp. 341-352, 1986.

[WOOL95] Wooldridge, M., Jennings, N., "**Intelligent agents: theory and practice**", *Knowledge Engineering Review* 10 (2), 1995.

A Use of the Axiom of the Excluded Middle in the logical analysis

This appendix shows the relationship between the set of beliefs potentially obtainable in the logical analysis and the kind of instances of the Axiom of the Excluded Middle that the agent is allowed to introduce in the analytic tableaux. For the sake of simplicity assume that I, the initial set of formulas (the formulas that are included in the left column of the initial tableau) is in CNF (beliefs are restricted to propositional logic, and negated and nested beliefs are not considered). Thus, I is a set of clauses $\{c_1, c_2, ..., c_n\}$.

It will be said that a formula f is *derivable* from the initial set of clauses I iff it is possible to perform a logical analysis, using the analytic tableaux method as described in section 3.3, with the initial tableau T_0 containing I in its left column, in such a way that f appears in the left columns of all open tableaux (at least one[19]). Recall that the open tableaux represent classes of doxastic alternatives (each tableau represents all those worlds that contain the formulas of the left column of the tableau and do not contain any of the formulas of its right column). That means that, if at a certain point of the logical analysis all the open tableaux contain[20] a certain formula f, this formula is contained in all the doxastic alternatives so, *via* the Kripke semantics, it is a belief of the agent at that point. Therefore, the set of potential beliefs is equal to the set of derivable formulas.

The rest of this appendix is a brief study of how the set of derivable formulas changes when different constraints are applied over the set of allowed instances of AEM. Each constraint defines a different *derivability* relation.

- No instances of AEM are allowed

  Let's define $\models_0$ as the derivability relation obtained with the logical analysis when the agent is not allowed to add to the open tableaux any instance of AEM. The set of derivable formulas (the set of potential beliefs) can be characterized in the following way:

  $(I \models_0 f)$ iff

  - I does not contain a formula and its negation, and
  - I does not contain the negation of f, and
  - $f = c_k$ or there exist in I some clauses of the form $(f \vee c_p)$ s.t. if f is removed from these clauses the result is an inconsistent set.

  The first and the second conditions are trivial; if I contained a formula and its negation, T_0 would automatically close, so the set of derivable beliefs from I would be empty. If I contains a certain formula, p, then $\neg p$ cannot be derived, because as soon as it is generated, it creates a tableau that contains p (that is contained in all tableaux) and $\neg p$, and therefore is immediately closed. These conditions will be called TC in the rest of this report. The third condition states that f must be a clause of I or a subclause such that, if it is removed

[19] There must be at least one open tableau, otherwise *any* formula would be derivable from an inconsistent set.

[20] All unqualified references to the tableaux in the rest of the report are supposed to refer to their left columns.

from every clause in which it appears, generates an inconsistent set. This condition is also quite obvious. Imagine that $I = \Delta \cup \{(f \vee c_1), ..., (f \vee c_p)\}$, where the formulas in Δ do not contain f. f can be $\models_0$-derived from I iff the tableau containing $\Delta \cup \{c_1, ..., c_p\}$ is closed, and that will only happen if that set is inconsistent.

Therefore, the set of $\models_0$-derivable formulas is simply a subset of the subformulas of I that are (classical) logical consequences of I[21].

This relation $\models_0$ seems to have only *negative* properties, such as the following:

- There is a certain *paraconsistency* because if I is inconsistent, it is not possible to derive every formula from it (only -some- subclauses can be derived). That is a good property but, unfortunately, this paraconsistency does not seem to have any clear motivation, because it is not very obvious why some subclauses can be obtained while similar clauses can't be derived.

  For instance, consider the set I=$\{(a \vee b \vee c \vee \neg d), (a \vee b \vee \neg c \vee \neg d), (a \vee b \vee d)\}$. It is easy to check that $(a \vee b \vee c)$ is a logical consequence of I (I is equivalent to $(a \vee b)$), but it is not $\models_0$-derivable from I. Surprisingly enough, $(a \vee b \vee c)$ is $\models_0$-derivable from the set I'=$I \cup$ {d} (that is equivalent to $((a \vee b) \wedge d)$).
- There is also *monotonicity*; if a conjunction is derivable from I, it is also derivable from any superset of I. The only way to *revoke* a derivable formula is to add to I a formula and its negation (thus getting rid of *all* derivable formulas), or to revoke a formula by adding its negation to I (so the only way to get some trivial non-monotonicity is to use the TC conditions, making I inconsistent).
- There is *neither logical omniscience nor perfect reasoning*, but the set of potential beliefs is *too* restricted (it is not even possible to derive all the subclauses of I that are logical consequences of I, as has been shown in the previous example).
- The set of derivable formulas is *not closed under provable subformulas*[22] (if q_1 is derivable from I, q_2 is a subclause of q_1, and q_2 is a logical consequence of I, it may be that q_2 is not derivable from I). This can be seen using again the same example ($(a \vee b \vee c \vee \neg d)$ is derivable from I, $(a \vee b \vee c)$ is a subclause of $(a \vee b \vee c \vee \neg d)$ and it is a logical consequence of I, but it is not derivable from I).

If no instance of AEM is allowed the set of derivable formulas is very small (a subset of the subclauses that are logical consequences of the initial set). In the rest of this appendix it is shown how this set can be enlarged if the agent is allowed to put some instances of AEM in the open tableaux of the logical analysis.

- Every instance of AEM is allowed

  Let's define the relation $\models_1$ just like $\models_0$, but in this case the agent can add *any* instance of AEM in any tableau. It can be seen that

[21] If f is $\models_0$-derivable, the tableau $I \cup \{\neg f\}$ would close in the classical analytic tableaux method, simply applying the same rules that had been applied to proof the derivability of f.

[22] If it were closed under provable subformulas, it would contain (forgetting TC) *all* provable subclauses, and it doesn't.

$(I \models_1 f)$ iff TC and $(I \models f)$

Thus, if there is no restriction imposed on the use of AEM, the agent can obtain (excluding the details of TC) all those formulas that are (classical) logical consequences of I[23]; in particular, if I is inconsistent, any formula is $\models_1$-derivable. Therefore, the agent would already have fallen into the pit of logical omniscience and perfect reasoning.

Recall that one of the aims of this research is to give to the agent a *sensible* way to derive some formulas from its initial set of beliefs, but avoiding logical omniscience and perfect reasoning. $\models_1$ must therefore be weakened, and that will be done by constraining the instances of the AEM that can be added in the tableaux of the logical analysis.

- Atomic instances of AEM are allowed

  Let's call $\models_2$ to the relation obtained when the agent is allowed to use AEM, but only in the *atomic* case; it can only add to the tableaux formulas of the type $(p \vee \neg p)$, being p an atom that appears in I. The $\models_2$-derivable formulas can be characterized as follows:

  $(I \models_2 f)$ iff TC and

  - $(I \models_0 f)$, or
  - $f = (p \vee \neg p)$, being p an atom appearing in I, or
  - f, if f is a literal, I is inconsistent and I contains $\neg f$ but does not contain f

  The first and the second clause deserve no comment; the third one is just saying that some literals that were not $\models_0$-derivable are $\models_2$-derivable, when I is inconsistent (it is just a minor detail without apparent significance; e.g. consider the (inconsistent) set $I=\{(\neg a \vee b), a, \neg b, (p \vee \neg q)\}$; q cannot be $\models_0$-derived because it is a literal that does not appear in I, but it is $\models_2$-derivable because the agent could add now the formula $(q \vee \neg q)$, analyse it and close the subtableau containing $\neg q$. So, surprisingly enough, if atomic instances of AEM are allowed, the set of derivable sentences is almost equal to the one obtained when no instances at all were allowed. Therefore, the agent should be less restrictive and allow a greater set of instances of AEM.

- Subformula instances of AEM are allowed

  $\models_3$ is the relation obtained when the agent can add to the tableaux formulas of the type $(s \vee \neg s)$, being s a subclause of a clause in I. The following result holds:

  $(I \models_3 f)$ iff TC and

  - $(I \models f)$ and f is a subclause of a clause in I, or
  - $f = (s \vee \neg s)$, being s a subclause of a clause in I

  This result is also quite obvious, because the agent can $\models_3$-derive any subclause s that is derivable from I by using the formula $(s \vee \neg s)$, analysing it and closing the subtableau

[23] If the agent wants to $\models_1$-derive a formula f from a set I, being f a logical consequence of I, it only needs to add the formula $(f \vee \neg f)$ in the tableau containing I and analyse it; it will obtain a tableau containing I and $\neg f$, and if the agent analyses this tableau exhaustively it will close all its subtableaux-because the set $I \cup \{\neg f\}$ is inconsistent- and the agent will finally believe f.

containing I and $\neg s$. Thus, it is possible to $\models_3$-derive (forgetting TC) all those subclauses of I that are logical consequences of I. It is possible to be less restrictive in the use of AEM without reaching $\models_1$, as shown with $\models_4$.

- Formulas composed by known atoms can be used in AEM instances

  Let's define $\models_4$ as the derivability relation obtained when the agent can add to the tableaux formulas of the type $(f \vee \neg f)$, being f any formula composed by atoms that appear in I. The following result holds:

  $(I \models_4 f)$ iff TC and $(I \models f)$ and f is a formula composed by atoms appearing in I

  Thus, $\models_4$ provokes logical omniscience and perfect reasoning, restricted to formulas composed by *known* atoms (this result is as obvious as the one obtained with $\models_1$). If one makes the (reasonable) assumption that every atom of the relevant vocabulary appears somewhere in I, this relation would already be equivalent to $\models_1$.

After this sketchy analysis of the influence of some instances of AEM in the potential beliefs of the agent, the following questions could be asked:

- Are the instances of AEM constrainable in other interesting ways?

- Is there any way to fill the gap between $\models_3$ (provable subclauses) and $\models_4$ (almost logical omniscience)? Is there any way to obtain *some* of the formulas classically derivable from I without obtaining *all* of them?

Surprisingly, a partial answer for the second item arose when another question was considered: How does the set of potential beliefs change (if it changes at all) if other tautologies are used in the logical analysis, instead of AEM? This point is taken in appendix B.

B Using other tautologies

In this appendix the following question is considered: do the potential beliefs obtainable by the agent in the logical analysis change (significantly) if the agent uses other classical tautologies, different from the Axiom of the Excluded Middle (AEM)? This appendix does not provide a full answer to this question, but the following example shows that maybe some interesting results can be obtained along this line.

The aim of this appendix is to compare the different uses of AEM (shown in appendix A) with the same uses of another tautology (QUO), namely $(p \Rightarrow (\neg p \Rightarrow q))$. Formulas are restricted in the same way as in the previous appendix, so the agent will add to the open tableaux instances of the formula $(\neg p \vee (p \vee q))$. Let's examine the same cases as in appendix A.

- $\models_0$: No instance of the tautology is allowed

  Obviously, in this case one obtains the same $\models_0$-derivable formulas that were obtained in appendix A.

- $\models_1$: Every instance of the tautology is allowed

  $\models_4$: Every instance of the tautology (composed by known atoms) is allowed

In these cases you also obtain the same results that the ones obtained using AEM[24] (logical omniscience and perfect reasoning).

- $\models_2$: Every *atomic* instance of the tautology is allowed

  Recall that, using AEM, in this case the following result held:

  $(I \models_2 f)$ iff TC and

  - $(I \models_0 f)$, or
  - $f = (p \vee \neg p)$, being p an atom appearing in I, or
  - f, if f is a literal, I is inconsistent and contains $\neg f$ but does not contain f

  Replacing AEM by QUO, the set of $\models_2$-derivable formulas can be characterized as follows:

  $(I \models_2 f)$ iff TC and

  - $(I \models_0 f)$, or
  - $f = (\neg p \vee (p \vee q))$, being p and q atoms appearing in I, or
  - f, if f is a literal, I is inconsistent and contains $\neg f$ but does not contain f, or
  - $f = (p \vee q)$, being p and q atoms of I and $(I \models p)$

  The first three clauses are not very surprising, but an unexpected result appears in the last clause; using QUO, it is possible to obtain disjunctions of atoms, if the first atom is a logical consequence of the initial set of formulas (obviously that implies that the disjunction is also a logical consequence of I; it could not be otherwise, because it is impossible to derive, using the analytic tableaux method, any formula that is not a logical consequence of the initial set of formulas). This result holds because, if p is a logical consequence of I, the tableaux containing I and $\neg p$ will eventually close and the tableau containing only $(p \vee q)$ will remain open. This result seems particularly interesting, because $(p \vee q)$ *does not have to be* a subclause in I, so it is the first time that, using a limited version of the analytic tableaux method, one can obtain derivable formulas that are not directly included in the original clauses, so it seems a step forward towards the answer of one of the questions at the end of appendix A: is there any way to obtain *some* of the formulas classically derivable from I (that are not subclauses in I) without obtaining *all* of them? It has just been shown that the answer is yes (*e.g.*, the formula $(\neg A \vee C)$ was not $\models_2$-derivable from the set {A, $(\neg A \vee B)$, $(\neg B \vee C)$} using AEM, but it is $\models_2$-derivable using QUO). Now other questions should be addressed:

  - How could one find the tautologies that should be used in the logical analysis in order to be able to derive a *sensible* set of potential beliefs?
  - Could the agent use a *set of tautologies*, depending on the applications? (maybe in one application the power of AEM,$\models_2$ would be enough, but in other application a larger set of beliefs should be considered, *e.g.* QUO,$\models_3$)

[24] Notice that, in these cases, AEM can be simulated using QUO, simply instanciating q with a contradiction.

Just to finish the comparison between AEM and QUO,

- $\models_3$: Every instance of QUO composed by subclauses of I is allowed

  Using AEM the result was:

  $(I \models_3 f)$ iff TC and

  - $(I \models f)$ and f is a subclause of a clause in I, or
  - $f = (s \vee \neg s)$, being s a subclause of a clause in I

  Using QUO the potential beliefs are the following:

  $(I \models_3 f)$ iff TC and

  - $(I \models f)$ and f is a subclause of a clause in I, or
  - $f = (\neg s_1 \vee (s_1 \vee s_2))$, being s_1, s_2 subclauses of clauses in I, or
  - $f = (s_1 \vee s_2)$, being s_1, s_2 subclauses of clauses in I and $(I \models s_1)$

  With this relation it is also possible to derive formulas that are not subclauses of the initial set of formulas, as in the previous case. *E.g.*, the formula $(A \vee (B \vee C))$ is $\models_3$-derivable from the set $\{A, (B \vee C)\}$ using QUO, but it is not $\models_3$-derivable using AEM.